A macro- and micro-evolutionary investigation of African *Camponotus* ants

Brigitte N. Eick

Supervisors:

Dr. C. O'Ryan (Department of Molecular and Cell Biology)

Dr. H. Robertson (South African Museum)

Prof. E. Harley (Division of Chemical Pathology)

Thesis Presented for the Degree of

Doctor of Philosophy

Department of Molecular and Cell Biology

University of Cape Town

November 2002

DECLARATION

This thesis reports the results of original research I conducted under the auspices of the Department of Molecular and Cell Biology, University of Cape Town, between 1998 and 2002. All the assistance that I received has been acknowledged. This work has not been submitted for a degree at any other university.

Signed by candidate

Brigitte Natalie Eick

Camponotus than the cytochrome oxidase II gene, based on almost all measures of phylogenetic utility. The primary hypothesis proposed to account for this observation is that these two mitochondrial genes are evolving under different evolutionary constraints. Specifically, the cytochrome oxidase II gene displays greater rate heterogeneity than the cytochrome *b* gene, thereby decreasing its utility for

phylogenetic analyses. Combining sequence data from both genes resulted in more robust phylogenetic hypotheses, with the combined topologies displaying greater congruence with the cytochrome *b* topologies than those based on cytochrome oxidase II sequence data. The morphological data produced a topology that was congruent with that obtained from molecular data, and provided increased support for certain nodes in the context of a combined molecular-morphological framework. The hypothesis that subgeneric classifications within *Camponotus* do not accurately reflect phylogenetic relationships was supported by the molecular phylogenies. An exception to this hypothesis was the monophyly of the subgenus *Myrmosericus*, based on cytochrome *b* data. The morphological and behavioural data provided support for a monophyletic group comprising the four species assigned to the subgenus *Myrmopiromis*. However, although these four species associated together in a group based on combined cytochrome oxidase II and cytochrome *b* sequences, this group was paraphyletic in the combined molecular topology, with two species in subgenus *Myrmopsamma* also falling within this group.

The recently rediscovered species, *Camponotus bifossus* (subgenus *Colobopsis*), whose major workers are characterized by phragmotic heads, associated as a sister species with *Camponotus. sp.* 12 (subgenus *Myrmespera*) with significant nodal support in all phylogenetic analyses. Furthermore, *Polyrhachis schistacea* grouped within the *Camponotus spp.* included in phylogenetic analyses, supporting the hypothesis of *Camponotus* paraphyly. However, when additional *Polyrhachis* species were included in the analysis, *Polyrhachis* formed a strongly supported monophyletic group. Due to the unresolved placement of two *Camponotus spp.* in this topology, however, the monophyly of *Camponotus* could not be confirmed.

Microsatellite markers developed in two other *Camponotus* species provided sufficient resolution to elucidate some aspects of the fine-scale colony genetic structure in the fynbos species *Camponotus klugii*. Colonies were found to be monodomous, with each nest founded by an independent queen, in contrast to Skaife's (1961) hypothesis that colonies of *Camponotus klugii* are polydomous. Colony kin structure was found to conform to the classic expectation of a social insect colony, comprising workers produced by a single, once-mated queen. Notably, evidence was found to support the occurrence of queen replacement in one of the sampled nests. This may be as a result of constraints on nest-site availability in the fynbos, although further ecological studies are required to confirm this hypothesis. No

evidence of clonal genetic structure was found in queenless nests, contrary to the hypothesis of Skaife (1961) that workers of this species may reproduce by thelytokous parthenogenesis in the absence of a queen.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my three supervisors, Dr Colleen O'Ryan, Prof Eric Harley, and Dr Hamish Robertson, who were always on hand to answer my often perplexing queries, and offer advice whenever needed. In particular, I would like to thank Dr Colleen O' Ryan for her concern for her students' intellectual development, and her ability to share her excitement about science and the process of scientific discovery. I would like to thank Prof. Harley for his cool, calm and rational approach to all seemingly insurmountable problems, and his faith that things will always work out. Last, but not least, I would like to thank Hamish Robertson for his ability to communicate his passion and enthusiasm for ants, and his willingness to give freely of his valuable time. A special thank you to Dr Terry Hedderson for his willingness to wade through first drafts of my phylogenetic work, and his invaluable commentary.

I would like to acknowledge Miranda Waldron from the EM unit at UCT for her help with scanning in SEM negatives, and Margie Cochrane from the SA museum for her help with the SEM microscopy work. I would like to thank Dr Keith Goodnight and Dr Tamsin Burland for responding to my relatedness enquiries, and Dr Per Kryger for his responses to my social insect queries.

The South African National Research Foundation (NRF) and the University Research Council (URC) provided financial support for the work detailed in this thesis.

A huge thank you to Jacqueline Bishop who was there for me when I needed her most, and who gave freely of her time and provided constant support, encouragement, constructive criticism, and words of PhD wisdom, as well as providing invaluable advice in the use of the Relatedness 5.0.7 software package. A special thank you to Paula de Coito for her willingness to read through all that I had written, and for the many hastily scheduled consultations concerning English grammar, as well as her enthusiasm and support. To my

friend Jessica Cunningham – thank you for all your support and encouragement, and for reading through multiple drafts of this thesis from the land of no sun. To Janet Kelso, Johann Visagie, Clare van der Willigen, Carlos Ruiz, Heidi de Wet, Ben de Wet and my brother Sven, your willingness to listen to all my worries, and provide calming words of wisdom has been my saving grace during this trying period of my life. And finally thank you to Dr David Jacobs, whose patience with my concerns, passion for scientific research and willingness to freely give of his time meant so much to my during the last stages of my PhD.

This thesis is dedicated to my mother, who has always encouraged my love of science, and provided me with all the opportunities I could ever have desired.

ABBREVIATIONS

bp	Base pairs
CFU	Colony forming units
CTAB	hexadecyltrimethylammonium bromide
dNTP	Deoxynucleotide
ddNTP	Dideoxynucleotide
kb	Kilobases
^{32}P	Phosphorus-32
PCR	Polymerase chain reaction
RNA	Ribose nucleic acid
RPM	Revolutions per minute
rRNA	Ribosomal RNA
tRNA	Transfer RNA
T_a	Annealing temperature
UCT	University of Cape Town
DNA	Deoxyribose nucleic acid
IUPAC	International union of pure and applied chemistry

TABLE OF CONTENTS

PART III

APPENDICES

Chapter 1

INTRODUCTION

Formicidae, the largest family within the insect order Hymenoptera, contains more than 9 000 described extant ant species, distributed among 296 genera in 16 subfamilies (Bolton, 1995). If persistence of a phyletic line through geological time is used as a measure of ecological success, then ants are indisputably one of the most successful taxonomic groups, as they have remained abundant and morphologically unchanged for over 50 million years (Wilson, 1987). Furthermore, their high species diversity, wide geographic range, high population densities and ability to occupy unique adaptive zones can be seen as further indicators of the phenomenal ecological success of this group. Their remarkable ecological dominance has been attributed to the fact that ants are the only eusocial predators that both live and forage primarily in soil and rotting vegetation on the ground, thereby occupying a unique environmental niche from which they appear to be able to exclude all potential competitors. Furthermore, the success of ants has been attributed to the competitive advantages that a highly social organisation confers over solitary individuals with respect to competition for food, territorial disputes and efficiency of labour (Wilson, 1987; Hölldobler and Wilson, 1990).

Camponotus Mayr (1861), the focus of this thesis, is the largest genus within the Formicidae, comprising more than 1500 described species and subspecies, in addition to 22 fossil species (Bolton, 1995). This genus is remarkably ecologically tolerant, evidenced by its worldwide distribution with species occupying every major biogeographical region (Bolton, 1995). Based on species diversity, geographical range, diversity of adaptations and local abundance, *Camponotus*, along with *Pheidole* and *Crematogaster*, is considered one of the most prevalent ant genera in the world (Wilson, 1976).

Colony structure in this genus appears to be simple, with colonies predominantly headed by single, once-mated queens (Akre *et al.*, 1994).

However, despite the multitude of species within this genus, very few genetic studies of colony structure have been undertaken thus far.

This thesis focuses on using molecular markers to address two different aspects of the biology of ants within the genus *Camponotus*. At a higher level, the phylogenetic utility of two mitochondrial genes for reconstructing and resolving relationships among a representative sample of African *Camponotus* species is evaluated. The resulting phylogenies are then used to evaluate a number of hypotheses pertaining to evolutionary relationships within this genus. This work is presented in Part I of this thesis. At a finer scale of genetic resolution, microsatellite markers developed in a number of *Camponotus spp.* are used to evaluate the sociogenetic structure of *Camponotus klugii*, an arboreally nesting species endemic to the Cape fynbos biome. These findings are reported within Part II.

The introduction to Part I, presented within this chapter, provides an overview of *Camponotus* taxonomy and systematics, an introduction to insect molecular phylogenetics, and a review of the current topical issues in systematic biology pertinent to this thesis. The introduction to the study on fine-scale genetic structuring in *Camponotus klugii* is provided in Part II of this thesis.

Part I

Section A: *Camponotus* taxonomy and systematics

The taxonomy of *Camponotus* is far from complete, and globally very few regional or national faunas have been comprehensively reviewed. This can be attributed both to insufficient taxonomic guidelines for this genus, as well as the unwieldy size of *Camponotus*. In an attempt to facilitate taxonomic handling of this group Emery published a comprehensive plan in 1896 dividing *Camponotus* into 26 subdivisions or 'maniples'[1], which were assigned to three major cohorts. These designations were largely based on structural and geographical considerations, although these were only explicitly defined by Emery in a later study (Emery, 1920). Subsequently, Ashmead (1905), Wheeler (1922), Forel (1912, 1914) and Santschi (1921) proposed additions and modifications to Emery's subgeneric taxonomic classifications. However, there remained no general consensus among researchers regarding delimitation of subgenera (Creighton, 1950). In 1925, Emery published a revised systematic catalogue for *Camponotus*, dividing the genus into 38 subgenera. Since this publication, eight further subgenera have been recognised, increasing the number of subgeneric divisions within *Camponotus* to 46 (Bolton, 1995).

There is, however, general dissatisfaction with these subgeneric divisions amongst ant systematists. They have been criticised on the basis that they are obscure, with there being no clear delimitation between the different subgenera due to intergradation of characters, which in turn manifests as artificial groupings (Arnold, 1924; Creighton, 1950; Brady *et al.*, 1999). Subgeneric diagnostic features include the extent and degree of pilosity on the scape, head and mesosoma; the shape of the mesosoma and petiole; colour of the head and mesosoma; number of mandibular teeth and nesting biology (see Appendix I). Many of these characters such as colour and hair distribution display intraspecific variation (H. Robertson pers. com.), indicating a high degree of evolutionary plasticity. Some characters also appear to be

[1] The subgeneric taxonomic rank was not yet recognised.

prone to convergent evolution, with others showing great evolutionary lability, decreasing their utility as diagnostic markers.

A case in point is illustrated by the subgenus *Colobopsis*. Species in this subgenus are characterised by the presence of sharply truncated, plug-shaped heads in the soldier and queen castes (Creighton, 1950). Workers use these modified heads to block arboreal nest entrances inhabited by these ants, a phenomenon known as 'phragmosis' (Wheeler, 1927). However, all degrees of phragmotic development are evident in *Camponotus*, with a variety of convergent mechanisms involving different relative proportions of the clypeus and anterior face resulting in a sharply truncated phragmotic head shape. This indicates that this character has evolved independently in several different groups of *Camponotus* (J. Longino pers. com.). Furthermore, phragmotic heads have evolved independently in other ant genera, indicating that monophyletic classifications based on this apparently homoplasious character, such as recognition of the subgenus *Colobopsis*, are highly questionable (Brady *et al.*, 1999).

The lack of an underlying evolutionary basis for many of the subgeneric classifications was highlighted by a preliminary study by Brady *et al.* (1999) based on molecular data. This study revealed that many of the subgeneric classifications did not appear to accurately reflect monophyletic groupings. The monophyletic status of this genus as a whole has been questioned, as *Camponotus* appears to be closely related to two other genera: *Polyrhachis* (477 species in Old World tropics) and *Dendromyrmex* (seven species in Neotropics). All known species in these genera, with the exception of two (Shattuck, 1999), are characterised by the absence of the metapleural gland in workers (Hölldobler and Engel-Siegel, 1985). This character is one of the strongest synapomorphies uniting ants as a whole, and therefore the absence of this character in these genera suggests monophyletic status for this group. It has been proposed that within this clade, *Camponotus* may represent a paraphyletic assemblage from which *Polyrhachis* and *Dendromyrmex* arose.

apomorphic characters, namely the presence of thick, blunt hairs on the gaster and an angulate cross-section to the hind tibiae.

Given the paucity of non-homoplasious morphological characters that distinguish species and subgenera in *Camponotus*, a molecular-based approach to phylogenetic reconstruction in this complex genus is warranted. The generation of molecular sequence data for resolving and clarifying phylogenetic relationships within the Formicidae has been viewed with excitement, although there are to date (November 2002) only a handful of published molecular phylogenies available (Baur *et al.*, 1993, 1995, 1996; Crozier *et al.*, 1995; Ayala *et al.*, 1996; Chenuil and McKey, 1996; Wetterer *et al.* 1998; Gadau *et al.*, 1999; Sameshima *et al.*, 1999; Brady *et al.*, 1999; Sauer *et al.*, 2000; Parker and Rissing, 2002).

The paucity of ant molecular phylogenies is rather surprising, especially in light of the problems associated with using morphological characters for reconstructing ant phylogenies. The current taxonomy and systematic relationships of many groups within the Formicidae are considered unreliable, due to a paucity of synapomorphic morphological characters with which to reconstruct phylogenetic relationships (Baroni Urbani *et al.*, 1992; Bolton 1994). Furthermore, there is considerable ambiguity surrounding the phylogenetic relationships of the major ant clades (Bolton, 1994).

In the following section, I provide an introduction to the use of molecular markers in phylogenetic reconstruction, focusing on the particular problems and challenges faced by insect molecular systematists.

Section B: Insect molecular phylogenetics

For the vast majority of lesser-studied taxonomic groups, such as arthropods, the task of selecting loci for phylogenetic analysis is not straightforward, as genomic information is limited (Rokas *et al.*, 2002). Without the availability of detailed information for the group of interest, the common practice is to use one or more gene regions that have already been applied with success for similar taxonomic ranks and closely related groups.

The choice of molecular marker for reconstructing phylogenetic relationships is of the utmost importance, as the properties of the data set may influence the outcome of the analysis to a greater degree than the tree-building algorithm or weighting scheme employed (Simon *et al.*, 1994). Selecting an appropriate gene for phylogenetic analysis requires matching the levels of sequence variation to the desired taxonomic level of study. Consequently, there should be sufficient sequence diversity to resolve taxonomic affinities, but minimal artifacts due to saturation (Graybeal, 1994; Swofford *et al.*, 1996). A further consideration when choosing a marker is to focus on markers that have been broadly applied, in order to facilitate synthesis of phylogenetic relationships across diverse taxa, as well as increase our understanding of molecular evolution through comparative studies (Caterino *et al.*, 2000). This practice is facilitated by the design of conservative primers that recognise their target sequences across broad taxonomic ranks. A suite of insect primer sequences targeting mitochondrial genes were compiled for phylogenetic applications by Simon *et al.*, (1994). However, using genetic markers that have proven phylogenetic utility in other taxa does not guarantee that these markers will display phylogenetic utility for the taxa of interest, due to the arbitrary nature of many taxonomic categories across groups.

In insects, as in vertebrates, mitochondrial DNA has been the marker of choice for reconstructing phylogenies (Simon *et al.*, 1994). The properties that make genes from this genome popular phylogenetic markers include the following: (i) insect mitochondrial DNA is maternally inherited as a single, non-recombining haploid genome (but see Dowton and Campbell, 2001); (ii) it is

present in multiple copies, facilitating DNA extraction and amplification from minute amounts of starting material; (iii) it displays extensive interspecific polymorphism providing a large pool of phylogenetically informative characters and (iv) it appears to evolve at a slightly faster rate than single-copy nuclear DNA with high codon bias (DeSalle *et al.*, 1987; Sharp and Li, 1989; Simon *et al.*, 1994).

Animal mitochondrial DNA occurs as a single, circular molecule of approximately 16 000 bp. This genome encodes 13 protein-coding genes, 22 tRNAs and two rRNA subunits, and contains a non-coding control region containing the origin of replication, referred to as the 'A-T' rich region in insects (Avise, 1994). With the exception of the tRNA genes, the gene order in general seems to be conserved within insects, although this gene order differs to that found in vertebrates (Simon *et al.*, 1994). Mitochondrial DNA-specified proteins generally encode components of the five respiratory complexes required for functioning of the mitochondria, with the organellar translation system involved in translating the mitochondrial proteins partially composed of the mitochondrial large subunit and small subunit ribosomal RNA genes (Gray *et al.*, 1999).

Insect mitochondrial DNA is characterised by extreme A+T richness, with a trend of increasing A+T richness observed from basal to apical orders (Simon *et al.*, 1994; Frati *et al.*, 1997). Hymenopteran mitochondrial DNA in particular exhibits one of the highest proportion of A+T nucleotides of any organism (Dowton and Austin, 1997). The highest A+T content is found in those groups considered to be relatively recently diverged in the hymenopteran phylogeny, namely bees, chalcidoids, scelionids and some endoparasitioid braconids (Whitfield and Cameron, 1998). Honeybee mitochondrial DNA is characterised by an A+T content of 84.9%, compared to the overall A+T content of 78.6% for the mitochondrial genome of *Drosophila yakuba* (Clary and Wolstenholme, 1985; Crozier and Crozier, 1993). The overall composition bias of honeybee mitochondrial proteins is slightly less A+T biased (83.3%) than the entire genome when averaged across all codon positions. However, third codon positions as a class are characterised by extreme A+T bias (92.1%).

The high A+T bias observed in insect mitochondrial DNA has been attributed to directional mutation pressure, which assumes that not all substitutions in the mitochondrial DNA are equiprobable (Jermiin *et al.*, 1994). Various hypotheses have been proposed to explain the occurrence of directional mutation pressure in mitochondria. These include incorporation of specific nucleotides by a defective mitochondrial DNA γ-polymerase, which is prone to oxidative damage (Asakawa *et al.*, 1991; Graziewicz *et al.*, 2002), as well as the highly asymmetrical mode of replication of mitochondrial DNA observed in various organisms. This leaves one strand exposed as a single strand for a longer period of time than the second strand, thereby effectively preventing double-stranded DNA editing by DNA endonucleases (Tamura, 1992; Jermiin *et al.*, 1994).

High levels of base composition bias may be a possible confounding factor when analysing sequences from insect mitochondrial genomes (Dowton and Austin, 1997; Simon *et al.*, 1994). With only two nucleotides essentially available for substitution, convergence will tend to be high. Furthermore, extreme under-abundance of a particular character state will increase the tendency for those sites to saturate prematurely, thereby obscuring phylogenetic signal (Irwin *et al.* 1991; Meyer, 1994). Skewed base composition can also violate the assumptions of parsimony (Perna and Kocher, 1995; Eyre-Walker, 1998), introducing systematic error in the reconstruction of ancestral states (Collins *et al.*, 1994). Furthermore, if the nucleotide bias varies between lineages, taxa may be incorrectly grouped together on the basis of shared nucleotide composition rather than common evolutionary history (Lockhart *et al.*, 1994; Martin, 1995). Phylogenetic reconstruction methods that take these biases into account, such as maximum likelihood, and weighted parsimony analyses, are therefore advocated (Simon *et al.*, 1994; Eyre-Walker, 1998; Rokas *et al.*, 2002).

An additional factor that should be taken into consideration when using mitochondrial DNA sequences amplified by universal primers for phylogenetic reconstruction, is the risk of amplification of nuclear homologues of the mitochondrial gene of interest. These nuclear mitochondrial pseudogenes

(Numts) have been documented in a wide variety of eukaryotes, including insects (reviewed by Bensasson *et al.*, 2001; see Sunnucks and Hale, 1996; Bensasson *et al.*, 2000; Stone *et al.*, 2001; Soucy and Danforth, 2002 for examples found in insects). However, there are no published reports in the literature or on the pseudogene web site (http://www.pseudogene.net) of the presence of any pseudogenes in the Formicidae (D. Bensasson pers. com.). Numts evolve under different evolutionary constraints compared to their mitochondrial counterparts, and may be characterised by frameshift deletion or insertion events (reviewed by Bensasson *et al.*, 2001). Their different evolutionary patterns and paralogous mode of inheritance compared to the authentic mitochondrial gene can lead to incorrect phylogenetic hypotheses when inadvertently included in systematic studies (Zhang and Hewitt, 1996; Sorenson and Quinn, 1998; Bensasson *et al.*, 2001).

To date, the mitochondrial genes of choice for studying phylogenetic relationships in insects have been the cytochrome oxidase I and II subunits (Beckenbach *et al.*, 1993; Crespi *et al.*, 1998; Caterino *et al.*, 2000). The evolution of the cytochrome oxidase I and II subunits have been studied in a diverse array of organisms, and structure-function relationships within the cytochrome oxidase protein complex have been well characterised.

Cytochrome *c* oxidase is a transmembrane complex of proteins composed of three mitochondrially encoded subunits (I to III) and four nuclear-encoded subunits (Frati *et al.*, 1997). The cytochrome oxidase II protein subunit contains two putative transmembrane domains and four copper binding sites, with these sites comprising the signature pattern for this protein due to the absence of phylogenetically conserved domains or sequence motifs. The cytochrome oxidase II gene has been sequenced across a wide variety of insect taxa, with homologous sequences available for nearly all orders, and has shown good utilty in resolving congeneric species relationships (Beckenbach *et al.*, 1993; Brown *et al.*, 1994; Emerson and Wallis, 1995; Spicer, 1995). The highly conserved cytochrome oxidase I gene has also been widely sequenced, although different regions of this large gene have been amplified across taxa complicating comparative studies (Simon *et al.*,

1994; Caterino *et al.*, 2000). Thus the majority of existing ant molecular phylogenies have been based on cytochrome oxidase I and II gene sequences (e.g. Ayala *et al.*, 1996; Chenuil and McKey, 1996; Wetterer *et al.*, 1998; Brady *et al.*, 1999; Gadau *et al.*, 1999; Sameshima *et al.*, 1999; Sauer *et al.*, 2000).

Cytochrome *b*, the vertebrate mitochondrial marker of choice (Meyer, 1994) has been less commonly utilised in insect phylogenetic studies. This is surprising, as this gene has been extensively characterised with regard to structure-function relationships, enhancing its utility for evolutionary investigations. Furthermore, it has provided good resolution for vertebrate relationships at several taxonomic levels (e.g. Graybeal, 1994).

Cytochrome *b* is the only mitochondrially-encoded protein that makes up Complex III of the mitochondrial oxidative phosphorylation system, and is involved in the transfer of electrons from dihydorubiquinone to cytochrome *c*. Variable positions within this protein occur within the eight transmembrane domains and the amino and carboxy terminals of the protein, whereas portions of the inner and outer membrane segments are biologically active and therefore more conserved. Thus this protein is a mosaic of evolutionarily conserved and variable regions, facilitating its use in phylogenetic studies at various taxonomic levels (Irwin *et al.*, 1991; Yoder *et al.*, 1996).

The paucity of cytochrome *b* - based insect phylogenies appears to be due to historical accident rather than lack of phylogenetic utility of this gene (Simmons and Weller, 2001), as it has been successfully utilised to resolve lower-level relationships in ants, wasps, sawflies and moths (Crozier *et al.*, 1995; Stone and Cook, 1998; Nyman *et al.*, 2000; Simmons and Weller, 2001; Parker and Rissing, 2002).

Challenges faced when using mitochondrial markers for reconstructing phylogenies include unequal base frequencies, rate inequalities, third codon position saturation, insufficient variation at replacement sites and the presence of psuedogenes (Yoder *et al.*, 1996; Zhang and Hewitt, 1996).

Furthermore, the peculiar problems associated with insect mitochondrial DNA, such as extreme A+T bias, has resulted in a shift in focus in recent years to utilising nuclear protein-coding genes for phylogenetic inference in insects.

Nuclear genes for which conserved primers are available include elongation factor 1 alpha (EF-1α), dopa decarboxylase (DDC), phosphenolpyruvate carboxykinase, glucose 6 phosphate dehydrogenase and opsins (Cho *et al.*, 1995; Soto-Adames *et al.*, 1994; Friedlander *et al.*, 1998; Mardulyn and Cameron, 1999). Nuclear protein-coding genes have traditionally only been used for high-level systematic studies due to the perception that they are too conserved to be used at lower taxonomic levels. However, the inherent variation in rate of evolution among codon positions has resulted in these typically more conserved genes successfully resolving lower level taxonomic relationships. EF-1α and DDC have proven particularly useful for resolving lower level relationships in the Noctuid moth subfamily Heliothinae and the saturniid moth tribe Attacini (Cho *et al.*, 1995; Friedlander *et al.*, 1998; Fang *et al.*; 1997, 2000). In these lepidopteran taxa, both EF-1α and DDC are present as single copies, display a relatively unbiased nucleotide composition compared to mitochondrial DNA and are not interrupted by introns (Mitchell *et al.*, 1997; Fang *et al.*, 2000). Using nuclear genes for systematic purposes, however, is considered more complex than using mitochondrial DNA due to recombination between alleles, low target copy number for amplification, the presence of paralogous copies, the presence of large introns and lineage sorting (Moore, 1995; Swofford *et al.*, 1996).

Once an appropriate set of markers has been decided upon, and the character data generated, the next step in phylogenetic reconstruction is analysis of the data.

Section C: Phylogenetic inference

The field of phylogenetic inference is well established, and Swofford *et al.* (1996) provide an excellent conceptual framework for understanding the practical and theoretical distinctions among alternative methodologies. These authors provide a comprehensive review of the available analytical techniques, the underlying assumptions of these techniques, as well as recommendations for phylogenetic analyses. Furthermore, many general reviews of the methods for inferring phylogenies are available (e.g. Hillis *et al.*, 1993; Slowinski and Page, 1999; Brocchieri, 2001). Therefore, in this section, I will focus on topical issues in systematic biology that are pertinent to my analysis.

To combine or not to combine

The easy access of most systematists to the tools of molecular biology has resulted in the generation of multiple molecular data sets to infer the phylogeny of interest. The generation of multiple data sets has further been fuelled by an increased awareness that reliance on a single data set may often result in insufficient phylogenetic resolution, or misleading phylogenetic inferences (Cao *et al.*, 1994; Cummings *et al.*, 1995; Baker and DeSalle, 1997; Baker *et al.*, 2001).

A gene tree may not accurately represent the phylogeny of the species from which the gene was sampled for two reasons. Firstly, the gene tree itself may be inaccurately inferred as a result of finite sample size, or inconsistency of tree estimators and the attendant systematic error (Swofford *et al.*, 1996; Slowinski and Page, 1999; Waddell *et al.*, 2000). Secondly, even if the correct gene tree is inferred, it may not accurately reflect the species phylogeny due to ancestral polymorphism, gene duplication, hybridisation or horizontal gene transfer (reviewed by Wendel and Doyle 1998).

The general finding that incongruent topological estimates from various molecular partitions[2] are the rule, rather than the exception (e.g. Baker and DeSalle, 1997; Cunningham 1997a, 1997b; Wiens and Hollingsworth, 2000; Baker *et al.*, 2001) has had two major ramifications. Firstly, it has fuelled the debate in the literature over whether different data sets should be combined or not. Secondly, attempts to understand and reconcile the underlying cause of the observed incongruence have led to increased emphasis being placed on identifying and incorporating the evolutionary dynamics of processes generating the patterns of variation into the analyses.

Two diametrically opposing philosophical schools of thought concerning the analysis of multiple data sets exist. At one extreme is the view that data sets should always be analysed separately and never combined. This is based on the premise that the strongest evidence that the correct phylogeny is being inferred is provided by topological congruence from multiple independent data sets (Miyamoto and Fitch, 1995). At the other end of the spectrum is the total evidence approach, whose advocates insist that data sets should always be combined prior to analysis, in order to maximize the explanatory power and descriptive efficiency of the character data, and thus allow resolution of conflicts by the combined data itself (Miyamoto, 1985; Kluge 1989). Bull *et al.* (1993) advocate a less extreme approach, intermediate to the 'always separate' and 'always combine' approaches known as the 'conditional combination' approach. This 'conditional combination' approach involves assessing the various partitions for significant incongruence, and only if the there is a lack of significant support for incompatible groupings, with any observed incongruence attributable to random error through inadequate sampling, is combined analysis advocated (reviewed by de Queiroz *et al.*, 1995 and Huelsenbeck *et al.*, 1996).

Combining genes that share a common evolutionary history has many potential benefits. The effects of unstructured homoplasy may be minimised in a simultaneous analysis of all available data (Thornton and DeSalle, 2000).

[2] In the context of this thesis the word 'partitions' refers to subsets of characters evolving under demonstrably different sets of rules (Bull *et al.*, 1993).

Furthermore, genes evolving at different rates may interact positively to resolve different levels of a phylogenetic tree (Wilgenbusch and de Queiroz, 2000; Baker *et al.*, 2001). By increasing the number of characters in an analysis, phylogenetic estimation may become more consistent[3] given an appropriate model of substitution (Steel and Penny, 2000). Combining genes may also result in a dramatic increase in resolving power compared with individual analysis of each gene (Buckley *et al.*, 2002). Combined analysis may also reveal novel relationships not discovered in the separate analysis of the different partitions due to hidden character congruence (Baker and DeSalle, 1997; McCracken *et al.*, 1999; Baker *et al.*, 2001). However, a combined phylogenetic analysis makes two fundamental assumptions: (i) that the combined partitions share a common evolutionary history, and (ii) that the chosen method of analysis is appropriate for each of the individual partitions. If a test for homogeneity among data sets fails, this indicates that one or both of these assumptions have been violated (Bull *et al.*, 1993; de Queiroz, 1993).

Testing for incongruence

A number of statistical tests to measure incongruence between data sets or trees produced from them have been proposed (Templeton, 1983; Kishino and Hasegawa, 1989; Rodrigo *et al.*, 1993; Farris *et al.*, 1994; Huelsenbeck and Bull, 1996; Shimodaira and Hasegawa, 1999; Waddell *et al.*, 2000; Buckley *et al.*, 2002). The more routinely applied or statistically robust of these incongruence tests are introduced below.

The most widely applied test of character incongruence is the incongruence length difference (ILD) test of Farris (1994). In essence, this parsimony-based test measures the degree to which homoplasy is increased by combined parsimony analysis over the levels already present in the individual data sets. The ILD test has been shown in some instances to be a significant indicator of phylogenetic accuracy (Cunningham, 1997b). However, criticisms of this test have become increasingly common in the recent phylogenetic literature.

[3] In a phylogenetic context, consistency is defined as the ability of an estimation method to converge to the correct tree as more characters are added (Swofford *et al.*, 1996).

Firstly, the perception of the sensitivity of this test has been questioned by various studies, with recommendations that *P*-values < 0.05 and as low as 0.001 should not preclude data set combination (Sullivan, 1996; Cunningham 1997a, 1997b; DeSalle and Brower, 1997; Mitchell *et al.*, 2000; Darlu and Lecointre, 2002). Secondly, the ILD test measures the increase in homoplasy as a result of combining partitions relative to the levels of homoplasy present within each data set, thereby assuming that as homoplasy increases (resulting in a significant test result), phylogenetic accuracy decreases. However, this has been shown to not necessarily always be the case (Sanderson and Donoghue, 1989; Yoder *et al.*, 2001). Lastly, recent studies have demonstrated that statistically incongruent partitions may contribute phylogenetic signal that emerges in the context of a combined analysis of the partitions, resulting in more robust phylogenetic hypotheses (Remsen and DeSalle, 1998; Yoder *et al.*, 2001).

Generally, it appears that parsimony tests of incongruence are more likely to confuse true topological incongruence with systematic error than probabilistic methods. The reason for this is that it is more difficult to account for potentially misleading features of the substitution process in a parsimony framework than it is in a model-based framework (Huelsenbeck and Bull, 1996; Yoder *et al.*, 2001; Dowton and Austin, 2002; Buckley *et al.*, 2002).

Huelsenbeck and Bull's (1996) maximum likelihood (ML) test of topological incongruence provides a more intuitive and statistically powerful test to evaluate incongruence between topologies produced by different data partitions. It does this by evaluating whether the same phylogenetic tree underlies two data partitions. However, this test is computationally time-consuming because it requires parametric bootstrapping to generate the distribution of the test statistic.

Two additional likelihood-based tests, the Kishino-Hasegawa (KH) test (Kishino and Hasegawa, 1989) and Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa, 1999) may be used to assess whether topological incongruence between data sets is significant (Goldman *et al.*,

2000). The KH test is only valid when comparing *a priori* specified topologies (Swofford *et al.*, 1996; Goldman *et al.*, 2000), whereas the SH test requires *a priori* decisions to be made regarding the number of topologies to be included in the calculation of confidence limits, which can drastically affect the size of the *P*-values obtained (Goldman *et al.*, 2000).

Recently, a promising Bayesian method for assessing phylogenetic incongruence was outlined by Buckley *et al.* (2002) using a similar conceptual approach to that of Huelsenbeck and Bull (1996). This Bayesian approach involves attempting to estimate the uncertainty associated with the topology from one partition, and determining whether the topology from the second partition lies within that region of uncertainty. This Bayesian approach does not require time-consuming parametric bootstrapping, in contrast to the ML-based test of Huelsenbeck and Bull (1996). Furthermore, unlike the non-parametric SH test, no *a priori* topologies need to be specified.

Analysis of heterogeneously evolving data sets

Much of the debate in the earlier sections regarding combination of different data sets is unwarranted if different underlying evolutionary histories are inferred as being responsible for the observed incongruence. Not only does combining data sets with different evolutionary histories violate one of the fundamental assumptions of combined analysis - that the same underlying topology is being reconstructed - it may also obscure interesting biological phenomena (Bull *et al.*, 1993; Mason-Gamer and Kellogg, 1996; Huelsenbeck and Crandall, 1997; Wendel and Doyle, 1998).

How the data should be dealt with if heterogeneity is identified as the cause of incongruence, however, is less obvious, and is illustrated by examples from the literature below. Phylogenetic methods make specific assumptions about the evolutionary processes generating the observed data, and when these assumptions are violated, which may well be the case if heterogeneously-evolving partitions are combined and analysed using a homogenous model of evolution, these methods may be less informative or may positively mislead

phylogenetic estimation (Felsenstein, 1978; Swofford *et al.*, 1996; Huelsenbeck and Crandall, 1997).

Bull *et al.* (1993) demonstrated that when combining simulated data sets generated from the same topology but evolving at different rates, the combined estimate of phylogeny was less accurate than when the slower evolving data set was analysed individually. Consequently, Bull *et al.* (1993) advocate that heterogeneously evolving data sets should not be combined. However, the Bull *et al.* (1993) study has been criticised on the grounds that unrealistic models of evolution for the two partitions were assumed, specifically with regard to how rate heterogeneity was modelled, with a uniformly high rate of evolution assumed for one partition, and a uniformly low rate of evolution for the other partition (Sullivan, 1996). Most real data sets have been shown to contain at least some degree of among-site rate variation, with evolutionary rates continuously distributed across sites (Yang *et al.*, 1994; Sullivan *et al.*, 1995; Rokas *et al.*, 2002). A study on sigmodontine rodents using 12S and cytochrome *b* sequences by Sullivan *et al.* (1995), demonstrated that these data partitions could be evolving heterogeneously, as combination was strongly rejected based on three different tests. However, when these gene sequences were combined, a well-supported phylogeny in which all well-corroborated relationships were recovered was inferred. Sullivan *et al.* (1995) recommend that provided data partitions share a common evolutionary history and there is some overlap in the distribution of evolutionary rates between the partitions, the potential will exist for phylogenetic signal in each of the partitions to be additive in the context of a combined analysis. A phylogenetic study of New Zealand cicada genera using five genes by Buckley *et al.* (2002), found that although each of their data partitions appeared to have evolved under the same underlying topology, combining the heterogenous data sets and analysing them assuming a single model of evolution for all partitions obscured underlying incongruence, and did not result in more strongly supported estimates of phylogeny.

Simultaneous partitioned analysis

In some cases, the phylogenetic estimate obtained from combined analysis of heterogeneously evolving data sets may be improved by applying differential weighting to accommodate different underlying evolutionary processes, or by specifying a broader model of evolution that is appropriate for all partitions (Chippindale and Wiens, 1994; Swofford *et al.*, 1996; Reed and Sperling, 1999). However, differential weighting in the context of parsimony lacks an explicit basis for judging which weighting scheme provides the best explanation for the data (Huelsenbeck *et al.*, 1994). In contrast, likelihood has an explicit basis for assessing the ability of different models to explain the data because likelihood scores are comparable across models (Swofford *et al.*, 1996). However, even in an explicit model-based framework, attempts to characterise all the available data using a single set of model parameters have not been found to improve phylogenetic estimates from heterogeneously evolving combined partitions (Wilgenbusch and De Queiroz, 2000; Buckley *et al.*, 2002). An attractive solution to this problem would be to specify a model of evolution for each partition that best captures the evolutionary parameters of that particular partition (Lio and Goldman, 1998; Caterino *et al.*, 2001). Unfortunately, current implementations of ML and Bayesian methods do not allow for sophisticated searches of parameter space with partitioned models (Waits *et al.*, 1999; Yang 1996a; Buckley *et al.*, 2002), although these abilities are currently being incorporated into phylogenetic analysis software such as PAUP (Wilgenbusch and De Queiroz, 2000). This promises to be an exciting future area of research in phylogenetic analysis of multiple molecular data sets.

Molecular versus morphological data

The debate about which type of data - molecular or morphological - is intrinsically superior for phylogeny reconstruction, based on the observation that these two data types rarely produced congruent estimates of phylogeny, has died down in recent years. The advantages and disadvantages of both these types of data are well characterised (reviewed by Hillis, 1987), and

studies that utilise both sources of data are likely to provide better descriptions and understanding of biological diversity than those that focus on one type of data (Moritz and Hillis, 1996; McCracken *et al.*, 1999). Combined analysis of molecular and morphological data has been criticised on the grounds that the morphological characters may be swamped by the more numerous molecular characters (Miyamoto, 1985; Hillis, 1987). However, studies have shown that morphological characters may make a significant positive contribution towards phylogeny reconstruction in a combined molecular-morphological context, even when the morphological parsimony-informative characters are heavily outnumbered by molecular characters (Baker *et al.*, 2001; Weiblen, 2001). Morphological and molecular characters may also complement each other by resolving relationships at different hierarchical levels (Pennington, 1996, but see Hedges and Maxson, 1996).

Combined analysis of molecular and morphological data has, until recently, only been feasible in a parsimony framework. However, Markov models for the modelling of morphological evolution have recently been developed (Lewis, 2001), thereby facilitating combined phylogenetic estimation of molecular and morphological characters in a ML framework. These methods have as yet not been implemented in phylogenetic software packages like PAUP* (Swofford, 1998).

Combination and congruence in practice

Most systematists follow the recommendation of Swofford (1991) that a multifaceted approach to phylogeny reconstruction should be taken, in which partitions are analysed both separately and in combination. Partitioned analysis of data can provide insight into the evolutionary process generating the observed variation (Huelsenbeck *et al.*, 1996), which then in turn can be taken into consideration in the context of a combined analysis in order to improve phylogenetic estimation (Huelsenbeck *et al.*, 1994; Chippindale and Weins, 1994). It is generally accepted that statistical tests of incongruence do not necessarily provide a definitive answer as to whether data should be analysed in combination or not, and should rather be used as a vehicle for

exploring the nature of the various data partitions, and how these partitions are likely to interact (Johnson and Soltis, 1998; McCracken *et al.*, 1999).

Assessing phylogenetic utility

The increase in quantity of DNA character information from a variety of genes over the past decade has led to an increased emphasis on discriminating among different types of character information, and incorporating these differences into the tree reconstruction process (Chippindale and Weins, 1994; Collins *et al.*, 1994, Yang, 1996a; Baker *et al.*, 2001). This follows on from the principle that some characters are less reliable than others, and may mislead phylogenetic inference unless these less reliable characters are recognised as such and accounted for. In practice the reliability of systematic data is usually assessed directly from the data at hand. Methods for quantifying utility are varied, and limited only to the ingenuity of the researcher, though common methods for nucleotide sequence data include (i) examining properties of the data prior to phylogenetic analyses e.g. base composition bias, transition/transversion ratios, saturation curves, divergence values (Graybeal, 1994; Friedlander *et al.*, 1998); (ii) examining properties of the data specific to a given phylogenetic analysis e.g. resolution, homoplasy and support measures; (iii) estimating the pattern of rate heterogeneity for the different partitions (Yang, 1998); (iv) assessing congruence of comparable data types (Johnson and Soltis, 1998) and (v) assessing topological congruence with a 'known' or strongly-supported phylogeny (Cho *et al.*, 1995; Spicer, 1995; Mitchell *et al.*, 1997; Mardulyn and Cameron, 1999).

Model-based optimality criteria

Model based approaches to phylogenetic inference have increasingly dominated phylogenetic methodology over the last decade (Steel and Penny, 2000). The preference for one optimality criterion[4] over another can be considered a philosophical decision, as there is no way of determining with

[4] An optimality criterion is a criterion that defines how well data fit a particular hypothesis. In the context of phylogenetic reconstruction this equates to how well the data fit a particular phylogenetic tree (Swofford *et al.*, 1996).

absolute certainty the effectiveness of any criterion at recovering the true phylogeny (Lewis, 1998). However, a vast body of literature exists exploring and comparing the relative performance of various optimality criteria in recovering known simulated phylogenies or artificially-generated phylogenies (e.g. Bull *et al.*, 1993; Hillis, 1995). Within this context, ML methods have been found to consistently outperform the other optimality criteria when data have been simulated under more complex conditions (Gaut and Lewis, 1995; Huelsenbeck, 1995a, 1995b; Yang, 1996a), thereby making it the most popular model-based criterion for analysis of real sequence data.

For purposes of characterising the underlying process of molecular evolution, one of the greatest advantages of using a likelihood framework in which to analyse sequence data is undoubtably the ability to allow the character data to falsify the evolutionary model (Goldman, 1993). Various models of nucleotide substitution can be statistically compared using likelihood ratio tests to evaluate whether a more complex, parameter-rich model provides a significantly better fit to the data than a simpler model (Huelsenbeck and Crandall, 1997). Although a wealth of models are currently available for modelling sequence evolution (reviewed by Lio and Goldman, 1998), any model of nucleotide substitution is necessarily a simplification of the actual evolutionary process generating the observed variation. Thus it is unlikely that any single model will be able to accommodate every observed nuance in the substitution processes that have generated any particular set of sequences (Posada and Crandall, 2001; Sullivan and Swofford, 2001). However, an advantage of a statistical based method of inference, such as ML, is that a perfect model is not a pre-requisite for obtaining the correct phylogeny (Yang *et al.*, 1994; Sullivan and Swofford, 2001), with ML shown to be particularly robust to model assumption violations (Huelsenbeck, 1995a; Sullivan and Swofford, 2001). Furthermore, it is well documented that more realistic models of nucleotide substitution lead to more accurate phylogenetic estimates of species relationships, as well as a better understanding of the underlying forces and mechanisms resulting in the pattern of variation observed in the sequences (Yang *et al.*, 1994; Swofford *et al.*, 1996; Huelsenbeck and Rannala, 1997; Lewis, 1998).

By incorporating more realistic models of sequence evolution into the phylogenetic reconstruction process, the result is an improved understanding of the biological processes shaping evolution at the molecular level, as well as an improved ability to infer from sequence data the history of life (Lio and Goldman, 1998).

Bayesian approaches to phylogenetic reconstruction

One of the most significant new developments in phylogenetics in the past decade is the application of Bayesian statistical methods to phylogenetic estimation (Huelsenbeck and Ronquist, 2001; Lewis, 2001). The Bayesian paradigm offers a unique perspective to the phylogeny problem. All other methods of phylogenetic inference, namely maximum likelihood (ML), maximum parsimony (MP) and minimum evolution (ME), attempt to optimise an objective function. ML estimates of phylogeny involve maximizing the probability of observing the data given a model of nucleotide substitution, topology and branch lengths. Both MP and ME methods attempt to minimise the number of reconstructed character transformations on a tree (Swofford *et al.*, 1996). However, the objective of Bayesian inference is to calculate the posterior probabilities of trees based on the joint probabilities of the tree branch lengths and the model of substitution (Huelsenbeck and Bollback, 2001). The fundamental distinction between ML and Bayesian inference in particular is that a Bayesian approach provides probabilities for a hypothesis given the data, rather than the probability of the data given the hypothesis (Lewis, 2001).

The central idea in Bayesian analysis is that the parameters of the statistical model should be treated as random variables, with inferences about model parameter values based on the probability of the parameter conditional on the observations. This conditional or posterior probability can be calculated by combining the likelihood function with a prior using Bayes theorem[5] (Huelsenbeck and Bollback, 2001). Bayesian phylogenetic inference involves calculating the joint posterior probability distribution of all parameters of the

[5] $Pr[Tree \mid Data] = \frac{Pr[Data \mid Tree] \times Pr[Tree]}{Pr[Data]}$ where the vertical bar should be read as "given".

phylogenetic model (topology, branch lengths, nucleotide substitution parameters), with the inference of any single phylogenetic parameter then based on the marginal distribution of that parameter. This allows estimation of the uncertainty associated with any parameter from the phylogenetic model.

Evaluating the posterior probability of a phylogenetic tree requires evaluation of a sum over all possible trees and integration over the space of branch lengths and substitution model parameters, which is analytically intractable for even small phylogeny problems (Larget and Simon, 1999). The breakthrough that solved the computational aspects of Bayesian approaches to phylogenetic reconstruction was the development of novel Markov Chain Monte Carlo (MCMC) algorithms (Rannala and Yang, 1996; Mau and Newton, 1997; Yang and Rannala, 1997; Larget and Simon, 1999; Mau *et al.*, 1999). MCMC is a sampling technique that efficiently evaluates the posterior probability of the tree by concentrating its sampling effort on drawing samples from the distribution of interest.

The end result of a Bayesian analysis is the joint posterior probability of topologies, branch lengths and substitution parameters. The posterior probability of any clade is simply the sum of posterior probabilities of all trees that contain that clade. This makes the posterior probability of a clade a more intuitive measure of support than bootstrap *P*-values, by virtue of being a more direct estimate of confidence in a hypothesis (Buckley *et al.*, 2002).

Bayesian phylogenetic inference is rapidly increasing in popularity amongst systematists due to the ease with which complex models of evolution can be incorporated into the analysis, the computational efficiency of the analysis, and the ability to accommodate phylogenetic uncertainty (e.g. Buckley *et al.*, 2002; Jow *et al.*, 2002; Whittingham *et al.*, 2002; Wilcox *et al.*, 2002).

Research Objectives of Part I

In light of the poor state of *Camponotus* systematics, and the paucity of ant molecular phylogenetic studies, the first molecular phylogenetic study of representative African *Camponotus* species was undertaken in order to:

(i) evaluate the comparative utility of cytochrome *b* and cytochrome oxidase II for resolving species level relationships in *Camponotus*, using a variety of methodological approaches
(ii) assess the validity of current morphological-based subgeneric classifications using molecular data
(iii) test the hypothesis of monophyly of the *Camponotus fulvopilosus* species group
(iv) determine the phylogenetic affinity of the recently rediscovered species *Camponotus bifossus*
(v) to test the hypothesis of *Camponotus* monophyly

Furthermore, an additional objective of this thesis was to identify and score putative phylogenetically informative morphological characters, in order to obtain an independent estimate of phylogeny and increase the total number of characters available with which to evaluate phylogenetic hypotheses.

Organisation of this thesis

This thesis is divided into three parts, and comprises six chapters in total. Part I consists of Chapters 2 to 5, where Chapter 2 details the Materials and Methods section for phylogenetic analyses. Chapter 3 reports the methodological approaches used in the phylogenetic analysis of molecular and morphological data generated in this study. The results and discussion of these analyses are presented in Chapter 4, which is sub-divided into five separate sections due to the diverse range of analyses conducted. Part A deals with the evolutionary characteristics of the two genes. Parts B and C address phylogenetic analyses of the single and combined molecular data sets respectively. Part D presents the phylogenetic analyses of morphological and behavioural characters, as well as that of the combined molecular-morphological data sets. Lastly, Part E presents the results of hypotheses testing. Chapter 5 is a synthesis of the phylogenetic results, and addresses the research objectives outlined in Chapter 1 Part I. Part II of this thesis is dealt with in Chapter 6, containing the Introduction, Research objectives, Materials and Methods, Results and Discussion of the microsatellite analysis of fine scale genetic structuring in *Camponotus klugii*. Finally, Part III contains the concluding remarks, and a record of all the scientific literature cited within this thesis.

PART I

Chapter 2

GENERATION OF CHARACTER DATA

In this chapter, the various methodological procedures and techniques used to obtain molecular and morphological character data for subsequent phylogenetic analyses are outlined.

I DNA sequence data

Sampling

Five to ten major workers of each ant species were obtained from the wet ant collection of the South African Museum, Cape Town (Table 2.1). All specimens were preserved in 70% or 96% ethanol with the exception of *Camponotus klugii* workers. Individuals from this species were collected from Silvermine Nature Reserve, Cape Peninsula, and frozen at -20°C for detailed colony genetic analyses (details are provided in Chapter 6: Materials and Methods).

Nucleic acid extraction and quantification

(i) Nucleic acid extraction

Total genomic DNA was extracted from at least two workers of each species using a modified CTAB extraction protocol (Villesen *et al.*, 1999). Prior to extraction, individual ants were soaked for two hours (with agitation) in a phosphate-buffered saline solution to ensure complete removal of ethanol, and then dried for ten minutes at room temperature. The abdomen and head of each individual were removed with a razor blade, and only the thorax and legs used for DNA extraction. The head region of certain insect taxa (including Hymenoptera) have been reported to contain compounds that when extracted act as inhibitors in the polymerase chain reaction (PCR) (Dowton and Austin, 1998). The abdomen was discarded in order to minimize possible contamination by endosymbiotic enteric bacteria (Tek Tay *et al.*, 1997). Ants

were pulverized with an eppendorf pestle in a 1.5ml sterile eppendorf tube after the addition of liquid nitrogen. Samples were incubated at 60°C for two hours after the addition of 350µl hexadecyltrimethylammonium bromide (CTAB) extraction buffer (1% hexadecyltrimethyl ammonium bromide, 0.75M NaCl, 50mM Tris pH 8.0, 10mM EDTA) and proteinase K to a final concentration of 50µg/ml. DNA was purified using a chloroform-isoamyl extraction followed by ethanol precipitation (Sambrook *et al.*, 1989). Pellets were air-dried, resuspended in 75µl sterile distilled water and stored at -20°C. Blank extractions with no tissue were included in order to control for any possible contaminants introduced during the extraction procedure.

(ii) DNA Quantification

The concentration of extracted DNA was assessed using spectrophotometry. A diode-array spectrophotometer was used to take absorbance readings of 1 in 100 dilutions in water of each sample at wavelengths of 260nm and 280nm. The ratio of the two absorbance readings (λ260: λ280) was calculated to obtain an estimate of the purity of the extractions: pure extractions of DNA free of protein contamination should have λ260: λ280 ratios of ≥ 1.8 (Sauer *et al.*, 1998). The concentration of DNA (in ng/µl) was calculated using the following equation:

Concentration DNA (ng/µl) = λ260nm X 50 X dilution factor

The concentration of extracted DNA ranged from 90ng/µl to 200ng/µl. Samples were diluted to a working stock of 50ng/µl and stored at -4°C in order to minimise possible contamination of stock extractions, and to avoid repeated cycles of freezing/thawing which may degrade DNA (Lahiri and Schnabel, 1993) and lead to precipitates (Qiagen Bench Guide, 2001).

Table 2.1. Details of samples examined in this study. Numbers following species names were arbitrarily given in this study to designate unclassified species within a subgenus (according to Bolton, 1995).

Species Name	*Subgenus*	Date Collected	Sampling Locality	Accession No. Prefix SAM-HYM-
Camponotus sp.24	*Tanaemyrmex*	7 Dec 1995	Tanzania, Mkomazi Game Reserve	C008575
Camponotus sp.2	*Tanaemyrmex*	27 Feb 1999	South Africa, Western Cape, Anysberg Nature Reserve	C013023
Camponotus sp.11	*Tanaemyrmex*	27 Feb 1999	South Africa, Western Cape, Anysberg Nature Reserve	C013020
C. acvapimensis	*Tanaemyrmex*	30 Nov 1995	Tanzania, Mkomazi Game Reserve	C008589
C. detritus	*Myrmopiromis*	9 Mar 1988	Namibia, Namib Naukluft Park	C001035
C. brevisetosus	*Myrmopiromis*	30 Oct 1991	South Africa, Kwazulu-Natal, Vernon Crookes Nature Reserve	C006709
C. storeatus	*Myrmopiromis*	15 Sept 1995	South Africa, Western Cape, Kogmanskloof	C017759
C. fulvopilosus	*Myrmopiromis*	22 Feb 1999	South Africa, Western Cape, Anysberg Nature Reserve	C013038
C. rufoglaucus	*Myrmosericus*	27 Feb 1999	South Africa, Western Cape, Anysberg Nature Reserve	C013019
C. cinctellus	*Myrmosericus*	30 Nov 1991	Tanzania, Mkomazi Game Reserve	C008571
C. cuneiscapus	*Myrmespera*	17 Oct 1963	South Africa, Northern Cape, Sandwerf, 24 miles from Calvinia	C008487
Camponotus sp.19	*Myrmespera*	7 Mar 1995	Namibia, Gross Spitzkoppe	C007980
Camponotus sp. 10	*Myrmespera*	9 Dec 1989	Tanzania, Mkomazi Game Reserve	C008582
C. nasutus	*Myrmespera*	10 Feb 1996	Namibia, Haasenhof Guest Farm	C0011434

Table 2.1. continued

Species Name	*Subgenus*	Date Collected	Sampling Locality	Accession No. Prefix SAM-HYM-
Camponotus sp.7	*Myrmotrema*	28 Feb 1999	South Africa, Western Cape, near Cogman's Kloof	C013045
*Camponotus sp.*12	*Myrmotrema*	16 May 1996	Tanzania, Gonja Forest Reserve, Western Cape	C009538
C. mystaceus	*Myrmopsamma*	11 Aug 1989	South Africa, Western Cape, Kromrivier Farm, Cedarberg	C001580
*Camponotus sp.*14	*Myrmopsamma*	29 Mar 1996	South Africa, Western Cape, Kleinmond Nature Reserve	C008798
C. sericeus	*Orthonotomyrmex*	26 Nov 1995	Tanzania, Mkomazi Game Reserve	C008581
C. chrysurus	*Myrmopelta*	26 Nov 1995	Tanzania, Mkomazi Game Reserve	C008569
C. bifossus	*Colobopsis*	5 Oct 2001	South Africa, Western Cape, Kleinmond	C017760
C. klugii	*Myrmamblys*	14 May 1998	South Africa, Silvermine Nature Reserve	C015003
Polyrhachis schistacea	n.a.	10 Dec 1995	Tanzania, Mkomazi Game Reserve	C008596

Primer selection

Universal insect mitochondrial primers were screened in *Camponotus* for their ability to bind and amplify the cytochrome oxidase II and cytochrome *b* genes (Simon *et al.*, 1994) (Table 2.2). One novel primer designed for this study, as well as primers taken from other studies were given standardised names based on the nomenclature guidelines of Simon *et al.* (1994) using a trinomial system of 'gene name - DNA strand - nucleotide position of the 3' base'. The 3' nucleotide position of all primers with non-standardised names was determined by performing pairwise sequence alignments between the primer sequences and the entire mitochondrial genome sequence of *Drosophila yakuba* (Genbank Accession Number NC_001322).

The sequence of the novel cytochrome oxidase II primer is as follows:

C2-J-3156: 5' GAT TTA ATA ATT TTT TTT CAT GA 3'

Table 2.2. Primer pair combinations screened for amplification in *Camponotus*.

Primers	Gene segment amplified	Source
F: TL2-J-3037 (A-tLEU) R: C2-N-3665 (H3665)	Cytochrome *c* oxidase subunit II	F: Simon *et al.*,1994 R: Chiotis *et al.*, 2000
F: C2-J-3033 (L3034) R: C2-N-3665 (H3665)	Cytochrome *c* oxidase subunit II	F: pers. com. R. Johnson R: Chiotis *et al.*, 2000
F: C2-J-3156 (GECOIIF) R: C2-N-3665 (H3665)	Cytochrome *c* oxidase subunit II	F: This study R: Chiotis *et al.*, 2000
F: TL2-J-3037 (A-tLEU) R: A8-N-3914	Cytochrome *c* oxidase subunit II	F: Simon *et al.*,1994 R: Simon *et al.*,1994
F: TL2-J-3037 (A-tLEU) R: C2-N-3389 (Marilyn)	Cytochrome *c* oxidase subunit II	F: Simon *et al.*,1994 R: Simon *et al.*,1994
F: CB-J-10933 (CB1) R: CB-N-11367 (CB2)	Cytochrome *b*	F: Simon *et al.*,1994 R: Simon *et al.*,1994
F:CB-J-10602 (CB7) R: CB-N-11367 (CB2)	Cytochrome *b*	F: Tek Tay *et al.*, 1997 R: Simon *et al.*,1994
F: CB-J-10602 (CB7) R:TS1-N-11683 bee (TRs)	Cytochrome *b*	F: Tek Tay *et al.*, 1997 R: Simon *et al.*, 1994
F: CB-J-10602 (CB7) R: TS2-N-11687 (tRs2)	Cytochrome *b*	F: Tek Tay *et al.*, 1997 R: Chiotis *et al.*, 2000
F: CB-J-10933 (CB1) R:TS1-N-11683 bee (TRs)	Cytochrome *b*	F: Simon *et al.*, 1994 R: Simon *et al.*, 1994
F: CB-J-10933 (CB1) R: TS2-N-11687 (tRs2)	Cytochrome *b*	F: Simon *et al.*, 1994 R: Chiotis *et al.*, 2000
F: N4-J-8944 (ND4) R: N1-N-12595 (ND1)	NADH dehydrogenase subunit IV & VI, cytochrome *b*, NADH dehydrogenase subunit I	F: Simon *et al.*,1994 R: Simon *et al.*,1994

J = majority strand (majority of genes transcribed off this strand in insects); N = minority strand; number following J or N refers to nucleotide position of the 3' base of primers with respect to the *Drosophila yakuba* mitochondrial genome (Clary & Wolstenhome, 1985). C2 = Cytochrome *c* oxidase subunit II, CB = Cytochrome *b*, N1 = NADH dehydrogenase (ND) subunit 1, N4 = ND subunit 4; A8 = ATP8; TL2 = tRNA leucine; TS1 = tRNA serine, TS2 = tRNA serine. F = forward primer, R = reverse primer. Names in parentheses are the common names of the primers provided by the various authors.

DNA amplification

(i) Primer Screening

Primers were synthesised by the DNA Synthesis Laboratory at the Department of Molecular and Cellular Biology (UCT). Aliquots were prepared by diluting the primer stock in sterile distilled water pH 8.0 to a concentration of 50μM. Primer aliquots were stored at -20°C. Primers were screened for their ability to bind to and amplify *Camponotus* mitochondrial DNA by performing PCRs using low stringency conditions. Amplification conditions were as follows: an initial denaturation step at 94°C for 3 minutes, followed by 35 cycles at 94°C for 30 seconds, annealing at 40°C or 45°C for 45 seconds followed by an extension step at 72°C for 1 minute. This was followed by an extension step of 5 minutes at 72°C to ensure extension of all incomplete length PCR products. Cycling was performed on a Hybaid thermocycler. PCRs were performed in 0.5ml thin-walled eppendorfs (Whitehead, S.A.) using 20μl reaction volumes overlayed with mineral oil. The reaction mixture consisted of the following reagents: each dNTP at a final concentration of 0.2mM, BIOTAQ DNA polymerase (Whitehead Scientific, S.A.) at a final concentration of 0.02U/μl, and 0.2μM of forward and reverse primers. Magnesium-free 10X reaction buffer (Whitehead Scientific, S.A.) was added to a 1x final concentration (16mM $[NH_4]_2SO_4$; 67mM Tris-HCl (pH 8.8); 0.01% Tween-20). Four different $MgCl_2$ titrations (1mM, 2mM, 3mM, 4mM) were tested. Approximately 100ng of template DNA was used per reaction. A PCR 'blank' tube where the template DNA was replaced by sterile distilled water was included in each experiment to control for possible PCR contamination of reagents. The success of amplification reactions was checked by agarose gel electrophoresis.

(ii) Agarose gel electrophoresis

Ten microlitres of each PCR reaction was mixed with Type III loading dye (Sambrook *et al.*, 1989) and loaded onto a 1.5% agarose (Type D1-LE, Whitehead Scientific, S.A.) gel with ethidium bromide added to a final concentration of 0.5μg/ml to enable visualization of PCR product. Gels were electrophoresed at 70 V in 0.5 X TAE. Lambda DNA (AEC Amersham, S.A.) digested with *Pst* I (Roche, S.A.) was used as a molecular size marker to estimate PCR fragment size.

(iii) Primer optimisation

Primer pairs that amplified product of the expected size in at least 80% of the species screened were optimised by adjusting the annealing temperature, temperature cycling parameters and Mg^{2+} concentration (see Table 2.3 for final optimised conditions). Reaction volumes were increased to 50μl to facilitate purification for sequencing reactions, and overlayed with mineral oil. The final concentrations of PCR reaction buffer, dNTPs, primer and template were the same as detailed above. Five microlitres of each reaction was checked by gel electrophoresis. Negative controls were included in each experiment.

Table 2.3. Optimised magnesium concentrations and cycling profiles for primer pairs. Thermal cycling was preceded by a 3 minute denaturation step at 94°C and followed by a 5 minute extension step at 72°C

Primer Pair	Mg^{2+} concentration in reaction (mM)	Number of cycles and optimised cycling parameters
A-tLeu-H3665	2.0	5 cycles: 94°C 30 seconds, 40°C 45 seconds, 72°C 30 seconds 30 cycles: 94°C 30 seconds, 45°C 45 seconds, 72°C 30 seconds
L3032-H3665	2.5	5 cycles: 94°C 30 seconds, 40°C 45 seconds,72°C 30 seconds 30 cycles: 94°C 30 seconds, 45°C 45 seconds, 72°C 30 seconds
GECOIIF-H3665	2.5	35 cycles: 94°C 30 seconds, 50°C 45 seconds, 72°C 30 seconds
CB7-CB2	1.5	35 cycles: 94°C 30 seconds, 52°C 45 seconds, 72°C 1 minute 30 seconds
CB7-tRs2	2.5	35 cycles: 94°C 30 seconds, 44°C 45 seconds, 72°C 1 minute 30 seconds

(iv) Reamplification

Samples that yielded faint bands after extensive optimisation were subjected to the following re-amplification procedure: 40µl of the PCR reaction were loaded onto a 1.5% agarose gel and the product band excised under long wavelength UVA light (λ = 365 nm) with a sterile razor blade. The gel slice was centrifuged through a 1000µl plugged filter tip (Whitehead Scientific, S.A.) for 5 minutes at 13 000 rpm. One microlitre of the eluant was then used as template and subjected to 15 to 20 cycles of PCR under optimised conditions specific to that primer pair. A gel slice corresponding to the PCR 'blank' reaction, as well as a gel slice excised from an unused lane were subjected to the same procedure as described above in order to ensure the absence of contamination.

Nucleic acid sequencing

All sequences in this study were generated by cycle sequencing using the ABI Prism® BigDye™ Terminator Cycle Sequencing Ready Reaction kit v1.0 (Applied Biosystems, Perkin Elmer S.A.) with modifications (see (ii) below). Sequencing was either performed directly on the PCR product, or the PCR fragment was first cloned and subsequently sequenced. Primers used for generating PCR product were also used as sequencing primers. At least two individuals from each species were sequenced to check for intra-specific variation in the gene fragment sequences.

(i) Purification of template for cycle sequencing

Excess primers, contaminating salts and dNTPs that could potentially interfere with cycle sequencing were removed by loading 45µl of PCR product onto Qiaquick® PCR purification columns (Southern Cross Biotechnology, S.A.) and following manufacturer's instructions (Qiaquick® Spin Handbook). Purified PCR product was eluted in 40µl sterile distilled water pH 8.0. The concentration of eluted PCR product was estimated by running 2µl of the

Qiagen-purified product on a 1.5% agarose gel and comparing the intensity of the band to pGEM®f(+) double-stranded DNA Control Template (ABI Prism® BigDye™ Terminator Cycle Sequencing Ready Reaction kit v1.0, Applied Biosystems) loaded on the gel at 20ng, 40ng, 60ng and 80ng.

(ii) Cycle Sequencing

Cycle sequencing reactions were performed in 10µl reaction volumes using 0.2ml thin-walled PCR tubes (Whitehead Scientific, S.A.). Each reaction contained 3.2pmoles of forward or reverse primer, 20 to 50ng of template PCR product depending on the length of the fragment to be sequenced, and Terminator Ready Reaction Premix at 2X concentration (1/4 dilution). Forward and reverse sequences were obtained for each PCR product. Thermal cycling was performed on a Hybaid PCR Sprint with cycling parameters as recommended for the DNA Thermal Cycler 480 (ABI Prism® BigDye™ Cycle Sequencing Ready Reaction Protocol manual). Salts and unincorporated dye terminators were removed from sequencing reactions by ethanol/sodium acetate precipitation or G50 Sephadex column purification (ABI Prism® BigDye™ Cycle Sequencing Ready Reaction Protocol manual). Sequenced products were electrophoresed on a MegaBASE 1000 DNA Sequencer (Amersham-Pharmacia Biotech) or an ABI Prism 3100 Genetic Analyzer (Applied Biosystems, U.S.A.).

Cloning of PCR products

PCR products that did not sequence successfully using the direct sequencing approach were cloned using the pGEM®-T Easy TA cloning system (Promega, S.A.) with modifications. Ligation reactions were performed in 5µl reaction volumes with the following reagents: 2.5µl 2 X Rapid Ligation Buffer, 25ng pGEM®-T Easy Vector, 0.5µl T4 DNA ligase (1.5U), Qiagen-purified PCR template and water to volume. A 3:1 insert:vector ratio was used as recommended (Promega Technical Manual Part # TM042). Ligation reactions were incubated overnight at 4°C to ensure the maximal number of

transformants. One microlitre of the ligation reaction was added to 25µl of JM109 High Efficiency Competent Cells and incubated on ice for 20 minutes. The competent cells were then transformed by heat-shocking them for 50 seconds in a waterbath at 42°C. The cells were placed on ice for 2 minutes and recovered by adding 450µl SOC medium to the tubes and incubating them for 1.5 hours at 37°C with gentle shaking (~150 rpm). Fifty microlitres of transformation culture was spread on LB/ampicillin/IPTG/X-Gal plates which were incubated overnight at 37 °C.

Typical transformation efficiencies calculated using the JM109 competent cells were around 1 X 10^8 cfu/µg DNA. Recombinant clones and one vector control were transferred to 2 ml eppendorfs containing 200µl luria broth with ampicillin (50µg/ml) and incubated with shaking at 37°C for 3 hours. One microlitre of a 1/10 dilution of each culture was used as template for a PCR reaction using the pUC/M13-40F (5' GTTTTCCCAGTCACGAC 3') and pUC/M13R (5' CAGGAAACAGCTATGAC 3') primers. Fifty microlitre reaction volumes contained final concentrations of the following reagents: 1µM of each primer, 0.2mM of each dNTP, 0.02U/µl BIOLINE *Taq* polymerase (Bioline, Whitehead Scientific, S.A.), 1.5mM $MgCl_2$ and 1X magnesium-free reaction buffer (final concentration: 16mM $[NH_4]_2SO_4$; 67mM Tris-HCl (pH 8.8); 0.01% Tween-20). Thermal cycling was performed on a PCR Hybaid Sprint with the following temperature profile: 94°C for 5 minutes, followed by 35 cycles at 94°C for 30 seconds, 50°C for 30 seconds and 72°C for 45 seconds. This was followed by an extension step at 72°C for 5 minutes. Five microlitres of PCR cocktail was electrophoresed on a 1.5% agarose gel The presence of an insert was assessed by comparing the size of the PCR product from a recombinant clone with that of a non-recombinant clone containing vector but no insert. If the vector did contain insert, the remaining 45µl of PCR product was purified using Qiagen PCR purification columns as described previously. The M13 primers used in PCR amplification were also used as sequencing primers in separate reactions. Vector sequences were edited out from the

chromatogram by using Chromas 2.21 (Technelysium Pty Ltd) before exporting the sequences in FASTA format[1].

Sequence editing and alignment

Sequence chromatograms were manually edited in Chromas 2.21 (Technelysium (Pty) Ltd). Nucleotides that showed irregular spacing or had competing background peaks were coded with lowercase letters. If confirmed by the complementary strand, the letters were converted to uppercase. If the strands contained contradictory nucleotides an IUPAC ambiguity code was used. Edited sequences were exported in FASTA format.

To confirm the identities of the sequenced fragments, sequences were checked against those in the National Centre for Biotechnology Information (NCBI) nucleotide databases using BLAST® (Basic Local Alignment Search Tool) searches. Specifically, a translated Blast Search (Blastx) was implemented using the invertebrate mitochondrial translation table (Appendix II). A Blastx search was used, as even after only small amounts of evolutionary change between sequences, simple nucleotide substitution scores contain less information to deduce homology than do encoded protein sequences (States *et al.*, 1991).

Once the identities of the sequences were confirmed as ant mitochondrial cytochrome *b* or cytochrome oxidase gene fragments, alignment of homologous sequences was performed using Clustal X (Thompson *et al.*, 1997). Default settings were used for the alignment of both the cytochrome *b* and cytochrome *c* oxidase gene fragments. Both cytochrome *b* and cytochrome oxidase II are protein-coding genes, therefore no indels or gaps that could potentially complicate the alignment were expected. Aligned sequences were imported into MacClade 4.0 (Maddison and Maddison, 2000)

[1] A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column.

and sequences colour-coded by amino acids (invertebrate mitochondrial code Translation Table V; Appendix II) to further proofread the edited sequences. All nucleotides dictating amino acid changes were confirmed by rechecking the original chromatograms. Missing data was coded with a '?'. Sequences were deposited online in Genbank using the SEQUIN submission tool (NCBI) under the following accession numbers: AY094505-AY094540.

Outgroup sequences and alignment with ingroup

Partial cytochrome *b* (405 bp) and cytochrome oxidase subunit II (538 bp) sequences for the two outgroup taxa chosen for this study, namely *Oecophylla* and *Formica*, were downloaded from Genbank using the following accession numbers: AF190328 (*Oecophylla smaragdina* cytochrome oxidase subunit II); AF191158 (*O. smaragdina* cytochrome *b*); AF190329 (*Formica lugubris* cytochrome oxidase subunit II) and AF191159 (*F. lugubris* cytochrome *b*) (Johnson and Crozier, unpublished). These sequences were aligned with one of the aligned ingroup sequences (and therefore the whole ingroup alignment) in MacClade 4 (Maddison and Maddison, 2000).

II Morphological characters

Specimens examined

Dry mounted worker specimens corresponding to the species collections listed in Table 2.1 were examined under a light microscope. Representative specimens of the two outgroup genera, *Oecophylla* and *Formica* were also examined.

Character selection and coding

An investigation was undertaken in order to identify informative characters for cladistic analysis. Emphasis was placed on attempting to identify discrete qualitative characters. However, quantitative characters that had discrete distributions amongst the species examined were also included. Thirteen morphological characters and one behavioural character were selected (Table 2.4) and scored (Table 2.5) across the workers of 25 species (see Appendix I). Examples of some of the morphological characters identified were recorded by scanning electron microscopy (Model JSM-5200, JEOL, Tokyo, Japan) and are presented in Figures 2.1 to 2.8.

Binary coding was employed for 12 of the 14 characters, and the remaining two characters were coded as multistate characters. An ordered transformation series was inferred for both these characters (see characters 3 and 9, Table 2.4) and they were therefore coded as ordered (additive) and arranged in a hypothesised transformation series (Farris, 1970). Arguments against ordered transformation series of characters have been proposed on the grounds that ordered characters represent hypotheses of character transformation that should be tested rather than assumed by cladistic analysis (Kitching *et al.*, 1998; Poe *et al.*, 2000). However, a counter-argument is that although unordered multistate characters may appear superficially to avoid hypotheses of transformation, they in reality provide a questionable alternative theory of transformation. By allowing any state to transform directly into any other simply amounts, the tendency exists for the most commonly occurring

states to be placed towards the base of the tree with the other states derived from them i.e. the "common equals primitive" criterion (Prendini, 2000). Therefore, characters 3 and 9 were treated as ordered in all subsequent analyses. Characters were polarized by means of outgroup comparison (Nixon and Carpenter, 1993).

Table 2.4. Characters and character states used for cladistic analysis of worker *Camponotus* and outgroups. Character states were scored 0 to 2. Unless otherwise indicated, characters were scored from major workers.

General

1. Polymorphic workers: absent (0); present (1)

Behavioural

2. Arboreal nesting: no (0); yes (1)

Head

3. Number of mandibular teeth: more than 8 (0); 7-6 (1); 5 (2)
4. Apical mandibular tooth: less than 1 ½ X length of the second tooth (0); greater than or equal to 1 ½ X length of the second tooth (1)
5. Clypeal psammophore: absent (0); present (1)
6. Shape of clypeus: without clypeal carina and margin without broad angulate projecting shelf (0); with clypeal carina and with broad, angulate projecting shelf (1)
7. Shape of base of scape: not flattened or expanded (0); flattened and expanded (1)
8. Dense pitting on head, each pit consisting of 5 to 6 punctate impressions and distance between pits frequently less than diameter of pit: absent (0); present (1)
9. Shape of back of head of minor worker in lateral view: broadly rounded; posterior margin meets nuchal carina at 90° (0); intermediately rounded; posterior margin meets nuchal carina at 55-70° (1); narrowly rounded; nuchal carina meets occiput at 45-55°(2)

Mesosoma

10. Shape of dorsolateral margin of propodeum: rounded (0); angulate (1)
11. Shape of hind tibiae: rounded (0); angulate (1)

Metasoma

12. Shape of petiole: scale-like (0); nodiform (1)
13. Thick blunt hairs on gaster: absent (0); present (1)
14. Metapleural gland: present (0); absent (1)

Table 2.5. Distribution of 14 character among 22 species of *Camponotus* and outgroups. Refer to Table 2.4 for character list.

Oecophylla	0	1	0	1	0	0	0	0	0	0	0	1	0	0
Formica	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Polyrhachis schistacea	0	0	2	1	0	0	0	0	0	1	0	1	0	1
Camponotus sp. 24	1	0	1	0	0	1	0	0	2	0	0	0	0	1
Camponotus sp. 2	1	0	1	0	0	1	0	0	1	0	0	0	0	1
Camponotus sp. 11	1	0	1	0	0	1	0	0	2	0	0	0	0	1
C. acvapimensis	1	1	1	0	0	0	0	0	0	0	0	0	0	1
C. detritus	1	0	1	0	0	0	0	0	0	0	1	0	1	1
C. brevisetosus	1	0	1	0	0	0	0	0	0	0	1	0	1	1
C. storeatus	1	0	1	0	0	0	0	0	0	0	1	0	1	1
C. fulvopilosus	1	0	1	0	0	0	0	0	0	0	1	0	1	1
C. rufoglaucus	1	0	1	0	0	1	0	0	0	0	0	0	0	1
C. cinctellus	1	0	1	0	0	1	0	0	0	0	0	0	0	1
C. cuneiscapus	1	0	2	1	0	0	1	0	0	0	0	0	0	1
Camponotus sp. 19	1	0	2	1	0	0	1	0	0	0	0	0	0	1
Camponotus sp. 10	1	0	2	1	0	0	0	0	0	0	0	0	0	1
C. nasutus	1	0	2	1	0	0	0	0	0	0	0	0	0	1
Camponotus sp. 7	1	1	1	0	0	0	0	1	0	0	0	0	0	1
Camponotus sp. 12	1	1	1	0	0	0	0	1	0	0	0	0	0	1
C. mystaceus	1	0	2	1	1	0	0	0	0	0	0	0	0	1
*Camponotus sp.*14	1	0	2	1	1	0	0	0	0	0	0	0	0	1
C. sericeus	1	0	2	1	0	0	0	0	0	1	0	1	0	1
C. chrysurus	1	1	1	0	0	0	0	0	0	0	0	0	0	1
C. bifossus	1	1	1	0	0	0	1	1	0	0	0	0	0	1
C. klugii	1	1	1	0	0	0	0	0	0	0	0	0	0	1

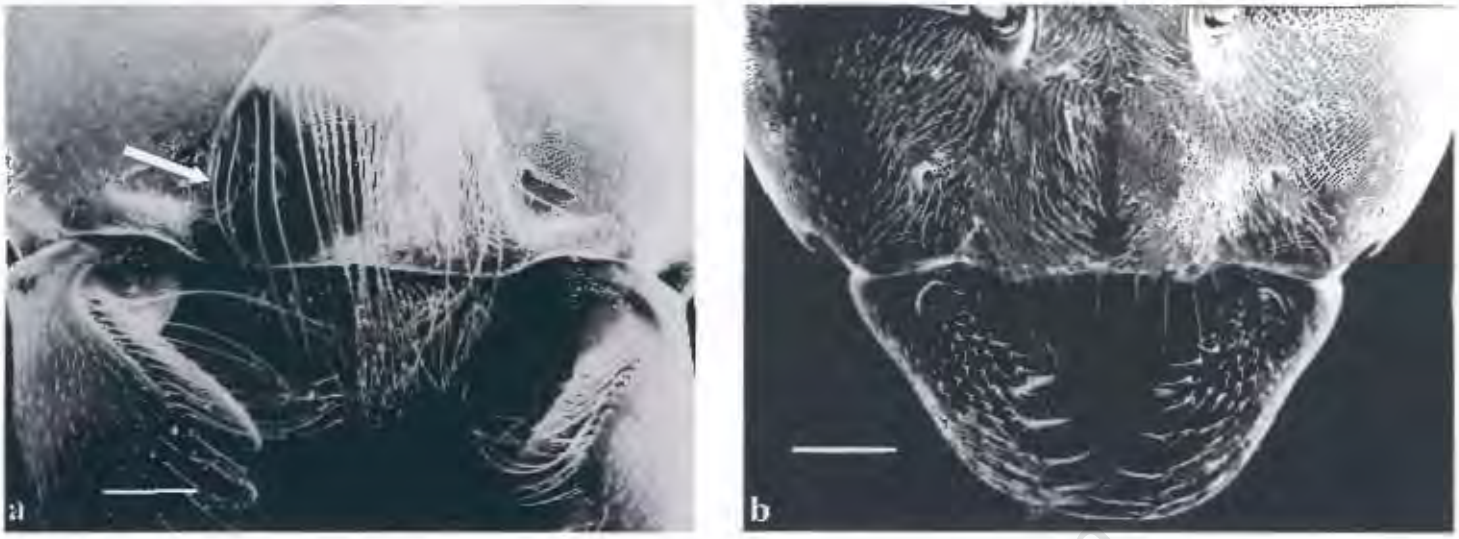


Figure 2.1. Scanning electron micrographs of *Camponotus spp.* showing character 5: clypeal psammophore (Table 2.4). (a) Clypeal psammophore present (*C. mystaceus*), arrowed and (b) clypeal psammophore absent (*C. sericeus*). Scale bar = 0.5 mm.

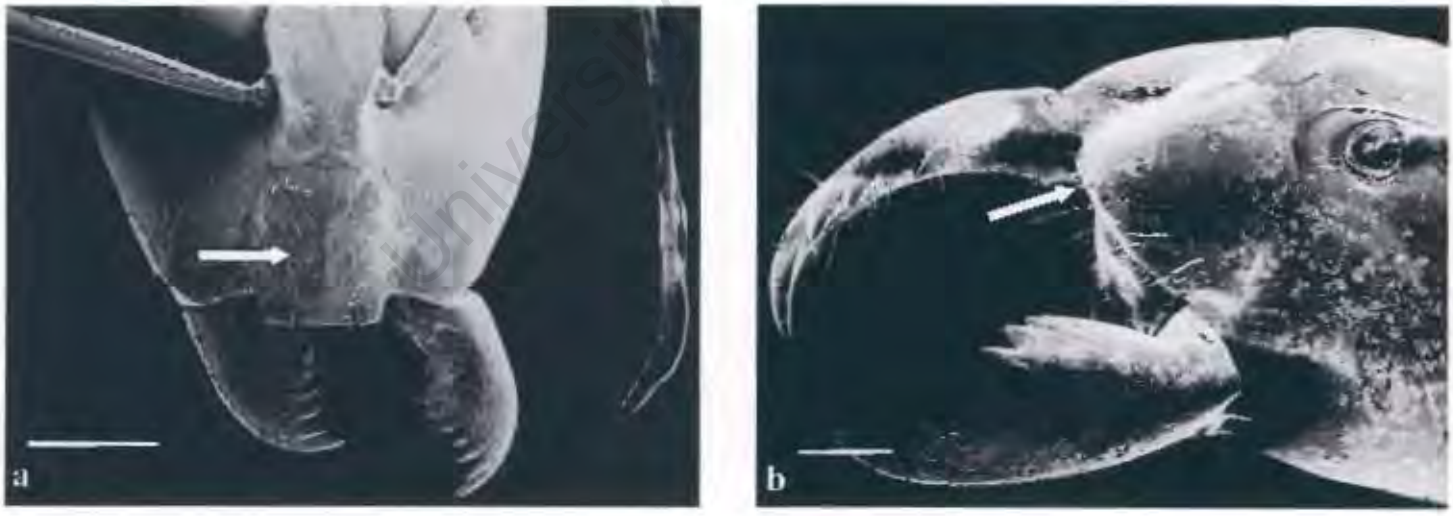


Figure 2.2. Scanning electron micrographs of *Camponotus spp.* showing character 6: shape of clypeus (Table 2.4). (a) With clypeal carina and broad, angulate projecting shelf (*C. sp.* 24), arrowed and (b) without clypeal carina or broad angulate projecting shelf (*Camponotus sp.* 10), arrowed. Scale bar = 1 mm.

Figure 2.3. Scanning electron micrographs of *Camponotus spp.* showing character 9: shape of back of head of minor worker in lateral view (Table 2.4). (a) Narrowly rounded (*Camponotus sp.* 11) and (b) broadly rounded (*C. bifossus*). Scale bar = 0.5 mm.

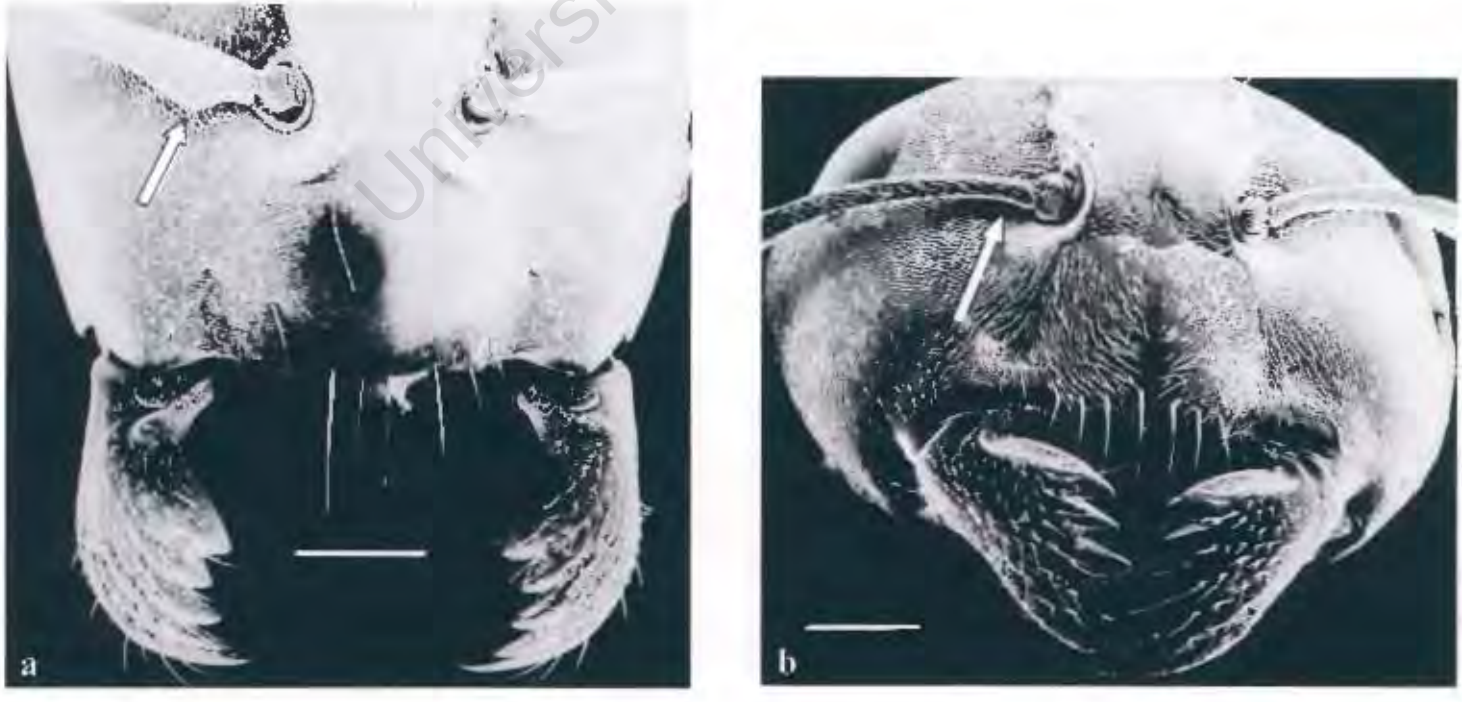


Figure 2.4. Scanning electron micrographs of *Camponotus spp.* showing character 7: shape of base of scape (Table 2.4). (a) Base of scape flattened and expanded (*C. cuneiscapus*), arrowed and (b) base of scape not flattened or expanded (*C. sericeus*), arrowed. Scale bar = 0.5 mm.

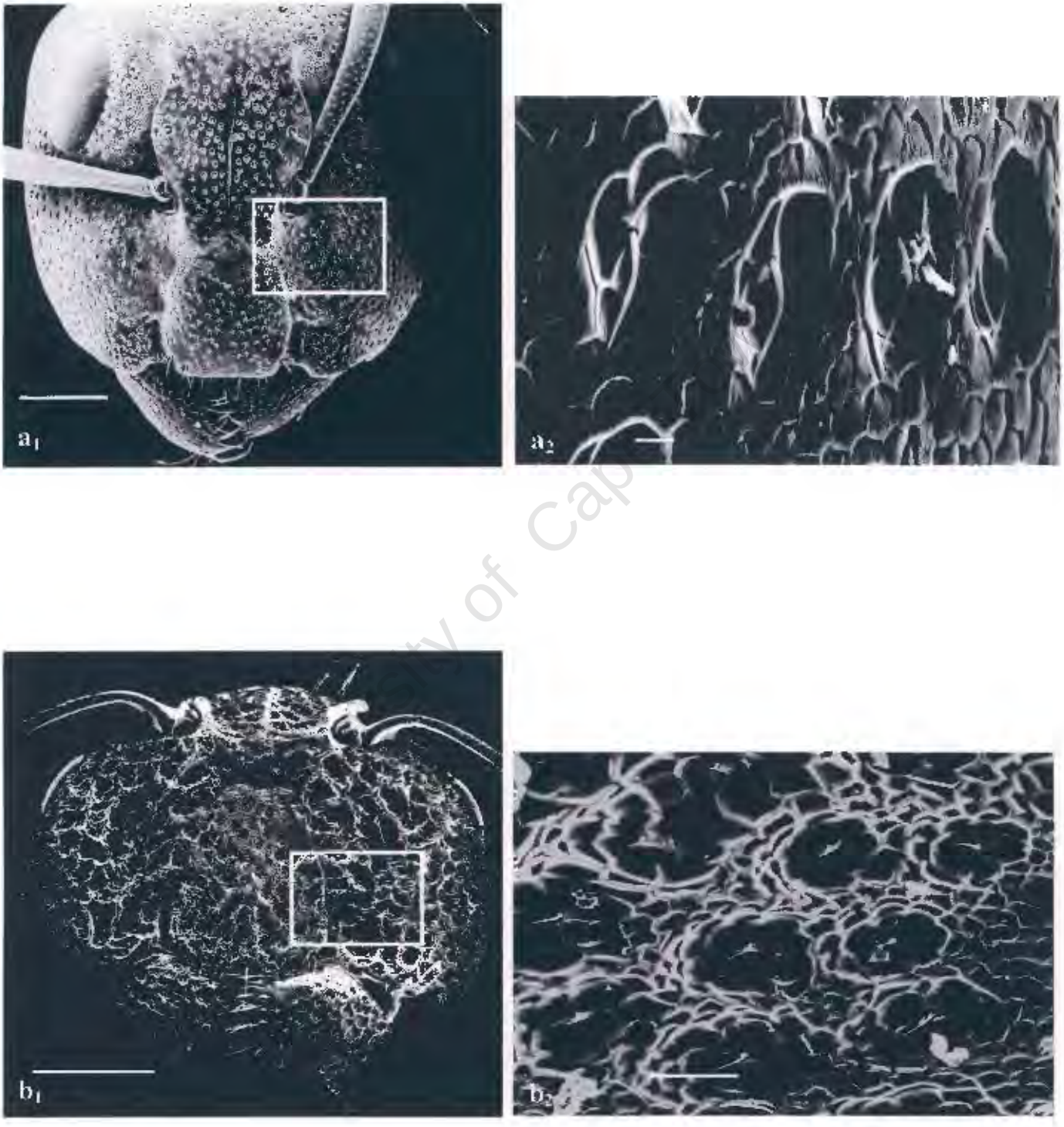


Figure 2.5. Scanning electron micrographs of *Camponotus spp.* showing character 8: dense pitting on head (Table 2.4). (a_1) Dense pitting on head of *C. sp.* 12 worker; (a_2) close-up of pitting in a_1 enclosed by box, (b_1) dense pitting on head of *C. bifossus* (b_2) close-up of pitting in b_1 enclosed by box. Scale bar a_1, b_1 = 0.5 mm. Scale bar a_2, b_2 = 10 µm and 50 µm respectively.

Figure 2.6. Scanning electron micrograph of *Camponotus sp.* showing character 10: shape of dorsolateral margin of propodeum (Table 2.4). (a) Propodeum of *C. sericeus* showing the angulate shape of the dorsolateral margin, arrowed. Scale bar = 0.5 mm.

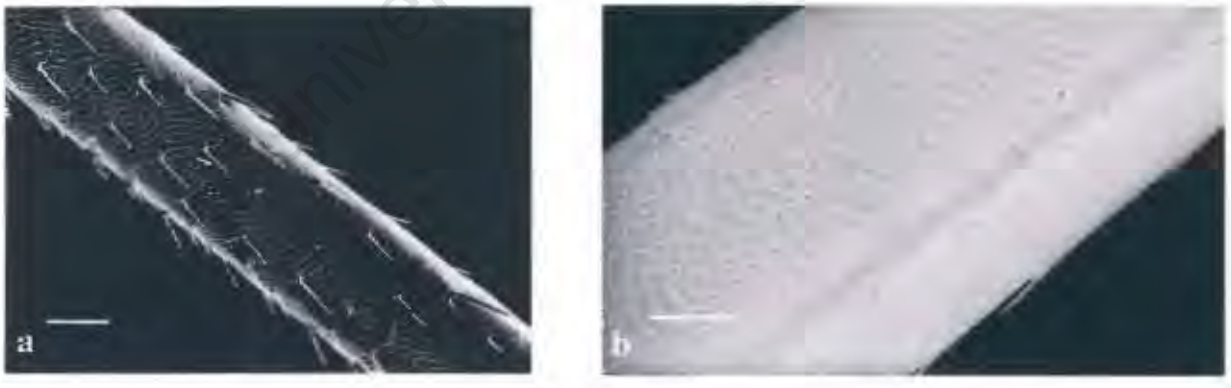


Figure 2.7. Scanning electron micrographs of *Camponotus spp.* showing character 11: shape of hind tibae (Table 2.4). (a) Hind tibia rounded (*C. sp.* 11) and (b) hind tibia characterised by angulate quadrate ridges (*C. storeatus*). Scale bar = 0.1 mm.

Figure 2.8. Scanning electron micrograph of *Camponotus sp.* showing character 12: thick blunt hairs on gaster (Table 2.4). (a) Dorsal view of *C. brevisetosus* showing presence of thick blunt hairs on the gaster. Scale bar = 0.5 mm.

Chapter 3

PHYLOGENETIC RECONSTRUCTION: MATERIALS AND METHODS

Reconstruction of phylogenies from molecular and morphological data is a complex undertaking. Our understanding of molecular and morphological evolution is by no means comprehensive, yet when constructing phylogenies we are forced to make certain assumptions, either implicit or explicit, about the way sequences or morphological characters evolve. Furthermore, there has been extensive debate in the phylogenetic literature over which optimality criterion is 'correct', although the choice of nucleotide substitution model or weighting of characters may play as large a role in the final outcome as the choice of analytical method (Simon *et al.*, 1994). In this chapter, the multifaceted methodological approach taken in order to explore the character data, design weighting schemes, choose appropriate nucleotide substitution models and generate phylogenetic hypotheses, is outlined.

I Data set construction and rationale

Hardware and software

Analyses were either performed on a Macintosh PowerPC running Mac OS 9.1 or a Flexi PII/ZXi Personal Computer running Windows® 98 depending on the requirements of the software listed in Table 3.1.

Table 3.1. Software used for phylogenetic analyses.

Program	Reference	Operating System	Application (this study)	Source
PAUP* 4.0 beta version 8 for Macintosh	Swofford, 1998	Mac OS 9.1	Phylogenetic inference	Sinauer Associates, Sunderland, Massachussetts
MEGA version 2.1	Kumar *et al.*, 2001	Windows	Sequence data characterisation	Free download: http://www.megasoftware.net
Modeltest version 3.06	Posada and Crandall, 1998	Mac OS 9.1	Evaluation of best fit nucleotide substitution model for the data	Free download: http://zoology.byu.edu/crandall_lab/Modeltest.htm
TreeRot version 2.0	Sorenson, 1999	Mac OS 9.1	Bremer Support and Partitioned Bremer Support Indices	Free download: http://www.mightyduck.bu.edu/TreeRot
6-P Parsimony spreadsheet	Cunningham, 1997a	Mac OS 9.1/ Microsoft Excel	Determination of stepmatrix for 6-P parsimony	Email request to author: cliff@duke.edu (Assoc. Prof. Cliff Cunningham)
MrBayes version 2.01	Huelsenbeck and Ronquist, 2001	Mac OS 9.1	Bayesian inference of phylogeny	Free download: http://morphbank.ebc.uu.se/mrbayes
MacClade 4.0	Maddison and Maddison, 2000	Mac OS 9.1	Conversion to NEXUS format and data exploration	Sinauer Associates, Sunderland Massachussetts

Data sets

A multifaceted approach to phylogenetic reconstruction was taken, with data sets analysed both separately and in combination (Swofford, 1991). A total of 20 data sets were analysed. Six data sets containing sequence data were constructed (Table 3.2). This was done to accommodate the fact that some of the taxa would not amplify for one of the two genes targeted despite multiple attempts at amplification with different primer sets (see Appendix III).

Table 3.2. Details of nucleotide data sets analysed in this study

Data set	Partial gene sequence	bp	No. taxa
A:	cytochrome *b* (cyt *b*)	660	19
B:	cytochrome *c* oxidase II (coii)	482	21
C:	cytochrome *b* (reduced)	660	15
D:	cytochrome oxidase II (reduced)	482	15
E:	cyt *b* + coii (reduced)	1142	15
F:	cyt *b* + coii (all)	1142	25

The combined nucleotide data sets were constructed in MacClade (Table 3.1) using the 'Import Sequences' Option under the 'Taxa' Menu. The taxon-reduced data sets C, D and E were constructed to allow taxonomically equivalent comparisons of cytochrome *b* and cytochrome oxidase II gene tree topologies with each other and the combined topology. Data set F was constructed to generate hypotheses of relationships for all species based on all available sequence data. This data set contained ten taxa for which character information for only one of the two mitochondrial amplicons was

available. The percentage nucleotide data missing for these ten taxa is indicated in Table 3.3.

Table 3.3. Taxa with missing data for combined sequences.

Species	% Missing sequence data
Camponotus detritus	57.8
Camponotus brevisetosus	57.8
Camponotus rufoglaucus	42.2
Camponotus cuneiscapus	42.2
*Camponotus sp.*19	57.8
*Camponotus sp.*12	42.5
Camponotus mystaceus	76.2
*Camponotus sp.*14	65.3
Camponotus sericeus	42.5
Camponotus chrysurus	42.2
Formica lugubris	26.1
Oecophylla smaragdina	26.0

The aligned nucleotide sequences comprising data sets A to F were translated to their amino acid sequences in MacClade using the invertebrate mitochondrial code (Appendix II) to obtain six amino acid data sets. Amino acid and nucleotide sequence alignment files were converted to Nexus file format (Maddison *et al.*, 1997) in MacClade. Morphological character data was entered into MacClade, and converted to Nexus file format for analysis using PAUP*.

Six data sets combining DNA sequences with morphological data were manually constructed using the text editor of PAUP*, after performing congruence tests in PAUP* (described on page 58). These tests were performed in order to evaluate whether the two data partitions (molecular and morphological) could be combined.

Since all sequences from individuals' from each species were identical or differed by < 1%, with differences due to single point mutations, only one representative per species was included in the analyses. By including only

one representative individual per species, the implicit assumption is made of species monophyly (de Queiroz *et al.*, 2002).

Data partitions

The following character sets were designated as data partitions: cytochrome *b* (n = 660, CYTB), cytochrome *c* oxidase II (n = 484, COII) and morphological characters (n = 14, MORPH). Additionally, three data partitions corresponding to the three codon positions of each gene were defined: 1st codon positions cytochrome *b* (n = 220, 1stposCYTB); 2nd codon positions cytochrome *b* (n = 220, 2ndposCYTB), 3rd codon positions cytochrome *b* (n = 220, 3rdposCYTB); 1st codon positions cytochrome *c* oxidase II (n = 161, 1stposCOII), 2nd codon positions cytochrome oxidase II (n = 162, 2ndposCOII) and 3rd codon positions cytochrome oxidase II (n = 161, 3rdposCOII). For the combined gene data sets (E and F), all 1st codon positions (n = 380, 1stposALL), 2nd codon positions (n = 381, 2ndposALL) and 3rd codon positions (n = 381, 3rdposALL) were defined.

II Data set characterisation

Basic sequence statistics and evolutionary distances

MEGA (Table 3.1) was used to calculate nucleotide and amino acid sequence composition, percent codon usage, amino acid variation and sequence divergence. Base compositional bias (C) was determined by the method of Prager and Wilson (1988). This value ranges from 0 to 1, with 0 indicating no base composition bias and 1 indicating complete bias i.e. fixed for a single nucleotide. Uncorrected pairwise sequence divergences were calculated, as well as corrected divergence estimates using the Tamura and Nei (1993) model of sequence evolution. Tamura-Nei distance corrections are advocated when there is a strong A+T bias in the data (Kumar *et al.*, 2001). Furthermore, the number of parsimony informative (PI) sites, variable sites and constant sites were obtained from PAUP*.

Patterns of sequence variation

The null hypothesis of base frequency stationarity among sequences was evaluated using the χ^2 heterogeneity test as implemented in PAUP*. Base frequency stationarity was separately evaluated for all nucleotide characters, as well as for only parsimony informative characters at each of the three codon positions (Waddell, 1999; Buckley *et al.*, 2001a). For each data set, χ^2 heterogeneity tests were performed with the ingroup sequences only, as well as with ingroup and outgroup sequences combined. This was done in order to assess whether the outgroup sequences could have a potentially confounding effect on subsequent phylogenetic analyses by violating the assumption of base frequency homogeneity across taxa (Cameron and Mardulyn, 2001).

The possible effect of base composition bias on phylogeny reconstruction was further evaluated by transforming the data using the LogDet transformation (Lockhart *et al.*, 1994). This transformation corrects for distortions in phylogenetic signal caused by compositional bias, as it does not assume stationarity of base composition throughout the tree (Swofford *et al.*, 1996). Minimum Evolution (ME) bootstrap consensus trees (Rzhetsky and Nei, 1992) constructed using LogDet corrected distances were compared to trees obtained from other model-based methods. This was done in order to evaluate whether non-stationarity of base composition, as indicated by the χ^2 test heterogeneity test, warranted special attention in subsequent phylogenetic analyses (Flook *et al.*, 1999).

Site-to-site estimates of the actual number of changes along the length of the cytochrome *b* and cytochrome oxidase II sequences were obtained using MacClade. The complete cytochrome *b* and cytochrome oxidase II data sets were optimised on the best MP topology recovered from each respective data set under equal weights (Whitfield and Cameron, 1998).

Patterns of nucleotide variability corresponding to possible functional constraints were examined by plotting the inferred number of nucleotide changes against a window of three amino acids. The ant cytochrome *b*

Homoplasy in the two genes was examined by generating charts of the distribution of the number of steps per character in MacClade, using the most parsimonious or one of the most parsimonious unweighted parsimony trees as a reference. The greater the proportion of sites within a gene with a high number of nucleotide changes, the more likely these variable sites are to be too saturated to provide useful phylogenetic signal. The frequency distribution of the observed number of substitutional changes for each gene was evaluated at first plus second and third codon positions (Mardulyn and Cameron, 1999; Mardulyn and Whitfield, 1999).

Phylogenetic signal analyses

The *g*1 (skew) statistic, used to detect phylogenetic signal in the data (Hillis and Huelsenbeck, 1992), was calculated in PAUP*. For each molecular data set, 10 000 trees were drawn at random from a set of all possible trees, and the tree length frequency was plotted. Data sets with phylogenetic signal produce strongly left-skewed tree length distributions as evidenced by *g*1 values significantly less than zero.

Relative frequency of types of base changes

A graphical representation of the relative frequency of base changes between the four nucleotides was evaluated on the best parsimony tree estimate for each nucleotide data set using the 'Chart State Changes and Stasis' option in MacClade (Whitfield and Cameron, 1998). These plots were used to estimate substitution bias between different nucleotide pairs in order to design specific weighting schemes for parsimony analyses (Dowton and Austin, 1998).

III Tree reconstruction

Outgroups and Rooting

Maddison *et al.* (1992) recommend that the taxa included as outgroups in phylogenetic analysis should not be too distantly related to the ingroup taxa. Therefore, *Formica lugubris* and *Oecophylla smaragdina* were designated as outgroup taxa in this study. Both these taxa are classified in the same subfamily as *Camponotus* and *Polyrhachis* (Formicinae), with *Formica*, *Camponotus* and *Polyrhachis* comprising the *Formica* genus group (Agosti, 1991). *Oecophylla* has previously been used as an outgroup in molecular phylogenies of *Camponotus* (Brady *et al.*, 1999; Sameshima *et al.*, 1999). Rooting was performed subsequent to all analyses using these two outgroups.

Model-based analyses : Bayesian analyses

Bayesian phylogenetic analyses were performed using MrBayes (Table 3.1). MrBayes implements a variant of Markov chain Monte Carlo (MCMC) to approximate the posterior probability density of parameters of nucleotide evolution. In MrBayes, each step in a Markov chain involves random modification of the tree topology, branch lengths or one of the parameters in the substitution model. Each new step is either accepted or rejected according to an acceptance probability calculated using the Metropolis-Hastings-Green algorithm, a variant of MCMC, which, if repeated many thousands or millions of times, allows the Markov chain to visit regions of tree space in proportion to their posterior probability (Huelsenbeck *et al*, 2001; Lewis, 2001).

The most general time reversible model of nucleotide evolution, the General Time Reversible (GTR) model (Yang, 1994; see Lio and Goldman, 1998 for a review of models of molecular evolution) was used for all analyses based on

results from Modeltest[2]. Rate heterogeneity was incorporated either by assuming a discrete gamma distribution[3] with four rate classes, and a proportion of invariant sites p_{inv} (Γ + I model of Gu *et al.*, 1995), or by assuming site specific rate parameters (SSRs). For analyses incorporating rate heterogeneity using site specific rates, three sites were defined for single gene data sets (A, B, C, D), corresponding to the three codon positions (SSR_3). For the combined nucleotide data sets (E and F) both gene-specific (two rates corresponding to the cytochrome *b* and cytochrome oxidase II partial sequences, hereafter referred to as the SSR_2 model) and codon-specific rates (six codon positions corresponding to three codon partitions from each of the two partial gene sequences, hereafter referred to as the SSR_6 model) were estimated during the analyses.

Parameter values were not defined *a priori*, but instead treated as unknown variables with uniform priors using the following default prior settings: rate matrix (0 to100), branch lengths (0 to 10), gamma shape parameter (0 to10) and the proportion of invariable sites (0 to1). Site specific rates were set to an initial value of 1.0. An uninformative prior was used for the topology, and a 'flat dirichlet' distribution assumed for the base frequency parameters, giving all combinations of base frequencies equal prior probability (J. Huelsenbeck pers. com.). A random tree generated by MrBayes was used as a starting tree for each Markov chain. For all analyses, four Markov chains were run for one million generations, comprising one cold chain and three incrementally heated chains. Tree sampling was performed every 50 generations, thereby generating 20 000 sample points. To ensure that the Markov chain had reached stationarity, the fluctuating value of the likelihood was graphically monitored. Samples collected prior to stationarity should be discarded as 'burn-in', as they contain no useful information about parameter values

[2] Described on page 58.

[3] The gamma distribution is an attractive model for describing among-site rate variation for two reasons (reviewed by Yang, 1996b). Firstly, the distribution of rates among sites is described by a single parameter α, equal to the inverse of the squared coefficient of variation of the substitution rate, that is comparable across data sets provided the same number of rate categories are used. This allows generalizations to be made regarding the relative amount of among-site rate variation among genes. Secondly, by setting the scale parameter β to $1/\alpha$, a distribution with a mean rate of 1 is obtained, with the consequence that a wide variety of site-rate distributions can be accommodated ranging from a nearly homogenous distribution to a highly variable distribution (Swofford *et al*, 1996; Buckley *et al*, 2001b).

(Huelsenbeck and Ronquist, 2001). Five thousand trees (corresponding to 250 000 generations) were discarded as burn-in. This included sample points in the apparently stationary region of the chain, following the recommendation of Leache and Reeder (2002) that it is better to be cautious and discard useful samples, rather than inadvertently include burn-in samples when estimating Bayesian posterior probabilities. A 50% majority-rule consensus of the remaining 15 000 trees was then used to generate posterior probability approximations for each clade, with the percentage of samples containing a particular clade representing that clade's posterior probability (Huelsenbeck and Ronquist, 2001).

If ≥ 95% of sampled trees contained a given clade, this clade was considered to be significantly supported by the data. Clades contained in 90% - 94% of all sampled trees were considered to be strongly supported (Leache and Reeder, 2002; Wilcox *et al.*, 2002).

To ensure that the analyses were not trapped in local optima, a number of precautions were taken. Each run was repeated at least twice for each data set with different starting trees to ensure convergence. Convergence was accepted if the log likelihood values approached similar mean values. The consensus topology and clade posterior probability values from the independent runs were also examined to ensure that similar log likelihood values were not obtained from incongruent topologies. Furthermore, the use of incrementally heated Markov chains enhances parameter space exploration. The random exchange of parameter values between heated chains and the cold chain acts to decrease the distance between optimal peaks in the parameter space, thereby avoiding entrapment in local optima (Huelsenbeck and Ronquist, 2001). Three incrementally heated Markov chains with default settings were run for each analysis (Leache and Reeder, 2002).

Model-based analyses: Maximum likelihood (ML)

(i) Model choice

The initial best-fit model of nucleotide evolution for each of the six nucleotide data sets was evaluated using Modeltest. Modeltest utilises likelihood ratio tests (LRTs) to evaluate hierarchical ML models of sequence evolution on an initial neighbour-joining tree. The significance of the increased fit (higher log likelihood) of a more parameter-rich or complex model compared to a less parameter-rich model is then evaluated using the likelihood ratio test statistic δ. Statistical testing is desirable in this context because although, overall, deterministic models inevitably improve the likelihood score, they also increase sampling variance, which may result in decreased accuracy (Cunningham *et al.*, 1998).

The LRT test statistic δ is defined as $\delta = -2(\ln L_1 - \ln L_0)$, where $\ln L_0$ is the natural logarithm of the likelihood under the simpler, constrained model, and $\ln L_1$ is the natural logarithm of the likelihood under the more complex, parameter-rich model. The significance of the improvement in fit of the more complex model is then evaluated using the associated *P*- values from a χ^2 distribution with q degrees of freedom (q = difference in number of free parameters between the two models), which is the appropriate likelihood ratio test statistic for nested models (Frati *et al.*, 1997; Huelsenbeck and Crandall, 1997).

Modeltest uses a '*from the bottom up*' approach for deciding which model represents the best compromise between the model goodness of fit and parameter minimization. The simplest model is tested first and then a series of more complex, nested hypotheses are evaluated to test for important parameters to include in the final model. A Bonferroni correction for multiple tests (Sokal and Rohlf, 1995) was therefore applied following Huelsenbeck and Crandall (1997). Parameter values evaluated by Modeltest on a neighbour joining tree using the best fit model of evolution (i.e. nucleotide equilibrium frequency values (π); instantaneous transition rate matrix values

and rate heterogeneity parameter values, accommodated by p_{inv} and α), were fixed in PAUP* in an initial ML run.

(ii) Search strategy

Tree reconstruction was implemented in PAUP* (Table 3.1). ML heuristic searches were performed with 10 random stepwise addition replicates and TBR branch swapping with the 'MulTrees' option in effect. Zero length branches were collapsed. A successive approach was used in which model parameters were re-estimated on the optimal ML tree using maximum likelihood, with this process reiterated until the tree topology and parameter values stabilized (Swofford *et al.*, 1996; Leache and Reeder, 2002). Furthermore, model parameters were estimated on the Bayesian consensus topology, best maximum parsimony (MP) tree (or strict consensus of the best MP trees) and a minimum evolution LogDet tree. If these parameters appeared to differ greatly from those estimated on the ML tree obtained using Modeltest parameters, these parameters were then used in a maximum likelihood run to evaluate whether these changes made any difference to the tree topology.

Parsimony analyses

(i) Tree statistics

All tree statistics were calculated with uninformative sites excluded as these inflate tree statistic values (Sanderson and Donoghue, 1989). Cladogram length, ensemble consistency indices (CI) (Kluge and Farris, 1969) and retention indices (RI) were recorded for each tree. The CI and RI values provide an indication of the measure of fit between the data and the topology. Ensemble indices are the summation over all characters of all the individual indices calculated for each character comprising the data set. The CI value provides an indication of the amount of homoplasy in the data and is defined as M/S where M = Σm and m is the minimum amount of change that a

character may show on the tree, and S = Σs where s is the amount of change required by the character on the tree being evaluated. If homoplasy is absent the consistency index is 1, with the CI value decreasing towards 0 as homoplasy increases. The RI value measures the amount of similarity interpreted as synapomorphy, where RI = (G – S)/(G - M) with S and M defined as above, and G = Σg, where g is the greatest number of steps a character can exhibit on any cladogram.

(ii) Characters

Nucleotides were treated as unordered characters with four alternative states, and amino acids were treated as unordered characters with 20 alternative states. All morphological characters with the exception of characters 3 and 9 were analysed as unordered characters (Table 2.4, Chapter 2). As a transformation series was inferred for characters 3 and 9, they were treated as ordered in all phylogenetic analyses.

(iii) Weighting schemes

Both *a priori* (hypothesis and tree independent) and *a posteriori* (hypothesis and tree dependent) weighting schemes were applied to the data matrices (Kitching, 1998) to determine which relationships were robust across a variety of weighting schemes (Funk *et al.*, 1995; Funk, 1999; Dowton and Austin, 2002). Both these approaches attempt to identify the characters least likely to have experienced homoplasy, and to give these characters greater weight in the analysis (Simon *et al.*, 1994).

a) Uniform weighting

All data sets were initially analysed using 'unweighted' parsimony with all characters assigned equal weight.

b) Transition downweighting

Differential weighting of various classes of characters may be justified based on observations that these classes of characters are evolving under different constraints relative to other classes (Funk *et al.*, 1995). Transitions appear to accumulate much more rapidly than transversions as sequences diverge, and are therefore expected to be more prone to homoplasy, decreasing their phylogenetic utility (Bielawski and Gold, 1996; Swofford *et al.*, 1996). To account for this, transversions are often given greater weights in parsimony analyses (Swofford *et al.*, 1996; Yoder *et al.*, 1996; Voelker and Edwards, 1998).

As an objective guide to weighting, estimates of the transition/transversion ratio (Ti/Tv) for each nucleotide data set was calculated by maximum likelihood optimisation on the Bayesian consensus tree topology, using a GTR+ I+ Γ model (Yoder *et al.*, 2001). The α shape parameter of the gamma distribution and proportion of invariant sites as well as the base frequencies were simultaneously computed. Transition bias estimates were obtained for first, second and third codon positions as well as all codon positions combined (Table 3.4).

Table 3.4. Ti/Tv ratio estimates for all codon positions of the nucleotide data sets

	Cytochrome *b*		Cytochrome oxidase		Combined	
	Data set A (cyt *b* all)	Data set C (cyt *b* reduced)	Data set B (COII all)	Data set D (COII reduced)	Data set E (combined reduced)	Data set F (combined all)
All	4	4	23	6	3	4
1st codon	1	1	7	2	2	2
2nd codon	2	2	3	2	2	2
3rd codon	14	21	14	43	10	7

Two Ti/Tv weighting schemes were implemented in separate analyses to counteract the transition bias. In one analysis, a cost matrix downweighting transitions relative to transversions according to the transition bias estimated from all sites was applied to all characters included in the analysis. In another analysis, three separate cost matrices to downweight transitions were constructed based on the Tv/Ti estimates for the three codon positions, and each codon position was then assigned its codon-specific step matrix.

The most extreme form of transition downweighting, namely transversion parsimony where all transitions are assigned a weight of zero, was also applied to all nucleotide data sets. Transversion weights were applied to all characters, and to third codon positions only. Transitions in first and second codon positions are unlikely to be saturated for species-level divergences, and therefore potentially phylogenetically informative information may be lost by assigning these transitions zero weights.

c) A-T Tranversion downweighting

The resolving power of parsimony analysis is greatly enhanced by differential weighting of character types according to their empirically determined frequency of occurrence (Hillis *et al.*, 1994). A+T bias will tend to result in underestimation of branch lengths in parsimony analysis unless this bias is accounted for by A↔T transversion downweighting (Dowton and Austin; 1997; Whitfield and Cameron, 1998, Dowton *et al.*, 2002). In order to explore the potential bias the very high A+T content, particularly at third codon positions, might exert on the parsimony analyses, a cost matrix downweighting A↔T transversions was applied to all nucleotide data sets.

Cost matrices were constructed to downweight A↔T tranversions two-fold (changes between A and T assigned unit weight, all other changes assigned a weight of two). This value was estimated from parsimony reconstructions of the frequencies of character change on the most parsimonious tree for each data set using MacClade. A↔T transversions were observed to occur, on

average, two times more frequently than any other type of character change with the exception of T↔C transitions in both cytochrome *b* and cytochrome oxidase II sequences. Figure 3.1, included in this section for ease of reading, graphically depicts the pattern of reconstructed character change for the large cytochrome *b* and cytochrome II oxidase data sets.

The A↔T cost matrix was applied to all characters, as well as characters at third codon positions only, as these characters show the greatest A+T bias in hymenopteran mitochondrial DNA (Simon *et al.*, 1994; Buckley *et al.*, 2001b).

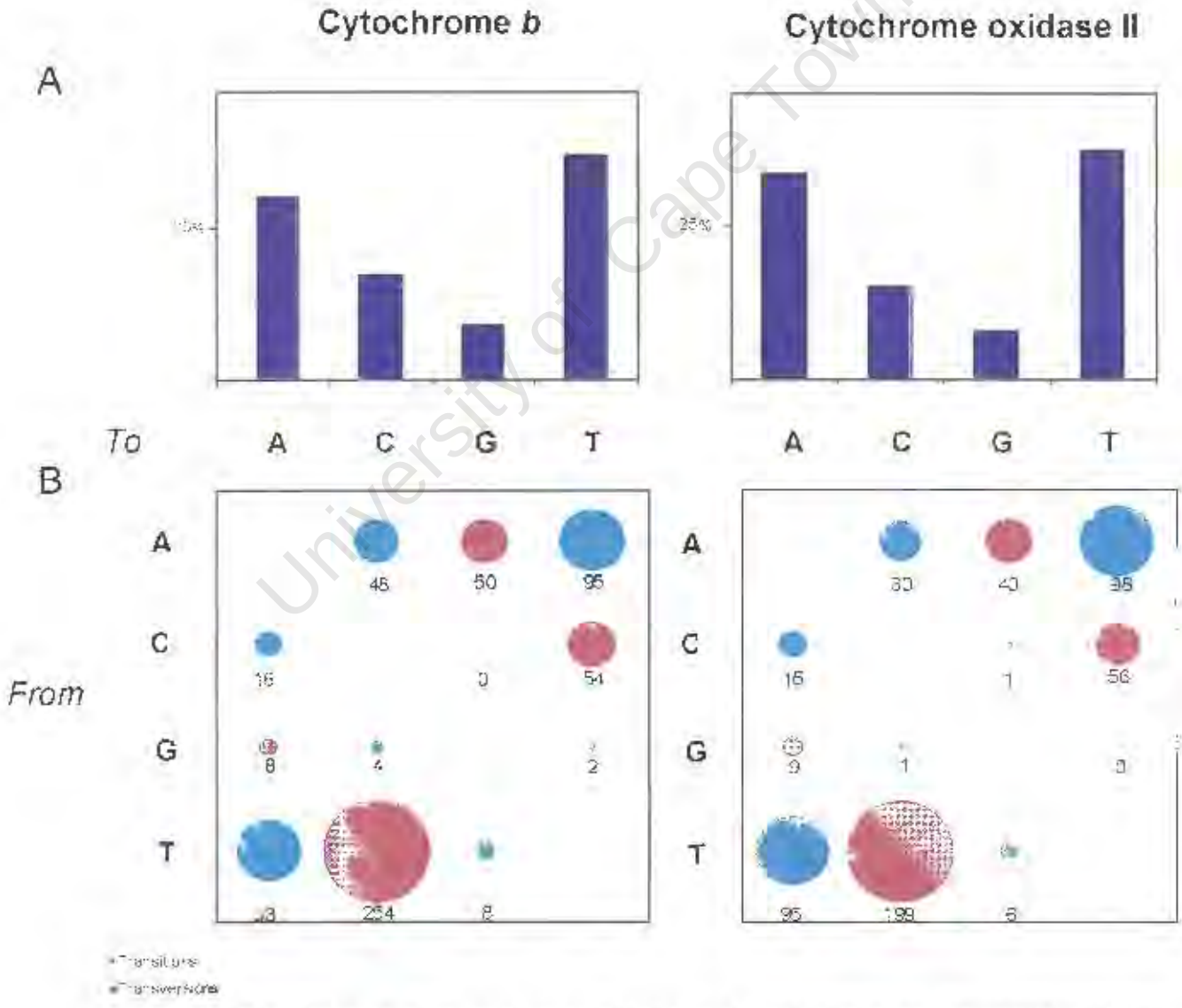


Figure 3.1. Patterns of reconstructed character change based on the unweighted MP tree topology for data sets A and B. (A) Average base frequencies among taxa for all codon positions. (B) Inferred direction of base change using parsimony reconstruction. Numbers below bubbles are the average number of changes inferred.

d) T-C downweighting

T↔C transitions predominated in both genes, consistent with previous studies based on insect mitochondrial genes (Brown *et al.*, 1994; Frati *et al.*, 1997; Muraji and Nakahara, 2001). There was substantial asymmetry in the reciprocal number of T→C and C→T transitions reconstructed by maximum parsimony (Figure 3.1). This can be attributed to the distorting effect of base compositional bias on parsimony character state reconstructions. This bias results in common-to-rare state changes being overestimated relative to rare-to-common state changes given moderate to high evolutionary rates, even though the underlying mutational process may be symmetrical (Collins *et al.*, 1994). Asymmetrical step matrices based on these biased character state reconstructions, are therefore likely to result in poor estimates of phylogeny.

To avoid the above-mentioned potential pitfall, yet still incorporate the empirical observation that T↔C transitions were the predominant mutation type, the number of T→C and C→T transitions were summed and an average rate for these transition types inferred relative to the occurrence of the other mutational types. T↔C transitions in both cytochrome oxidase II and cytochrome *b* sequences were found to occur, on average, six times more frequently than all other types of nucleotide changes. A symmetrical cost matrix was therefore constructed in which this particular transition type was downweighted six-fold relative to the other mutation types.

e) Successive weighting

The rationale behind successive weighting is that each character is weighted according to the degree of homoplasy it exhibits on the most-parsimonious tree, based on the cladistic concept of character consistency. Those characters that display high levels of homoplasy are given relatively lower weights compared to those characters exhibiting lower levels of homoplasy (Kitching *et al.*, 1998). The rescaled consistency index was used to reweight characters after an initial unweighted parsimony run. This value is the product

of the ci and ri values for that character (Swofford, 1991; Kitching *et al.*, 1998). Heuristic searches were performed with these new character weights in order to test the stability of resulting topologies as well as choose amongst equally parsimonious trees for some data sets (Farris, 1969; Orti and Meyer, 1997; but see Kitching *et al.*, 1998). This procedure was repeated until the weights assigned to each character in two successive iterations were identical i.e. the topology stabilised. Successive weighting was applied to all nucleotide data sets.

f) Six parameter parsimony

Six parameter (6-P) parsimony may be considered a special case of generalized parsimony, where a cost is assigned to the transformation from any character state to any other (Williams and Fitch, 1990; Cunningham, 1997a; Stanger-Hall and Cunningham, 1998). For nucleotide data, six substitution classes are considered: A↔T, A↔C, A↔G, C↔G, C↔T and G↔T. A specific weight is defined for each one of the six substitution classes based on their observed frequencies. This allows the heterogeneity of substitutions within, as well as between transitions and transversions to be accommodated. Six-parameter weighting has been shown to increase accuracy and congruence among data sets in certain instances (Cunningham, 1997a).

To obtain appropriate step matrices, R-matrices representing the relative substitution rates among the four nucleotides were estimated under the GTR model from the single most parsimonious unweighted tree, or one of the most parsimonious unweighted trees using maximum likelihood optimisation. The resulting R-matrix was then copied into an Excel spreadsheet kindly supplied by Cliff Cunningham (Table 3.1), and the resulting 6-P step matrix obtained by taking the negative of the natural logarithm of each frequency parameter (Mitchell *et al.*, 2000). The resultant weight matrix was then applied to all included characters in a parsimony analysis. For data sets with more than one partition i.e. the combined data sets E and F, the data were analysed both

with a single matrix estimated for the entire data set applied to all characters, and with independent 6-P weight matrices previously calculated for each data partition applied separately to the corresponding partitions (Cunningham, 1997a; Stanger-Hall and Cunningham, 1998; Buckley *et al.*, 2001a). Those step matrices that violated the triangle inequality assumption were adjusted using PAUP* (Cunningham 1997a; Buckley *et al.*, 2001a)[4]. Six-parameter weighting was applied to all nucleotide data sets. Cost matrices for the six nucleotide data sets are shown in Table 3.5.

(iv) Protein parsimony

Protein data sets, in addition to unordered parsimony analysis (Fitch, 1971), were also analysed by protein parsimony weighting in order to evaluate the contribution of nonsynonymous changes to the overall phylogenetic signal (Friedlander *et al.*, 1998; Fang *et al.*, 2000). A protein parsimony step matrix (Protpars) was constructed in MacClade using the invertebrate mitochondrial genetic code, and a heuristic search with this protein parsimony step matrix applied to all amino acid characters implemented. The step matrix consisted of transformation scores of "1" or "2" determined by the minimum number of nonsynonymous changes separating particular codons. This method uses the genetic code to designate which pairs of amino acids are adjacent, and allows changes only among adjacent states. One of the major assumptions of this method is that the probability of a synonymous base change is much higher than the probability of a nonsynonymous base change (Felsenstein, 1996).

(v) Search strategy

All parsimony analyses were performed with uninformative sites excluded. Heuristic searches of tree space were conducted using 1000 random stepwise addition replicates and tree bisection and reconnection (TBR) branch swapping with the 'MulTrees' option in effect. ACCTRAN character optimisation was used, and zero length branches were collapsed.

[4] Correcting for the triangle inequality is important to ensure that changing directly from one nucleotide to another is never more expensive than passing through an intermediate nucleotide state (Cunningham, 1997a).

Table 3.5. Cost matrices for the six classes of substitution employed in the six parameter parsimony analyses.

Type of change	Cost (number of steps)					
	data set A (cyt *b* all)	data set B (COII all)	data set C (cyt *b* reduced)	data set D (COII reduced)	data set E (combined reduced)	data set F (combined all)
Transition types						
A↔G	2	2	2	2	2	2
C↔T	0	0	1	0	0	1
Transversion types						
A↔T	3	3	3	3	3	2
A↔C	3	3	3	3	3	3
T↔G	17	15	15	6	14	17
C↔G	4	4	1	4	3	1

Clade Robustness

(i) Non-parametric bootstrap analyses

To obtain estimates of nodal support for parsimony analyses, non-parametric bootstrap (1000 pseudoreplicates, three random sequence addition replicates, 50% majority rule) analyses were performed using a heuristic search strategy and TBR branch swapping. Support for nodes on ML trees was assessed using 100 heuristic ML replicates, each with a starting tree obtained by neighbour-joining (data sets A to E) with parameters fixed to those estimated by Modeltest for each data set, and TBR branch swapping. For data set F, the starting tree was obtained by random addition of sequences as 24 distances were undefined for the neighbour-joining tree due to the presence of missing data. Bootstrap values were classified into three support categories following de Queiroz *et al.*, (2002): weak support 50-69%; moderate support 70-89% and strong support 90-100%.

(ii) Bremer support

Bremer support or decay indices were calculated for parsimony topologies as an additional estimate of nodal support (Bremer, 1988). The Bremer support index for a node is defined as the additional steps required in the shortest tree that is inconsistent with a given node compared to the shortest unconstrained tree. Bremer support indices were obtained for the parsimony trees by executing command files generated by TreeRot (Table 3.1) in PAUP*. These command files contained constraint statements for each node in the shortest or strict consensus tree, as well as commands to search for trees inconsistent with each of the constraint statements in turn. Heuristic searches with 20 random addition replicates were performed for each constraint statement. The Bremer support values for each node were then determined by parsing the PAUP* log file from each run.

IV Incongruence testing

Character incongruence

Data partition heterogeneity was assessed using the Incongruence Length Difference (ILD) test of Farris *et al.* (1994,1995) implemented in PAUP* as the Partition Homogeneity Test (PHT). This test uses the incongruence metric I_{MF} of Mickevich and Farris (1981) as a distance measure D, where I_{MF} is defined as the number of homoplasious steps required by each individual data set or partition to explain the shortest trees recovered from analysis of the combined data. This index ranges from 0 (shortest tree recovered from each data set is identical) to 1 (no homoplasy observed in either data set and both data sets produce unique tree topologies). Thus the amount of homoplasy or character conflict within each data set is a vital component of the character conflict between data sets as measured by I_{MF} (Johnson and Soltis, 1998).

The ILD test explicitly evaluates the null hypothesis that characters are randomly distributed between data sets with respect to the phylogenetic

information they contain. An observed D for the two data partitions is calculated, and a null distribution is then generated by randomising data into partitions of sizes equal to the original partitions. D is calculated for the random partitions some number of times, represented by W. The number of replicates for which D from the random partitions is less than for the observed partition is calculated and designated S. The P-value for determining the probability of rejecting the null hypothesis of data partition homogeneity is equal to 1-S/W (Yoder *et al.*, 2001).

In order to generate a null distribution for the various partitions being compared, ILD tests used 200 or 1000 replicates depending on time constraints. Starting trees were obtained by simple stepwise addition sequences of taxa and TBR branch swapping. The partitions compared included first versus second versus third codon positions within each gene, cytochrome *b* versus cytochrome oxidase II sequences and molecular versus morphological data partitions. Prior to implementation of the ILD test, uninformative characters were removed. Removal of these characters is particularly important when there is a large difference in the number of variable characters between the original data partitions (which is fairly common when comparing molecular and morphological data), to ensure that the ILD test does not overestimate the amount of incongruence between data partitions (Lee, 2001).

Topological incongruence

The Shimodaira-Hasegawa (SH) test of alternative tree topologies was implemented in a maximum likelihood framework to statistically evaluate alternative hypotheses of phylogenetic relationships within *Camponotus* (Shimodaira and Hasegawa, 1999). This test, unlike the Kishino-Hasegawa test of alternative tree topologies (Kishino and Hasegawa, 1989) and parsimony-based Templeton tests (Templeton, 1983) allows *a priori* and *a posteriori* hypotheses to be compared. The SH test simultaneously compares multiple tree topologies and corrects the corresponding *P* values to

accommodate the fact that multiple tests have been performed. In addition, it can be used to compare *a posteriori* hypotheses as it readjusts the expectation of the null hypothesis (that two trees are not different) accordingly, a process known as 'centering' (Buckley *et al.*, 2001a). This test was implemented in PAUP* using 1 000 bootstrap replicates and RELL parameter optimisation of the GTR + I + Γ model (test *posNPncd* of Goldman *et al.*, 2000).

V Relative contribution of data partitions in combined analyses

Partitioned Bremer support (PBS) provides an indication of the relative contribution of different data partitions in a combined parsimony analysis to the Bremer support value for each node. The partial Bremer index for each data partition is calculated by subtracting the number of steps for that partition in the most parsimonious tree(s) from the number of steps for that partition in the shortest tree(s) with the node in question. Values of the partial Bremer index may be positive or negative, but the sums of the partial decay indices for a given node equals the overall decay index if the partitions specified are mutually exclusive, but together comprise all of the characters in the original analysis (Baker and DeSalle, 1997; Baker *et al.*, 1998).

TreeRot was used to generate command files for PAUP* containing constraint statements for each node in the shortest or strict consensus tree as well as commands to search for trees not consistent with each of the constraint statements in turn. The command file was then executed in PAUP* and heuristic searches with 20 random addition replicates were performed for each constraint statement. The partitioned Bremer support values were subsequently obtained by parsing the PAUP* log file in TreeRot. In the combined gene analyses, partial decay indices were calculated for the three codon positions of each data partition as well as the two genes in total. In the combined molecular-morphological analyses, molecular and morphological partitions were designated, and PBS values for these two partitions were evaluated. For the molecular comparisons, PBS values were standardized for differences in size between data partitions (large data sets could exhibit more

support simply because they contain more characters) by dividing the PBS values by the number of parsimony informative characters in each data partition (Baker *et al.*, 1998). For the molecular-morphological comparisons, the ratio of the PBS for each data partition was compared to the ratio of the minimum number of steps for each partition on the combined tree to evaluate which partition was providing more support relative to its size. This was done to compensate for the potentially larger information content of the DNA characters (four possible states) compared to the morphological and ecological characters (Baker *et al.*, 1998; Remsen and DeSalle, 1998; Baker *et al.*, 2001).

VI Hypothesis testing

Constraint analyses

Support for the monophyly of *Camponotus* and subgeneric classifications was assessed by constraint analysis. Constraint topologies were constructed in MacClade using the Tree Editor utilities. ML searches were then performed using the optimal sequence evolution model for that particular data set with the 'Topological constraints enforced' option in effect, keeping topologies compatible with the constraint topology. The optimal topologies resulting from constraint searches were then compared with the appropriate unconstrained topologies using the SH test.

Testing for a molecular clock

The existence of a molecular clock for the nucleotide sequences was evaluated using a likelihood ratio test (Felsenstein, 1981). This hypothesis is satisfied if DNA substitutions follow a Poisson process and the mean rate of substitution has remained constant in different lineages. A ML heuristic search was performed for each nucleotide data set with and without the 'molecular clock enforced' option activated in PAUP*, using model parameters estimated by Modeltest for each data set. The clock-like model is a special case of the more generalised model, with rates of substitution amongst branches

constrained to be the same rather than free to vary, and is therefore the null hypothesis.

The significance of the difference in log likelihoods of the two trees was evaluated using the likelihood ratio test statistic $\delta = 2 \log \Lambda$, with s - 2 degrees of freedom where s is the number of taxa. The statistic is evaluated against critical values of the χ^2 distribution (Huelsenbeck and Rannala, 1997).

VII Phylogenetic informativeness of morphological characters

The Permutation Tail Probability (PTP) test was implemented in PAUP* to compare the distribution of steps obtained from 1000 random reassignments of the morphological character states with the number of steps required to optimise each morphological character, as well as the entire morphological data set, on an independent molecular phylogeny (Lee *et al.*, 1996). The molecular phylogeny used was the 50% majority rule Bayesian consensus topology recovered after Bayesian analysis of the complete combined molecular data set under a GTR + I + Γ model of sequence evolution (Bayes+I^{all} topology) (see Chapter 5 Part C for details).

VIII Congruence of phylogenetic methods and models

The congruence of maximum parsimony with Bayesian inference was evaluated by assessing the number of shared nodes and nodal support for various MP analyses. For each nucleotide data set, the Bayesian tree inferred using a GTR + I+ Γ model of nucleotide evolution was used as the reference tree. The congruence of ML methods with Bayesian inference was assessed by comparing parameter value estimates for models of nucleotide evolution (Leache and Reeder, 2002). Estimates of nucleotide substitution parameters, branch lengths, topological estimates and posterior probability values were also compared between Bayesian analyses implementing GTR + I + Γ or GTR + SSR models of sequence evolution. The correlation between posterior probability values under different models of sequence evolution was assessed using a Spearmans Rank Correlation test using SPSS v9.0 (SPSS Inc. 1998).

Chapter 4

RESULTS AND DISCUSSION

In this chapter, the results of the diverse range of phylogenetic analyses implemented are presented, and some discussion of these results is provided.

Part A: Sequence Evolution

Base Composition bias and sequence statistics

The sequence data showed patterns of variation typical of insect mitochondrial DNA, namely A+T richness and unequal distribution of varied sites among character partitions (Table 4.1). Base composition bias was most pronounced at third codon positions of both genes, with the C value for these positions approximately twice that displayed at the first and second codon positions. This is consistent with previous studies (Simon *et al*, 1994; Dowton and Austin, 1997; Whitfield and Cameron, 1998; Mardulyn and Whitfield, 1999; Cameron and Mardulyn, 2001). The pronounced A+T bias at third codon positions effectively renders those sites two-state sites, with the implication that homoplasy will occur more frequently at these sites. The overall bias was slightly larger for cytochrome oxidase II (C = 0.34) than for cytochrome *b* (C = 0.29), with the values obtained for cytochrome oxidase II at first, second and third codon positions identical to those documented for another ant species (*Lasius sp.*) (Liu and Beckenbach, 1992).

Table 4.1. Patterns of nucleotide substitution for the two partitions of mitochondrial sequence data (mtDNA).

MtDNA Subset	Base Pairs	Nucleotide (%) A	T	C	G	A+T	Bias[A]	No. Variable sites (% of total in category)	No. of parsimony-informative (PI) sites (% of total in category)	No. of constant sites
Cyt *b*										
All	660	32.3	39.8	18.4	9.5	72.1	0.29	310	258	350
1st pos	220	37.0	30.5	18.1	14.3	67.5	0.23	74 (24%)	58(22%)	146 (42%)
2nd pos	220	25.0	42.7	20.2	12.2	67.7	0.24	34 (11%)	22 (9%)	186 (53%)
3rd pos	220	34.7	46.3	16.9	2.0	81.0	0.41	202 (65%)	178 (69%)	18 (5%)
COII										
All	484	35.8	39.8	16.0	8.3	75.6	0.34	229	187	255
1st pos	161	39.8	30.0	16.7	13.5	69.8	0.26	58 (25%)	45(24%)	103 (40%)
2nd pos	162	27.0	43.8	18.9	10.3	70.8	0.28	31 (14%)	21(11%)	131 (51%)
3rd pos	161	40.8	45.7	12.4	1.1	86.5	0.49	140 (61%)	121(64%)	21 (9%)

[A] Base composition bias C = 2/3 Σ [c*i* – 0.25], where c*i* is the frequency of the *i*th base (Prager and Wilson ,1988).

Cytochrome oxidase II and cytochrome *b* displayed similar patterns of codon variability, with the heterogeneity in variability among codon positions clearly evident (Table 4.1 and Figure. 4.1). Third codon positions of both amplicons were highly variable followed by first and then second codon positions. This observed pattern of variability is exactly what is predicted when the degree of selective constraints on the various codon positions is considered. First position substitutions tend to result in more conservative amino acid substitutions than substitutions at second codon positions, and are therefore expected to display more variability than second codon sites. Third position changes largely result in silent substitutions with no change reflected at the amino acid level, and are therefore expected to be the most variable (Irwin *et al.*, 1991).

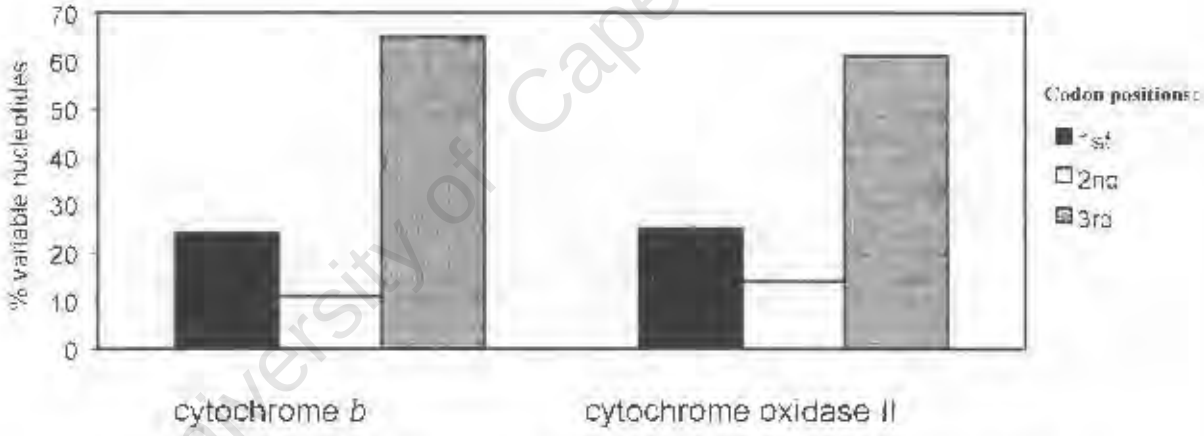


Figure 4.1 Proportion of variable nucleotides at each codon position for the cytochrome *b* and cytochrome oxidase II amplicons. Percentages reflect the number of nucleotides observed to vary divided by the total number of variable nucleotides observed.

Variability was also observed at the amino acid level, with 60 parsimony informative (PI) amino acid sites observed for all cytochrome *b* sequences and 49 PI sites observed for all cytochrome oxidase II sequences.

The inferred pattern of nucleotide variability across the portions of the cytochrome oxidase II and cytochrome *b* genes sequenced is presented in Figures 4.2 A and B. Patterns of variability across the ant cytochrome *b*

fragment sequenced are consistent with those documented by Irwin *et al*, (1991). Transmembrane domain 2 (TM2) displays low variability, transmembrane domain 5 (TM5) displays high variability while the third and fourth transmembrane regions display intermediate levels of variability, as measured by the number of nucleotide substitutions. The low levels of variability observed for TM2 are due to strong functional constraints against mutational change in this region in which one of the heme-ligating histidines is situated (Howell, 1989; Irwin *et al.*, 1991).

The region of cytochrome oxidase II sequenced in this study shows a wider range of variability than is present in the cytochrome *b* region, with a greater proportion of sites evolving at both high (≥ 18 nucleotide substitutions; 0.28 versus 0.21) and low (≤ 4 nucleotide substitutions; 0.09 versus 0.01) rates in cytochrome oxidase II compared to the cytochrome *b* fragment. In particular, the putative mitochondrial matrix and intracristal domains of cytochrome oxidase II seem to be comprised of both rapidly and slowly evolving sites, although whether this pattern holds true across phylogenetically diverse taxa has not been explicitly evaluated (Howell, 1989).

Nucleotide sequence divergence

Uncorrected sequence divergence estimates among ingroup taxa ranged from 0.2% to 22.1% with a mean of 18.3% for cytochrome *b*, and 0% to 21.5% with a mean of 14.8% for cytochrome oxidase II. Tamura-Nei (1993) corrected distances values ranged from 0.2% to 27.8% with a mean of 22% for cytochrome *b* and 0% to 26.8% with a mean of 17.6% for cytochrome oxidase II (Appendix IV). Corrected distances between the ingroup taxa and the two outgroups, *Formica lugubris* and *Oecophylla smaragdina*, ranged from 17.0% to 40.5% (cytochrome oxidase II) and 21.4% to 35.7% (cytochrome *b*). Amino acid divergence estimates were high for both genes, ranging from 0.4% to 19.5% with a mean divergence of 13.9% for ingroup comparisons of cytochrome *b* amino acid sequences, and 0% to 20.0% with a mean divergence of 11.3% for cytochrome oxidase II ingroup sequences

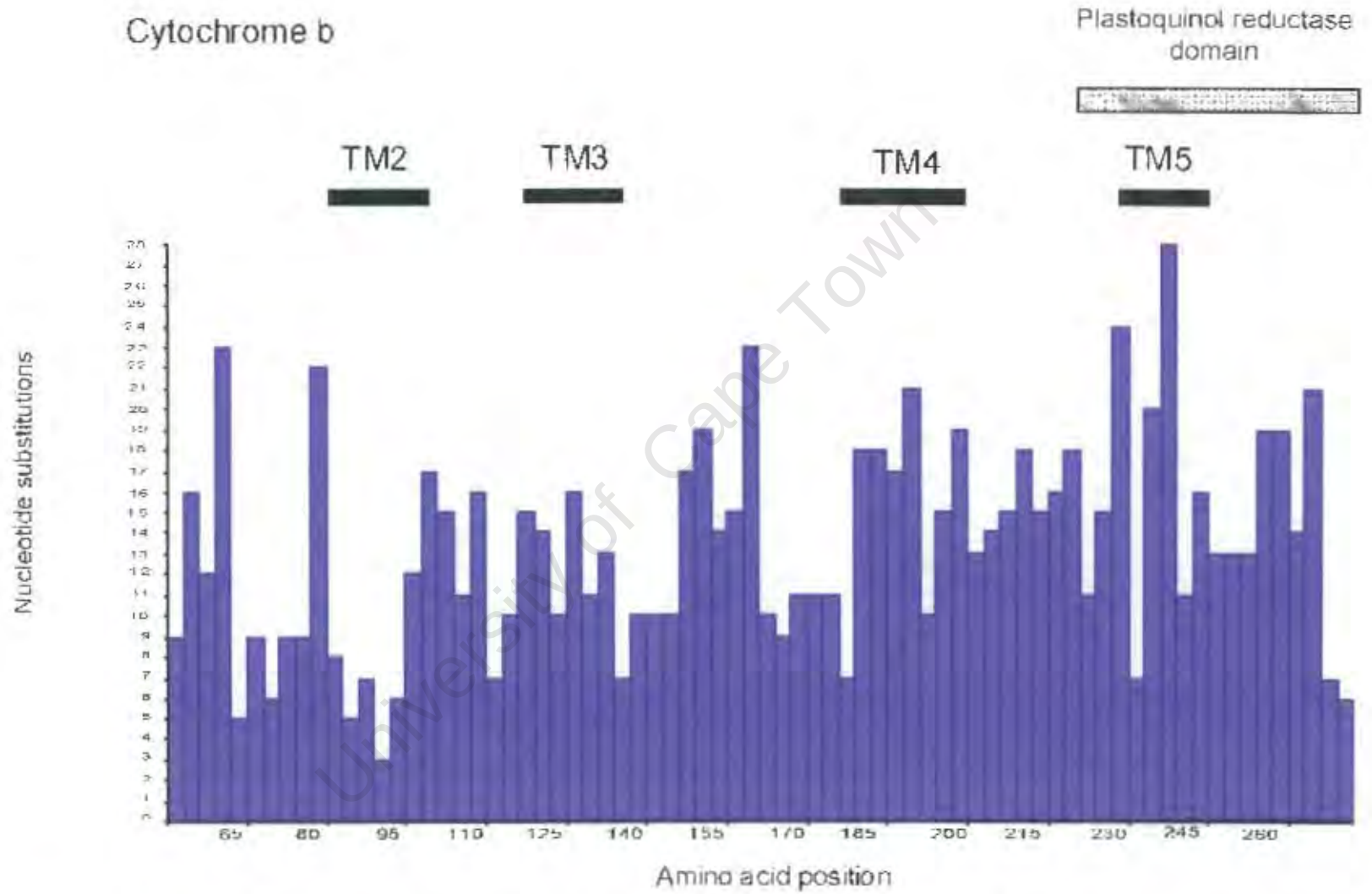


Figure 4.2A. Pattern of variation of nucleotide substitution across the cytochrome *b* gene fragment. Variability is illustrated by plotting the number of inferred nucleotide substitutions in a moving window of three amino acids numbered relative to the human cytochrome *b* protein. The location of the transmembrane and functional domains are indicated at the top of the figure.

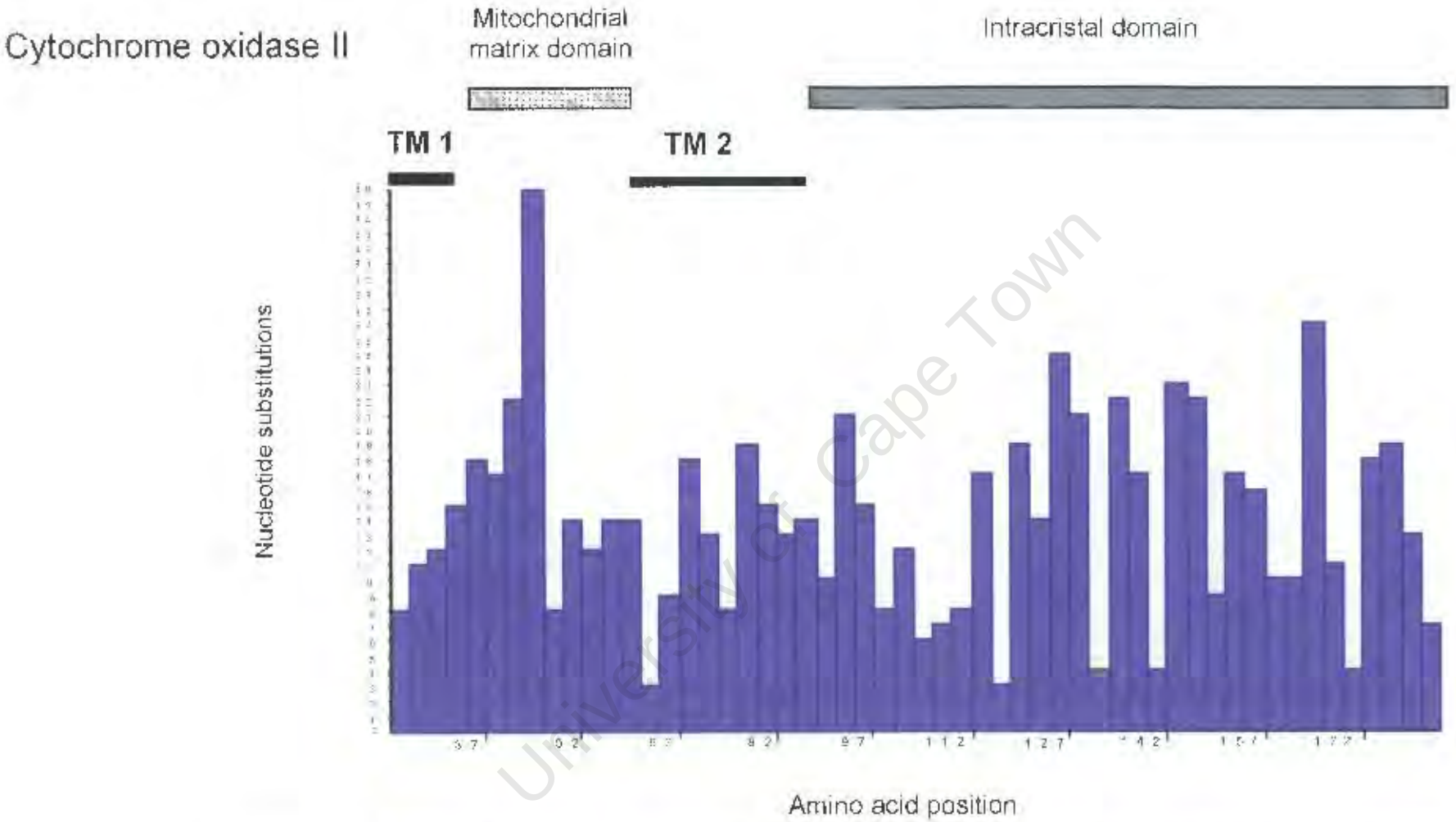


Figure 4.2B. Pattern of variation of nucleotide substitution across the cytochrome oxidase II gene fragment. Variability is illustrated by plotting the number of inferred nucleotide substitutions in a moving window of three amino acids numbered relative to the human cytochrome oxidase II protein. The location of the transmembrane and functional domains are indicated at the top of the figure.

Codon usage

Codon usage for both mitochondrial proteins reflects the typical bias against codons ending in guanine previously documented in insect mitochondrial protein-coding genes (Liu and Beckenbach, 1992). Within the framework of this study, only 2% of the cytochrome *b* codons and 1% of the cytochrome oxidase II codons terminated in guanine. At the amino acid level, the effect of amino acid occurrence varying according to base composition can be expressed in terms of the ratio of the total occurrences of "G+C" (P, A, R, G) and "A + T" (F, I, M, Y, N, K) amino acids (Crozier and Crozier, 1993). For the cytochrome *b* gene of *Camponotus*, the ratio of "GC" to "AT" amino acid occurrences was 0.34, whereas this ratio was 0.25 in the cytochrome oxidase II gene. This indicates a greater effect of base composition bias on amino acid composition in the cytochrome oxidase II gene fragment compared to the cytochrome *b* gene fragment.

Base frequency stationarity

The hypothesis of base composition homogeneity could not be rejected for either cytochrome *b* or cytochrome oxidase II when all codon positions were considered using all sequence sites (Table 4.4). When parsimony informative sites were evaluated separately, however, base composition homogeneity was rejected for both cytochrome *b* and cytochrome oxidase II fragments. This appears to be due to the A+T-rich third codon positions, which were the only codon positions that rejected the hypothesis of homogeneity, whether all sites were considered, or parsimony informative sites only. Results obtained with the two outgroup taxa excluded followed the same trends detailed above. Therefore, these two outgroup taxa were included in all subsequent phylogenetic analyses.

Heterogeneity in base composition at third codon positions appears to be prevalent in hymenopteran mitochondrial genes, with these sites rejecting the hypothesis of base composition homogeneity (Brady *et al.* 1999; Mardulyn and Whitfield, 1999; Cameron and Mardulyn, 2001; Weiblen, 2001). The ILD test, however, could not reject the hypothesis of congruence of first, second and third codon positions for either the cytochrome *b* or cytochrome oxidase II sequences (data set A: $P = 0.52$; data set B: $P = 0.95$).

Table 4.2. Results of χ^2 base composition homogeneity tests with outgroups included.

	All Sites			Pi Sites Only		
	χ^2	*P*	d.f.	χ^2	*P*	d.f.
Cyt *b*						
All	49.7	0.64	54	100.8	0*	54
1st	17.2	1.00	54	60.7	0.25	54
2nd	7.6	1.00	54	61.0	0.24	54
3rd	82.4	0.01*	54	80.4	0.01*	54
COII						
All	57.6	0.56	60	138.6	0*	60
1st	12.5	1.00	60	38.1	0.98	60
2nd	9.6	1.00	60	27.6	1.00	60
3rd	141.9	0*	60	147.0	0*	60

*Indicates significance at $P < 0.05$

Mutational saturation

Third codon positions of both genes appear to be saturated with respect to both transitional and transversional mutations (Figures 4.3 B and C). This saturation is apparent both in ingroup-outgroup comparison as well as within-ingroup comparisons. Transitions at the first codon position of cytochrome *b* appear to be reaching saturation, in contrast to transition mutations at cytochrome oxidase II first codon positions (Figure 4.3 A). The relationship between observed and expected sequence divergences appears linear for first and second position transversion mutations for both genes, and second

codon transitions for both genes, indicating that saturation has not yet occurred at these positions (plots not shown).

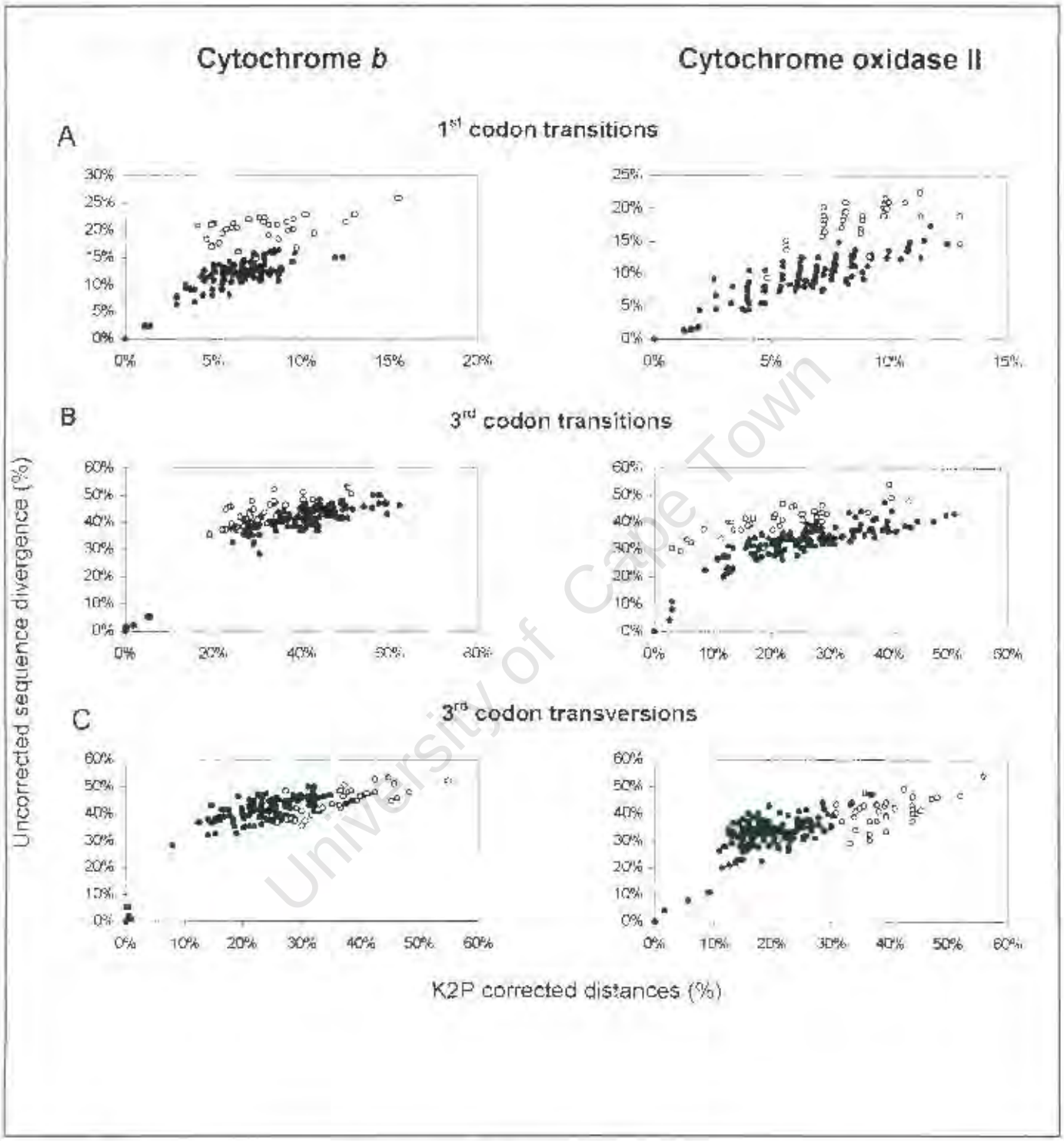


Figure 4.3. Saturation plots for cytochrome *b* and cytochrome oxidase II gene fragments. The K2P corrected sequence divergence values are plotted on the x-axis, with uncorrected divergence values plotted on the y-axis. A) first codon transitions, B) third codon transitions and C) third codon transversions. Solid circles represent comparisons within the ingroup taxa, open circles represent ingroup-outgroup comparisons.

The observation that third codon positions of both genes are saturated is further supported by frequency distribution plots of the number of steps per nucleotide site (Figure 4.4). For both cytochrome *b* and cytochrome oxidase II first and second codon positions, only a small number of characters required more than three steps on the most parsimonious trees, with the majority of characters appearing invariant. Those that did vary required only one, two or three steps. When only third codon positions were considered, the distribution shifted, with the majority of third codon nucleotides requiring three or more steps on the most parsimonious tree, while 10% of cytochrome *b* third codon positions and 24% of cytochrome oxidase II third codon positions require six or more steps on the most parsimonious tree (Figure 4.4).

Based on the mutational saturation plots and frequency distribution plots, the third codon positions of both gene fragments appear to be saturated, and may therefore be misleading in subsequent phylogenetic analyses (Mardulyn and Whitfield, 1999).

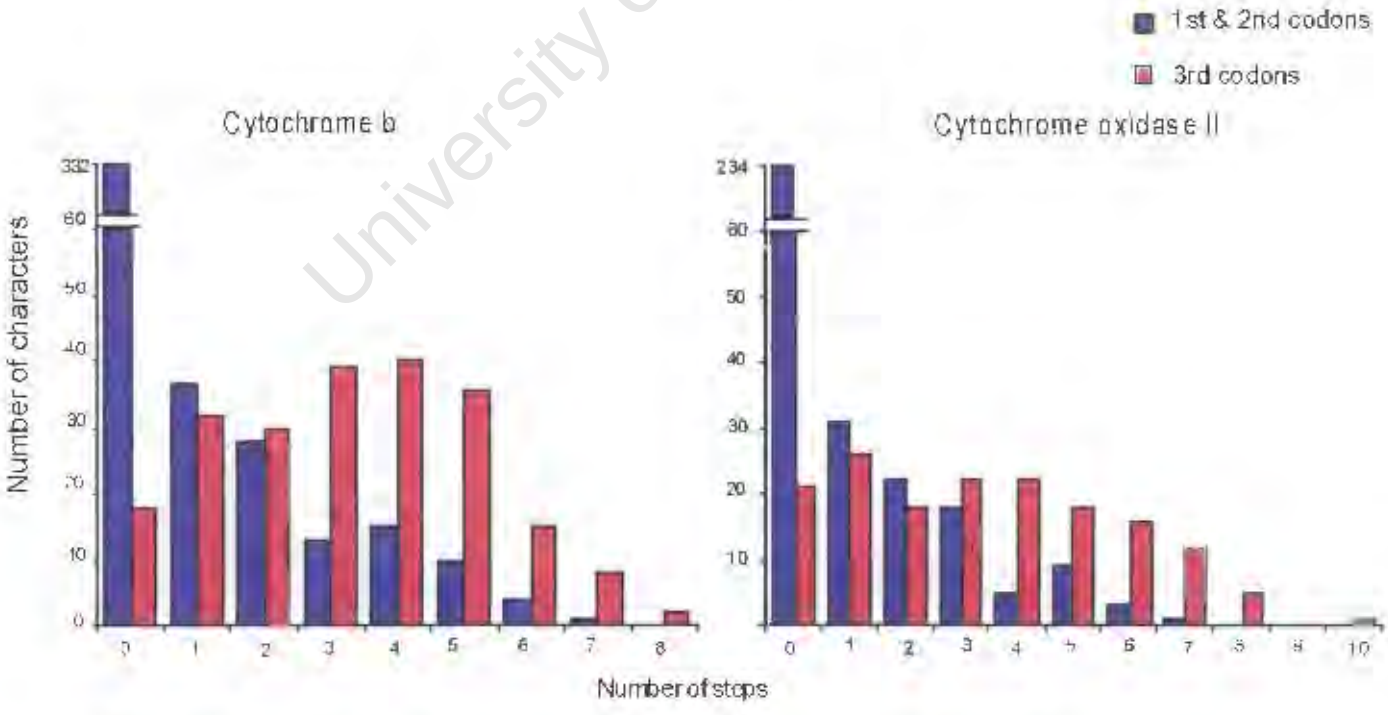


Figure 4.4. Distribution of the number of changes per character for cytochrome *b* and cytochrome oxidase II sequences, as reconstructed by parsimony.

Phylogenetic signal

There was significant non-random structure in both the cytochrome *b* and cytochrome oxidase II datasets, reflected by significantly left-skewed *g*1 values (Table 4.3). The third codon positions of cytochrome *b* had the most negative *g*1 values of all three codon positions in this amplicon, indicating strong phylogenetic signal in this data partition despite apparent homoplasy and mutational saturation of this set of sites (Figures. 4.3 and 4.4). The third codon positions of cytochrome oxidase II also provided significant phylogenetic signal, although the *g*1 value was less negative than that for the first and second codon positions. This indicates that the apparent mutational saturation at these sites could adversely affect the resolving power of these cytochrome oxidase II third codon characters in subsequent phylogenetic reconstructions.

Table 4.3. Results of *g*1 skew tests for cytochrome *b* and cytochrome oxidase II sequences.

	Cytochrome *b*		Cytochrome oxidase II	
Sites	Dataset A	Dataset C	Dataset *B*	Dataset D
ALL	-1.00*	-1.18*	-0.73*	-1.02*
1st codon	-0.79*	-0.68*	-0.73*	-1.11*
2nd codon	-0.68*	-0.62*	-1.20*	-1.56*
3rd codon	-0.95*	-1.2*	-0.67*	-0.35*

* indicates significance at $P < 0.01$ (Hillis and Huelsenbeck, 1992).

Part B: Phylogenetic analyses of single gene partitions

I *Tree reconstruction: complete cytochrome b data set*

(i) Model-based analyses

Bayesian analyses

For all cytochrome *b* Bayesian runs, the mean log likelihoods of the sample points reached equilibrium after approximately 1 X 10^4 generations, indicating that after this time model parameters were being sampled according to their posterior probabilities (Huelsenbeck *et al.*, 2001)(Figure 4.5).

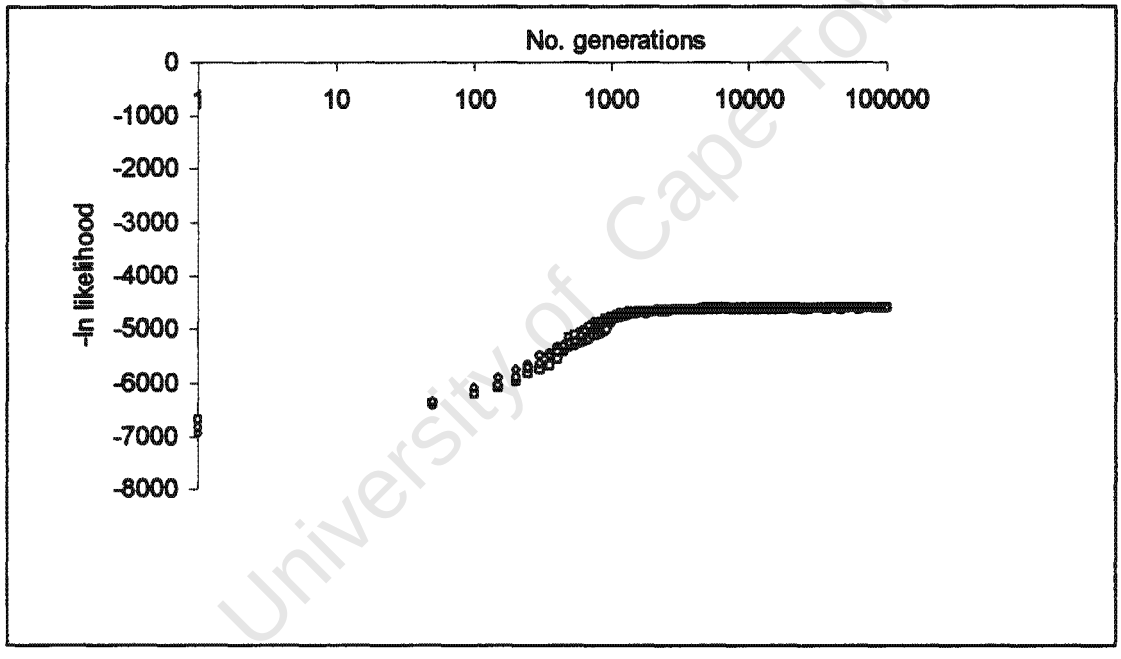


Figure 4.5. Log plot of -ln *L* values for cytochrome *b* from three independent Bayesian runs using a GTR + I + Γ model. The results of the three separate analyses are superimposed to illustrate convergence in log likelihood values after 1 x 10^3 to 1 x 10^4 generations. Points are only plotted up to the 100 000^{th} generation as stationarity had been reached well before this point (refer to text).

The three runs incorporating a gamma distribution (Γ) and invariant sites (I) to model rate heterogeneity all produced an identical consensus tree topology (hereafter referred to as BayesΓ+I^{cytb} tree) with similar posterior probability support values for all nodes and mean ln *L* = -4588 (Figure. 4.6). The Bayesian runs modelling rate heterogeneity using three site specific rates

(SSR_3) corresponding to the three codon positions each converged on the same consensus tree with highly congruent posterior probability values, hereafter referred to as the BayesSSR_3^{cytb} tree, with mean ln *L* = -4591 (not shown). This tree was very similar in topology to the Bayes$\Gamma+I^{cytb}$ tree, except for the placement of *C. klugii*. In the BayesSSR_3^{cytb} tree, *C. klugii* is sister to the clade containing ((*C. fulvopilosus* +*C. storeatus*), (*C. bifossus* + *C. sp.* 12), *C.sp.*7), while in the Bayes$\Gamma+I^{cytb}$ tree, its position is unresolved.

Camponotus bifossus and *C. sp.* 12 form a clade (Figure 4.6, node 2) that is significantly supported in both Bayesian analyses (bpp[1] = 1). Other significantly supported clades include (*C. sericeus* (*C. acvapimensis* + *C. chrysurus*)) (Figure 4.6, node 10) and (*C. sp. 11* (*C. rufoglaucus* + *C. cinctellus*)) (Figure 4.6, node 6). There is significant support for a sister-group relationship between *Polyrhachis schistacea* and *C. nasutus* (Figure 4.6, node 12), and for a sister group relationship between this clade and *C. cuneiscapus* (Figure 4.6, node 13). A sister group association between *C. fulvopilosus* and *C. storeatus,* both of the subgenus *Myrmopiromis*, was strongly favoured in both Bayesian analyses (Figure 4.6, node 1).

The posterior probability support values for some nodes differ substantially between the two Bayesian consensus trees. The monophyly of the subgenus *Myrmosericus* represented by *C. rufoglaucus* and *C. cinctellus* (Figure 4.6, node 5) has significant posterior probability in the BayesSSR_3^{cytb} tree, compared with the non-significant support for this clade in the Bayes$\Gamma+I^{cytb}$ tree. Similarly, the association of species in the subgenus *Myrmespera* (*C. sp.* 10, *C. nasutus*, *C. cuneiscapus*) with *P. schistacea* in a clade (Figure 4.6, node 14) is more probable in the Bayesian analysis assuming a GTR + SSR_3 model than the analysis based on a GTR + I + Γ model. However, this clade is not significantly supported in either analyses.

[1] bpp = Bayesian posterior probability

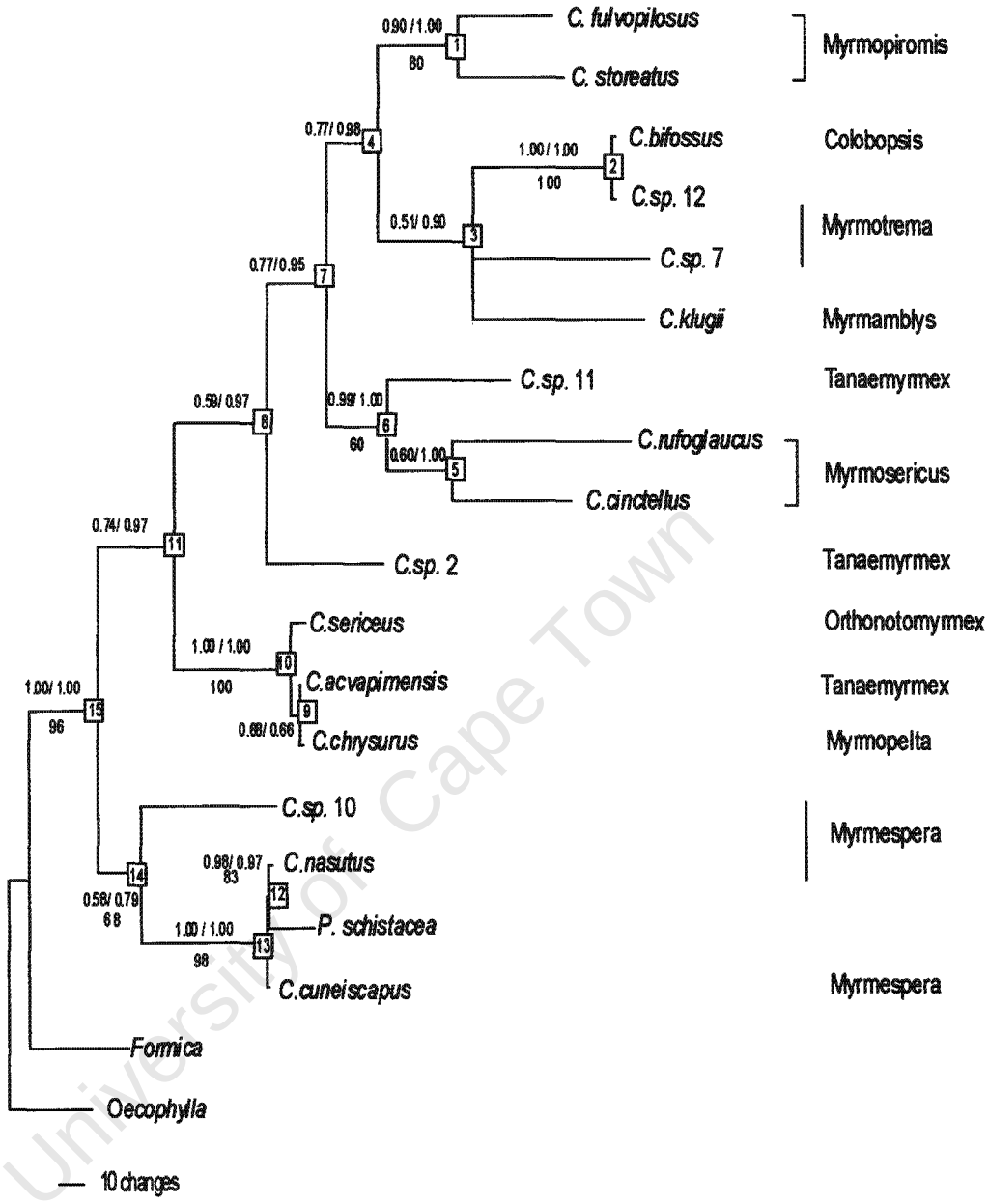


Figure 4.6. The 50% majority rule consensus tree from Bayesian analysis of cytochrome *b* sequences using a GTR + I + Γ model of nucleotide evolution (BayesΓ+I^{cytb}). The two numbers adjacent to nodes represent posterior probabilities from two separate Bayesian analyses using a GTR + I + Γ model and GTR + SSR_3 model respectively. Numbers below branches indicate non-parametric bootstrap proportions greater than 50%. Branch lengths are drawn proportional to the number of changes as indicated by the scale bar. Subgeneric classifications are indicated to the right of the tree, with closed brackets representing monophyletic groups.

Bayesian estimates of first, second and third codon position relative substitution rates indicated that third codon positions are the most rapidly evolving, followed by first and then second codon positions (Table 4.4). This is to be expected given the redundancy of the genetic code. The Bayesian analysis incorporating rate heterogeneity using a gamma distribution and invariant sites provided a better fit to the observed sequences than accounting for rate heterogeneity using three codon specific rates. This is reflected by the lower mean likelihood score for the latter topology (Table 4.4)

Maximum likelihood analyses

Hierarchical likelihood ratio tests indicated that the GTR + I + Γ model was the most appropriate model of nucleotide sequence evolution for the observed *Camponotus* cytochrome *b* sequences. One most likely tree topology (ln *L* = -4560) was obtained using the GTR + I + Γ model of nucleotide evolution with parameters estimated on an initial neighbour-joining tree using Modeltest[2] (hereafter referred to as ML^{cytb}). This topology was stable, as successive searches with parameter values re-estimated on the previous ML tree yielded the same optimal topology.

This tree was almost identical in topology to the $BayesΓ+I^{cytb}$ tree, except for the placement of the clade (*C. bifossus* + *C. sp.* 12), which form a sister group to (*C. fulvopilosus* + *C. storeatus*) in the ML tree, with *C. sp.* 7 and *C klugii* successively basal to this clade. However, this relationship was not statistically supported (bootstrap < 50%), making the ML^{cytb} tree fully compatible with the $BayesSSR_3^{cytb}$ tree, where a sister-clade relationship between (*C. fulvopilosus* + *C. storeatus*) and (*C. bifossus* + *C. sp. 12*) is not evident.

Branch support

The ML bootstrap consensus tree was less resolved than either of the two Bayesian consensus trees, with only eight out of 16 possible nodes resolved compared with 15 for the $BayesSSR_3^{cytb}$ tree and 14 for the $BayesΓ+I^{cytb}$

[2] $\pi A = 0.33$; $\pi C = 0.17$; $\pi G = 0.09$; $\pi T = 0.40$; $r_{AC} = 2\ 185\ 328$; $r_{AG} = 4\ 521\ 738$; $r_{AT} = 2\ 868\ 410$; $r_{CG} = 1\ 103\ 286$; $r_{CT} = 28\ 937\ 992$; $r_{GT} = 1$; $p_{inv} = 0.43$; $\alpha = 0.90$.

consensus tree. All clades that were significantly supported in both Bayesian analyses received bootstrap support (Figure 4.6, nodes 2, 6, 10, 12, 13, 15). However, no clear relationship between significant Bayesian posterior probabilities (bpp) and bootstrap values was evident, with bootstrap values ranging from 60% to 100% corresponding to bpp values ≥ 0.95.

Table 4.4. Nucleotide substitution model parameter means, variances and 95% credible regions (C.R.) for the cytochrome *b* complete dataset using the GTR model of nucleotide substitution with two different rate heterogeneity models. Values represent the average across three (GTR + I + Γ) or two (GTR + SSR_3) runs.

	Models of sequence evolution					
	GTR + I + Γ			GTR + SSR_3		
Parameter	Mean	Variance	95% C.R.	Mean	Variance	95% C.R.
-ln *L*	4588	30	4600, 4578	4591	28	4603, 4582
Branch lengths	5.34	0.83	3.89, 7.41	2.67	0.02	2.43, 2.95
[a]r_{GT}	1.00	0	1.00, 1.00	1.00	0	1.00, 1.00
r_{CT}	74.54	349.00	37.12, 99.44	61.01	304.50	32.48, 87.10
r_{CG}	3.51	4.55	0.73, 9.35	5.62	6.71	1.76, 11.69
r_{AT}	7.41	5.75	3.25, 12.24	11.21	11.14	11.85, 16.87
r_{AG}	15.29	33.66	6.87, 30.58	21.90	24.66	10.96, 34.76
r_{AC}	5.97	13.77	2.17, 11.24	21.36	11.97	5.27, 16.86
[b]πA	0.33	0	0.30, 0.36	0.32	0	0.29, 0.34
πC	0.17	0	0.16, 0.19	0.17	0	0.16, 0.19
πG	0.08	0	0.06, 0.10	0.09	0	0.07, 0.11
πT	0.40	0	0.37, 0.43	0.41	0	0.38, 0.43
[c]α	0.86	0.04	0.50, 1.28	-	-	-
[d]p_{inv}	0.42	0	0.33, 0.48	-	-	-
[e]1st codons	-	-	-	0.44	0	0.38, 0.50
2nd codons	-	-	-	0.13	0	0.10, 0.16
3rd codons	-	-	-	2.42	0	2.36, 2.49

[a] $R_{X\leftrightarrow Y}$, relative rate parameters
[b] π, estimated frequency of the nucleotide
[c] α, gamma shape parameter
[d] p_{inv}, proportion of invariant sites
[e] codon specific rate

The ML^{cytb}, $BayesSSR_3^{cytb}$ and $BayesI^{cytb}$ tree topologies did not represent statistically significant alternative explanations of the data, as evaluated using the Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa, 1999) (Table 4.5).

Table 4.5. Shimodaira-Hasegawa test results for comparison of Bayesian and ML tree topologies inferred from cytochrome *b* sequences.

Tree topology	-ln *L*	Difference -ln *L*	*P*
$BayesI^{cytb}$	4560.74	0.30	0.50
$BayesSSR_3^{cytb}$	4561.25	0.80	0.64
ML^{cytb}	4560.45	Optimal	-

Comparison of Bayesian and ML model parameter estimates

The base frequency estimates obtained using ML were almost identical to the mean base frequency estimates obtained from Bayesian analyses of the cytochrome *b* sequences using a GTR + I + Γ model. This is expected when using flat priors (Larget and Simon, 1999; Leache and Reeder, 2002). The p_{inv} and α values estimated by ML fell within the 95% credible region estimated by Bayesian analysis using the GTR + I + Γ model (Table 4.4). The rate matrix values estimated from the data using ML and Bayesian methods differed by orders of magnitude. However, the ranking order of the relative rate parameters was identical, with C↔T transitions predominating, followed by A↔G, A↔T, A↔C, C↔G and finally G↔T changes. The alpha parameter estimate of 0.86 produces an *L*-shaped distribution (Appendix V), indicating that most sites in the cytochrome *b* amplicon have low substitution rates, while a few sites have very high mutation rates (Yang, 1996a; Lewis, 1998). A large proportion of sites also appear to be invariable (p_{inv} = 0.43). Thus by applying a Γ and I model of rate heterogeneity with the parameter values set to 0.86 and 0.43 respectively, most sites will be considered incapable of accepting substitutions or very slowly evolving, with only a small number of sites evolving rapidly.

Bayesian estimates of mean branch lengths using the GTR+ I + Γ model were almost double that obtained using codon-specific rates, and had a much greater associated variance (Table 4.4).

LogDet transformation

The bootstrap LogDet/ME consensus tree topology was fully congruent with the ML^{cytb} and Bayesian trees, with those clades with high bootstrap support present on the ML^{cytb} and Bayesian trees also present in the distance tree. The congruence between the tree topology of the ME/LogDet consensus tree and the Bayesian and ML trees indicates that although base frequency homogeneity was rejected for cytochrome *b* when third codon positions and parsimony-informative sites were considered, it does not appear to be influencing topology reconstruction. However, this statement is made with the proviso that the LogDet transformation is adequately correcting for base composition heterogeneity across the tree.

(ii) Parsimony analyses

The results of various parsimony analyses are presented in Table 4.6. Numbered nodes correspond to nodes labelled on the BayesΓ+I^{cytb} tree (Figure 4.6). In general, nucleotide weighting schemes that assigned a greater cost to transitions increased congruence with the Bayesian reference topology relative to the unweighted MP tree. Three weighting schemes yielded topologies identical to the BayesΓ+I^{cytb} tree. These weighting schemes were: (i) downweighting transitions by a factor of four; (ii) analysing transversions only; and (iii) downweighting T↔C transitions six-fold. Trees recovered after downweighting A↔T transversions at all or third codon positions only, and successive weighting of characters by their rescaled consistency indices, were less congruent with the Bayesian topology than those produced under equal weights. Trees based on amino acid characters were less resolved than the nucleotide-based topologies, although the number of nodes that obtained bootstrap support were comparable. Applying a protein parsimony (Protpars) matrix to the amino acid data improved nodal resolution compared to unweighted protein parsimony analysis.

All bootstrap consensus parsimony tree topologies were compatible with the topologies in Figure 4.6 when considering nodes with medium to high bootstrap support. Generally, the parsimony bootstrap consensus trees were less resolved than the MP or strict consensus topologies, reflecting the lack of statistical support for certain nodes (Table 4.6).

The placement of *C. sp.* 7, *C. sp.* 2 and *C. klugii* was subject to a high degree of uncertainty, reflected by the lack of statistical support for grouping of these taxa with any others in any of the MP trees (nodes 3, 4 and 8). Certain clades were well supported in all parsimony analyses. These include the sister taxa relationships of (*C. bifossus* + *C. sp.* 12) (node 2), and (*C. fulvopilosus* + *C. storeatus*) (node 1). The sister group relationship between (*C. acvapimensis* + *C. chrysurus*) (node 9) was also supported by weak to medium bootstrap values and Bremer decay values of three or more in the majority of MP searches, with the association of *C. acvapimensis, C. chrysurus* and *C. sericeus* strongly supported in all analyses (node 10). The association of *P. schistacea* with *C. nasutus* received weak to moderate bootstrap support (node 12), whereas there was strong bootstrap support in all MP trees for an association between (*C. cuneiscapus,* (*P. schistacea* + *C. nasutus*))(node 13), with *C. sp.* 10 sister to this grouping with weak to medium support (node 14).

CI values ranged from 0.34 to 0.57,and RI values ranged from 0.40 to 0.69.

Table 4.6. Congruence of cytochrome *b* MP topologies with the BayesΓ+I^{cytb} topology.

Nodes in BayesΓ+I^{cytb} tree	Unweighted	4:1 Tv:Ti	1st 1:1 Tv:Ti 2nd 2:1 Tv:Ti 3rd 14:1 Tv:Ti	Ti = 0 all positions	Ti = 0 3rd codon positions	A↔T downweighted 2:1 ALL position	A↔T downweighted 2:1 3rd positions	T↔C downweighted 6:1	Successive weighting	6-P parsimony	Unweighted protein parsimony	ProtPars parsimony
1	●●	●●	●●	●●	●●●	●●	●	●●	●●●	●●	●●	●●●
2	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
3	+	+	+	+	-	+	-	+	●●●	-	-	-
4	+	+	+	+	+	+	-	+	●●	+	-	+
5	-	+	-	+	-	-	-	+	-	+	-	-
6	+	●	+	●	●	-	+	●	●●●	●	-	-
7	+	+	+	+	+	-	-	+	●	+	-	-
8	●	+	+	+	+	-	-	+	●●	+	-	-
9	●●	●●	●●	●	●●	●●	●●	●●	●●	●	●	●
10	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
11	+	+	+	+	+	-	-	+	●●	+	-	+
12	-	●●	●	●●	●	-	-	●●	-	●●	-	●
13	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
14	●	●●	●●	●●	●●	●	●	●●	-	●●	●●	●●●
15	●●●	●●●	●	●●●	●●●	●●●	●●	●●●	●●●	●●●	●●●	●●●
Length	1025	2366	5244	426	560	1772	1608	3581	224	1428	243	273
CI	0.39	0.38	0.35	0.34	0.39	0.37	0.41	0.40	0.57	0.37	0.57	0.52
RI	0.45	0.50	0.49	0.52	0.53	0.40	0.59	0.50	0.69	0.52	0.51	0.54
No. MP trees	5	2	2	1	12	1	2	2	1	2	8	1
No. congruent nodes	13/15	15/15	14/15	15/15	13/15	9/15	8/15	15/15	12/15	14/15	7/15	10/15
No. congruent nodes with bootstrap ≥ 50%	8/15	9/15	8/15	9/15	9/15	7/15	7/15	9/15	12/15	9/15	7/15	8/15

Note: Tree statistics given for each analysis are based on the bootstrap consensus topology.
●●● indicates strong bootstrap support (90% - 100%) and a Bremer support value ≥ 10
●● indicates medium bootstrap support (70-89%) and Bremer support of > 3
● indicates weak bootstrap support (50-69%)
+ indicates that the node was present in the MP tree/strict consensus tree but with bootstrap < 50%
- indicates that clade was not present in the MP/strict consensus tree or the bootstrap consensus tree

(iii) Phylogenetic relationships

The monophyly of *Myrmopyromis* (*C. fulvopilosus* + *C. storeatus*) was favoured in all analyses (Figure 4.6, node 1). This clade was present in ≥ 90% of all Bayesian topologies, and received moderate to strong ML and parsimony bootstrap support. The two representatives of the subgenus *Myrmosericus* (*C. rufoglaucus* + *C. cinctellus*) formed a monophyletic group in both Bayesian consensus trees, with this relationship also recovered in some of the MP searches (Figure 4.6, node 5). However, this relationship was only significantly supported when a GTR + SSR_3 model was assumed. In all analyses, species of subgenus *Myrmespera* formed a clade that also contained *P. schistacea* (Figure 4.6, node 14). This indicates that the genus *Camponotus* is not monophyletic, with *P. schistacea* arising from within the subgenus *Myrmespera*. However, robust support for this hypothesis is lacking, reflected by the non-significant posterior probability values, and weak to moderate levels of bootstrap support for this node. *Tanaemyrmex* is polyphyletic, with species of this subgenus widely distributed across the tree in all analyses. The association of *C. bifossus* with *C. sp.* 12 of subgenus *Myrmotrema* was significantly supported in all analyses, with Bayesian posterior probability values of 1.0 for both Bayesian analyses, and 100% bootstrap support estimates for ML and parsimony analyses (Figure 4.6, node 2). The position of *C. sp.* 7, which also belongs to this subgenus, was unresolved in all MP topologies. The hypothesis of ingroup monophyly was never falsified using *Formica* and *Oecophylla* as outgroups to root the tree, with the node to the outgroups strongly supported in all analyses (Figure 4.6, node 15).

II *Tree reconstruction: reduced cytochrome b data set*

(i) Model-based analyses

Bayesian analyses

Bayesian analysis of the reduced[3] cytochrome *b* data set using a GTR+ I + Γ model of sequence evolution resulted in a topology (hereafter referred to as BayesΓ+$I^{cytbsmall}$) that was fully congruent with a pruned BayesΓ+I^{cytb} tree (Figure 4.7). Thus the exclusion of certain taxa had no impact on the estimation of phylogenetic relationships among the remaining species. Nodal support for shared nodes between the two topologies (nodes 1, 2, 3, 4 and 12 in Figure 4.6 corresponding to nodes 1, 2, 3, 4 and 9 in the BayesΓ+$I^{cytbsmall}$ tree, Figure 4.7 A) was fairly stable. One exception was the large increase in posterior probability of a clade comprising (*C.sp.*7 + *C. klugii* + (*C. bifossus* + *C. sp.* 12) from 0.51 in the BayesΓ+I^{cytb} topology to 0.76 the BayesΓ+$I^{cytbsmall}$ topology. However, despite the large increase in the posterior probability of this clade in the BayesΓ+$I^{cytbsmall}$, it is still not significantly supported.

The Bayesian consensus topology using a GTR+ I+ Γ model and three codon-specific rates (Bayes$SSR_3^{cytbsmall}$) closely resembles the BayesΓ+$I^{cytbsmall}$ topology. The only difference is the placement of *C. sp.* 7, which forms a sister taxon to the clade (*C. fulvopilosus* + *C. storeatus*) in the Bayes$SSR_3^{cytbsmall}$. This is in contrast to it grouping with ((*C. bifossus* +*C. sp. 12*) + *C. klugii*) in the BayesΓ+$I^{cytbsmall}$ tree, though this clade was only observed in 53% of all trees sampled. Interestingly, for the reduced subset of taxa, a GTR + SSR_3 model fits the data better than a GTR + I + Γ model.

[3] The reduced data set comprised those 15 taxa for which both cytochrome *b* and cytochrome oxidase II sequence data was available; see Chapter 3 page 49.

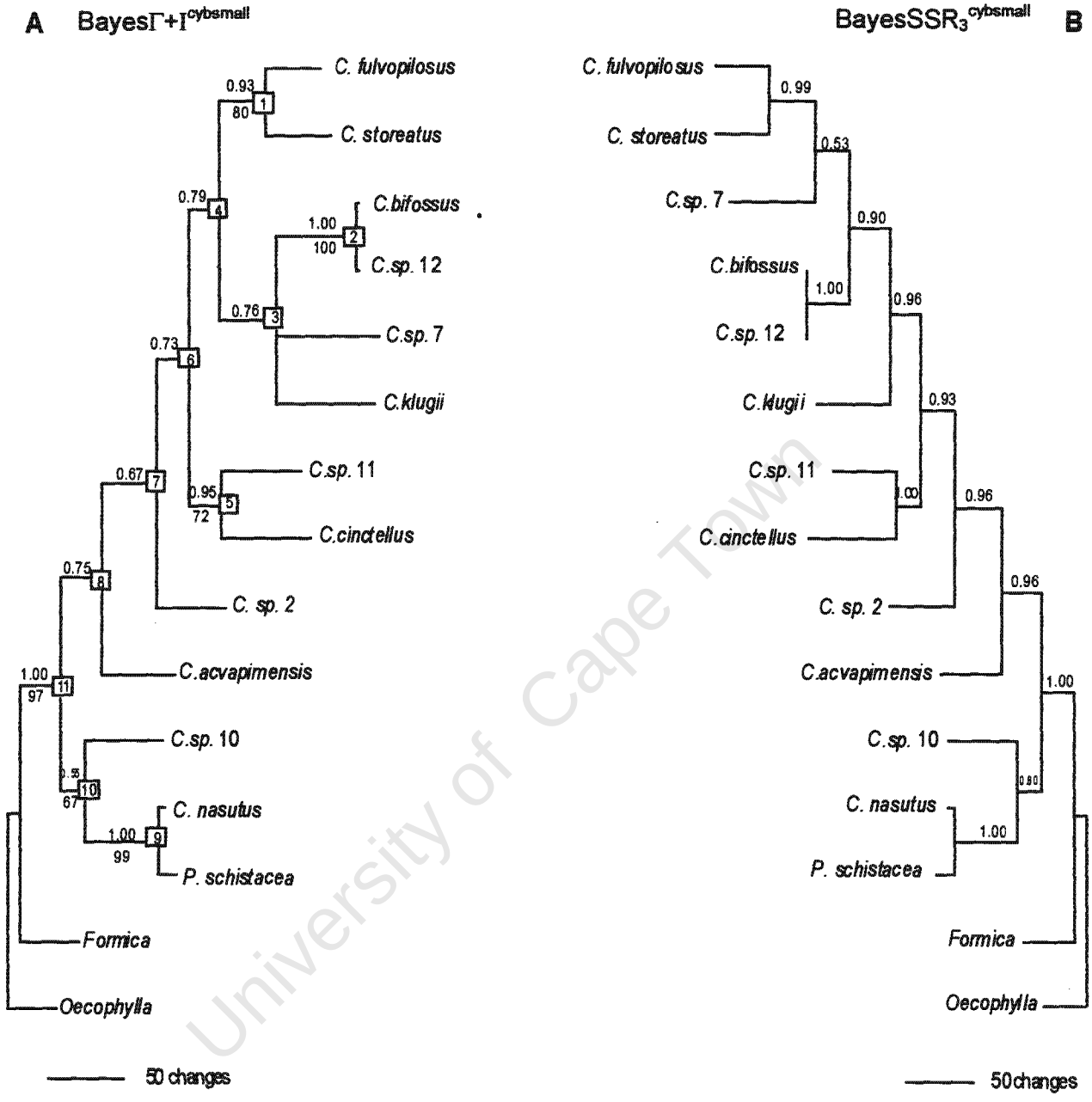


Figure 4.7. The 50% majority rule consensus trees from Bayesian analysis of cytochrome *b* sequences for a subset of taxa under (A) a GTR + I + Γ model, mean ln *L* = -4237 or (B) a GTR + SSR_3 model, mean ln *L* = -4209. Numbers above branches represent posterior probability values. Numbers below branches indicate non-parametric ML bootstrap proportions greater than 50% based on 100 pseudoreplicate data sets. Branch lengths are drawn proportional to the number of changes as indicated by the scale bar.

Maximum likelihood analyses

Maximum likelihood analysis of the reduced cytochrome *b* data set supported a single optimal tree with ln *L* = -4214 using parameter values estimated by Modeltest[4]. The ML tree ($ML^{cytbsmall}$) was topologically identical to the $Bayes\Gamma+I^{cytbsmall}$ tree when nodes with bootstrap support values of > 50% were considered. Reiterative ML searches using updated ML model parameter values did not result in different topologies.

Comparison of Bayesian parameter estimates between the reduced and complete data sets

All but one of the Bayesian parameter estimates for the reduced cytochrome *b* data set fell within the 95% credible region for each parameter estimated for the complete cytochrome *b* data set. The exception was the relative rate parameter for A↔T transversions estimated using the GTR + SSR_3 model. The Bayesian estimate of p_{inv} for the reduced data set (p_{inv} = 0.40) was similar to that estimated for the larger data set (p_{inv} = 0.42). Greater rate heterogeneity was inferred for the reduced data set sequences compared to the complete data set, with α decreasing in value from 0.86 to 0.73. This is indicative of the fact that this parameter is sensitive to taxon sampling and tree topology, consistent with the findings of Sullivan *et al.*, (1996).

The similarity of Bayesian and ML topologies and model parameter estimates between the complete and reduced data sets indicates that model-based phylogenetic reconstructions of relationships between species in this study, based on cytochrome *b*, are not overly sensitive to taxon sampling.

LogDet transformation

The LogDet/ME bootstrap distance tree of the reduced data set was fully congruent with the ML and Bayesian trees based on the same sequences.

[4] πA = 0.34; πC = 0.17; πG = 0.09; πT = 0.40; r_{AC} = 1 959 258; r_{AG} = 3 324 160; r_{AT} = 2 003 347; r_{CG} = 818 820; r_{CT} = 23 523 494; r_{GT} = 1; p_{inv} = 0.43; α = 0.86.

(ii) Parsimony analyses

Parsimony analyses of the reduced cytochrome *b* data set showed similar trends to those observed for the complete cytochrome *b* data set (Table 4.7). Numbered nodes correspond to nodes labelled on the BayesΓ+$I^{cytbsmall}$ tree (Figure 4.7 A). Weighting schemes that downweighted transitions relative to transversions increased congruence with the Bayesian reference topology. Three weighting schemes resulted in topologies almost identical to the reference topology: (i) 4:1 downweighting of transitions at all codon positions, (ii) differential downweighting of transitions at each codon position and (iii) downweighting T↔C transitions at all positions six-fold. Trees recovered after downweighting A↔T transversions showed decreased congruence with the reference topology compared to those produced under equal weights. Both topologies based on amino acid residues were poorly resolved.

In general, resolution of the bootstrap MP consensus trees was poor, with moderate to high bootstrap support values limited to nodes at the tip of the tree rather than more basal relationships, as observed during parsimony analysis of the complete cytochrome *b* data set. CI values varied from 0.37 to 0.61, compared with 0.34 to 0.57 in the complete cytochrome *b* data set. The slight increase in CI values is consistent with the findings of Sanderson and Donoghue (1989) that the consistency index is highly correlated with the number of taxa, with a decrease in the number of taxa associated with an increase in CI values. RI values for topologies based on the reduced cytochrome *b* data set were uniformly lower than those based on the complete data set, indicating that although certain characters may have become more consistent as reflected by higher CI values, there was a decrease in the number of synapomorphic characters.

All nodes shared among the complete and reduced data sets received comparable levels of bootstrap and Bremer support. Furthermore, the nodes that were supported by moderate to high bootstrap values were also present in the Bayesian and ML tree topologies.

Table 4.7. Congruence of cytochrome *b* (reduced data set) MP topologies with the BayesΓ+I^{cytbsmall} topology.

Nodes in BayesΓ+I^{cytbsmall} tree	Unweighted	4:1 Tv:Ti	1st 1:1 Tv:Ti 2nd 2:1 Tv:Ti 3rd 21:1 Tv:Ti	Ti = 0 all positions	Ti = 0 3rd codon positions	A↔T downweighted 2:1 ALL position	A↔T downweighted 2:1 3rd positions	T↔C downweighted 6:1	Successive weighting	6-P parsimony	Unweighted protein parsimony	ProtPars parsimony
1	●●●	●●	●●	●●●	●●●	●●	●	●●	●●●	●●	●●	●●●
2	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●
3	-	-	+	-	-	+	+	+	●●	+	-	-
4	+	+	-	-	+	+	-	+	●●	+	-	-
5	-	●●	●●	●●●	●●	+	-	+	●●	●●	-	-
6	+	●	+	+	●	-	-	●●	-	+	-	-
7	●	+	+	+	-	-	-	-	●●●	-	-	-
8	●	●	+	●	●●	-	+	●●	●●	-	-	-
9	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
10	●	●●	●●	●●	●●	●	●	●●	-	●●	●●	●●●
11	●●●	●●●	●	●●●	●●●	●●●	●●	●●●	●●●	●●●	●●●	●●●
Length	871	2010	6617	419	469	1499	1438	3037	182	1721	229	259
CI	0.43	0.42	0.37	0.37	0.43	0.41	0.45	0.44	0.61	0.38	0.55	0.55
RI	0.38	0.42	0.38	0.42	0.47	0.33	0.32	0.43	0.65	0.36	0.40	0.40
No. MP trees	5	1	1	4	4	1	1	1	1	3	1	10
No. congruent nodes	9/11	10/11	10/11	9/11	9/11	8/11	7/11	10/11	9/11	9/11	5/11	5/11
No. congruent nodes with bootstrap ≥ 50%	7/11	8/11	6/11	7/11	8/11	5/11	5/11	7/11	9/11	6/11	5/11	5/11

Note: Tree statistics given for each analysis are based on the bootstrap consensus topology.
●●● indicates strong bootstrap support (90% - 100%) and a Bremer support value ≥ 10
●● indicates medium bootstrap support (70-89%) and Bremer support of > 3
● indicates weak bootstrap support (50-69%)
+ indicates that the node was present in the MP tree/strict consensus tree but with bootstrap < 50%
- indicates that clade was not present in the MP/strict consensus tree or the bootstrap consensus tree

III *Tree reconstruction: complete cytochrome oxidase II data set*

(i) Model-based analyses

Bayesian analyses

As observed for the cytochrome *b* Bayesian analyses, the log likelihoods of the sample points for all cytochrome oxidase II Bayesian analyses reached stationarity after approximately 1 X 10^4 generations (plot not shown). The three runs incorporating gamma and invariant sites parameters to model rate heterogeneity all yielded an identical topology (hereafter referred to as BayesΓ+I^{coii}, mean ln *L* = -3556) with similar estimates of clade posterior probability support (Figure 4.8). The two Bayesian runs assuming a GTR + SSR_3 model also converged on an identical topology (hereafter referred to as BayesSSR_3^{coii}, mean ln *L* = -3630) with congruent probability estimates for clades (figure not shown).

The topologies from the two Bayesian analyses were almost completely congruent. *C. cinctellus* and *C.sp.*19 (Figure 4.8, node 1) associate as sister taxa in both topologies, although this association is only significantly supported under a GTR + I + Γ model. These sister taxa fall into a clade with *C. sp.* 2 and *C. sp.* 11 (Figure 4.8, node 2), with this relationship significantly supported under a GTR + SSR_3 model. *Camponotus klugii* and *C. sp.* 24 (Figure 4.8, node 9) associate as sister taxa in both trees. It is the position of this clade that is responsible for the topological disagreement between the two Bayesian topologies. This clade forms a sister group to the clade containing ((*C. cinctellus* + *C.sp.* 19) + *C.sp.* 2 + *C. sp.* 11) in the BayesSSR_3^{coii} tree, whereas the placement of this clade is unresolved in the BayesΓ+I^{coii} tree. However, this sister-group relationship lacks robust support in the BayesSSR_3^{coii} tree. The association between *C. sp.* 12 and *C. bifossus* observed in phylogenetic analyses of the cytochrome *b* data was reaffirmed by Bayesian analysis of the cytochrome oxidase II sequence data. The clade containing these two species (Figure 4.8, node 8) has significant posterior probability in both Bayesian analyses. The placement of *C. sp.* 7 is unresolved in both analyses, as it was in the cytochrome *b* analyses. A large paraphyletic

assemblage subtended by node 7 (Figure 4.8), comprising all species in subgenus *Myrmopiromis* (*C. storeatus, C. brevisetosus, C. detritus* and *C. fulvopilosus*) and the two species representatives of the subgenus *Myrmopsamma* (*C. mystaceus and C. sp.* 14), is present with significant posterior probability in both Bayesian consensus trees. Within this large group, two distinct sister groups are significantly supported (bpp = 1.0) in both Bayesian analyses. *Camponotus storeatus* and *C. brevisetosus* associate as sister species (Figure 4.8, node 6), with this clade forming a sister group to a group comprising ((*C.detritus*, *C. fulvopilosus*), *C. mystaceus*) *C. sp.* 14) (Figure 4.8, node 3). The position of this large paraphyletic assemblage in the tree is unresolved, though a large group (Figure 4.8, node 10) consisting of ((*C. bifossus* + C. *sp.* 12), *C. sp.* 7) and the paraphyletic group subtended by node 7 is significantly supported in both Bayesian trees. The position of *C. acvapimensis* is unresolved in both trees. *Polyrhachis schistacea* appears as a sister taxon to all included *Camponotus* species, with the exception of *C. sp.* 10 and *C. nasutus*. However, this relationship did not receive significant support in either of the two Bayesian topologies. *Camponotus sp.* 10 and *C. nasutus* form an unresolved polytomy near the base of the tree.

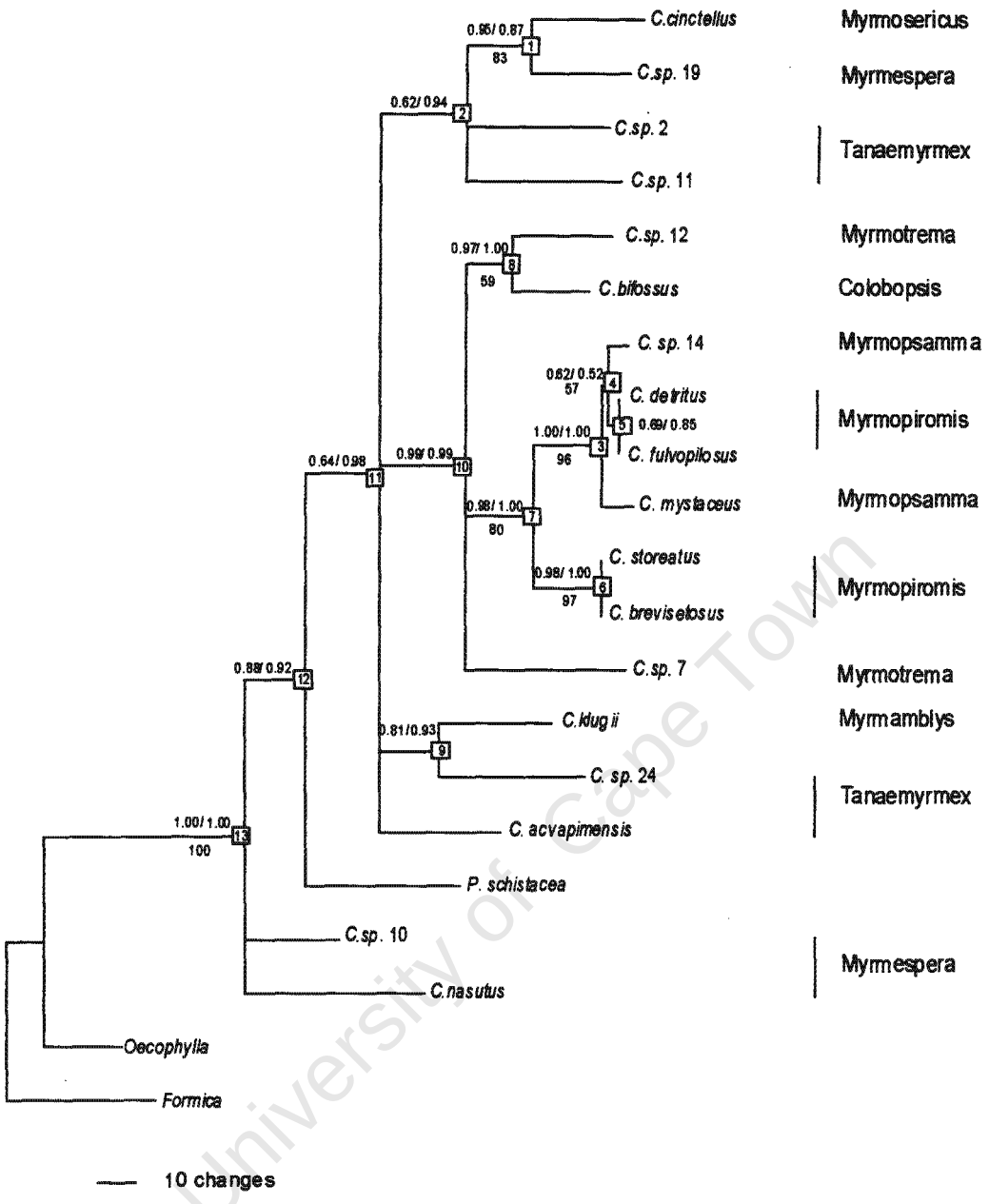


Figure 4.8. The 50% majority rule consensus tree from Bayesian analysis of cytochrome oxidase II sequences using a GTR+ I + Γ model of nucleotide evolution (BayesΓ+I^{coii}). The two numbers adjacent to nodes represent posterior probabilities from two separate Bayesian analyses using a GTR + I + Γ model and GTR + SSR_3 model respectively. Numbers below branches indicate non-parametric bootstrap proportions greater than 50%. Branch lengths are drawn proportional to the number of changes as indicated by the scale bar. Subgeneric classifications are indicated to the right of the tree.

The codon specific rate parameters for cytochrome oxidase II are very similar to those estimated for the cytochrome *b* data (see Tables 4.4 & 4.8), with third codon positions evolving at the fastest rate, followed by first and second codon sites. A GTR + I + Γ model of evolution provided a better fit to all cytochrome oxidase II sequences than a GTR + SSR_3 model, as indicated by the higher likelihood score of the GTR + I + Γ model (Table 4.8).

Maximum likelihood analysis

Modeltest results indicated that a GTR + I + Γ model of sequence evolution also best explained the cytochrome oxidase II sequences. Two most likely trees were obtained using Modeltest parameters[5] ($ML1^{coii}$ and $ML2^{coii}$; ln *L* = -3525) that differed from each other only with respect to the placement of *C. acvapimensis*. The strict consensus of these two ML trees was almost identical to the $BayesSSR_3^{coii}$ tree. The Bayesian tree was less resolved than the ML tree, with (*C. sp.* 2 + *C sp.* 11) forming a sister group in the strict consensus ML tree compared to their unresolved relationship in the $BayesSSR_3^{coii}$ tree. Iterative searches of tree space using model parameters estimated by ML on the strict consensus tree of $ML1^{coii}$ and $ML2^{coii}$ resulted in two ML topologies ($ML3^{coii}$ and $ML4^{coii}$). The strict consensus of these topologies differed from the strict consensus of the $ML1^{coii}$ and $ML2^{coii}$ tree topologies only with respect to the relationship of *C. sp.* 2 and *C. sp.* 11. *Camponotus. sp.* 2 was sister to a clade comprising (*C. sp.* 11 + (*C. cinctellus* + *C. sp.* 19)) in the strict consensus of $ML3^{coii}$ and $ML4^{coii}$, whereas (*C. sp.*2 + *C. sp.* 11) were sister taxa in the original strict consensus tree, and formed a sister group to the two sister species (*C. cinctellus* + *C. sp.* 19) (trees not shown).

Branch support

Bootstrap support for nodes was poor, with only seven out of 18 possible nodes supported by bootstrap values of more than 50%. Of the five nodes that received significant posterior probability support from both Bayesian analyses (Figure 4.8, nodes 3, 6, 7, 8, 10, 13), four were also supported by bootstrap

[5] $\pi A = 0.37$; $\pi C = 0.13$; $\pi G = 0.07$; $\pi T = 0.43$; $r_{AC} = 2\ 762\ 266$; $r_{AG} = 496\ 110$; $r_{AT} = 218\ 372$; $r_{CG} = 119\ 059$; $r_{CT} = 3\ 783\ 341$; $r_{GT} = 1$; $p_{inv} = 0.35$; $\alpha = 0.46$.

values ≥ 50%. However, as observed for the cytochrome *b* analyses, there was no clear relationship between significant Bayesian posterior probability values and bootstrap support, with bootstrap values ranging from 59% to 100% corresponding to bpp ≥ 0.95 (Figure 4.8).

Table 4.8. Nucleotide substitution model parameter means, variances and 95% credible regions (C.R.) for each likelihood model parameter of the cytochrome oxidase II complete dataset using the GTR model of nucleotide substitution with two different rate heterogeneity models. Values represent the average across three (GTR + I + Γ) or two (GTR + SSR_3) runs.

	Models of sequence evolution					
	GTR+ I + Γ			GTR+ SSR_3		
Parameter	Mean	Variance	95% C.R.	Mean	Variance	95% C.R.
-ln *L*	3556	32	3568	3630	31	3642, 3620
Branch lengths	10.68	9.61	6.40, 18.53	2.52	0.02	2.26, 2.81
[a]r_{GT}	1.00	0	1.00, 1.00	1.00	0	1.00, 1.00
r_{CT}	60.51	428.44	24.63, 97.52	63.16	338.31	29.96, 100.80
r_{CG}	8.43	26.72	1.22, 21.45	4.24	6.78	0.76, 10.49
r_{AT}	2.55	1.43	0.75, 5.42	9.60	8.21	4.47, 15.03
r_{AG}	33.87	187.54	14.09, 68.46	18.26	23.46	8.57, 30.57
r_{AC}	2.78	2.50	0.58, 6.62	7.62	8.54	3.15, 14.20
[b]πA	0.39	0	0.35, 0.42	0.36	0	0.33, 0.40
πC	0.13	0	0.11, 0.15	0.14	0	0.12, 0.16
πG	0.04	0	0.02, 0.05	0.08	0	0.06, 0.10
πT	0.45	0	0.42, 0.48	0.42	0	0.39, 0.45
[c]α	0.42	0.01	0.30, 0.60	-	-	-
[d]p_{inv}	0.37	0	0.27, 0.45	-	-	-
[e]1st codons	-	-	-	0.43	0	0.36, 0.51
2nd codons	-	-	-	0.15	0	0.11, 0.19
3rd codons	-	-	-	2.42	0	2.34, 2.51

[a] R $_{X\leftrightarrow Y}$, relative rate parameters
[b] π, estimated frequency of the nucleotide
[c] α , gamma shape parameter
[d] p_{inv}, proportion of invariant sites
[e] codon specific rate

None of the ML or Bayesian tree topologies was rejected as a significantly worse explanation of the data using the SH test (Table 4.9).

Table 4.9. Shimodaira-Hasegawa test results for comparison of Bayesian and ML tree topologies inferred from cytochrome oxidase II sequences.

Tree topology	-ln *L*	Difference -ln *L*	*P*
BayesΓ+I^{coii}	3526.80	1.57	0.310
BayesSSR$_3$^{coii}	3525.67	0.44	0.68
ML1^{coii}	3525.24	0.01	0.83
ML2^{coii}	3525.24	0.01	0.82
Strict consensus ML1^{coii} + ML2^{coii}	3525.24	0.01	0.82
ML3^{coii}	optimal	-	-
ML4^{coii}	3525.23	0	0.89
Strict consensus ML3^{coii} + ML4^{coii}	3525.23	0	0.89

Comparison of ML and Bayesian model parameter estimates

Base frequency parameter values estimated by Modeltest on a neighbour joining tree corresponded well with the Bayesian GTR+ I+ Γ base frequency estimates, with the exception of guanine. The ML estimate of this base frequency parameter (πG = 0.07) lay outside the Bayesian 95% credible intervals estimated using the GTR+ I + Γ model of sequence evolution, but not the GTR + SSR$_3$ model (Table 4.8). The ML estimates of the gamma shape parameter (α = 0.46) and the proportion of invariant sites (p_{inv} = 0.37) fell within the 95% credible interval of the Bayesian GTR + I + Γ analysis. This value of α, together with a p_{inv} estimate of 0.37 indicates extreme rate heterogeneity in the *Camponotus* cytochrome oxidase II sequences, with the majority of sites invariant or evolving at a slow rate and a few sites evolving extremely rapidly. The substitution rate matrix values estimated by ML differed by orders of magnitude from the Bayesian estimates. The two most predominant substitution types estimated by Bayesian and ML methods were C↔T and A↔G transitions.

LogDet Transformation

The bootstrap LogDet/ME majority rule consensus tree was poorly resolved (8/18 nodes). However, nodes supported by bootstrap values of more than 50% were all present in the ML and Bayesian topologies, and no novel clades were observed. Therefore heterogeneity of base composition for the cytochrome oxidase II third codon positions and parsimony informative sites does not appear to be biasing phylogenetic reconstruction, with the proviso that the LogDet transformation is correcting adequately for base composition heterogeneity across the tree.

(ii) Parsimony analyses

The results of parsimony analyses of the cytochrome oxidase II sequences using a variety of weighting schemes is presented in Table 4.10. Numbered nodes correspond to nodes labelled on the BayesΓ+I^{coii} tree (Figure 4.8). Assigning greater costs to transitions than transversions generally increased congruence to the BayesΓ+I^{coii} topology. Excluding transitions at third codon positions was found to be the most successful strategy in this regard. A$\leftrightarrow$T transversion downweighting did not appear to be an effective weighting strategy, as the resultant topologies showed decreased congruence with the BayesΓ+I^{coii} topology compared to topologies produced under equal weighting.

Unweighted protein parsimony performed poorly, with a total of 516 MP trees recovered. Furthermore, only three nodes were resolved in the strict consensus of these 516 MP trees. Applying a Protpars weight matrix to the amino acid characters reduced the number of MP trees to four, and increased congruence with the reference topology.

All MP topologies were compatible with the Bayesian and ML topologies when nodes with medium to strong bootstrap support were considered. Relationships recovered across all trees were the grouping of *C. sp.* 14, *C. detritus* and *C. fulvopilosus* in a group (node 4), supported by weak to moderate bootstrap values. The sister association between *C. storeatus* and

C. brevisetosus (node 6) was strongly supported in all nucleotide parsimony analyses, with bootstrap values of > 90% and Bremer decay indices of > 10. Furthermore, *C. detritus* and *C. fulvopilosus* (node 5) were associated as sister taxa in all nucleotide-based parsimony analyses. The sister position of *C. mystaceus* to the group subtended by node 4 (node 3) was supported by weak to medium support in ten of the 12 analyses. A sister-clade relationship of (*C. storeatus* + *C. brevisetosus*) to a group comprising *C. mystaceus* , *C. sp.* 14, *C. detritus* and *C. fulvopilosus* (node 7) received moderate support in most analyses. The monophyly of the ingroup (node 13) received strong support in all analyses.

CI values for the various parsimony topologies ranged from 0.29 to 0.57, with RI values ranging from 0.26 to 0.65.

(iii) Phylogenetic relationships

Phylogenetic analyses of cytochrome oxidase II sequences did not indicate support for any of the subgeneric classifications of *Camponotus*. Although there was robust support for a sister-group association between *C. storeatus* and *C. brevisetosus* (both in subgenus *Myrmopiromis*) across all analyses (Figure 4.8, node 6), the other two representatives of this subgenus included in this study (*C. fulvopilosus and C. detritus*) did not associate in a monophyletic group with *C. storeatus* and *C. brevisetosus*. Rather, there was strong support in all analyses for a group comprising (*C. mystaceus* (*C. sp.* 14 (*C. detritus* + *C. fulvopilosus*))). This group (Figure 4.8, node 3) formed a sister-group to (*C. storeatus* + *C. brevisetosus*), with this relationship (Figure 4.8, node 7) significantly supported in both Bayesian analyses and moderately supported by ML bootstrap analysis. *Camponotus bifossus* and *C. sp.* 12 associated as sister taxa with significant posterior probability in both Bayesian analyses (Figure 4.8, node 8), although bootstrap support for this clade was low under both MP and ML optimality critera. A sister-association between *C. sp.* 19 (subgenus *Myrmespera*) and *C. cinctellus* (subgenus *Myrmosericus*) was evident in all analyses (Figure 4.8, node 1). However, robust support for this association was only obtained under Bayesian and ML analyses using a

GTR + I + Γ model of sequence evolution. The node separating the two outgroup taxa from the ingroup species maximally supported in all analyses (Figure 4.8, node 13).

Table 4.10. Congruence of cytochrome oxidase II MP topologies with the BayesΓ+I^{coii} topology.

Nodes in BayesΓ+I^{coii} tree	Unweighted	23:1 Tv:Ti	1st 7:1 Tv:Ti 2nd 3:1 Tv:Ti 3rd 14:1 Tv:Ti	Ti = 0 all positions	Ti = 0 3rd codon positions	A↔T downweighted 2:1 ALL position	A↔T downweighted 2:1 3rd positions	T↔C downweighted 6:1	Successive weighting	6-P parsimony	Unweighted protein parsimony	ProtPars parsimony
1	+	+	+	+	+	+	+	+	●●	+	-	+
2	-	-	-	-	+	-	-	-	-	-	-	-
3	●●●	●●	●●	●	●●	●●	●●	●●●	●●●	●	-	-
4	●●	●●	●●	●●	●●	●	●	●●	●●	●●	●●	●●
5	●●	●	●	●	●●	●●●	●●	●●	●●●	●●	-	-
6	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
7	●●	+	+	+	●●	●●	●●	●●	●●●	●	-	+
8	●	+	+	+	●	+	+	+	●●	+	-	+
9	-	-	-	-	+	-	-	-	-	-	-	-
10	-	+	+	+	+	-	-	+	-	+	-	-
11	+	+	+	+	-	-	-	●	●●●	+	-	+
12	+	+	-	+	+	-	-	●	●●●	+	-	-
13	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
Length	823	9799	5366	388	489	1343	1242	2797	126	1245	214	228
CI	0.34	0.31	0.29	0.28	0.33	0.33	0.37	0.37	0.57	0.30	0.44	0.43
RI	0.30	0.28	0.27	0.29	0.38	0.27	0.26	0.40	0.65	0.32	0.56	0.57
No. MP trees	1	1	1	8	1	20	6	1	1	4	516	4
No. congruent nodes	10/13	11/13	10/13	11/13	12/13	8/13	8/13	11/13	10/13	11/13	3/13	7/13
No. congruent nodes with bootstrap ≥ 50%	7/13	5/13	5/13	5/13	7/13	6/13	6/13	8/13	10/13	6/13	3/13	3/13

Note: Tree statistics given for each analysis are based on the bootstrap consensus topology.
●●● indicates strong bootstrap support (90% - 100%) and a Bremer support value ≥ 10
●● indicates medium bootstrap support (70-89%) and Bremer support of > 3
● indicates weak bootstrap support (50-69%)
+ indicates that the node was present in the MP tree/strict consensus tree but with bootstrap < 50%
- indicates that clade was not present in the MP/strict consensus tree or the bootstrap consensus tree

IV *Tree reconstruction: reduced cytochrome oxidase II data set*

(i) Model-based analyses

Bayesian analyses

Bayesian analysis of the reduced[6] cytochrome oxidase II data set resulted in two topologies reflecting the different model parameters used to incorporate rate heterogeneity, with repeat runs for each model converging on identical topologies (Figure 4.9). The Bayesian topology based on a GTR + I + Γ model (Bayes$\Gamma+I^{coiismall}$) resembles a pruned version of the Bayes$\Gamma+I^{coii}$ topology. The clade comprising ((*C. bifossus* + *C. sp.*12), (*C. storeatus* + *C. fulvopilosus*), *C. sp.* 7) (Figure 4.9 A, node 3) represents a pruned version of the large clade subtended by node 10 in the Bayes$\Gamma+I^{coii}$ topology (Figure 4.8), indicating the robustness of the analyses to taxon removal. *Camponotus sp.* 2 and *C. sp.* 11 associate as sister taxa in the Bayes$\Gamma+I^{coiismall}$ tree (Figure 4.9 A, node 4), although this relationship is not significantly supported. The large clade subtended by node 5 in the Bayes$\Gamma+I^{coiismall}$ is equivalent to a pruned clade subtended by node 11 in the Bayes$\Gamma+I^{coii}$ tree. *Polyrhachis schistacea* is sister taxon to all *Camponotus* species included in the analysis, with the exception *of C. nasutus* and *C. sp.* 10, as observed in the ML and Bayesian topologies based on the complete cytochrome oxidase II data set. This relationship, however, is not strongly supported in either of the two Bayesian analyses of cytochrome oxidase II sequences for a reduced subset of taxa. Neither of the two nucleotide evolution models was able to resolve the placement of *C. nasutus* and *C. sp. 10.*

The two Bayesian topologies based on the pruned cytochrome oxidase II data set are not topologically congruent. The major reason for the topological discordance is the presence in the Bayes$SSR_3^{coiismall}$ topology of a clade comprising (((*C. bifossus* + *C. sp.* 12), (*C. storeatus* + *C. fulvopilosus*), *C. sp.* 7), *C. cinctellus*, *C. klugii*). This clade, though strongly supported (bpp = 0.93)

[6] The reduced data set comprised those 15 taxa for which both cytochrome *b* and cytochrome oxidase II sequence data was available; see Chapter 3 page 49.

in the BayesSSR$_3^{coiismall}$ topology, is not present in the BayesΓ+I^{coiismall} topology.

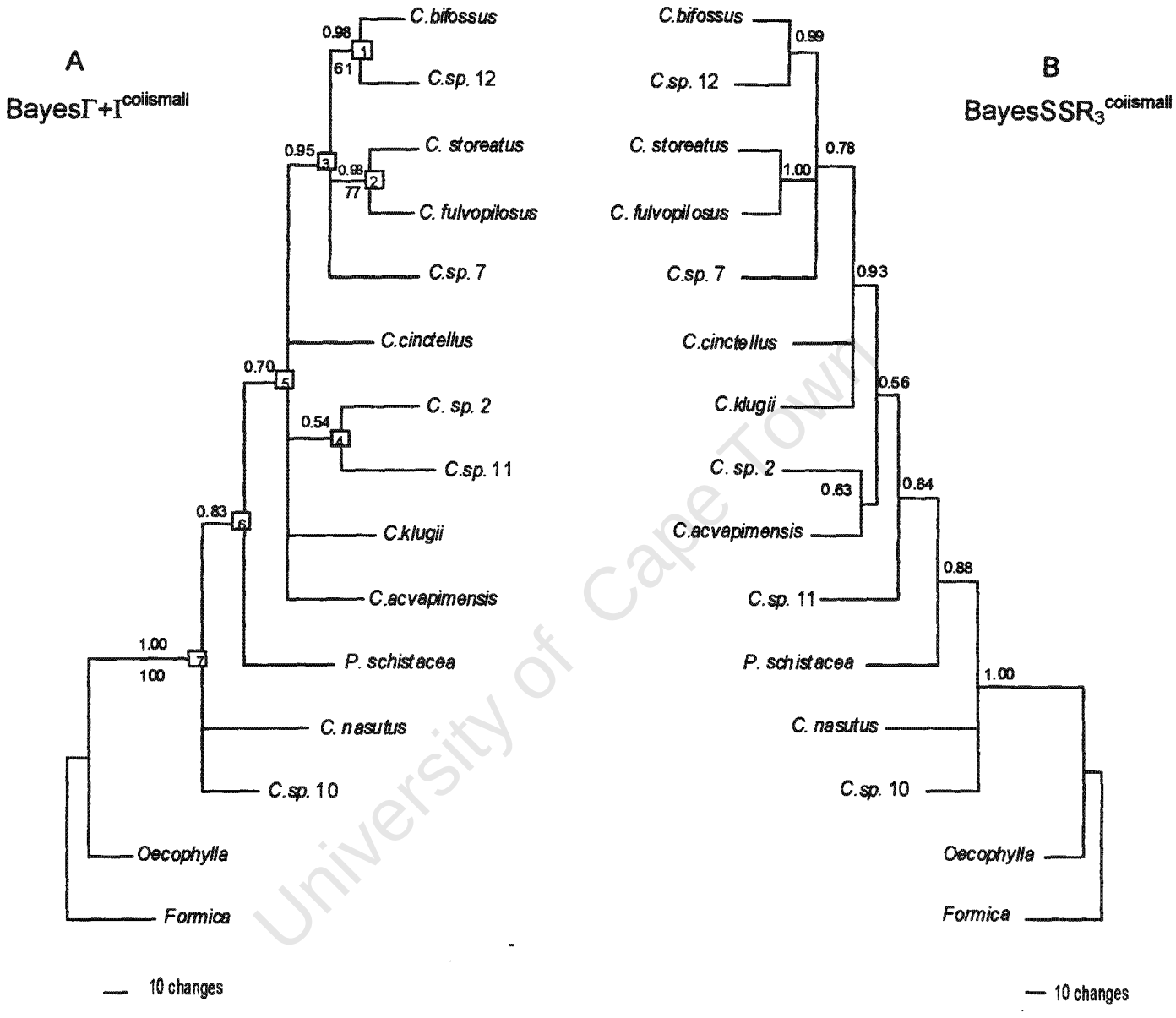


Figure 4.9. The 50% majority rule consensus trees from Bayesian analysis of cytochrome oxidase II sequences for a subset of taxa using (A) a GTR+ I + Γ model, mean ln *L* = -3178 or (B) a GTR + SSR$_3$ model, mean ln *L* = -3201. Numbers above branches represent posterior probability values. Numbers below branches indicate non-parametric ML bootstrap proportions greater than 50% based on 100 pseudoreplicate data sets. Branch lengths are drawn proportional to the number of changes as indicated by the scale bar.

The Bayesian analysis incorporating rate heterogeneity by assuming a gamma distribution and proportion of invariant sites fit the observed sequences better than rate heterogeneity modelled on three codon-specific rates. A GTR + I + Γ model provided a better fit to the cytochrome oxidase II sequences for a subset of taxa than a GTR + SSR_3 model, as was observed after Bayesian analyses of the complete cytochrome oxidase II data set.

Maximum likelihood analyses

Hierarchical likelihood ratio tests indicated that a GTR + I + Γ model best fit the cytochrome oxidase II sequences for a reduced subset of taxa. Two ML topologies with the same log likelihood value (ln *L* = -3155) were encountered in different areas of tree space (different 'tree islands') using model parameters estimated by Modeltest[7]. The two islands were not visited equally, with the second island only encountered in one of the ten searches. Repeating the ML search with parameter values estimated by ML on the most frequently encountered topology from the first search gave identical results, with two tree islands encountered, one of which was hit 9/10 times, the other only once. The strict consensus of the two trees from each island ($ML^{coiismall}$) was almost identical in topology to the $Bayes\Gamma+I^{coiismall}$ tree. However, only three nodes were supported by bootstrap values greater than 50% (Figure 4.9A).

Comparison of Bayesian model parameter values between the reduced and complete cytochrome oxidase II data sets

Values of Bayesian model parameters for the reduced data set fell within the 95% credible region estimated for the complete cytochrome oxidase II data set, with the exception of the branch length parameter. The mean estimate of branch length for the reduced data set based on the GTR+ I + Γ model was 4.86, compared to the mean value of 10.68 calculated for the complete data set using the same model. This may be due to the higher value of α estimated by Bayesian analysis for data set D (α = 0.62) compared to the low alpha value inferred for data set *B* (α = 0.42). This indicates that there is relatively

[7] $\pi A = 0.37$; $\pi C = 0.13$; $\pi G = 0.07$; $\pi T = 0.43$; $r_{AC} = 31.86$; $r_{AG} = 44.09$; $r_{AT} = 18.83$; $r_{CG} = 8.02$; $r_{CT} = 359.52$; $r_{GT} = 1$; $p_{inv} = 0.34$; $\alpha = 0.48$.

less rate heterogeneity in the reduced data set (data set D), with fewer sites evolving very rapidly, which in turn could result in decreased estimates of branch length.

LogDet transformation

LogDet transformation of the data and subsequent bootstrap analysis using the minimum evolution algorithm resulted in a consensus topology with only four resolved nodes, of which only two (nodes 2 and 7 in the BayesΓ+$I^{coilsmall}$ tree) (Figure 4.9) were supported by medium to strong bootstrap values.

(ii) Parsimony analyses

Results of parsimony analyses of the reduced cytochrome oxidase II data set are provided in Table 4.11. Numbered nodes correspond to nodes labelled on the BayesΓ+$I^{coilsmall}$ tree (Figure 4.9 A). Three weighting schemes resulted in MP topologies completely congruent with the BayesΓ+$I^{coilsmall}$ topology:
(i) downweighting all transitions by a factor of six, (ii) downweighting T↔C transitions by a factor of six and (iii) applying a six-parameter cost matrix.
As observed in all previous analyses, A↔T transversion downweighting decreased topological congruence with the reference tree, especially when applied to all codon positions. Protein parsimony using a Protpars cost matrix resulted in topologies more similar to the reference topology than unweighted protein parsimony. In general, only two nodes were resolved in the bootstrap consensus trees: the node associating *C. fulvopilosus* and *C. storeatus* (node 2), and the node connecting the outgroup taxa to the ingroup taxa (node 7). This particular node received strong support in all analyses, whereas node 2 received mixed support, ranging from bootstrap values of 50% to 91% and Bremer support indices of 1 to 8.

CI values ranged from 0.33 to 0.67, with RI values varying between a minimum of 0.12 and a maximum of 0.68. CI values were similar for the analyses based on the complete and reduced data sets. RI values were uniformly lower for the reduced data set, with the exception of those obtained for the successive weighting analysis. This was especially noticeable for the

protein parsimony topologies, where the RI values for the unweighted and Protpars-weighted topologies decreased from 0.56 to 0.25, and 0.57 to 0.38 respectively. These results indicate that the removal of certain taxa resulted in the loss of synapomorphic characters.

Table 4.11. Congruence of cytochrome oxidase II MP topologies (reduced data set) with the BayesΓ+I^{coiismall} topology.

Nodes in BayesΓ+I^{coiismall}	Unweighted	6:1 Tv:Ti	1st 2:1 Tv:Ti 2nd 2:1 Tv:Ti 3rd 43:1 Tv:Ti	Ti = 0 all positions	Ti = 0 3rd codon positions	A↔T downweighted 2:1 all position	A↔T downweighted 2:1 3rd codons	T↔C downweighted 6:1	Successive weighting	6-P parsimony	Unweighted protein parsimony	ProtPars parsimony
1	+	+	-	+	●●	-	+	●●	●●●	+	-	+
2	●●●	●●	-	-	●●	●●	●●	●●	●●●	●●	●●	●●
3	-	+	+	+	+	-	-	+	-	+	-	-
4	+	+	-	-	-	-	+	+	●●●	+	-	-
5	+	●	+	●	●	-	-	●	●●●	●	+	●
6	+	+	+	-	●	-	-	●	●●●	●	-	+
7	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
Length	702	1929	11 938	307	357	856	988	2362	93	944	179	177
CI	0.36	0.43	0.29	0.33	0.37	0.42	0.36	0.44	0.67	0.37	0.51	0.52
RI	0.14	0.41	0.12	0.24	0.34	0.34	0.12	0.40	0.68	0.30	0.25	0.38
No. MP trees	1	1	1	10	4	5	1	1	1	2	32	2
No. congruent nodes	6/7	7/7	4/7	4/7	6/7	2/7	4/7	7/7	6/7	7/7	3/7	5/7
No. congruent nodes with bootstrap ≥ 50%	2/7	3/7	1/7	2/7	5/7	2/7	2/7	5/7	6/7	4/7	2/7	3/7

Note: Tree statistics given for each analysis are based on the bootstrap consensus topology.
●●● indicates strong bootstrap support (90% - 100%) and a Bremer support value ≥ 10
●● indicates medium bootstrap support (70-89%) and Bremer support of > 3
● indicates weak bootstrap support (50-69%)
+ indicates that the node was present in the MP tree/strict consensus tree but with bootstrap < 50%
- indicates that clade was not present in the MP/strict consensus tree or the bootstrap consensus tree

Part C: Phylogenetic analyses of combined nucleotide data sets

I *Phylogenetic analyses of the combined cytochrome b and cytochrome oxidase II data set for a subset of taxa*

(i) Nucleotide substitution model

The best- fit model of nucleotide evolution for both the cytochrome oxidase II and cytochrome *b* sequences, as evaluated by likelihood ratio tests, was a GTR model with rate heterogeneity incorporated by a gamma distribution and a proportion of invariant sites. Bayesian estimates of parameter value estimates for the two genes for equivalent taxa were largely congruent. All cytochrome *b* parameter value estimates lay within the 95% confidence intervals estimated for the cytochrome oxidase II sequences, and vice versa. The only exceptions to this were the r_{AG} of the cytochrome oxidase II sequences, and the equilibrium nucleotide frequency of thymine for the cytochrome oxidase II sequences, which both fell just outside the upper 95% confidence estimates estimated for the cytochrome *b* sequences (Table 4.12).

Table 4.12. Nucleotide substitution model parameter means, variances and 95% credible regions (C.R.) for each likelihood model parameter of the reduced cytochrome *b* and cytochrome oxidase II data sets using a GTR + I + Γ model of nucleotide evolution. Values represent the average across three independent runs.

	Cytochrome *b*			Cytochrome oxidase II		
Parameter	Mean	Variance	95% C.R.	Mean	Variance	95% C.R.
- ln *L*	4237	26.76	4248, 4229	3178	24.84	-3189, 3170
Branch lengths	5.74	1.46	4.00, 8.71	4.85	1.47	3.16, 7.86
r_{GT}	1.00	0	1.00, 1.00	1.00	0	1.00, 1.00
r_{CT}	72.58	270.61	45.82, 96.12	62.94	334.80	31.92, 93.71
r_{CG}	2.52	2.24	0.39, 6.34	4.28	15.07	0.19, 14.14
r_{AT}	5.82	3.10	2.80, 10.20	4.50	3.20	1.57, 8.93
r_{AG}	11.96	19.30	5.80, 22.21	23.07	180.02	8.48, 55.82
r_{AC}	6.02	5.76	2.24, 11.55	5.57	6.97	1.62, 11.71
πA	0.34	0	0.30, 0.37	0.37	0	0.33, 0.41
πC	0.17	0	0.15, 0.19	0.14	0	0.12, 0.16
πG	0.08	0	0.06, 0.10	0.05	0	0.03, 0.08
πT	0.41	0	0.38, 0.43	0.44	0	0.41, 0.47
α	0.74	0.03	0.45, 1.12	0.62	0.03	0.37, 1.00
p_{inv}	0.40	0	0.32, 0.47	0.39	0	0.29, 0.48

(ii) Patterns of rate heterogeneity

The Bayesian estimate of α for the cytochrome oxidase II data was less than that for the cytochrome *b* sequences, indicating relatively greater rate heterogeneity in cytochrome oxidase II (Table 4.12). To further explore differences in patterns of rate heterogeneity for the two genes, α was estimated for each gene partition by ML optimisation on the MP tree topology obtained from combined analysis of the two genes with T↔C transitions downweighted 6:1. Alpha and p_{inv} were calculated for first, second and third codon positions of each gene, as well as all positions combined using a GTR model with the underlying R-matrices and nucleotide equilibrium frequencies assumed to be constant across the two partitions. This was done in order to examine the effects of among-site rate variation between the two genes in isolation from other factors. Additionally, SSR values for each codon position of the two genes were calculated, including all sites, or considering variable sites separately, in order to compare patterns of variation for SSR versus α estimates of among- site rate variation. Results are presented in Table 4.13.

The overall estimate of α was markedly lower for the cytochrome oxidase II sequences than for the cytochrome *b* sequences, indicating greater among-site rate variation in this gene compared to cytochrome *b*. The estimates of p_{inv} for the two genes differed only slightly, emphasising the large discrepancy in the value of α for the two genes. This difference was even more pronounced when the same underlying R-matrix was not assumed for both genes, with the alpha value estimated for the cytochrome oxidase II sequences (α = 0.44) half that of the cytochrome *b* sequences (α = 0.83). Thus, the cytochrome oxidase II sequences display extreme rate heterogeneity compared to the cytochrome *b* sequences.

The second codon positions of both genes displayed very low alpha values, indicating extreme among-site rate variation in substitutions at these positions. The cytochrome *b* first codon sites exhibited greater among-site rate variation than the cytochrome oxidase II first codon sites, with the

estimates for the third codon positions > 1 for both genes, indicating little among-site rate variation for substitutions at these positions. The SSR estimates for equivalent codon positions from each gene were highly congruent, with third codon positions showing the greatest average rate of substitution, followed by first codon positions and lastly second codon positions. Therefore the extreme among-site rate variability in second codon positions of both genes, reflected by low α values, was unaccounted for by the SSR rate estimates for these positions. The same trend was observed when variable sites only were considered, indicating that the unequal proportion of unvaried sites could inflate differences among SSR parameters. The patterns of rate heterogeneity as represented by α and SSR rates for first, second and third codon positions are consistent with those observed by Buckley *et al.*, (2001b) for mitochondrial protein-coding sequences from species of the New Zealand cicada genus *Maoricicada*.

Although the two genes do appear to be evolving at heterogeneous rates, the Bayesian estimates of α and p_{inv} for each data set fell within the 95% credible interval for the other data set, indicating that these data sets do not display such extreme rate heterogeneity that their combined analysis is precluded.

Table 4.13. Rate heterogeneity parameters for the two gene partitions estimated by ML optimisation under a GTR model on the MP topology with T↔C transitions downweighted 6:1.

Rate parameters	Character partitions							
	Cyt *b*				COII			
	1st	**2nd**	**3rd**	**All**	**1st**	**2nd**	**3rd**	**ALL**
α	0.59	0.12	1.86	1.00	1.71	0.16	1.11	0.66
p_{inv}	0.48	0.34	0	0.43	0.56	0.27	0	0.38
SSR (all sites)	0.43	0.11	2.43	n.a.	0.42	0.13	2.46	n.a.
SSR (variable sites)	0.68	0.40	1.18	n.a.	0.68	0.33	1.30	n.a.

(iii) Character incongruence

The ILD test rejected the null hypothesis of data set homogeneity for the reduced cytochrome *b* and cytochrome oxidase II sequences ($P = 0.03$) using a significance value of $P = 0.05$ as recommended by Farris (1994). However, the null hypothesis was not rejected for individual codon comparisons (1stposCYTBt vs. 1stposCOII, $P = 0.85$; 2ndposCYTB vs. 2ndposCOII, $P = 0.30$; 3rdposCYTB vs. 3rdposCOII, $P = 0.09$). Given the recommendation in the recent literature that *P*-values < 0.05 should not preclude data set combination (Sullivan, 1996; Cunningham 1997a & b; DeSalle and Brower, 1997, Mitchell *et al*, 2000; Darlu and Lecointre, 2002), and the inability of this test to reject homogeneity for the various codon positions, it was decided to proceed with analysis of the combined data set.

II *Tree reconstruction: combined cytochrome b and cytochrome oxidase II reduced data sets*

(i) Model-based analyses

Bayesian analyses

Markov chains for independent runs with different evolutionary models reached stationarity after approximately 1×10^4 generations, with independent runs for a particular phylogenetic model converging on identical consensus topologies with congruent posterior probability values for nodes. Bayesian analysis of the combined cytochrome oxidase II and cytochrome *b* sequence data for 15 taxa resulted in the same well resolved topology (Figure 4.10; hereafter referred to as the Bayes^{comsmall} topology) with high posterior probability support values (>0.80), regardless of whether a GTR + I + Γ, GTR + SSR_2 or GTR + SSR_6 model of nucleotide evolution was assumed. The highest likelihood score was obtained under a GTR + SSR_6 model (mean ln *L* = -7525), followed by a GTR + I + Γ model (mean ln *L* = -7536) and a GTR + SSR_2 model (mean ln *L* = -8242). As nodal support obtained under a GTR + SSR_6 and GTR + SSR_2 model was very similar, only Bayesian posterior probabilities under the GTR + SSR_6 model are indicated in Figure 4.10.

Rates calculated for corresponding codon positions of the two genes were very similar, with the mean estimated rate of each cytochrome oxidase II codon position falling within the 95% credible region estimated for the corresponding cytochrome *b* codon position, and vice versa (data not shown). The pattern of rate variation between codon positions was as predicted, with the highest substitution rate estimated for third codon positions, followed by first and then second codon positions. Rate estimates for the entire cytochrome oxidase II and cytochrome *b* amplicons were almost identical (0.96 for cytochrome oxidase II and 1.02 for cytochrome *b*), with the mean estimate of each gene lying within the 95% credible region of the other gene.

Maximum likelihood analyses

The single ML topology (ln *L* = -7515) obtained using the GTR+ I + Γ best-fit model indicated by Modeltest[8] was identical to the Bayes^{comsmall} topology. This ML topology was stable to parameter value changes, as iterative searches with parameter values re-estimated by maximum likelihood on the previous ML tree yielded an identical topology to that initially obtained.

Comparison of Bayesian and ML model parameter estimates

Base frequency parameter values, α and p_{inv} values were identical to the mean estimates for these parameters averaged across three independent Bayesian runs using the GTR + I + Γ model. The ML estimates of α (0.90) and p_{inv} (0.42) for the combined data set were very similar to that obtained for the cytochrome *b* data (α = 0.86, p_{inv} = 0.42). The most predominant substitutions as estimated by both ML and Bayesian methods were C↔T transitions, followed by A↔G transitions and A↔T transversions. Five nodes (Figure 4.10 nodes 1, 2, 9, 10, 11) were supported by bootstrap values of greater than 80%, complementing the significant posterior probability of these nodes.

[8] πA = 0.34; πC = 0.16; πG = 0.08; πT = 0.41; r_{AC} = 36 493; r_{AG} = 74 017; r_{AT} = 45 616; r_{CG} = 16 446; r_{CT} = 382 340; r_{GT} = 1; p_{inv} = 0.42; α = 0.90.

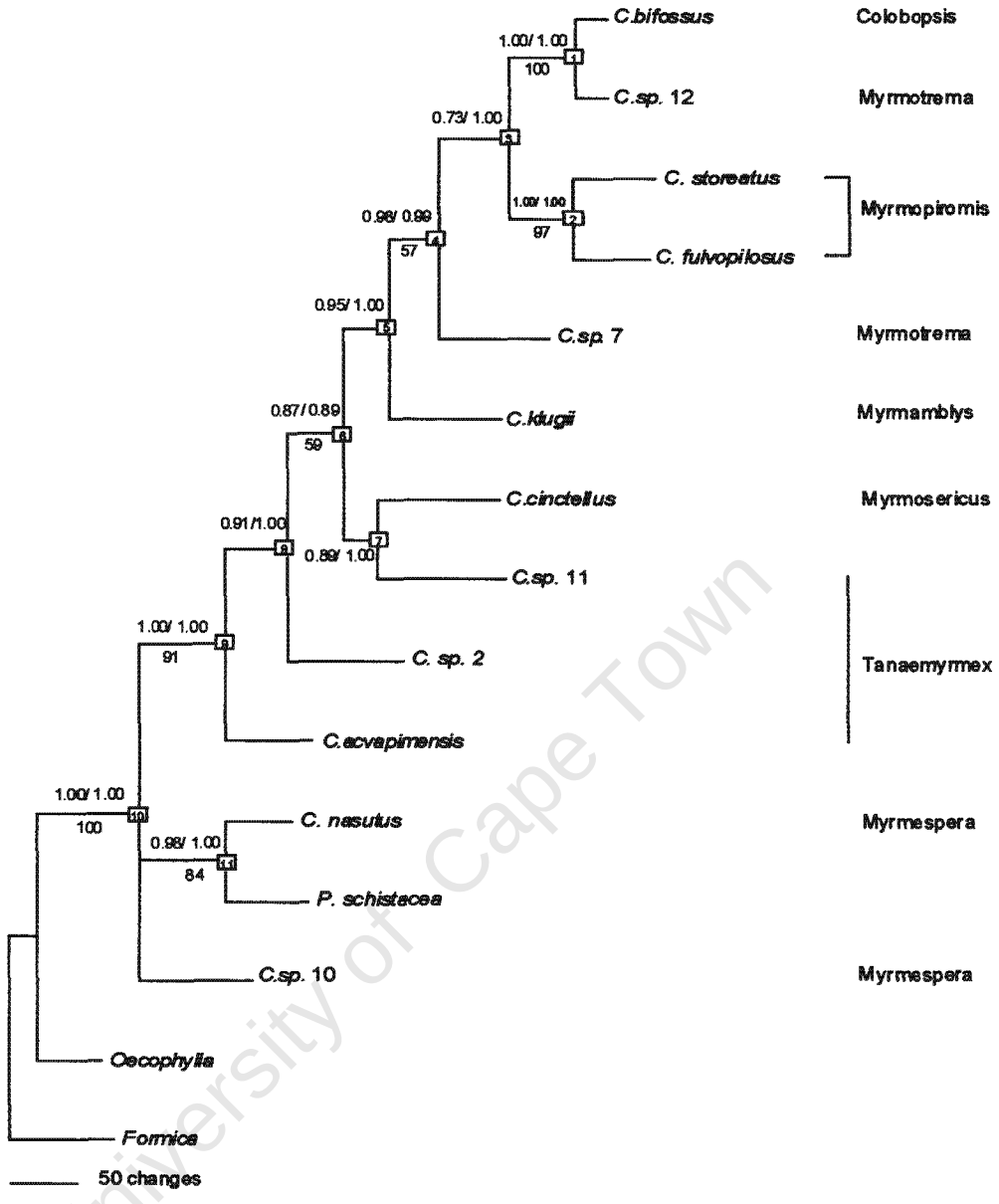


Figure 4.10. The 50% Majority rule Bayesian consensus topology based on the combined cytochrome oxidase II and cytochrome *b* data for a subset of taxa (Bayes^{comsmall}). Mean ln *L* = -7536 based on a GTR+ I + Γ model of sequence evolution. The two numbers adjacent to nodes represent posterior probabilities from two separate Bayesian analyses using a GTR + I + Γ model or GTR + SSR_6 model respectively. Numbers below branches indicate non-parametric ML bootstrap proportions greater than 50% based on 100 pseudoreplicate data sets. Subgeneric classifications are indicated to the right of the tree, with closed brackets representing monophyletic groups. Branch lengths are drawn proportional to the number of changes as indicated by the scale bar.

Comparison of combined Bayesian topology with cytochrome b and cytochrome oxidase II Bayesian topologies under a GTR + I + Γ model

Bayesian analysis of the combined data using a GTR+ I + Γ model produced a topology more similar to the BayesΓ+I^{cytbsmall} topology (Figure 4.7 A) than the BayesΓ+I^{coiismall} topology (Figure 4.9 A). Eight clades were shared between the cytochrome *b* and combined topologies, whereas only five clades were shared between the cytochrome oxidase II and combined topologies. The symmetric difference, or number of clades present in only one of the two trees being compared, was four for the Bayes^{comsmall} topology and the BayesΓ+I^{cytbsmall} topology, and twelve for the comparison of the Bayes^{comsmall} topology with the BayesΓ+I^{coiismall} topology.

(ii) Parsimony analyses

Results of parsimony analyses are presented in Table 4.14. Numbered nodes correspond to nodes labelled on the BayesΓ+I^{comsmall} tree (Figure 4.10). Parsimony analyses of the combined data resulted in well-resolved trees. The four nodes supported by a posterior probability of 1.0 and strong ML bootstrap values (Figure 4.10, nodes 1, 2, 9 and 11) all received bootstrap support of more than 90% and Bremer decay indices greater than ten in all nucleotide-based analyses (node 9 only received moderate support in the two amino-acid based analyses). Nucleotide weighting schemes, with the exception of downweighting A↔T transversions, improved congruence with the reference topology. Downweighting transitions appeared to be the most effective strategy for increasing congruence with the reference topology, as optimal topologies generated by four different transition downweighting schemes shared all nodes present in the reference topology. The topology recovered from parsimony analysis in which a T↔C downweighting matrix was applied to all characters, in addition to containing all nodes present in the Bayes^{comsmall} tree, was unique in that each branch present was statistically supported. Parsimony analysis of the amino acid data using a Protpars cost matrix resulted in a more resolved tree than unweighted analysis (seven versus four nodes resolved).

Consistency indices and RI values for the combined data set topologies were moderate to high, and ranged from 0.38 to 0.76 and 0.29 to 0.65 respectively.

Combining the data resulted in statistical support for almost all nodes present in the MP or strict consensus trees. This is in contrast to the previous single-gene analyses presented, where many nodes present in the optimal or strict consensus topology collapsed after bootstrap analysis. As expected, clades present in both cytochrome oxidase II and cytochrome *b* reduced data sets were also present in the combined analysis with increased bootstrap and Bremer support values relative to those in the individual data sets.

Table 4.14. Congruence of combined parsimony topologies based on a subset of taxa with the Bayes^{comsmall} topology.

Nodes in Bayes^{comsmall}	Unweighted	3:1 Tv:Ti	1st 2:1 Tv:Ti 2nd 2:1 Tv:Ti 3rd 10:1 Tv:Ti	COII: Tv/Ti 6:1 CYTB: Tv:Ti 4:1	COII 1st Tv:Ti 3:1 COII 2nd Tv:Ti 3:1 COII 3rd Tv:Ti 43:1 Cytb 1st Tv:Ti 1:1 Cytb 2nd Tv:Ti 2:1 Cytb 3rd Tv:Ti 21:1	Ti = 0 all positions	Ti = 0 3rd codon positions	A↔T downweighted 2:1 ALL position	A↔T downweighted 2:1 3rd positions	T↔C downweighted 6:1	Successive weighting	6-*P* parsimony (combined matrix)	6-*P* parsimony (COII+Cytb matrices)	Unweighted protein parsimony	ProtPars parsimony
1	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
2	●●●	●●●	●●●	●●●	●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
3	●	+	-	●	-	-	-	-	-	●	●●●	-	●	-	-	-
4	-	+	-	●●	-	●	●	-	-	●	-	●●	●	-	●	-
5	-	●	-	+	-	-	-	+	-	●	-	+	+	-	-	-
6	-	●●	●●	●●	●	●●	●	-	-	●●	●●●	●●●	●●	-	●	-
7	+	●	●	●	●	●	●	-	-	●●	●●●	●●	●●	-	-	-
8	●●	●	-	●	-	-	-	+	-	●	●●●	-	●	-	-	-
9	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●	●●	●●
10	●●●	●●●	●	●●●	-	●●	●●	●●●	●●●	●●●	●●●	●●	●●●	-	●	●
11	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
Length	1476	2771	6154	3944	16 018	623	844	2509	2234	5046	378	2038	2473	394	389
CI	0.43	0.42	0.38	0.42	0.38	0.38	0.42	0.40	0.40	0.45	0.76	0.59	0.42	0.55	0.57
RI	0.36	0.41	0.37	0.42	0.38	0.43	0.44	0.30	0.29	0.44	0.65	0.43	0.43	0.39	0.50
No. MP trees	1	1	1	2	1	2	8	1	2	1	1	1	1	20	13
No. congruent nodes	8/11	11/11	7/11	11/11	9/11	8/11	8/11	7/11	5/11	11/11	8/11	9/11	11/11	4/11	7/11
No. congruent nodes with bootstrap ≥ 50%	7/11	9/11	7/11	10/11	6/11	8/11	8/11	5/11	5/11	11/11	8/11	8/11	10/11	4/11	7/11

Note: Tree statistics given for each analysis are based on the bootstrap consensus topology.
●●● indicates strong bootstrap support (90% - 100%) and a Bremer support value ≥ 10
●● indicates medium bootstrap support (70-89%) and Bremer support of > 3
● indicates weak bootstrap support (50-69%)
+ indicates that the node was present in the MP tree/strict consensus tree but with bootstrap < 50%
- indicates that clade was not present in the MP/strict consensus tree or the bootstrap consensus tree

(iii) Phylogenetic relationships

Seven out of eleven nodes received significant support in both Bayesian analyses (Figure 4.10, nodes 1, 2, 4, 5, 9, 10, 11). Of these seven nodes, five received moderate-to-strong ML bootstrap support (Figure 4.10, nodes 1, 2, 9, 10, 11). The sister-association of *C. bifossus* and *C. sp.*12 was maximally supported in Bayesian, ML and parsimony analyses (node 1, Figure 4.10). The hypothesis of monophyly of subgenus *Myrmopiromis* was also significantly supported, with *C. storeatus* and *C. fulvopilosus* associating together in a clade with significant Bayesian posterior probability, and strong ML and MP bootstrap support (node 2, Figure 4.10). *Camponotus sp. 7* was sister taxon to a mixed group comprising ((*C. bifossus* + *C. sp.* 12) + (*C. storeatus* + *C. fulvopilosus*)) with significant Bayesian posterior probability, although this relationship was only weakly supported by ML and MP bootstrap values (node 4, Figure 4.10). *C. klugii* was sister to the group subtended by node 4, a relationship significantly supported in both Bayesian analyses (node 5, Figure 4.10). Furthermore, the combined data strongly favoured a sister association between *C. acvapimensis* and a large paraphyletic assemblage comprising species of subgenera *Colobopsis*, *Myrmotrema*, *Myrmopiromis*, *Myrmamblys*, *Myrmosericus* and *Tanaemyrmex* (node 9, Figure 4.10).

(iv) Partitioned Bremer support

Partitioned Bremer support analysis was performed using the combined data set with T↔C transitions downweighted by a factor of six, as bootstrap analysis of this weighted matrix resulted in a bootstrap consensus topology identical to the Bayes^{comsmall} tree. Nodes 1 to 11 in Table 4.15 therefore correspond to those in Figure 4.10.

Results of partitioned Bremer support analyses indicate that the bulk of the support for the nodes in the combined analysis parsimony tree are provided by the cytochrome *b* sequence data (82% of total ingroup support). In

particular, the third codon positions of cytochrome *b* provide 59% of the total ingroup support. The cytochrome oxidase II partition contributed only 18% of the total Bremer support, with the first codon positions contributing the majority of the support (20%). The third codon positions appear to support alternative groupings not found on the combined analysis tree, as indicated by the negative values for many of the nodes[9] (Table 4.15). Of the 11 nodes analysed, nine received greater support from the cytochrome *b* sequence data, whereas only two were more strongly supported by cytochrome oxidase II data. The greater support for node 11 provided by the cytochrome oxidase II data may be an artefact of missing data, as 291 bases are missing from the two outgroups relative to the ingroups for the cytochrome *b* data, whereas only seven bases are missing in the outgroup taxa relative to the ingroups in the cytochrome oxidase II alignment.

The ratio of the standardized Bremer support for COII:Cyt *b* was 0.34 for the ingroup taxa. This indicates that the low (non-standardized) PBS ratio of 0.22 calculated for the ingroup taxa is not simply an artefact of the smaller number of characters in the cytochrome oxidase II partition: even with data set size taken into account, the cytochrome oxidase II sequences contribute less to the overall support than do the cytochrome *b* sequences.

[9]Positive values within a combined analysis framework indicate that a given partition provides support for that particular node over the alternative relationships specified in the most parsimonious tree(s) not containing that node. Negative values in the combined analysis framework indicate that the length of a partition is shorter on the topology of the alternative tree(s) not containing a given node and therefore provide contradictory evidence for the relationship found in the simultaneous analysis tree (Baker and DeSalle, 1997, Baker *et al*, 1998).

Table 4.15. Overall and partitioned Bremer support for the combined analysis parsimony tree with T↔C transitions downweighted 6:1.

		Cytochrome oxidase II				Cytochrome *b*				
Node No.	Bremer Support	All	1st codons	2nd codons	3rd codons	All	1st codons	2nd codons	3rd codons	PBS Ratio COII:Cyt *b* All positions
1	93	13	17	1	-5	80	23	1	56	0.16
2	61	20	14	6	0	41	19	1	21	0.48
3	9	7	0	0	7	2	-5	-1	8	3.50
4	18	2	6	-1	-3	16	11	5	0	0.12
5	14	-15	7	0	-22	29	2	6	21	-0.39
6	41	10	0	0	10	31	-5	-1	37	0.32
7	20	-2	-2	0	0	22	0	0	22	-0.23
8	13	4	7	0	-3	9	2	-1	8	0.44
9	69	49	23	2	24	20	8	0	12	2.45
10	33	-21	2	-1	-22	54	15	6	33	-0.39
11	236	167	56	39	72	69	26	14	29	2.42
Total	607	234	130	46	58	373	96	30	247	0.62
% Total		38	21	8	9	62	16	5	41	
Total Ingroup	371	67	74	7	-14	304	70	16	218	0.22
% Total ingroup		18	20	2	4	82	19	4	59	

III *Topological incongruence between cytochrome b-based and cytochrome oxidase II-based phylogenies*

Phylogenetic reconstructions of species relationships based on separate analyses of the two different mitochondrial genes were not congruent, even though these two genes are presumed to share the same evolutionary history due to their position on the non-recombining mitochondrial genome. This statistical incongruence is reflected by the results of Shimodaira-Hasegawa tests presented in Table 4.16.

Table 4.16. Results of Shimodaira-Hasegawa test for competing hypotheses generated by the cytochrome *b*, cytochrome oxidase and combined reduced data sets.

	SH test		
Competing Topologies	Cyt *b* small	COII small	Combined small
Cyt *b*			
$Bayes\Gamma+I^{cytbsmall}$	optimal	0.057	0.510
$BayesSSR_6^{cytbsmall}$	0.776	0.038*	0.467
$ML^{cytbsmall}$	0.750	0.154	0.853
COII			
$Bayes\Gamma+I^{coiismall}$	0.008*	optimal	0.046*
$BayesSSR_6^{coiismall}$	0.004*	0.491	0.020*
$ML^{coiismall1}$	0.006*	0.733	0.033*
Combined			
$Bayes^{comsmall}$	0.755	0.145	1.000
MP tree with T↔C transitions downweighted 6:1[10]	0.777	0.145	optimal

* indicates significance at $P < 0.05$

[10] Although this MP tree contains all nodes present in the $Bayes^{comsmall}$ topology, the position of *Camponotus sp.* 10 is more resolved in the MP topology, where it is sister taxon to a group comprising *C. nasutus* and *P. schistacea*. In the $Bayes^{comsmall}$ tree, the position of *Camponotus sp.* 10 with respect to (*C. nasutus* + *P. schistacea*) is unresolved.

The cytochrome *b* data rejected the three cytochrome oxidase II-derived topologies, as did the combined data set. The cytochrome oxidase II data, in contrast, only rejected the cytochrome *b* Bayesian topology based on a GTR + SSR_6 nucleotide evolution model. The two topologies based on the combined cytochrome oxidase II and cytochrome *b* data were not rejected by any of the data sets.

In order to identify the source of incongruence between the cytochrome *b*- and cytochrome oxidase II-based topologies, posterior probability and ML bootstrap support for conflicting nodes in the different topologies were examined. Two well-supported nodes were present in all cytochrome *b* model-based topologies, but not present in any of the cytochrome oxidase II topologies.

The first node corresponds to node 5 in the BayesΓ+$I^{cytbsmall}$ tree in Figure. 4.7A, uniting *C. cinctellus* and *C. sp.*11 as sister species with significant posterior probability (bpp = 0.95 and bpp = 1.00 in the BayesΓ+$I^{cytbsmall}$ and Bayes$SSR_3^{cytbsmall}$ topologies respectively), and moderate bootstrap support in the $ML^{cytbsmall}$ tree (72%). The second node (Figure 4.7A node 9) groups *C. nasutus* and *P. schistacea* as sister species with a posterior probability of 1.00 in both Bayesian topologies, and a bootstrap value of 99% in the $ML^{cytbsmall}$ topology. No nodes with strong posterior probability support and moderate to strong bootstrap support were present in any of the cytochrome oxidase II model-based topologies that were not also present in the cytochrome *b* topologies.

To test whether the incongruence between topologies could be localised to these two clades, three constraint topologies were constructed in which 1) *C. cinctellus* and *C. sp.*11 were constrained to be monophyletic; 2) *C. nasutus* and *P. schistacea* were constrained to be monophyletic and 3) both these clades were constrained to be monophyletic. Maximum likelihood searches with these topological constraints enforced were then implemented using the cytochrome oxidase II sequences.

The cytochrome *b* data could not reject the optimal cytochrome oxidase topologies recovered when both topological constraints were enforced (Table 4.17). Enforcing constraint 1 resulted in the cytochrome *b* data rejecting the cytochrome oxidase II optimal constraint topologies. However, when *C. nasutus* and *P. schistacea* were constrained to be sister taxa, the cytochrome *b* data no longer rejected the optimal cytochrome oxidase II constraint tree. This indicates that it was the separation of these two species in the optimal Bayesian and ML topologies inferred from the cytochrome oxidase II sequences that was responsible for the significant topological incongruence between optimal trees obtained from each data set. The cytochrome oxidase II data did not reject any of the constraint topologies as significantly worse explanations of the data (results not shown).

It therefore appears that although the cytochrome *b* sequence data contains strong signal for an association between *C. nasutus* and *P. schistacea*, the cytochrome oxidase II sequence data is ambivalent about the association of these two taxa.

Table 4.17. Localisation of topological incongruence between cytochrome *b*-based and cytochrome oxidase II-based topologies, using SH tests.

Topology	-ln *L*	Difference -ln *L*	*P*-value
BayesΓ+I^{coiismall}	4245.96	31.59	0.004*
BayesSSR$_3$^{coiismall}	4244.95	30.59	0.006*
ML^{coiismall}	4242.83	28.46	0.008*
Constraint 1: (*C. cinctellus*+*C.sp.2*) optimal tree 1	4241.25	26.89	0.015*
(*C. cinctellus*+*C.sp.2*) optimal tree 2	4240.81	26.44	0.014*
Constraint 2: (*C. nasutus* + *P. schistacea*)	4230.21	15.84	0.103
Constraint 3: (*C. cinctellus*+*C.sp.2*) & (*C. nasutus* + *P. schistacea*)	4228.74	14.37	0.116
BayesΓ+I^{cytbsmall}	4214.80	0.43	0.794
BayesSSR$_3$^{cytbsmall}	4216.24	1.87	0.830
ML^{cytbsmall}	4214.37	optimal	-

* indicates significance at $P < 0.05$

(IV) Topological reconstruction: combined cytochrome b and cytochrome oxidase II sequences for all taxa

(i) Model-based analyses

Bayesian analyses

Stationarity of log likelihood values of Markov chain sample points for each independent analysis was achieved after approximately 1 X 10^4 generations. Bayesian analysis of all taxa included in this study, using a GTR + I + Γ model of sequence evolution, resulted in one topology with the highest likelihood (mean ln *L* = -8273) presented in Figure 4.11 (hereafter referred to as the BayesΓ+I^{all} tree). Two alternative topologies obtained using the GTR + SSR_6 model (mean ln *L* = -8345) or two gene-partition specific rates for cytochrome oxidase II and cytochrome *b* (GTR + SSR_2 model, mean ln *L* = -9138) were very similar to the optimal tree. The minor topological differences between the three trees were limited to poorly supported nodes (posterior probability support < 0.90) and concerned the placement of three taxa, namely *C. sp. 24*, *C. sp. 2* and *C.sp.* 10. Branch length estimates for the three topologies were also very similar. Estimates of codon specific rates and gene specific rates were identical to those estimated for the reduced combined data set (data set E). Despite the large proportion of missing data, the BayesΓ+I^{all} tree was well resolved, with 21 out of 22 possible nodes resolved. Nine of these nodes (Figure 4.11, nodes 4, 7, 8, 9, 11, 15, 17, 19, 21) received posterior probability support of 1.0 in all three Bayesian analyses.

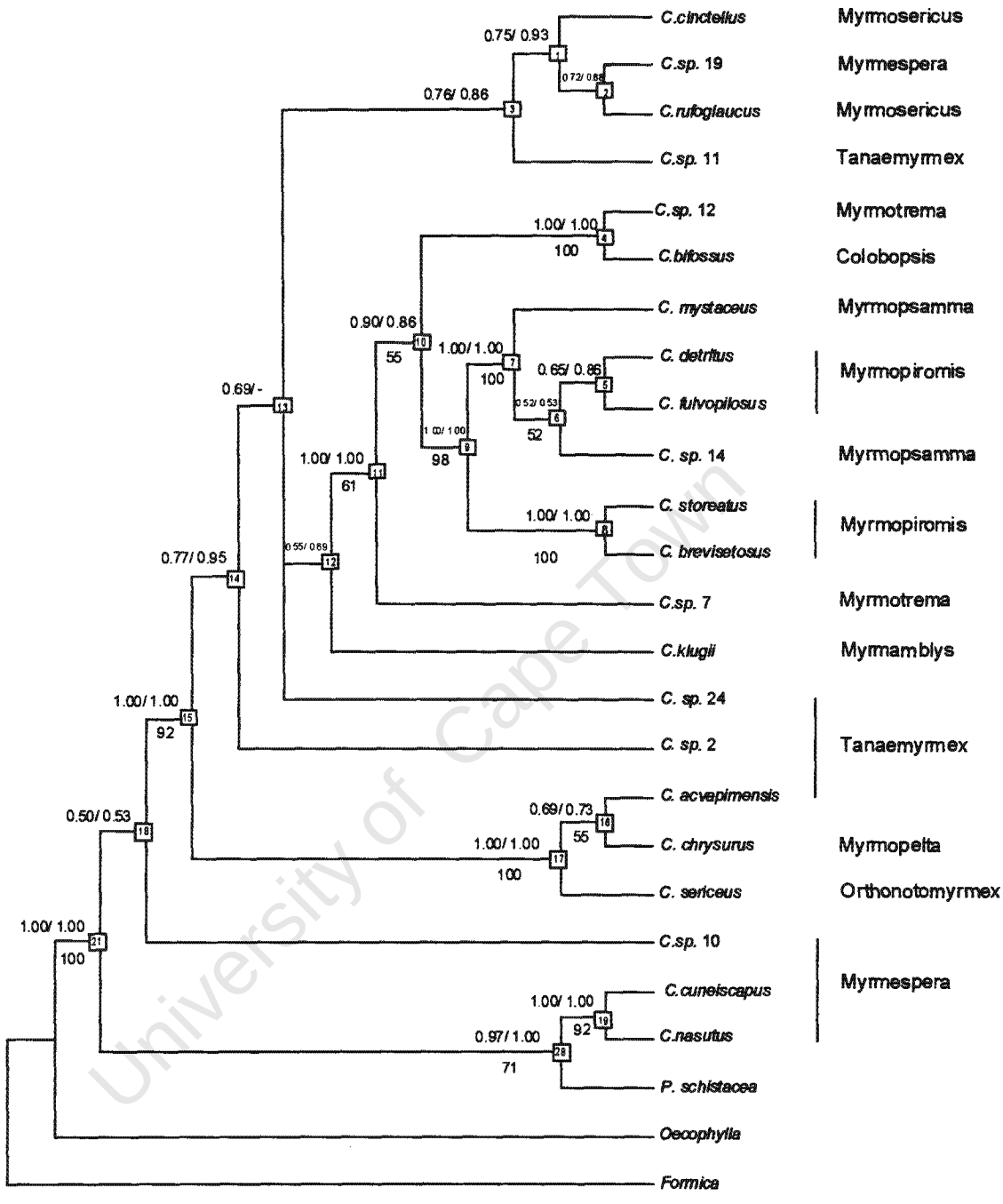


Figure 4.11. The 50% Majority rule Bayesian consensus cladogram estimated from the combined molecular data set for all taxa sampled based on a GTR + I + Γ model of nucleotide substitution (BayesΓ+I^{all} tree). The two numbers adjacent to nodes represent posterior probabilities from two separate Bayesian analyses using a GTR + I + Γ model or GTR + SSR_6 model respectively. Numbers below branches indicate non-parametric ML bootstrap proportions greater than 50% based on 100 pseudoreplicate data sets.

As was observed for the combined sequences for a subset of taxa, likelihood ratio tests implemented in Modeltest indicated that a GTR + I + Γ model of sequence evolution best explained the observed sequences. Base frequency estimates and p_{inv}[11] were essentially identical to those obtained for the reduced combined data set. Only the value of α differed between the two data sets, with a slightly lower value of α estimated for all combined sequences (α = 0.71) compared to that estimated for the reduced combined data set (α = 0.90). The relative rates[12] for the various substitutions, though numerically not equivalent, followed the same ranking order as for the taxon-reduced combined data set.

Maximum likelihood analyses

One most likely topology with ln *L* = -8239 (hereafter referred to as ML^{all}) was obtained after a heuristic search. This tree was very similar to the $Bayes\Gamma+I^{all}$ topology with two exceptions: (i) *C. cinctellus* and *C. sp.* 19 were sister in the ML^{all} tree with bootstrap support of 62%, whereas *C. sp.* 19 was reconstructed as sister taxon to *C. rufoglaucus* in the $Bayes\Gamma+I^{all}$ topology and (ii) *C. klugii* and *C. sp.* 24 appear as sister taxa in the ML^{all} tree, whereas no such relationship exists between these two species in the $Bayes\Gamma+I^{all}$ tree. However, this relationship did not receive any significant support. The ML topology was stable to changes in parameter values, as iterative searches with parameter values estimated by ML on the ML^{all} tree recovered an identical topology to that initially obtained.

(ii) Relative fit of substitution models

Assuming a single model of evolution for the nucleotide substitution process for both genes provided a significantly worse fit to the data than allowing each partition to evolve under its own model, as evaluated by LRTs. The likelihood for each partition was estimated separately under a GTR + I + Γ model on the unweighted MP topology based on the combined sequences; these two

[11] $\pi A = 0.35$; $\pi C = 0.16$; $\pi G = 0.08$; $\pi T = 0.41$, $p_{inv} = 0.41$; $\alpha = 0.71$.

[12] $r_{AC} = 141\ 727$; $r_{AG} = 357\ 031$; $r_{AT} = 208\ 694$; $r_{CG} = 99\ 289$; $r_{CT} = 1\ 930\ 050$; $r_{GT} = 1$.

values were then summed and compared to the likelihood score obtained when both partitions were combined (Huelsenbeck and Rannala, 1997; Waits *et al.*, 1999).

For the subset of taxa for which both gene sequences were available, allowing a single GTR + I + Γ model for the combined sequences (ln L = -7515) and comparing this value to the sum of the log likelihoods for each partition (ln L = -7362) indicated a significant improvement in fit associated with allowing a unique process for each partition ($\chi^2_{[10]}$ = 306, $P < 0.0001$). Results for the data set containing combined cytochrome *b* and cytochrome oxidase II sequences for all taxa were similar. Allowing each partition to evolve under its own unique GTR + I + Γ model (ln L = -8081) provided a significantly better fit to the data than when assuming a single process for both combined sequences (ln L = -8238; $\chi^2_{[10]}$ = 314, $P < 0.0001$).

For both combined data sets, the fit of a GTR + SSR_6 model to the data was significantly better than that of a GTR + SSR_2 model of nucleotide evolution, which assumes a single uniform rate for all characters in each gene partition (data set E: $\chi^2_{[4]}$ = 1436, $P < 0.01$; data set F: $\chi^2_{[4]}$ =1600, $P < 0.01$).

(iii) Congruence of clade posterior probabilities

There was a strong correlation between the posterior probability support for shared clades based on a GTR + I + Γ model and the posterior probability for clades based on a GTR + SSR_6 model of sequence evolution (Figure 4.12; r_s = 0.95, d.f. = 20, $P < 0.01$). This indicates that the probability density distributions for these two models of sequence evolution were similar, despite the different models of rate heterogeneity assumed, thereby increasing confidence in those clades that received strong support in both analyses.

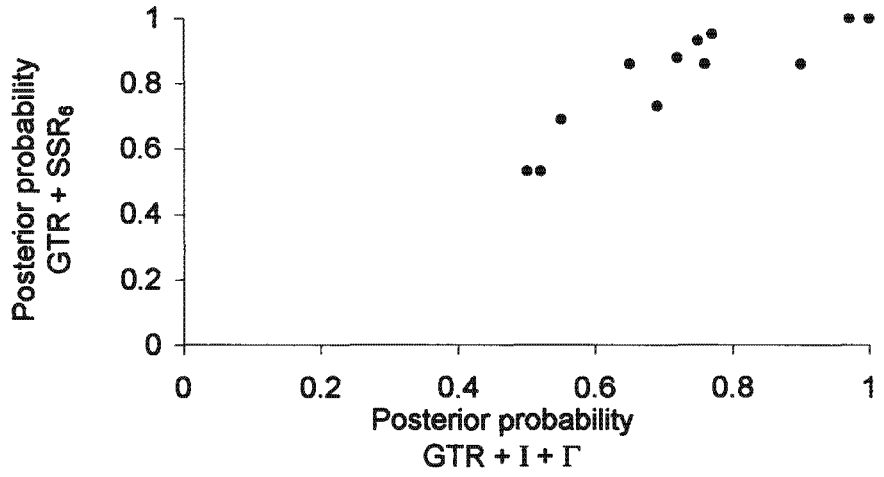


Figure 4.12. Plot of clade posterior probability support estimates from Bayesian analyses of the complete combined data set (data set F) implementing a GTR + I + Γ or GTR + SSR_6 model of sequence evolution.

There was a trend for the GTR + SSR_6 nodal support estimates to be greater than those obtained under a GTR + I + Γ model of evolution (SPSS v9.0 Wilcoxon Signed Ranks test, $Z = -2.58$, $P < 0.05$). This is consistent with the findings of Sullivan *et al.* (1997) that statistical support for nodes under a simplistic model of evolution may be artificially inflated due to underestimation of the actual number of multiple substitutions that have occurred on long branches.

(iv) Parsimony analyses

Results of parsimony analyses of the combined data are presented in Table 4.18. Numbered nodes correspond to nodes labelled on the BayesΓ+I^{all} tree (Figure 4.11). All nucleotide weighting schemes, with the exception of A↔T transversion downweighting, either improved or had no effect on congruence with the reference topology compared to the unweighted parsimony tree. No weighting strategy managed to retrieve the same topology as the Bayesian reference topology. Bootstrapped MP topologies were not well resolved, with only nine to 13 out of a possible 22 nodes resolved. The only node present in the MP trees not observed in the BayesΓ+I^{all} topology supported a sister

relationship between *C. nasutus and P. schistacea*, albeit with weak bootstrap values and low Bremer decay indices.

A large clade comprising all included species in the subgenera *Myrmopiromis* and *Myrmopsamma* was recovered by all MP searches (Table 4.18, node 9). Within this clade, the sister relationship of *C. storeatus* and *C. brevisetosus* was strongly supported in all analyses (Table 4.18, node 8). This clade formed a sister clade to a well-supported group comprising (((*C.detritus*, *C. fulvopilosus*), *C. sp. 14*), *C. mystaceus*), subtended by node 7. The other group that consistently appeared in the various MP analyses with bootstrap support of ≥ 98% was that containing *C. bifossus* and *C. sp. 12*, which was as expected given the strong support for this clade in all previous analyses based on the two mitochondrial genes. A clade comprising ((*C. acvapimensis*, *C. chrysurus*), C. *sericeus*) (Table 4.18, node 17) was also recovered in all parsimony analyses with varying levels of bootstrap support. Trees based on amino acid residues were poorly resolved, although they had higher CI and RI values compared to the nucleotide-based parsimony topologies. Multiple MP trees were recovered in the various parsimony analyses. This is consistent with the findings of Wiens and Reeder (1995) and Wiens (1998) that including large amounts of missing data in phylogenetic analyses increases the number of most parsimonious trees recovered.

CI values ranged from 0.30 to 0.52, and RI values from 0.27 to 0.52 for the complete combined data set.

Table 4.18. Congruence of combined (all) MP topologies with the BayesΓ+I^{all} topology.

Nodes in BayesΓ+I^{all}	Unweighted	Tv:Ti 4:1	1st, 2nd, 3rd Tv:Ti 2 1 Tv:Ti 2:1 Tv:Ti 7:1 Tv:Ti	COII: Tv/Ti 23:1 CYTB: Tv:Ti 4:1	COII 1st Tv:Ti 7:1 COII 2nd Tv:Ti 3:1 COII 3rd Tv:Ti 14:1 Cytb 1st Tv:Ti 1:1 Cytb 2nd Tv:Ti 2:1 Cytb 3rd Tv:Ti 14:1	Ti = 0 all positions	Ti = 0 3rd codon positions	A↔T downweighted 2:1 ALL position	A↔T downweighted 2:1 3rd positions	T↔C downweighted 6:1	Successive weighting	6-P parsimony (combined matrix)	6-P parsimony (COII+Cytb matrices)	Unweighted protein parsimony	ProtPars parsimony
1	-	+	-	+	-	+	-	-	-	+	●	-	+	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	+	-	-	-	-	-	-	-	-	-	●●	-	-	-	-
4	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●
5	●●●	●●	●●	●●	●●	-	●	●●	●●	●●	●●●	●●	-	-	-
6	●●	●●	●●	●●	●●	●●	●●	●	●	●●	●●	●	●●	●●	●●
7	●●	●●	●	+	+	●	●●●	●●	●●	●●	●●●	●●	●●	-	-
8	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
9	●●●	●●	+	+	+	●●	●●●	●●●	●●●	●●●	●●●	●●●	●●	●	●
10	-	+	+	+	+	-	-	-	-	+	-	-	-	-	-
11	-	+	+	+	+	●	-	-	-	●	-	-	●	-	-
12	-	+	+	+	+	+	-	-	-	+	●●	+	+	-	-
13	-	-	+	+	+	-	-	-	-	+	-	-	-	-	-
14	-	+	-	-	-	+	-	-	-	-	●	+	+	-	-
15	●	●●	+	-	-	●●	●●	-	-	●	●●	●	●	●●	●●
16	●●	●●	●●	●	●●	-	●	●●	●●	●●	●●	●●	-	●	●
17	●●●	●●●	●●	●	●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●	●
18	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
19	-	-	-	+	-	-	-	-	-	-	●●	-	+	-	-
20	●	●●	●●	-	●	●	●●	●	●	●	●●	●	●	-	●
21	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
Length	1848	4322	6322	12 913	11 221	765	1048	3031	2818	6385	359	2925	2551	451	474
CI	0.37	0.36	0.31	0.30	0.31	0.33	0.36	0.36	0.35	0.40	0.56	0.35	0.36	0.52	0.51
RI	0.39	0.43	0.35	0.27	0.34	0.48	0.50	0.37	0.32	0.46	0.67	0.40	0.48	0.49	0.52
No. MP trees	36	3	3	3	3	27	9	6	6	8	3	2	27	696	120
No. congruent nodes	12/21	16/21	15/21	17/21	14/21	13/21	12/21	10/21	10/21	16/21	16/21	13/21	14/21	8/21	9/21
No. congruent nodes with bootstrap ≥ 50%	11/21	11/21	9/21	7/21	8/21	10/21	12/21	10/21	10/21	12/21	16/21	11/21	10/21	8/21	9/21

Note: Tree statistics given for each analysis are based on the bootstrap consensus topology.
●●● indicates strong bootstrap support (90% - 100%) and a Bremer support value ≥ 10
●● indicates medium bootstrap support (70-89%) and Bremer support of > 3
● indicates weak bootstrap support (50-69%)
+ indicates that the node was present in the MP tree/strict consensus tree but with bootstrap < 50%
- indicates that clade was not present in the MP/strict consensus tree or the bootstrap consensus tree

Part D: Morphological & combined molecular-morphological phylogenetic analyses

I *Tree reconstruction: morphological and behavioural characters*

(i) Parsimony analysis

Four MP trees were obtained after unweighted parsimony analysis (Figure 4.13; strict consensus tree: length = 23, CI = 0.70, RI = 0.85). Only 12 out of 22 nodes were resolved, with the low Bremer and bootstrap support values reflecting the small number of characters and character states upon which this analysis was based. The one exception was the branch subtending all species in the subgenus *Myrmopiromis*, which had bootstrap support of 90% and a Bremer support value of 2. Morphological characters only provided weak support for the monophyly of all ingroup species relative to the two outgroups, in contrast to molecular data.

(ii) Phylogenetic relationships

Polyrhachis schistacea appeared as sister taxon to *C. sericeus* with moderate support, lending additional support to the hypothesis that *Camponotus* is paraphyletic. A clade comprising both species in the subgenus *Myrmotrema* (*C. sp.* 12, *C. sp.* 7) and *C. bifossus* was weakly supported. The monophyly of the subgenus *Myrmopsamma*, represented by *C. mystaceus* and *C. sp.* 14 obtained only weak supported in the bootstrap consensus tree. The subgenus *Tanaemyrmex* appeared to be polyphyletic, and there was no support for the monophyly of the subgenus *Myrmosericus*.

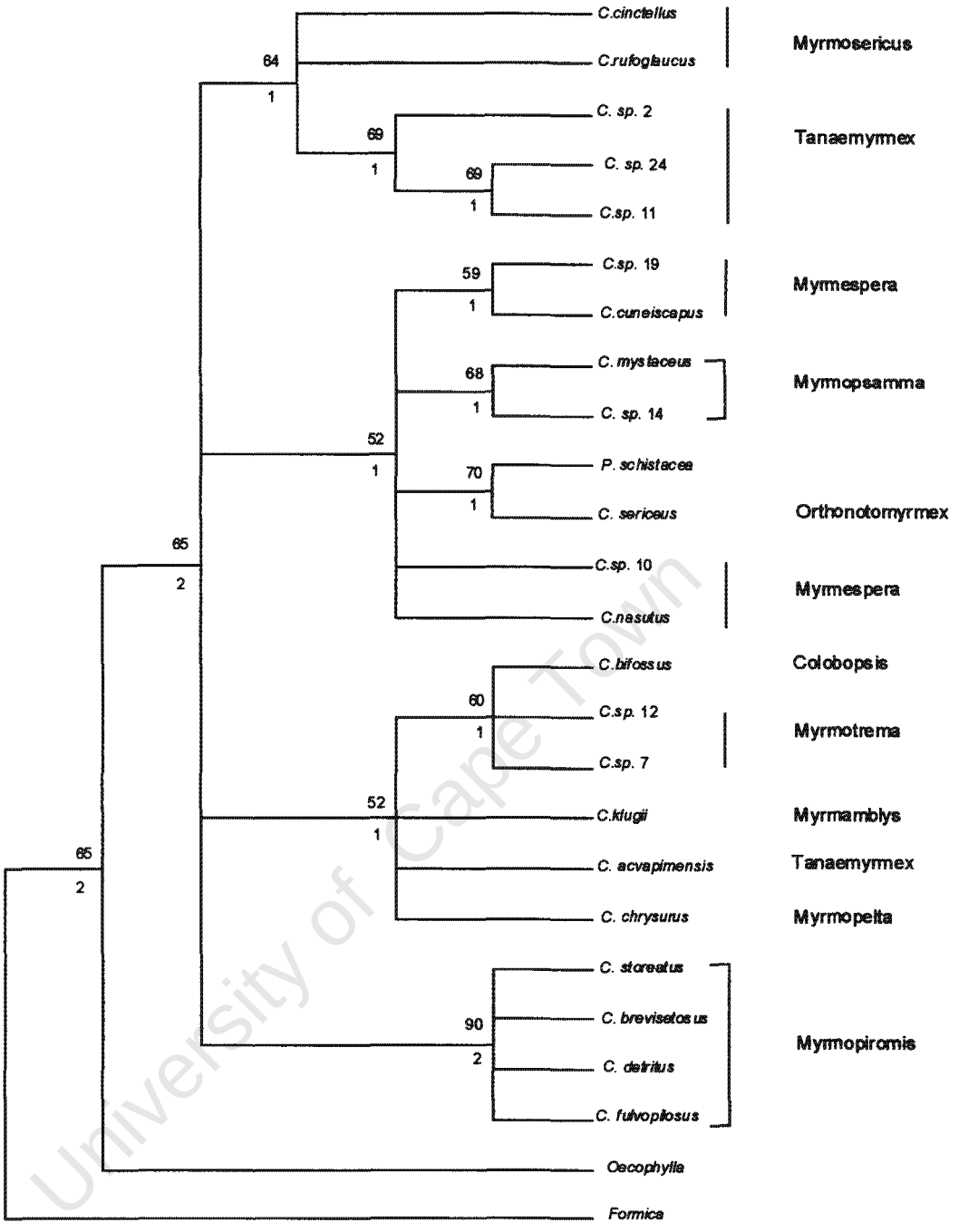


Figure 4.13. Strict consensus of four equally parsimonious trees based on morphological and behavioural characters. Numbers above branches represent non-parametric bootstrap values obtained from 1000 pseudoreplicates; numbers below branches are Bremer support values.

II *Combined molecular and morphological analyses*

The ILD test could not reject the hypothesis of character homogeneity for the morphological and molecular data, regardless of which molecular partition was used (results not shown). However, these non-significant results could also be as a result of the small size of the morphological partition relative to the molecular partition (Dowton and Austin, 2002).

Morphological characters contributed towards both topology and branch support in the combined analyses, despite the small size of the morphological partition (Table 4.19). For all molecular datasets, addition of morphological characters improved resolution relative to the molecular data alone, although novel relationships revealed in the combined analyses did not generally receive statistical support. The one exception to this was the novel association of *C. mystaceus* with *C. sp.* 14 (both in the subgenus *Myrmopsamma*) as sister taxa in analyses based on a combined cytochrome oxidase II + morphology matrix, as well as the cytochrome oxidase II + cytochrome *b* + morphology matrix. This association was supported by a bootstrap of 80% and a Bremer decay value of two in both analyses, even though this clade only received weak bootstrap (68%) and a Bremer decay value of one in the analysis based on morphological characters.

Both bootstrap support and Bremer decay indices for specific clades increased in the context of a combined molecular-morphological analysis. The interaction of the morphological data with the molecular data appears to be complex. Certain nodes received increased support in the context of a combined analysis of all data partitions, even though there was no increased support for these nodes when only one of the molecular partitions was analysed in conjunction with the morphological characters. Furthermore, the increased support for certain nodes observed when either cytochrome *b* or cytochrome oxidase II sequences were analysed in conjunction with morphological data, was not reflected in the analysis based on all available characters.

A clade uniting *C. detritus* and *C. fulvopilosus,* as well as a clade containing *C. sp.* 12 and *C. bifossus* received additional support when analysing morphological characters in conjunction with the cytochrome oxidase II sequences. Increased support for a clade comprising *C. detritus*, *C. fulvopilosus*, *C. mystaceus* and *C. sp.*14 was evident in the analysis based on morphological and cytochrome oxidase II characters, as well as the analysis based on all molecular characters and morphological characters. The association of the above clade with *C. brevisetosus* and *C. storeatus* received increased support when morphological characters were added to the cytochrome oxidase II sequences. The sister-group relationship of *C. fulvopilosus* and *C. storeatus*, as well as support for a node uniting *C. cuneiscapus*, *C. nasutus*, *C. sp.*10 and *P. schistacea* was strengthened when cytochrome *b* sequences and morphological characters were combined. The sister-group association of *C. acvapimensis* and *C. chrysurus* also received increased support when morphological characters were added to cytochrome *b* sequence data. This increased support was reflected in the consensus bootstrap topology based on a matrix combining the cytochrome *b*, cytochrome oxidase II and morphological characters. The association of *C. cuneiscapus* with *P. schistacea* and *C. nasutus* obtained increased support when morphological characters were added to a combined molecular matrix. Bootstrap analysis of this large combined matrix also resulted in increased support for the ingroup-outgroup node, and a node subtending a large clade comprising all ingroup taxa with the exception of *C. cuneiscapus*, *C. nasutus*, *C. debellator* and *P. schistacea*.

Relative to its size, the morphological partition provided marginally greater support than the molecular partitions alone. This was indicated by the increase in the PBS ratio of the morphological to molecular data relative to the minimum number of steps for each data partition. This effect may be due to the higher CI of the morphological data compared to the molecular data (Table 4.19) (Baker *et al.*, 1998).

Table 4.19. Effect of combining morphological with molecular data in a combined analysis framework. Results are presented for the combination of morphological data with the complete cytochrome *b*, cytochrome oxidase II and combined molecular data sets (data sets A, C, and F).

Molecular partition	No. nodes resolved in unweighted MP/strict consensus tree (mol)	No. nodes resolved in unweighted MP/strict consensus tree (com)	No. nodes B/S > in combined tree	No. nodes B/S < in combined tree	No. nodes B/S = in combined tree	No. nodes BM > in combined tree	No. nodes BM < in combined tree	No. nodes BM = in combined tree	Min. No. Steps Ratio mor:mol	PBS Ratio mor:mol	Difference	CI unweighted MP/strict consensus tree (mol)	CI unweighte MP/strict consensu tree (mor
Cyt *b*	14	15	3	0	5	3	0	5	0.03	0.04	+	0.44	0.56
COII	17	17	4	0	2	3	0	3	0.04	0.13	+	0.40	0.56
Cyt *b* + COII	9	14	5	1	4	7	1	2	0.02	0.15	+	0.35	0.70

Note: mol, molecules; mor, morphology; com = combined mor + mol data; B/S = bootstrap support; BM = Bremer support . Nodes were considered to have increased or decreased bootstrap support if this value differed by 2% and increased or decreased Bremer support if this value differed by 1. The 'Difference' column indicates which partition provides greater support relative to data set size, with a plus sign indicating more support from the morphological data.

III Phylogenetic informativeness of morphological characters

Results of randomisation tests of all morphological and ecological characters indicated that they were significantly more congruent with the molecular-based BayesΓ+I^{all} topology than would be expected by chance alone (PTP test, $P < 0.001$). When characters were examined individually, it was found that the characters that were the best predictors of the phylogenetic relationships suggested by molecular data, were the shape of the clypeal shelf, the shape of the hind tibae, the presence of thick blunt hairs on the gaster, the presence of a metapleural gland and whether the species nested arboreally (Table 4.20).

Table 4.20. Congruence of morphological characters with the BayesΓ+I^{all} topology based on 1000 permutation test replications.

Character description	*P* value
Number of mandibular teeth	0.09
Apical clypeal tooth length	0.21
Presence/absence of clypeal psammophore	1.00
Shape of clypeal shelf	0.03*
Shape of base of scape	1.00
Shape of petiole	1.00
Shape of dorsolateral margin of propodeum	1.00
Shape of hind tibae	0.01*
Presence/absence of thick blunt hairs on gaster	0.01*
Arboreal nesting	0.03*
Presence/absence of metapleural gland	0.03*
Dense pitting on head	0.08
Shape of back of head of minor worker	1.00
Polymorphic workers	0.07

* indicates significance at $P < 0.05$

Part E: Hypotheses testing

I *The molecular clock hypothesis*

The hypothesis that rates of substitution across branches were homogenous was rejected at the 95% confidence level for both cytochrome *b* and cytochrome oxidase II sequences. Rejection of homogenous rates of substitution was observed regardless of whether all sequenced taxa were included, or only a subset of taxa (cytochrome *b*: Dataset A: $\chi^2_{[17]} = 36$; $P = 0.005$; Dataset C: $\chi^2_{[13]} = 23$; $P = 0.04$; cytochrome oxidase II: Dataset B: $\chi^2_{[19]} = 36$; $P = 0.001$; Dataset D: $\chi^2_{[13]} = 27$; $P = 0.001$).

II *Subgeneric monophyly*

To evaluate whether the monophyly of the *Camponotus* subgenera included in this study was supported by molecular data, monophyly constraint topologies were constructed and ML heuristic searches performed with these topological constraints enforced. The most likely trees compatible with these constraints were then compared to the unconstrained ML and Bayesian topologies using Shimodaira-Hasegawa tests. All datasets, with the exception of the reduced cytochrome oxidase II data set (data set D), rejected the topologies with the various subgenera constrained to be monophyletic as significantly worse explanations of the data than the unconstrained topologies (results not shown). The failure of the reduced cytochrome oxidase II data set to reject the constraint topology may be due to the low resolving power of the cytochrome oxidase II sequence data, as commented on previously.

III *Camponotus monophyly*

Camponotus formed a paraphyletic group in all analyses, whether based on molecular or morphological data, with *P. schistacea* nested within the ingroup taxa. To evaluate whether *Camponotus* monophyly was a significantly worse

explanation of the data, ML heuristic searches were performed with *Camponotus* monophyly enforced. Shimodaira-Hasegawa tests comparing the resulting optimal constraint topology/topologies with the unconstrained ML and Bayesian topologies were all nonsignificant. This indicates that the cytochrome oxidase II and cytochrome *b* sequences, either separately or in combination, do not have the ability to differentiate between the two alternate hypotheses given the taxa sampled.

In order to further investigate the relationship between *Camponotus* and *Polyrhachis*, eight additional *Polyrhachis* species were included in a combined analysis. Cytochrome oxidase II and cytochrome *b* sequences from eight species (Appendix VI) were obtained from Genbank (Johnson and Crozier, unpublished data). These sequences were aligned to the combined sequences for the subset of taxa for which both gene sequences were available, which resulted in a matrix of 842 aligned nucleotides for 23 taxa. Hierarchical LRTs indicated that a GTR + I + Γ model best explained the observed sequences, therefore Bayesian analysis with sequences evolving under this model was implemented. Parameter values reached stationarity after 1 x 10^4 generations, with trees sampled during this phase discarded as burn-in. Three independent Bayesian runs all converged on the same consensus topology with similar nodal support (Figure 4.14). Bayesian parameter values for the GTR + I + Γ model averaged across the three runs were as follows: $\pi A = 0.36$; $\pi C = 0.17$, $\pi G = 0.05$; $\pi T = 0.41$; $r_{AC} = 5.20$; $r_{AG} = 33.35$; $r_{AT} = 5.39$; $r_{CG} = 6.68$; $r_{CT} = 51.96$; $r_{GT} = 1.00$; $p_{inv} = 0.38$; $\alpha = 0.77$.

Polyrhachis forms a significantly supported monophyletic group in the Bayesian consensus topology (Figure 4.14). The majority of *Camponotus* species also group together in a clade with significant posterior probability (bpp = 1.0). However, the position of *C. sp.* 10 and *C. nasutus* in the Bayesian consensus topology is unresolved. When nodes that appear in less than 50% of the posterior probability distribution are included, however, all *Camponotus* species do form a monophyletic group, in which *C. sp. 10* and *C. nasutus* are progressively basal to the well-supported *Camponotus* clade (bpp = 0.41 and

bpp = 0.26 respectively). However, given these non-significant support values, no strong statement regarding *Camponotus* monophyly can be made.

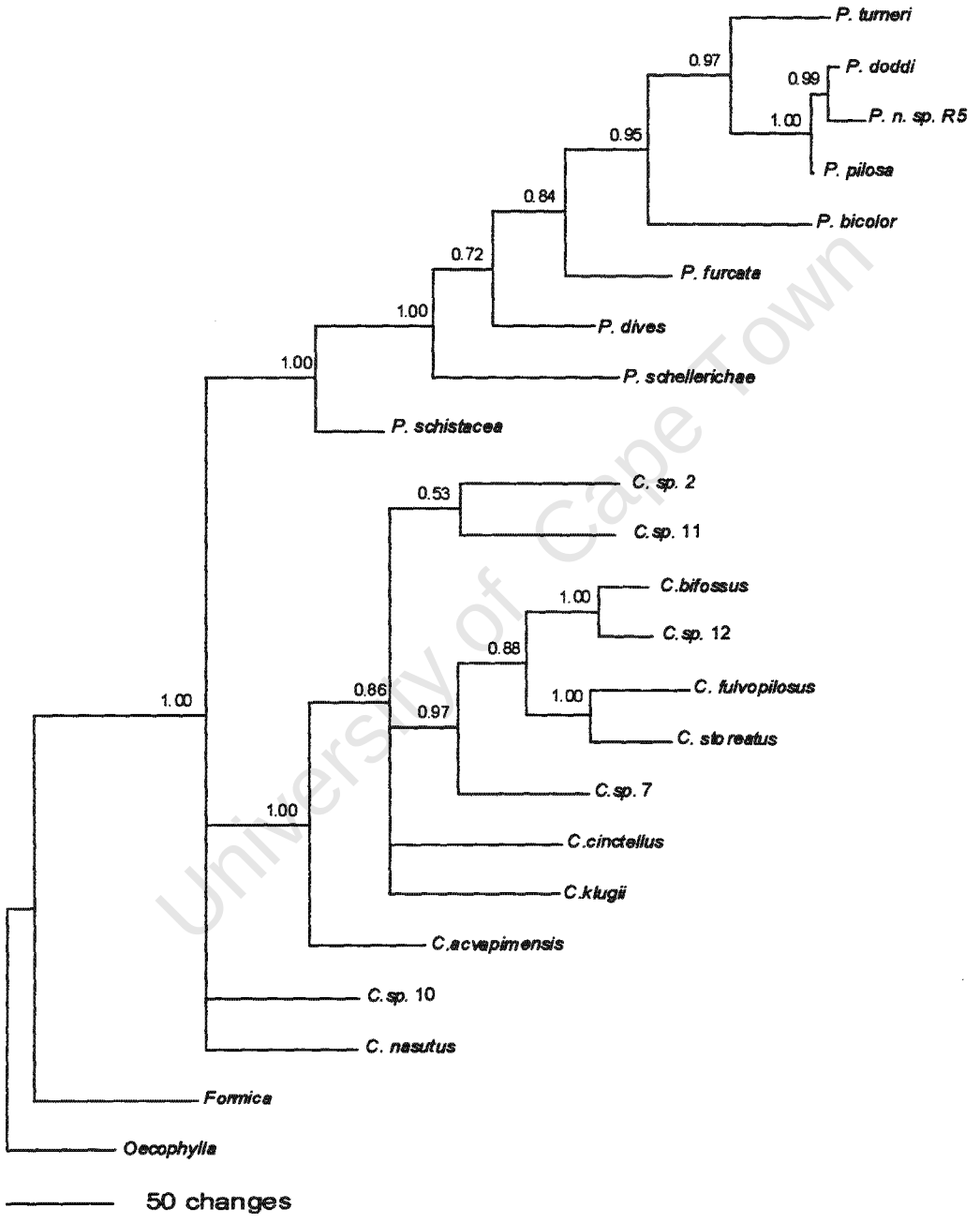


Figure 4.14. The 50% majority rule consensus tree from Bayesian analysis of combined cytochrome oxidase II and cytochrome *b* sequences for *Polyrhachis* and *Camponotus*, using a GTR + I + Γ model of sequence evolution (mean ln *L* = -7928). Numbers above branches indicate posterior probability support. Branch lengths are drawn proportional to the number of changes as indicated by the scale bar.

Identification of a cytochrome oxidase II pseudogene

A putative nuclear pseudogene was amplified from genomic DNA extracted from *C. rufoglaucus* using the insect universal primers TL2-J-3037 (Simon *et al.*, 1994) and C2-N-3665 (Chiotis *et al.*, 2000). A single band was observed after PCR amplification, which was identical in size to that expected for the mitochondrial amplicon. The sequence chromatogram was of excellent quality, with no sequence ambiguities (Appendix VII). Thus mis-identification of this gene as a pseudogene cannot be attributed to poor sequence quality. A nucleotide Blast search indicated homology to the cytochrome oxidase II sequences of other ant species in Genbank. However, closer examination of the Blast alignment results revealed numerous insertion-deletion events (indels). Alignment of this sequence to the *Camponotus* cytochrome oxidase II sequences elucidated in this study confirmed these observations.

Eight indels were observed relative to the cytochrome oxidase II sequences: four deletions (two of 2 bp, two of 4 bp) and four insertions (one of 1 bp, two of 2 bp and one of 4 bp). All of these indels would result in frameshift mutations. Uncorrected sequence divergence estimates were within the range observed for the mitochondrial cytochrome oxidase II sequences, ranging from 12% to 18% for ingroup comparisons, and increasing to 27% for comparisons with the outgroups. Base composition bias for the pseudogene was higher than that for the mitochondrial cytochrome oxidase II sequences (C = 0.40 versus C = 0.34), with the mitochondrial origin of this gene evident in its extreme A-T bias (80% A-T content). This sequence has been deposited in Genbank under Accession Number AF520584.

The clean sequence obtained from direct sequencing of the PCR product indicated that the pseudogene was preferentially amplified over the mitochondrial gene. Indeed, all attempts at amplifying the mitochondrial copy of COII for this species were unsuccessful. This is in contrast to other studies

where the pseudogene product was found to comprise only a small fraction of the total amplified product, with the mitochondrial gene copy predominating (e.g. Collura *et al.*, 1996; Williams and Knowlton, 2001).

The pseudogene amplified in *C. rufoglaucus* was clearly identifiable as such by the numerous indels resulting in frameshift mutations. However, pseudogenes may often be very similar in sequence to the mitochondrial gene due to similar rates of evolution or recent nuclear integration, and therefore remain undetected at this level. The following may serve as warning signs of a pseudogene: (i) more than one PCR product, (ii) consistent sequence ambiguities with both forward and reverse primers, (iii) nucleotide sequences radically different from those expected and (iv) unusual or contradictory tree topologies after phylogenetic analysis (Bensasson *et al.*, 2001). In this study, only single PCR products were obtained for cytochrome oxidase II amplifications, with no consistent sequence ambiguities observed with forward and reverse primers. Blast searches of the sequences indicated homology with other *Camponotus* and formicid species, and no indels were detected after sequence alignment, or stop codons upon translation to protein. Therefore, based on all the available evidence, it is highly unlikely that the other cytochrome oxidase II sequences generated in this study are pseudogenes. However, caution is warranted in future molecular phylogenetic studies of *Camponotus*.

Chapter 5

PHYLOGENETIC SYNTHESIS

Comparative phylogenetic utility of cytochrome *b* and cytochrome oxidase II

By almost all measures of phylogenetic utility, the performance of cytochrome *b* is superior to that of cytochrome oxidase II for resolving species-level relationships in *Camponotus*.

Although both mitochondrial genes display a high A+T bias that is most pronounced at third codon positions, this bias is greater for the cytochrome oxidase II sequences. Based on saturation curves, third codon positions of both genes appear to be saturated. However, the third codon positions of cytochrome oxidase II are more homoplasious than the third codon positions of cytochrome *b*, as inferred by the larger number of steps required to map these sites on the MP tree for those sequences. This observation is corroborated by results of *g*1 skew tests, which indicate that the third codon positions of cytochrome *b* provide the strongest phylogenetic signal compared to the first and second codon positions of this gene. In cytochrome oxidase II sequences on the other hand, the weakest signal is provided by third codon positions, and the strongest signal by the highly invariable second codon positions, followed by first codon positions. These findings are consistent with previous ant molecular studies, where high levels of homoplasy have been inferred for the third codon positions of cytochrome oxidase II (Chenuil and McKey, 1996; Wetterer *et al.*, 1998).

The cytochrome oxidase II sequences display a higher transitional bias than the cytochrome *b* sequences. An overall transition/transversion (Ti/Tv) ratio of 23 was estimated for all cytochrome oxidase II sequences compared to a Ti/Tv ratio of only four for all cytochrome *b* sequences. When taxon-equivalent datasets were compared, the cytochrome oxidase II sequences still displayed a greater Ti/Tv ratio than the cytochrome *b* sequences. This effect is particularly pronounced at the third codon positions of cytochrome oxidase II,

which have a Ti/Tv ratio of 43. This is double the ratio estimated for the third codon positions of cytochrome *b*. These results indicate that the cytochrome oxidase II nucleotides, and third codon positions in particular, are potentially more homoplasious than the cytochrome *b* nucleotides (Bielwaski and Gold, 1996; Swofford *et al.*, 1996).

Patterns of variation of nucleotide substitution across the two genes indicate that both genes contain rapidly and slowly evolving regions. However, there are proportionally more rapidly and slowly evolving regions in the cytochrome oxidase II region sequenced than in the longer cytochrome *b* fragment. This implies greater rate heterogeneity in cytochrome oxidase II.

Based on estimates of homoplasy in the two genes, phylogenetic analyses of sequences from the two genes supported the expectation of decreased utility of cytochrome oxidase II sequences compared to cytochrome *b* sequences. The greater amount of homoplasy in the cytochrome oxidase II sequences, as well as the lesser amount of character data obtained for this gene, is reflected in the level of resolution observed in various phylogenetic analyses. More nodes were resolved and significantly supported for model-based and parsimony analyses of the cytochrome *b* datasets, compared to phylogenetic analyses of cytochrome oxidase II sequences for the equivalent subset of taxa. In unweighted parsimony analysis of each individual gene for the same subset of taxa, cytochrome oxidase II yielded trees with lower consistency and retention indices than cytochrome *b*, indicating greater homoplasy in the cytochrome oxidase II data.

The cytochrome oxidase II Bayesian consensus topology under a GTR + I + Γ model had 12 nodes in conflict with the topology retrieved in combined Bayesian analyses, compared to the cytochrome *b* Bayesian consensus topology which had only four nodes in conflict with the combined topology. Furthermore, the results of partitioned Bremer analysis indicate that the bulk of support for the combined-analyses topology is provided by the cytochrome *b* gene, in particular the third codon positions of this gene. The third codon positions of cytochrome oxidase II provide contradictory signal compared to

that provided by all codon positions of cytochrome *b* and first codon positions of cytochrome oxidase II.

When evaluating topological incongruence between the two genes, the cytochrome oxidase II sequences were not able to reject most cytochrome *b* topologies, whereas the cytochrome *b* data were able to reject the cytochrome oxidase II unconstrained topologies. This indicates the presence of strong phylogenetic signal in the cytochrome *b* data, with no equivalent signal present in the cytochrome oxidase II sequences.

One possible explanation for the lesser utility of the cytochrome oxidase II gene compared with the cytochrome *b* gene may be the greater rate-heterogeneity inferred for the cytochrome oxidase II gene. The mean α value estimated by Bayesian methods for all cytochrome oxidase II sequences was almost half of that estimated for all cytochrome *b* sequences. However, it is important to bear in mind that this comparison is not strictly accurate, as the estimation of α is sensitive to both taxon sampling (Sullivan *et al.*, 1999) and topology (Sullivan *et al.*, 1996). Nevertheless, when α was estimated for taxon-equivalent datasets on the same tree topology, the same pattern was observed, where the estimate of α was much less for cytochrome oxidase II then for cytochrome *b*. This supports the conclusion that the cytochrome oxidase II gene shows greater among-site rate variation than the cytochrome *b* gene.

Genes that display extreme among-site rate variation are expected to be less phylogenetically informative than genes of equivalent length displaying moderate or high among-site variation (Yang, 1998). The low proportion of sites free to vary within the former class of genes are expected to be evolving very rapidly, whereas the remaining sites are not tolerant of multiple substitutions, resulting in the phenomenon where similarity at the tips of the tree may be due to homoplasy, not homology (Sullivan *et al.*, 1995; Yang, 1998). This effect may be exacerbated by the presence of base compositional bias, as this bias tends to accumulate at rapidly evolving sites (Sullivan *et al.*, 1995; Buckley *et al.*, 2001). Excessive among-site rate variation can exert a

major effect on estimation of phylogenetic relationships. It violates all methods of phylogenetic analyses, including model-based methods like ML, causing them to become biased due to inadequate correction for superimposed substitutions (Kuhner and Felsenstein, 1994; Yang *et al.*, 1994). In this study, cytochrome oxidase II sequences displayed greater rate heterogeneity as well as greater base compositional bias at third codon positions compared with cytochrome *b* sequences.

Despite the results discussed above which indicate that cytochrome *b* displays greater phylogenetic utility than cytochrome oxidase II, pairwise sequence divergence estimates for cytochrome *b* were generally higher than those for cytochrome oxidase II. Saturation plots also indicated that the first codon positions of cytochrome *b* are saturated in contrast to the first codon positions of cytochrome oxidase II. Both these factors imply lower phylogenetic utility of the cytochrome *b* gene relative to cytochrome oxidase II.

Saturation plots and divergence estimates are prominent features of molecular systematic studies, with characters showing rapid evolution and saturation typically downweighted or even eliminated in subsequent analyses (Simon *et al.*, 1994; Stanger-Hall and Cunningham, 1998; Reed and Sperling, 1999). However, these strategies are increasingly being criticised. One major criticism is that saturation as evaluated by saturation plots is a distance-based concept, and saturated or noisy data may still provide phylogenetic signal in character-based analyses (Yang, 1998; Baker *et al.*, 2001; Broughton *et al.*, 2001; Kjer *et al.*, 2001). Furthermore, Yang (1998) concluded that highly divergent sequences may be more informative about phylogenetic relationships than sequences of low divergence. He suggested that the lack of signal at low divergence values poses a greater problem for phylogenetic estimation than the accumulation of noise in highly divergent sequences, with saturation only a problem at uncorrected sequence divergences of 30% to 40%, rather than above 15% to 20% as previously suggested (Meyer, 1994). Importantly, Yang (1998) noted that the accuracy of phylogenetic reconstruction is not only dependent on the amount of evolution that has occurred in the sequences, but also on the pattern of substitution across the

branches of the phylogenetic tree and the number of branches in the tree (Yang, 1998). In particular, rapidly evolving third codon positions have been found to provide the bulk of the phylogenetic signal for species relationships (Yang, 1996a; Yoder *et al.*, 1996). This finding is consistent with the results of this study, in which the third codon positions of cytochrome *b* contribute the bulk of the support and signal for phylogenetic relationships.

The greater sequence divergence estimates for cytochrome *b* are consistent with the higher α value inferred for these sequences, with a greater proportion of sites tolerant of multiple substitutions compared to a gene displaying high among-site variation, such as cytochrome oxidase II (Yang, 1998).

A further caution against the utility of cytochrome oxidase II sequence data for resolving phylogenetic relationships in *Camponotus* is the discovery of a cytochrome oxidase II pseudogene. If unrecognised pseudogene sequences are included in phylogenetic analyses, the fundamental assumption of orthology of sequences will be violated, and significant errors may be introduced into phylogenetic analyses (Sorenson and Quinn, 1998; Bensasson *et al.*, 2001).

Heterogeneity, congruence and combinability

Although there was overlap in Bayesian parameter estimates for the GTR+ I + Γ model of sequence evolution for the equivalent cytochrome oxidase II and cytochrome *b* sequence datasets, results of a likelihood ratio test indicated that allowing each gene to evolve under its own model of nucleotide evolution provided a significantly better fit to the data than assuming a single model of evolution for both partitions. This indicates substantial heterogeneity in the evolutionary dynamics of these two genes, presumably due to different evolutionary constraints on the two gene products. It also implies that the best estimate of phylogenetic relationships in a combined analysis framework would be obtained by allowing each partition to evolve under its own best-fit model.

Phylogenetic analyses of the cytochrome *b* and cytochrome oxidase II sequences for the same taxa resulted in statistically incongruent topological estimates. As these two genes are linked on the non-recombining mitochondrial genome, this finding indicates that at least one of the two genes is providing misleading signal. Incongruent topological estimates from mitochondrial genes are mainly attributed to the genes evolving under different evolutionary constraints (Bull *et al.*, 1993; Simon *et al.*, 1994), with differences in the pattern of variation identified as a major factor causing topological incongruence (Sullivan *et al.*, 1995). The cytochrome *b* and cytochrome oxidase II gene fragments sequenced in this study appear to be evolving heterogeneously, and display different patterns of variation. Therefore the topological incongruence between cytochrome oxidase II- and cytochrome *b*- based topologies is not unexpected.

A further factor contributing towards the observed incongruence between the two mitochondrial genes may be the smaller size of the cytochrome oxidase II fragment sequenced (Wilgenbusch and de Queiroz, 2000). Both gene fragments sequenced in this study contain an equal percentage of phylogenetically informative characters per number of characters sequenced (~ 39% for cytochrome *b* and cytochrome oxidase II; see Table 4.1). However, as only 484 base pairs of cytochrome oxidase II were sequenced per taxon compared with 660 for the cytochrome *b* gene, there were fewer phylogenetically informative sites in the cytochrome oxidase II sequences. Thus the prediction was that topologies resulting from analysis of the cytochrome oxidase II fragment were likely to be less resolved than those obtained from cytochrome *b*. This prediction was corroborated in this study, with cytochrome *b* topologies more resolved than those based on cytochrome oxidase II sequence data for the same subset of taxa. Furthermore, incongruence between individual gene-based topologies could be localised to a particular clade absent from the cytochrome oxidase II-based topologies, but present with strong support in topologies inferred from cytochrome *b* sequences. When constraint analysis was performed, the cytochrome *b* sequences were not able to reject the cytochrome oxidase II optimal constraint topology with this node constrained as a significantly worse

explanation of the data. Furthermore, the cytochrome oxidase II sequences were not able to reject the optimal constraint topology as a significantly worse explanation of the data than the unconstrained optimal topology. These findings indicate that it is lack of signal in the cytochrome oxidase II data set, rather than strongly conflicting signal, that may contribute to the observed incongruence. Thus by obtaining additional sequence data from the cytochrome oxidase II gene, cytochrome oxdiase II topologies may be further resolved, which in turn may result in increased congruence between the topologies supported by the two mitochondrial genes. However, if the evolutionary dynamics of these two mitochondrial genes differ substantially, especially with respect to rate heterogeneity, as argued above, increasing the number of cytochrome oxidase II characters sampled will not necessarily resolve the problem of topological incongruence between the two mitochondrial genes.

Despite the lesser utility of cytochrome oxidase II sequences for resolving phylogenetic relationships on their own, combining the cytochrome oxidase II data with the cytochrome *b* sequences resulted in more robust hypotheses of evolutionary relationships for taxa. These hypotheses are characterised by a greater proportion of significantly supported nodes present in the Bayesian topologies and higher bootstrap values for nodes in the ML and MP topologies. This indicates that weak signal present in the cytochrome oxidase II sequences was amplified in the context of a combined analysis. These results are consistent with the observation of Chippindale and Weins (1994) that combining partitions that share the same history, even if they evolve under different evolutionary constraints, enhances the accuracy of phylogenetic estimation by increasing the signal above the noise. Results of partitioned Bremer support indicated that although the third codon positions of cytochrome oxidase II supported alternative groupings not present in the combined topology, the first codon positions of this gene provided a large proportion (20%) of the total support for ingroup relationships in combined analysis.

Efficacy of parsimony weighting schemes

Given the complex evolutionary model inferred for the evolution of the cytochrome *b* and cytochrome oxidase II sequences and the high levels of sequence divergence, maximum parsimony performed surprisingly well, with various weighting schemes recovering the same topology obtained using Bayesian inference with a GTR+I +Γ model of nucleotide evolution.

Parsimony analyses, based on differential weighting of character transformation types according to their empirically determined frequency of occurrence, proved to be highly effective in increasing resolution relative to the unweighted parsimony tree. Topologies derived from the differentially weighted sequences generally produced topologies highly congruent with the Bayesian topology for that data set. Pyrimidine transition downweighting in particular was a highly effective weighting scheme in all analyses. The least successful weighting strategy was A↔T transversion downweighting, which resulted in trees with very few nodes in common with the Bayesian and other MP topologies. This contradicts the hypothesis that downweighting of conversions between A and T may generally increase resolution when analysing high A+T content genes by parsimony (Dowton and Austin, 1997). The lack of utility of A↔T downweighting in resolving species-level relationships is in agreement with the finding of Dowton and Austin (1998) that this downweighting scheme was not useful for recovering relationships among recently derived Braconid wasp lineages. The performance of six parameter parsimony was comparable to that of transition downweighting, which is not unexpected given that transitions were assigned lower weights than all transversion types in the six parameter matrices for all nucleotide datasets.

Applying a protein parsimony (Protpars) cost matrix to the amino acid data resulted in more resolved topologies than unweighted parsimony analysis, indicating that this weighting scheme is able to extract more signal from protein data. Parsimony analysis of the amino acid residues of cytochrome *b* using a Protpars matrix, which concentrates on nonsynonymous changes, resulted in the recovery of all nodes strongly supported by nucleotide data,

indicating that nonsynonymous substitutions contribute strongly to the phylogenetic signal provided by this gene. This is in contrast to Protpars parsimony analysis of the cytochrome oxidase II sequences, which only recovered some of the strongly supported nodes based on nucleotide parsimony analysis.

Many clades were robust to a variety of parsimony weighting schemes, giving increased confidence in the validity of these groups.

Congruence of phylogenetic methods

Bayesian and ML topologies were largely congruent, although the Bayesian 50% majority rule consensus topologies were substantially more resolved than the ML bootstrap topologies. Base frequency parameters and the rate heterogeneity parameters α and p_{inv}, estimated by ML on a neighbour-joining tree, were highly congruent with those estimated during Bayesian analyses and fell within the 95% Bayesian credible region. Bayesian and ML estimates of the relative rate values for the six possible nucleotide transformation types under the GTR model differed by orders of magnitude, with ML optimisation generally inferring much higher substitution rates for all transformation types. C↔T (r_{CT}) transitions followed by G↔T (r_{GT}) transversions were always the most frequent substitution type estimated for both genes by ML and Bayesian methods based on GTR + I + Γ and GTR + SSR models. The rarest substitution type inferred for both genes under the two different models were G↔T (r_{GT}) transversions, with the relative rate of this substitution type set to 1.0 and all other parameters scaled relative to it. The high variances and large confidence intervals for the rate parameter estimates, as well as branch length estimates, suggest considerable uncertainty in these estimates, indicating that additional sequence information is required.

Bayesian consensus topologies based on GTR + SSR or GTR + I + Γ models were remarkably homogenous, despite the large discrepancy in branch length estimates between the two models. The GTR + SSR mean branch length estimates for both genes were consistently lower than those inferred using a

GTR + I + Γ model. The relative underestimation of branch lengths by the SSR model relative to the I + Γ model is consistent with the observations of Buckley *et al.*, (2001). Buckley *et al.* (2001) hypothesised that this relative underestimation may be due to the assumptions made by an SSR model that each site within a given rate class is equally likely to accept a substitution. This could result in an underestimation of the number of multiple substitutions within rate classes displaying extreme among site variation. The second codon positions of both genes, and the first codon position of cytochrome *b*, fall within this category. Very low values of α were estimated for these positions, indicating extreme rate variation. This clearly violates the assumption of homogenous rates for each category.

The underestimation of the actual number of substitutions that have occurred along a branch may manifest in 'long-branch' attraction, even under ML estimation, if there is a poor fit between the model of evolution and the data (Felsenstein, 1978; Hendy and Penny, 1989). However, only minor topological differences were observed between the models when optimal topologies were compared, and nodes with significant posterior probabilities were recovered by both models under Bayesian analyses, indicating that long branch attraction was not a problem for this particular set of taxa. The significant correlation between the posterior probability support estimates obtained when rate heterogeneity was accounted for by a Γ + I model or a SSR model indicates the robustness of the topological relationships to model assumptions.

Congruence of morphological characters with the combined molecular phylogeny

The congruence of certain morphological and behavioural characters with the combined molecular phylogeny for all taxa indicates the potential these non-molecular characters have for providing an independent estimate of phylogenetic relationships in *Camponotus*. Future research could focus on identifying additional non-molecular characters suitable for phylogenetic analysis (e.g. larval characters), and then evaluating the utility of these characters in light of the molecular phylogeny. Further characterisation of species in the field may also yield additional ecological character information that could be incorporated into phylogenetic analyses.

Phylogenetic relationships and hypotheses testing

Subgeneric divisions

Given the taxa included in this study, the majority of *Camponotus* subgeneric classifications do not appear to accurately reflect monophyletic groupings.

The monophyly of subgenus *Tanaemyrmex* was rejected by all molecular data. This is consistent with the predictions of Sauer *et al.* (2000), that further molecular analysis of species assigned to this large and heterogeneous subgenus would indicate the necessity for a major systematic revision of this subgenus. Furthermore, the monophyly of subgenera *Myrmotrema* and *Myrmespera* were rejected in constraint analyses of combined and single gene sequences.

The sister association of *C. sp.* 12 (subgenus *Myrmotrema*) with *C. bifossus* (subgenus *Colobopsis*) rather than *C. sp.* 7 (subgenus *Myrmotrema*), is strongly and consistently supported by all character data. The low level of cytochrome *b* sequence divergence between these two species suggests a recent speciation event. The close evolutionary association of *C. bifossus*, whose major workers are characterised by phragmotic heads, and *C. sp. 12*,

whose major workers lack phragmotic heads, indicates that the phragmotic head of *C. bifossus* is a recent adaptation. This emphasises the need for caution when using this character as a synapomorphy for the subgenus *Colobopsis*. It is recommended that *C.bifossus* be placed in subgenus *Myrmotrema*, based on the strongly supported sister relationships between *C. bifossus* and *C.sp.*12 of subgenus *Myrmotrema.*

There are two exceptions to the finding that subgeneric subdivisions in *Camponotus* do not accurately reflect monophyletic groups. Firstly, cytochrome *b* data supports the monophyly of the subgenus *Myrmosericus*, with the clade containing the two included species in this subgenus, *C. cinctellus* and *C. rufoglaucus*, recovered with significant posterior probability in Bayesian analysis assuming a GTR + SSR model of nucleotide evolution. Secondly, there was strong statistical support for the association of *C. fulvopilosus* and *C. storeatus* (subgenus *Myrmopriomis*) in both cytochrome *b* and combined analyses, with the monophyly of the four species in this subgenus (*C. fulvopilosus, C. storeatus, C. detritus, C. brevisetosus*) supported by morphological characters. Molecular data grouped these species together in a significantly supported group with *C. mystaceus* and *C. sp. 14*, (subgenus *Myrmopsamma*). The paraphyletic association of species assigned to these two subgenera may be an artifact of missing data. No cytochrome *b* sequence data was available for either *C. mystaceus* or *C. sp. 14*, and only incomplete cytochrome oxidase II sequences were obtained for both these taxa. Furthermore, combined molecular-morphological parsimony analysis supported the monophyly of *C. sp. 14* and *C. mystaceus*. Additional sequence data is required to address the hypotheses that subgenera *Myrmopiromis* and *Myrmopsamma* describe monophyletic groups.

The sister association of *C. detritus* with *C. fulvopilosus*, and the low levels of sequence divergence between these two taxa supports the hypothesis of Robertson and Zacharides (1997) that *C. detritus* may have recently speciated from small populations of *C. fulvopilosus.* These authors suggest that speciation of the Namib species, *C. detritus*, may have occurred by means of a founder event, whereby a small population of *C. fulvopilosus*

became separated from the larger population by a newly formed, inhospitable dune system. If this hypothesis of a recent founder-type speciation event is correct, then a paraphyletic gene tree pattern is predicted, with *C. detritus* nesting within *C. fulvopilosus* (Avise, 2000). A phylogeographic study of these two species is required in order to explicitly test this hypothesis.

Camponotus monophyly

Both molecular and morphological data indicated that *Camponotus* was paraphyletic, with *P. schistacea* grouping with the ingroup camponotine species. However, topologies with *Camponotus* constrained to be monophyletic were not significantly different from the optimal topologies, thus precluding a decisive conclusion concerning *Camponotus* monophyly or paraphyly.

Phylogenetic analyses of a combined molecular data set containing additional *Polyrhachis* species did not resolve this issue. *Polyrhachis* formed a significantly supported monophyletic group, with the majority of *Camponotus* species also confined to a strongly supported clade. However, the unresolved position of *C. nasutus* and *C. sp. 10* in this topology prevented a reliable assessment of *Camponotus* monophyly.

Future Studies

In order to further elaborate on the phylogenetic hypotheses presented in this study, further character information and taxon sampling is required. The cytochrome *b* and cytochrome oxidase II sequences were equivocal about the placement of certain species (e.g. *C. sp. 10*, *C. nasutus* and *C. sp. 7*), whether the character data was analysed in a combined framework or individually. Furthermore, deeper nodes were often poorly supported, indicating the need for additional molecular data, preferably from a more conservative gene. The difficulties experienced in amplifying the target gene fragments in some taxa may be due to mutation of the primer binding sites. This hypothesis is supported by the very high sequence divergence values

reported in this study, thus further supporting the use of a more slowly evolving gene for elucidating phylogenetic relationships within *Camponotus*.

Cytochrome oxidase I is the most conserved of the mitochondrial protein-coding genes, with universal insect primer sequences available for this gene (Simon *et al.*, 1994). This gene is thus a good potential candidate for obtaining additional molecular character data with which to address camponotine systematics. However, sequence data obtained from a nuclear gene such as EF-1α may be of greater utility than sequence data from an additional mitochondrial gene for resolving phylogenetic relationships in *Camponotus* (Cummings *et al.*, 1995; Mitchell *et al.*, 2000). This is because adding characters from independent, unlinked genes has been shown to have greater benefit for phylogenetic inference than obtaining additional characters for a single gene or linked genes. This has been attributed to endemic biases in any one gene being diluted in the context of a combined analysis. Additionally, different genes may provide signals for groupings where other genes are uninformative.

The presence of a cytochrome oxidase II pseudogene in *C. rufoglaucus* indicates that translocation of mitochondrial gene copies to the nucleus is a distinct possibility in this genus of ants. To avoid this potential problem when additional taxa are incorporated, a pure mitochondrial extraction or mitochondrial-enriched extractions could be used as template (Sunnucks and Hales, 1996; Bensasson *et al.*, 2000). However, working with small organisms like ants makes it difficult to obtain a pure mitochondrial purification of sufficient quantity for subsequent experiments (Bensasson *et al.*, 2000). Furthermore, nuclear pseudogenes have been amplified from mitochondrial-enriched extractions (Williams and Knowlton, 2001) as well as 'pure' mitochondrial extractions (Collura and Stewart, 1995). The use of reverse transcribed mRNA as a template for PCR is therefore recommended, to eliminate the possibility that a nuclear pseudogene could be amplified when using universal primers (Collura *et al.*, 1996; Williams and Knowlton, 2001).

Summary

In summary, the molecular sequences generated in this study are informative concerning phylogenetic relationships in the ant genus *Camponotus*. The diverse array of methodological approaches utilised in this study provided insight into the relative performance of the two mitochondrial genes in phylogenetic reconstruction. Cytochrome *b* sequences proved especially informative in reconstructing evolutionary relationships among *Camponotus spp.*, yielding well resolved and robustly supported phylogenies. The cytochrome oxidase II sequences, in contrast, appeared to be most informative when analysed in combination with the cytochrome *b* sequences. Morphological and behavioural characters identified in this study also showed promise for phylogenetic reconstruction of relationships in this complex genus.

The arbitrary nature of the majority of current subgeneric classifications based on morphological characters is confirmed by molecular data, thereby validating the concerns of ant systematists regarding the subgeneric classification criteria. Given the taxa sampled in this study, the hypothesis of *Camponotus* monophyly cannot be rejected. However, Bayesian analysis of a data set comprising cytochrome *b* and cytochrome oxidase II sequence data for 12 *Camponotus spp.* and nine additional species of *Polyrhachis* suggests that *Camponotus* may indeed be monophyletic, and indicates a monophyletic origin for *Polyrhachis*. Additional taxon sampling of both *Camponotus* and *Polyrhachis* is required to resolve this issue.

PART II

Chapter 6

FINE-SCALE GENETIC STRUCTURE OF *CAMPONOTUS KLUGII*

The remarkable dominance of ants in most ecological systems is attributable in a large part to their highly social organisation, with all known ant species classified as eusocial. Thus, elucidating the social structure of ant colonies has been a major focus of biologists attempting to understand how this highly social lifestyle evolved, and how it is maintained. This exciting field of study has been revolutionised in recent years by the application of microsatellite markers to fine-grained genetic analyses of social insect colonies. Although *Camponotus* is the largest genus of ants, occupying a diverse variety of ecological niches, very few molecular studies of the social organisation of species in this genus have been undertaken. This chapter presents the results of the first microsatellite-based analyses of the colony structure of the endemic fynbos species, *Camponotus klugii.*

Introduction

Background

Eusocial insect societies, characterised by reproductive division of labour, overlapping generations and cooperative care of the young, have intrigued evolutionary biologists and lay people alike for decades. Of particular interest is the perception of these societies as 'superorganisms', with all individuals working together in harmony for the common good (Hölldobler and Wilson, 1990). However, from Darwin's perspective, these societies appeared to pose a potentially fatal challenge to his theory of natural selection. He viewed the existence of an apparently altruistic sterile worker caste as "by far the most serious special difficulty, which my theory has encountered" (Darwin, 1859). In essence, the question that puzzled evolutionary biologists was how altruistic behaviour displayed by sterile workers, whereby they forfeit their own reproduction in order to aid reproduction by the queen, could be propagated when these workers by definition can produce no offspring of their own?

This apparent difficulty was resolved by the seminal work of W.H. Hamilton, whose inclusive fitness theory provided a quantitative framework for the conditions favouring the spread of an altruistic gene (Hamilton, 1963).

In its simplified form, Hamilton's Rule can be written as follows:

$$rb - c > 0$$

where r is the degree of relatedness between the altruist and the recipient of the altruistic act, b is the benefit in terms of extra offspring produced by the recipient of the altruistic act, and c is the cost of the altruistic behaviour to the altruist in terms of offspring not produced. Altruistic behaviour will be favoured when this inequality is satisfied i.e. when the fitness gain to the recipient multiplied by the relatedness of the altruist and recipient is greater than the fitness loss to the altruist. This theory became more commonly known as kin selection theory (Maynard Smith, 1964) due to the obvious way that close kinship between recipient and altruist would result in a high value of '*r*'.

The haplodiploid hypothesis for the origin of eusociality

Kin selection theory provides an appealing explanation as to why there have been multiple independent origins of eusociality within the order Hymenoptera. All members of this order, namely bees, wasps and ants, have an unusual method of sex determination, that of 'haplodiploidy'. Female reproductives are capable of laying two types of eggs; (i) fertilised, diploid eggs that develop into females and (ii) haploid, unfertilised eggs that develop into males (Bourke and Franks, 1995). Thus male hymenopterans almost always only have a haploid set of chromosomes. As a consequence of this, all sperm produced by males are genetically identical. This haplodiploid sex determination system leads to unusual coefficients of relatedness when compared to those expected for a diploid-diploid organism. This is best illustrated by considering a social insect colony headed by a single, once-mated, outbred queen (Table 6.1).

Table 6.1. Relatedness levels within a hymenopteran colony in which a single queen mates with a single male (after Bourke and Franks, 1995).

Relationship	Life-for-life[1] Relatedness (*r*)
Queen, daughter	0.5
Female, son	0.5
Father, daughter	1.0
Queen's mate, queen's son	0.0
Sister, sister	0.75
Sister, brother	0.25
Brother, brother	0.5
Queen, grandson	0.25
Female, nephew	0.375

Hamilton proposed that the unusually high levels of relatedness (r = 0.75, Table 6.1) between hymenopteran sisters might have facilitated the evolution of altruism by kin selection, and thus account for the numerous independent origins of eusociality within this order (Hamilton, 1964). Furthermore, the haplodiploidy hypothesis could explain why hymenopteran workers are female. Males are on average as equally related to their siblings (r = 0.5 Table 6.1) as they are to their offspring (r = 0 with the mate's sons and r = 1 with daughters, Table 6.1), therefore there would be no genetic predisposition to worker behaviour in males.

The haplodiploid hypothesis as the sole explanation for the multiple independent origins of eusociality within the Hymenoptera has, however, been questioned (Hölldobler and Wilson, 1990). As Bourke and Franks (1995) and others (Hölldobler and Wilson, 1990, Queller and Strassmann, 1998) have noted, not all haplodiploid species are eusocial[2], nor are all eusocial insects haplodiploid[3]. Therefore factors aside from haplodiploidy must be involved in the evolution of eusociality. Within the Hymenoptera, other traits such as nest-building, the possession of a sting and maternal care may have contributed to

[1] Life-for-life relatedness is a compound measure that incorporates both regression relatedness as well as a term for relative sex-specific reproductive values. In haplodiploids, the sex-specific reproductive value of females is twice that of males, as by definition females have twice the number of chromosomes as males, which they pass on to both sexes in the next generation, whereas males only pass on their genes to females (Bourke and Franks, 1995).

[2] e.g. certain mites, thrips, whiteflies, scale insects and beetles within the genera *Micromalthus* and *Xylosandrus*, and some solitary species of bees and wasps (Hölldobler and Wilson, 1990).

[3] e.g. termites (Bourke and Franks, 1995).

the frequent occurrence of eusociality within this group, and the presence of female-only workers (Bourke and Franks, 1995). However, although the exact contribution of haplodiploidy to eusocial evolution is uncertain, kin selection theory remains vital to explain eusocial evolution (Bourke and Franks, 1995).

Kin structure and conflict

Kin selection theory also provides a theoretical framework for understanding how various social behaviours and life history strategies of the eusocial Hymenoptera have evolved. Potential kin conflict over reproduction is a universal feature of hymenopteran societies, whereby conflict is predicted to occur due to the asymmetries of relatedness between various parties in a social insect colony and the reproductive individuals. Following on from this, the primary determinant of potential conflict within eusocial hymenopteran colonies is the kin structure of the colony itself, and the impact of this kin structure on relatedness values among individuals in a colony and the future reproductive brood (Ratnieks and Reeve, 1992).

Kin structure in ants

There are four main factors that will affect the relatedness structure of a colony: (i) the number of, and relatedness among queens, (ii) the effective paternity frequency, (iii) the level of worker reproduction and (iv) the level of inbreeding within a colony (Ross, 1993; Seppa, 1994; Bourke and Franks, 1995).

The number of queens per colony varies widely both between and within ant species, with polygyny the predominant social structure (Keller, 1995). Ant species also vary with respect to effective paternity frequency, although multiple paternity[4] appears to be rare in ants with the notable exceptions of leafcutter ants in the genera *Atta* and *Acromyrmex* (Boomsma and Ratnieks, 1996; Fjerdingstad *et al.*, 1998; Villesen *et al.*, 1999; Fjerdingstad and

[4] Although many ant queens have been observed to copulate with multiple males, genetic studies have shown that multiple mating does not necessarily result in multiple paternity, therefore the term 'multiple paternity' is preferred over 'multiple mating' (Boomsma and Ratnieks, 1996).

Boomsma, 2000) and the harvester ant *Pogonomyrmex occidentalis* (Cole and Wiernasz, 2000). However, the generality of this conclusion is questionable as it is based on an analysis of only 19 ant species for which sufficient data was available to estimate paternity frequency. Indeed, paternity frequency may also vary within populations of the same ant species (Boomsma and Ratnieks, 1996; Sanetra and Crozier, 2001).

Complete worker sterility in ants is limited to a few genera, with workers of most species able to produce male eggs due to the presence of functional ovaries (Bourke and Franks, 1995). Worker reproduction of male haploid eggs appears to be widespread among ants (e.g. Evans, 1993; Herbers and Mouser, 1998; reviewed by Bourke, 1988). Inbreeding in ants appears to be rare (e.g. Ross, 1993; Chapuisat and Crozier, 2001; Hammond *et al.*, 2001 but see Chapuisat *et al.*, 1997). Therefore, the key determinant of relatedness structure within ant colonies is the effective number of reproducing individuals within the colony.

The ant genus *Camponotus* is thought to be almost exclusively monogynous (Akre *et al.*, 1996; Seppa and Gertsch, 1996), therefore the primary determinant of intracolony genetic variation within colonies of this genus is expected to be determined by the number of males that mated with the queen. The potential for kin conflict over reproduction and sex allocation will therefore be discussed within this context.

(i) Conflict over haploid male production

In a colony where workers are produced by a once-mated single queen, workers are more related to their own male offspring (r = 0.5, Table 6.1) and to those of their sisters (r = 0.375, Table 6.1) than to queen-produced males (r = 0.25, Table 6.1). The queen, however, is more related to her own sons (r = 0.5 Table 6.1) than her grandsons (r = 0.25, Table 6.1). Therefore potential queen-worker conflict over male production is predicted, with workers expected to have greater influence over male production than queens because they rear and tend the brood. In contrast, within colonies in which the

queen is multiply-mated, workers are on average more closely related to the queen's male offspring (r = 0.25, Table 6.1) than to their sisters sons, most of whom will be half-nephews (r = 0.125, Table 6.1) rather than nephews (r = 0.375, Table 6.1). Although each worker should still favour producing her own eggs, other workers should preferentially destroy these eggs and/or act aggressively towards laying workers in order to concentrate the colony's effort and resources on queen produced male eggs – a hypothesis known as 'worker policing' (Ratnieks, 1988). In this scenario, the interests of the workers and the queen coincide, thereby reducing potential within-colony conflict over reproductive allocation.

Support for these kin selection predictions comes from empirical studies that have shown that worker policing is widespread in the honeybee genus *Apis*, which is characterized by extreme polyandry (Ratnieks and Visscher, 1989), and occurs facultatively in the wasp *Dolichovespula saxonica*. In this wasp, worker policing occurs in colonies with multiply mated queens, but is absent from colonies headed by singly-mated queens, as predicted by kin selection theory (Foster and Ratnieks, 2000). Evidence for worker policing in ants is rare, with worker policing only documented in some species of primitive Ponerine ants (reviewed by Monnin and Ranieks, 2001).

Worker-laid male eggs have been detected in monogynous, monandrous colonies of *Myrmica punctiventris* (Herbers and Mouser, 1998), *Crematogaster smithii* (Heinze *et al.*, 2000) and *Protomognathus americanus* (Fotizik and Herbers, 2001), indicating worker control of male production. However, an increasing number of studies have shown that males produced in monogynous, monandrous colonies of various hymenopteran species are queen-derived, indicating unexpected queen control of haploid male production (Foitzik *et al.*, 1997; Foster *et al.*, 2000; Paxton *et al.*, 2001; Green and Oldroyd, 2002; Palmer *et al.*, 2002). This may be due to queen policing of worker eggs, or workers selecting to refrain from producing male eggs if it impacts negatively on colony productivity. Another possibility is that workers are not able to differentiate between queen-produced and worker-produced

male eggs (Bourke and Franks, 1995). Further research is required to address these alternate hypotheses.

(ii) Conflict over sex allocation

Kin selection theory also predicts queen-worker conflict over sex ratio allocation in a population of Hymenoptera consisting of monogynous, monandrous colonies (Trivers and Hare, 1976). Queens are predicted to prefer equal investment in their daughters and sons, to whom they are equally related (r = 0.5, Table 6.1), whereas workers will only maximize their inclusive fitness by preferentially rearing sisters (r = 0.75, Table 6.1) over brothers (r = 0.25, Table 6.1) at a ratio of 3:1 in the reproductive brood. At this sex investment ratio[5], although reproductive females will only have one-third the mating success of males, their three-fold higher relatedness to their sisters, when compared to their brothers, will perfectly balance their reduced mating success. Hence, if there is worker control of the colony, the stable sex allocation ratio should be female biased. In the case of multiple paternity, two different outcomes are predicted depending on whether all queens display the same mating frequency, or whether queens differ from each other with respect to mating frequency (Bourke and Franks, 1995). In both instances, the queen's relatedness to her offspring remains unchanged, as in the single mating scenario. If all queens display the same elevated mating frequency, worker-worker relatedness is predicted to decline uniformly across colonies, resulting in a decrease in the inclusive fitness benefits workers may derive from rearing sisters. If workers are capable of manipulating the sex ratio allocation, then the population female bias is expected to be reduced when compared to the monandrous case. If queens differ in the number of times they have been mated, and assuming that workers are capable of assessing queen mating frequency and controlling sex allocation, then each colony should produce the sex that the workers within it are relatively more related to. That is, workers in colonies headed by singly mated queens should prefer a female-biased sex ratio, whereas workers in colonies headed by multiply

[5] defined as the total energy invested in females divided by the total energy invested in males (Bourke and Franks, 1995).

mated queens are expected to favour male production. This is expected to result in a bimodal distribution of sex ratios within the population.

Empirical evidence for facultative sex ratio biasing due to variation in queen mating frequencies comes from research conducted by Sundstrom and collegues on the ants *Formica truncorum* and *Formica exsecta*. Colonies of both these species showed a bimodal sex distribution, with colonies headed by single-mated queens producing mainly females, and those headed by multiply mated queens producing mainly males. These findings are therefore consistent with worker control of sex allocation (Sundstrom, 1994; Sundstrom *et al.*, 1996).

Furthermore, there is empirical evidence for a significantly female-biased sex ratio in monogynous ant species, indicating worker control of sex allocation, though not at the 3:1 ratio predicted by Trivers and Hare (1976) (Boomsma, 1989). One reason for the lower-than-expected bias may be that polyandrous and polygynous species were inadvertently included in the analysis. However, it is also possible that there may be shared queen-worker control of sex allocation in simple family colonies, with neither party able to dominate. This would then result in a ratio intermediate to their respective equilibrium values (Bourke and Franks, 1995).

Relatedness

Following on from the above kin selection arguments, it is evident that an estimate of the relatedness between altruist and recipient is fundamental to evaluating any kin selection hypothesis. Relatedness is a measure of genetic similarity between individuals. However, the exact interpretation of how this genetic similarity should be defined has been extensively debated in the population genetic literature (Michod and Hamilton; 1980; Bourke and Franks, 1995 and references therein). In the framework of kin selection theory, Hamilton felt that relatedness was best defined as a regression coefficient (Hamilton 1970, 1972). This regression coefficient is obtained by regressing the gene frequency among potential recipients of a social action, across

groups of social interactants, against the average frequency of the gene among random potential donors of this social action (Bourke and Franks, 1995). In this context, relatedness can be interpreted as the probability of allele sharing between individuals, above the background probability set by the allele's frequency in the population (Queller and Goodnight, 1989; Crozier and Pamilo, 1996). In other words, the concept of relatedness is concerned not with average gene frequencies, but with how individuals with frequencies other than the population average associate with one another (Bourke and Franks, 1995).

When relatedness is measured within a familial structure, such as a monogynous ant colony, the regression estimate of relatedness will measure genetic similarity due to ties of pedigree. It should be noted that an implicit assumption made by basing estimates of relatedness on genotypic data obtained from neutral genetic markers is that relatedness at neutral loci estimates relatedness at the loci involved in social actions. This assumption can be justified on the basis that kinship makes average relatedness between any two individuals the same across all their autosomal loci (Bourke and Franks, 1995).

Genetic markers

To obtain accurate estimates of relatedness between individuals, highly variable molecular markers are essential. The low heterozygosity of allozyme markers in haplodiploid social insects (Hedrick and Parker, 1997) has led to a range of DNA-based markers being utilised to estimate genetic relatedness. These range from random amplified polymorphic DNA markers (Hasegawa, 1995) to multilocus DNA fingerprints of minisatellite or microsatellite repeat sequences (e.g. Ross *et al.*, 1999; Gadau *et al.* 1996; Satoh *et al.*, 1997).

Microsatellites, however, are arguably the marker of choice in social insect studies due to their co-dominant inheritance, apparent selective neutrality and high variability, even within social insect colonies. This is reflected in the wealth of microsatellite libraries developed in recent years for eusocial

species (e.g. see Table 6.2 for a list of microsatellites developed for ants) (Queller *et al.*, 1993).

Table 6.2. Overview of ant species in which microsatellite primers have been developed.

Subfamily	Species	No. loci	Reference
Dolichoderinae	*Linepithema humile*	19	Krieger and Keller, 1999
	Linepithema humile	4	Tsutsui *et al.*, 2000
	Linepithema humile	20	Ingram and Palumbi, 2002
Formicinae	*Camponotus consobrinus*	5	Crozier *et al.*, 1999
	Camponotus ligniperdus	5	Gertsch *et al.*, 1995
	Formica exsecta	15	Gyllenstrand *et al.*, 2002
	Formica lugubris	5	Chapuisat, 1996
	Lasius niger	4	Gertsch, P *
	Myrmecocystus mimicus	5	Kronauer *et al.*, 2002
	Petalomyrmex phylax	14	Dalecky *et al.*, 2002
	Polyrhachis doddi	3	Johnson, R.N *
Myrmicinae	*Acromyrmex echinatior*	5	Ortius-Lechner *et al.*, 2000
	Apterostigma collare	2	Villesen *et al.*, 1999
	Atta colombica	2	Fjerdingstad *et al.*, 1998
	Cataulacus mckeyi	11	Debout *et al.*, 2002
	Crematogaster smithii	2	Heinze *et al.*, 2000
	Cyphomyrmex longiscapus	2	Villesen *et al.*, 2002
	Leptothorax acervorum	4	Bourke *et al.*, 1997
	Leptothorax nylanderi	3	Foitzik *et al.*, 1997
	Leptothorax spinosior	10	Hamaguchi *et al.*, 1993
	Myrmica punctiventris	3	Herbers and Mouser, 1997
	Myrmica tahoensis	3	Evans, 1993
	Myrmicocrypta ednaella	2	Villesen *et al.*, 1999
	Pheidole pallidula	22	Fournier *et al.*, 2002
	Pogonomyrmex barbatus	10	Volny and Gordon, 2002
	Solenopsis invicta	8	Krieger and Keller, 1997
	Trachymyrmex cf. zeteki	4	Villesen *et al.*, 2002
Ponerinae	*Diacamma ceylonense*	6	Gopinath *et al.*, 2001
	Diacamma cyaneiventre	8	Doums, 1999
	Gnamptogenys striatula	12	Giraud *et al.*, 1999
	Platythyrea punctata	10	Schilder *et al.*, 1999
	Rhytidoponera metallica	8	Chapuisat *et al.*, 2000
	Rhytidoponera sp. 12	5	Tek Tay and Crozier, 2000
Prionomyrmecinae[6]	*Prionomyrmex macrops*	14	Sanetra and Crozier, 2000
Pseudomyrmicinae	*Pseudomyrmex pallidus*	9	Peters, 1997

*Unpublished microsatellite sequences available in Genbank (http://www.ncbi.nlm.nih.gov)

Microsatellites are tandem repetitive DNA sequences of two to six base pairs, randomly dispersed throughout the eukaryotic genome. They are generally abundant, with a microsatellite motif estimated to occur every few thousand nucleotides in the genome of the wasp *Vespula rufa* (Thoren *et al.*, 1995).

[6] Prionomyrmecinae (= Northomyrmeciinae) (Baroni Urbani, 2000).

These markers display high mutation rates, ranging from 10^{-2} to 10^{-5}, resulting in multiple alleles present at any particular microsatellite locus within populations (Jarne and Lagoda, 1996). The high variability displayed by microsatellite loci is thought to result from the interaction of mutation and genetic drift, and to a lesser extent, selection and recombination (Jarne *et al.*, 1998). Mutation of microsatellite loci is thought to result primarily as a consequence of DNA polymerase slippage during replication (Toth *et al.*, 2000).

Variation at microsatellite loci is readily assayed by PCR amplification, using primers complementary to the unique sequences flanking specific repetitive arrays, followed by electrophoretic sizing of the PCR products (Glenn *et al.*, 1996). One practical disadvantage of using microsatellite markers is the high cost of constructing and screening species-specific microsatellite libraries. For example, primers developed in one species may amplify polymorphic microsatellite loci in closely related species within the same genus or subfamily (Schilder *et al.*, 1999; Gertsch *et al.*, 1995; Chapuisat, 1996). However, the pattern of cross-species amplification is not predictable based solely on phylogenetic affinity, with amplification failures most likely due to species-specific random mutations in the microsatellite flanking regions. Microsatellite primers developed in *Camponotus ligniperdus* (Gertsch, 1995) failed to amplify microsatellite loci in seven other *Camponotus* species (P. Gertsch and J. Gadau, unpublished data; referenced in Gadau *et al.*, 1996), and primers developed in *Platytherea punctata* failed to amplify microsatellite loci in four closely related *Platytherea* species (Schilder *et al.*, 1999).

The utility of microsatellites in social insect studies is not limited to estimating genetic relatedness. Pedigrees within the colony can also be reconstructed with genotypic data, allowing important aspects of the mating biology and sociogenetic structure of the colony e.g. queen mating frequency (the number of patrilines), queen turnover (the number of matrilines) and nest usurpation to be elucidated (e.g. Estoup *et al.*, 1994; Bourke *et al.*, 1997; Paxton *et al.*, 2001). In addition, the genotypic composition of nests may be used to infer

whether colony structure is polydomous, with several nest fragments comprising one large colony, or monodomous, where each nest is equivalent to an independent colony (Herbers and Mouser, 1998; Foitzik and Herbers, 2001). Furthermore, allelic frequency data from microsatellite loci can be used to infer population genetic structure, in keeping with the more traditional application of microsatellite markers in evolutionary studies (e.g. Chapuisat *et al.*, 1997; Goropashnaya, 2001; Fournier *et al.*, 2002).

Microsatellites and Camponotus

Species within *Camponotus* are thought to be almost exclusively monogynous, with only a handful of species known to be polygynous (see Curtis, 1985; Carlin *et al.*, 1993; Gertsch *et al.*, 1995; Crozier *et al.*, 1999; Gadau *et al.*, 1998; Sanada *et al.*, 1997; Satoh *et al.*, 1997 for examples of polygynous species). The effective mating frequency of species within this genus is largely unknown as genetic studies have been limited to a handful of species. DNA-based studies of mating frequency in *Camponotus ligniperdus* indicated a low frequency of double-mating (Gadau *et al.*, 1996; Gadau *et al.* 1998), and there was evidence for multiple mating by *C. herculeanus* queens based on allozyme and microsatellite genotypic data (Seppa and Gertsch, 1996). However, the remaining species investigated, namely *C. consobrinus* (Crozier *et al.*, 1999; Fraser *et al.*, 2000), *C. floridans* (Gadau *et al.*, 1996; Gadau *et al.*, 1998) and *C. gigas* (Gadau *et al.*, 1996) all had colony genetic profiles consistent with singly mated queens.

Camponotus klugii (= *C. werthi* Forel; synonymy determined by H. Robertson, unpublished), more commonly know as the black sugar ant, is endemic to the Cape Fynbos biome. The Cape Fynbos biome coincides with the Cape Floristic Kingdom, the smallest of the six Floral Kingdoms of the world. This Kingdom is recognised as one of the most biologically diverse areas on earth, and is largely contained in South Africa's Western Cape Province. Fynbos, an evergreen, sclerophyllous shrubland is the predominant vegetation type in the Cape Fynbos biome (Cowling *et al.*, 1997). The most characteristic plant types of fynbos are proteas, ericoid shrubs, geophytic herbs and restiod

grasses. Over 70% of the plant species found in fynbos are endemic, and this vegetation type also supports many endemic insect species.

Camponotus klugii nests in empty cavities in dry, dead fynbos plants produced by the boring activity of beetle and caterpillar larvae, rather than constructing nests themselves (Skaife, 1961; H. Robertson pers. com.). Nests are generally small in size, rarely exceeding 100 individuals, and generally consist of a single queen and her worker offspring, although queenless nests with workers and brood present are a common phenomenon in this species.

Skaife (1961) and Curtis (1985) suggested that these queenless nests might represent satellites of queenright[7] nests, with workers transporting brood and other workers across to inhabit these nests due to space limitations in the original nest, consistent with a polydomous colony structure. Alternatively, these nests may represent orphaned colonies, with brood present due to worker egg laying. Intriguingly, Skaife (1961) found that when he transported a queenless colony from the field to his laboratory, removed the brood and maintained this colony artificially, eggs were produced that developed into workers. Furthermore, three additional queenless colonies maintained in his laboratory were observed to produce winged males. These findings imply that this species is capable of both arrhenotokous parthenogenesis (parthenogenetic production of males from unfertilised haploid eggs), as well as thelytokous parthenogenesis (parthenogenetic production of females from unfertilised diploid eggs).

Although arrhenotoky is a key feature of the hymenopteran haplodiploid sex determination system, obligate or facultative thelytoky has only been formally demonstrated by rearing experiments for five species. These are *Pristomyrmex pungens* (Tsuji 1988) and *Messor capitatus* (Grasso *et al.* 2000) both in the subfamily Myrmicinae; *Cerapachys biroi*, Ceraphachyinae (Tsuji and Yamauchi, 1995; Ravary and Jaisson, 2002); *Cataglyphis cursor*, Formicinae (Cagniant, 1979) and *Platythyrea punctata*, Ponerinae (Heinze and Hölldobler, 1995). Crozier and Pamilo (1996), however, speculate that it

[7] Refers to a colony containing one or several queens (as opposed to queenless).

is likely that facultative thelytoky might occur sporadically in many more species.

The predicted genetic signature of thelytokous parthenogenesis is a clonal structure, with all individuals sharing identical genotypes to the extent that entire populations are fixed for a particular homozygous or heterozygous genotype. This has been recently demonstrated in a microsatellite-based study of the thelytokous ant *Platythyrea punctata* (Schilder *et al.*, 1999), and the Cape parasitic honeybee *Apis mellifera capensis* (Kryger, 2001).

Research Objectives of Part II

In light of the paucity of molecular-based social studies on *Camponotus* species in general, and Skaife's (1961) seminal research on the fynbos ant *Camponotus klugii,* a microsatellite study of *Camponotus klugii* was undertaken in order to:

(i) assess the suitability of microsatellite primers developed in *Camponotus consobrinus* and *Camponotus ligniperdus* for amplifying polymorphic loci in *Camponotus klugii*

(ii) determine the sociogenetic structure of *Camponotus klugii* colonies using informative microsatellite loci

(iii) evaluate the hypothesis of Skaife (1961) and Curtis (1985) that *Camponotus klugii* colonies are polydomous, comprising many nest fragments

(iv) evaluate the hypothesis of Skaife (1961) that in queenless colonies of *Camponotus klugii*, workers can be produced by thelytokous parthenogenesis.

Materials and Methods

Sample collection and molecular analyses

(i) Sample collection

Seven nests of *Camponotus klugii* were collected from a single, isolated *Protea repens* bush in Silvermine Nature Reserve, Cape Town (34°05' S; 18°25' E) in May 1998, representing the total complement of nests present on that bush. Nests were located in the hollowed-out bases of dead protea inflorescences. No nests on neighbouring vegetation were found. Of these seven nests, three were incipient colonies consisting of a foundress queen only, and presumed to be of recent origin due to the absence of eggs and brood. Of the remaining four nests, two were queenright with a single queen and workers present, whereas two consisted of workers only, with no queen present. No males were identified in any of the nests. Whole colonies were stored at -20°C, and subsequently censused. DNA from all individuals from each nest was extracted for microsatellite analysis.

(ii) Nucleic acid extraction and quantification

(a) DNA extraction

Total genomic DNA was extracted from the thorax and legs of each individual ant using a phenol-chloroform extraction method (Evans, 1995). Prior to extraction, the head and abdomen of each ant was removed using a sterile razor blade, and the frozen thorax of each individual ant was pulverized in a 1.5ml eppendorf with an eppendorf pestle after addition of liquid nitrogen. Blank extractions without ant tissue were included in each batch of extractions to control for possible DNA contamination introduced during the extraction procedure.

(b) DNA quantification

The DNA concentration was quantified using diode array spectrophotometry as outlined in Chapter 2. Samples were diluted in Tris-EDTA pH 8.0 to a working concentration of 50 ng/μl, and stored at 4°C.

(iii) Microsatellite amplification

Eight workers from each of the four nests containing workers were screened for microsatellite amplification at thirteen microsatellite loci (Table 6.3). When PCR amplification resulted in visible product, but no allelic variation was observed amongst the eight randomly chosen workers, an additional twenty chosen workers were screened.

Primers were tested over a range of magnesium chloride concentrations and annealing temperatures until a repeatable microsatellite motif for the test samples was produced. Prior to PCR amplification, 12.5pmoles of the forward primer for each locus was radioactively end-labelled with 20μCi of γ^{32}P-dATP (AEC Amersham (Pty) Ltd, United Kingdom) using T4 polynucleotide kinase as per manufacturers instructions (AEC Amersham, South Africa). In order to ensure maximal end-labelling, incubation was extended from 30 minutes to 90 minutes at 37°C. DNA amplification reactions were performed in 0.2ml thin-walled tubes in 10μl reaction volumes containing the following reagents: 0.2U BIOTAQ DNA polymerase (Bioline, Whitehead Scientific), 1 X Bioline reaction buffer (final concentration: 16mM $[NH_4]_2SO_4$; 67mM Tris-HCl (pH 8.8); 0.01% Tween-20), 0.2mM dNTPs, 1 to 4mM $MgCl_2$, 1.25μM end-labelled forward primer, 1.25μM cold reverse primer and 100ng template DNA. PCR reaction mixtures were overlayed with mineral oil and thermal cycling was performed on an MJ Research Inc. PTC-100™ thermal cycler. The cycling profile consisted of an initial denaturation step for 3 minutes, followed by 35 cycles at 94°C for 30 seconds, T_a for 45 seconds and extension at 72°C for 45 seconds, followed by a final extension step at 72°C for 10 minutes. Each PCR reaction was stopped by the addition of 4μl of formamide loading dye (95% formamide, 20mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol) (Sambrook *et*

amplification experiments to check for contamination of PCR reagents. Samples were stored at -20°C prior to electrophoresis.

Table 6.3. Characteristics of microsatellite loci screened for amplification in *Camponotus klugii.*

Locus	Source species	Reference	No. alleles amplified in source species	Repeat Type
Camp4	*Camponotus ligniperdus*	Gertsch *et al.*, 1995	1	$(AT)_4A(AT)_3G(TA)_3(CA)_4$
Camp6	*C. ligniperdus*	Gertsch *et al.*, 1995	8	$(AC)_{33}$
Camp8	*C. ligniperdus*	Gertsch *et al.*, 1995	4	$(GT)_7(TG)_4TC(TG)_2A(GT)_2$
Ccon12	*Camponotus consobrinus*	Crozier *et al.*, 1999	8	$(GA)_3GG(GA)_4AA(GA)_9$
Ccon20	*C. consobrinus*	Crozier *et al.*, 1999	11	$(TC)_9$
Ccon42	*C. consobrinus*	Crozier *et al.*, 1999	13	$(GA)_{11}$
Ccon70	*C. consobrinus*	Crozier *et al.*, 1999	38	$(GA)_2AA(GA)_{27}$
Ccon79	*C. consobrinus*	Crozier *et al.*, 1999	28	$(GA)_5(AG)_2GGGAA(GA)_{12}$
FL12	*Formica lugubris* B	Chapuisat, 1996	5	$(TC/AG)_{12}$
FL20	*F. lugubris* B	Chapuisat, 1996	12	$(TC/AG)_{17}$
FL21	*F. lugubris* B	Chapuisat, 1996	4	$(TC/AG)_{16}$
FL29	*F. lugubris* B	Chapuisat, 1996	4	$(TC/AG)_{12}$
FL43	*F. lugubris* B	Chapuisat, 1996	1	$(TG/AC)_7$

(iv) Polyacrylamide gel electrophoresis and allelic scoring

Three microlitres of PCR product per sample was loaded on a standard 6% denaturing polyacrylamide gel (Sambrook *et al.*, 1989) immediately after heating at 94°C for five minutes followed by snap cooling on ice. Samples were electrophoresed in 1 X TBE buffer at 69 W for 2.5 to 4 hours depending on the size of the alleles. Allele sizes were scored manually by comparison with an M13mp18-AT sequencing ladder. This size ladder was prepared by end-labelling the -40 universal primer (5' GTT TTC CCA GTC ACG AC 3') as described for the microsatellite primers, and performing dideoxy sequencing using the Sequenase Version 2.0 Sequencing Kit (USB, South Africa). After electrophoresis, the gel was transferred to Whatmann No.1 filter paper, vacuum-dried for one hour and then exposed to Cronex Medical X-Ray film

(Du Pont). Depending on the strength of the signal, as assessed using a Geiger-Mueller counter, gels were either exposed for 3 to 4 hours at room temperature, or overnight at -70°C.

(v) Genotyping of *C. klugii* individuals

All individuals were genotyped at seven microsatellite loci found to be polymorphic in *C. klugii*, using the optimal annealing temperatures and $MgCl_2$ concentrations detailed in Table 6.4. Reaction volumes and reagents were as previously described. Samples that had been previously genotyped were included in each experiment to facilitate scoring and ensure reproducibility.

Table 6.4 Optimal annealing temperatures and magnesium concentrations for the seven microsatellite found to be polymorphic in *C. klugii*. Number of alleles and allelic size ranges are based on the genotypes of 168 individuals.

Locus	T_a (°C)	$MgCl_2$ (mM)	No. alleles amplified	Size range (bp)
Camp4	58	2.0	4	205 – 211
Camp8	55	2.0	2	121 – 123
Ccon12	53	3.0	7	165 – 181
Ccon20	57	1.5	6	289 – 299
Ccon42	53	1.5	3	262 – 266
Ccon70	59	1.5	4	159 – 165
Ccon79	57	1.5	9	330 – 346

Genetic data analyses

Statistical Significance

Statistical significance was set to $\alpha = 0.05$ for all statistical tests. *P* values for multiple applications of the same test were corrected for Type I errors using a modified Bonferroni correction unless otherwise noted. Specifically, the modified Bonferroni correction procedure of Holm (1979) as described in Legendre and Legendre (1998) was applied, in order to avoid a high probability of Type II errors often observed when correcting using the strict Bonferroni procedure.

I Nest genetic structure

i) Genotype reconstruction

Within each nest, individuals were sorted into genotypes across the seven microsatellite loci. Nests were classified as monogynous if they exhibited no more than two worker genotypes per locus, with these two genotypic classes sharing at least one allele (Crozier *et al.*, 1999). For all monogynous nests, putative queen genotypes were reconstructed assuming Mendelian segregation of alleles. Queen alleles were defined as those found in half the worker offspring for a heterozygous queen, or all the worker offspring for a homozygous queen (Palmer *et al.*, 2002). At each locus, the common allele shared by all the workers was assumed to be from a single, haploid father (Queller *et al.*, 2000). For loci at which all workers in a nest were fixed for a particular homozygous or heterozygous genotype, it was not possible to unambiguously assign the paternal and maternal alleles. Putative queen genotypes were reconstructed for workers from both queenright nests in addition to the queenless nests, as the multilocus genotype of the queens found in the queenright nests were not consistent with those queens being the mother of the workers present in the nest (see Results).

ii) Genetic variation

Genetic variation was quantified using two measures: the average number of alleles per locus, and the expected heterozygosity (H_e) for each locus (Avise, 1994). Expected heterozygosities for each locus were calculated from the observed frequencies of alleles in queens according to the following formula based on the assumption of Hardy-Weinberg equilibrium:

$$H_e = 1 - \sum_{i=1}^{k} p_i^2$$

where p_i is the frequency of the i^{th} of k alleles (Avise, 1994).
The average heterozygosity for all loci was calculated by summing the allele frequencies according to the following formula:

$$1 - \frac{1}{m} \sum_{l=1}^{m} \sum_{i=1}^{k} p_i^2$$

where m represents the number of loci (Weir, 1996).

Standard errors of the locus-specific heterozygosity estimates as well as overall heterozygosity over all loci were calculated according to Nei and Roychoudhury (1974) in order to gauge the effect of small sample size.

iii) Genotypic Differentiation

Genotypic differentiation between nests was tested for using a log-likelihood based (*G*) exact test implemented in Genepop (Raymond and Rousset, 1995; web version at http:///wbiomed.curtin.edu.au/genepop). This test is performed on contingency tables under the null hypothesis that the genotypic distribution is identical across populations. Significance values were calculated using Fisher's method of combining exact test probabilities (Zar, 1996). Default settings were used for the Markov chain parameters: 1000 dememorization steps, 100 batches and 1000 iterations per batch. Both pairwise nest genotypic differentiation, as well as genotypic differentiation across the whole population, defined as the four nests containing workers, was tested for at each and all of the seven microsatellite loci.

iv) Linkage Disequilibrium and Hardy Weinberg equilibrium

Linkage disequilibrium between loci and deviation from Hardy-Weinberg equilibrium were tested for using Fisher's Exact tests (Zar, 1996) implemented in the software package Genepop. A Markov chain method was used to estimate exact probabilities for contingency tables constructed for linkage disequilibrium or deviation from Hardy-Weinberg using the following parameters: 1000 dememorisation steps; 400 batches and 1000 iterations per batch. Significance values were calculated using Fisher's method of combining exact test probabilities.

a) *Linkage disequilibrium*

Linkage disequilibrium was tested for using two input data files: (i) genotype data from all queens and (ii) genotype data from all 168 individuals scored. Individual workers within social insect colonies cannot be considered independent samples, due to expected genotypic correlation with other workers as a result of close kinship ties (McCauley and Goff, 1998). If these individuals are included in tests for linkage disequilibrium between pairs of loci, non-random allelic associations of unlinked loci may be incorrectly inferred (Hartl and Clark, 1997). However, the second data set was tested for linkage disequilibrium with the reasoning that if non-significant results were obtained even when groups of potential kin-related individuals were included, this could be taken as strong evidence that there was no linkage disequilibrium among loci.

b) *Hardy Weinberg Equilibrium*

Population-level departures of allele frequencies from those expected under Hardy Weinberg equilibrium were tested for using queen genotypic data. Individual colonies comprising the offspring of one singly mated queen are not expected to conform to Hardy-Weinberg proportions. Within these colonies, alleles are not distributed into genotypes at random, but according to the laws of Mendel (Hartl and Clark, 1997; McCauley and Goff, 1998).

II Queen mating frequency

(i) *G*-test goodness of fit

To test the hypothesis that the workers from each nest were the offspring of a single, once-mated queen, proportions of observed genotypes in each nest were compared with those expected under the hypotheses of monogyny and monandry using a goodness-of-fit *G*-test (Zar, 1999; Green and Oldroyd, 2002; Palmer *et al.*, 2002). Proposed matings and genotype distributions for each nest are presented in Appendix VIII.

The log likelihood ratio is defined as

$$\Sigma f_o \ln (f_o/f_e)$$

where f_o is the observed count and f_e is the expected count under the hypothesis being tested. Twice the log likelihood ratio, *G*, approximates the χ^2 distribution using the same number of degrees of freedom as would be used for chi-square testing (Zar, 1999). Loci were pooled either to ensure that the expected frequency for each category was ≥ 3, or when observed frequencies in any category were zero, as the logarithm of zero is undefined (Zar, 1999). All possible combinations of informative loci were evaluated for each nest.

(ii) Error detection

There are two possible sources of error when inferring queen mating frequency from worker genotypes: (a) error caused by non-detection and (b) error caused by limited sampling.

a) *Non-detection error*

The non-detection error is the probability that the offspring of one male are genetically indistinguishable from the offspring of a second male, due to the of sharing identical alleles at all loci by inseminating males. Assuming that the genetic loci are unlinked, and that the population is in Hardy-Weinberg

equilibrium with random mating, the non-detection error can be quantified by the following formula:

$$\Pi \Sigma q_i^2 \quad \text{(Boomsma and Ratnieks, 1996)}$$

where q_i is the frequency of the i^{th} allele, with the sum of squared allele frequencies for each locus multiplied over all loci. The greater the heterozygosity per locus and the greater the number of loci, the smaller the non-detection error. Queen allele frequencies were used to calculate the non-detection error at the population level. The non-detection error within each nest was estimated by multiplying the putative paternal alleles of the major patriline for each nest across all loci. When putative queen and male alleles could not be distinguished from each other due to all worker offspring being fixed for a particular genotype, the sum of the allele frequencies occurring in the workers at this locus was multiplied by the allele frequencies at the other loci. This results in a conservative estimate of the nest non-detection error (Palmer *et al.*, 2002).

b) *Non-sampling error*

The non-sampling error is the probability that a particular paternal genotype will not be detected due to limited sampling of progeny. As sample size increases, this error becomes negligible. The probability of not sampling a patriline due to limited sample size was therefore calculated as $(1-p)^n$, where p is the proportion of workers representing a particular patriline, and n is the sample size of workers collected (Boomsma and Ratnieks, 1996). This error was calculated for each nest assuming non-sampling of a patriline represented by 25% of workers in a nest, and non-sampling of a patriline represented by only 10% of workers.

(iii) Relatedness

Within the context of this study, relatedness was measured as a regression coefficient using Queller and Goodnight's (1989) index of relatedness *R*. This value can be interpreted in terms of the degree to which individuals show allelic identity by descent. The value of *R* can range from -1 to +1, and is calculated using the following formula:

$$R = \frac{\Sigma\Sigma\Sigma(Py - P^*)}{\Sigma\Sigma\Sigma(Px - P^*)}$$

where P_x is the frequency of the current allele in the individual of interest, P_y is the frequency of the current allele in all individuals to whom the relatedness of P_x is being estimated, and P^* is the frequency of the allele in the population, with all putative relatives of the current P_x individual excluded. The numerator and denominator are first summed over all alleles, loci and individuals in the population before taking a ratio of the sums to obtain the estimate of relatedness *R*.

Queller and Goodnight's (1989) relatedness index of has several advantages over previous regression methods of calculating relatedness. It allows estimation of relatedness for a single group of individuals or a single pair of individuals, as well as correcting the bias introduced by sampling a small number of groups. This index also gives more weight to rare alleles and therefore informative loci, in contrast to multidimensional regression coefficients, which lose information by averaging over loci (Queller and Goodnight, 1989).

The software package RELATEDNESS 5.0.7 (Queller and Goodnight, 1989) was used to calculate *R* for a number of categories based on genotypic data from all seven microsatellite loci. All 168 genotyped individuals were used as the reference population. Symmetrical estimates of within-nest worker-worker relatedness (R_{WW}), as well as the average nest worker-worker relatedness

were calculated. Symmetrical relatedness estimates of average worker-worker relatedness for all pairwise combinations of nests (n = 6) were also estimated. For these relatedness calculations, allele frequencies were bias-corrected by group, removing all the putative relatives of the current individual X from the calculation of the population allele frequency P*. By removing these related individuals from the calculation, a bias towards the allele frequencies observed in the current X individual, which would result in an underestimate of relatedness, is prevented. To obtain average estimates of R within and across all nests, nests were equally weighted. Estimates of queen-mate relatedness (R_{QM}), as well as male-male relatedness (R_{MM}) and queen-worker relatedness (R_{QW}) were obtained with frequency bias correction by group and groups weighted equally. Standard errors of R were calculated by jackknifing over loci, as recommended by Queller and Goodnight (1989) when the number of groups sampled is small relative to the number of loci.

Because the jacknifing procedure obtains a normal distribution for the error around the mean, significant deviation of R from the predicted value of 0.75 for worker-worker relatedness under a once mated single queen, and significant deviation of R from zero for queen-mate relatedness was tested using two-tailed t-tests (Zar, 1999; Chapuisat *et al.*, 1997). Confidence intervals (95% CI) are reported for all worker-worker relatedness estimates.

Population genetic effective queen mating frequency

Due to the haplodiploid sex determination system in the Hymenoptera, under the assumption of random mating, the female offspring of a single, once mated queen are expected to have on average 75% of their genes identical by descent. Mating with more than one male reduces this pedigree relatedness, with half sisters only expected to share 25% of their genes.

where R_s is the average genetic relatedness among sisters of the same matriline. This number can be interpreted as the number of mates per queen that have resulted in the observed genetic diversity among the queen's daughters, given no inbreeding, no variance of mate numbers among queens, nor relatedness between inseminating males (Pedersen and Boomsma, 1999). $M_{e,g}$ was calculated for the sampled population of *Camponotus klugii* based on the average population relatedness estimate of within-nest worker-worker relatedness.

Inbreeding

The inbreeding coefficient for the sampled population (*F*) was calculated using RELATEDNESS 4.2 (Queller and Goodnight, 1989). This program implements an identity-by-descent method to estimate inbreeding coefficients, where the results can be interpreted as actual identities above random, divided by the maximum possible identities above random. Groups were weighted equally, and standard errors were estimated by jackknifing over loci. The significance of the deviation of the estimate from zero was assessed using a two-tailed *t*-test with degrees of freedom equal to the number of loci minus one (Hammond *et al.*, 2001).

Results

Suitability of markers

Amplification was unsuccessful at loci FL12, FL21 or FL43. Amplification was successful at Camp6, FL20 and FL29; however, all genotyped workers (n = 28) were monomorphic at these three loci. A total of 168 individuals were therefore genotyped at the seven polymorphic loci listed in Table 6.4. The number of alleles ranged from two to nine for each locus, resulting in an average of five alleles per locus for the individuals of *Camponotus klugii* sampled in this study (Table 6.4).

Higher levels of polymorphism were detected in *C. klugii* at loci originally identified in *Camponotus consobrinus* (Ccon loci), than at loci initially identified in *Camponotus ligniperdus (*Camp loci). An average of 5.8 alleles per locus were amplified using the Ccon primers, versus 2.3 alleles per locus using the Camp primers (taking the monomorphic Camp6 locus into account).

Expected heterozygosity values in *C. klugii* ranged from 0.11 to 0.86 with an average expected heterozygosity of 0.62 (Table 6.5). This falls within the range of heterozygosity values found in other ant species (Evans, 1993; Chapuisat, 1996; Krieger and Keller, 1999). Furthermore, it indicates a sufficient level of resolution to elucidate nest genetic structure. Average expected heterozygosity for the two loci amplified in *C. klugii* using the *C. ligniperdus* primers Camp4 and Camp8, (H_e = 0.36) was substantially lower than the average expected heterozygosity for the five loci amplified in *C. klugii* using *C. consobrinus* primers (H_e = 0.73).

Table 6.5. Allelic variation and expected heterozygosity estimates based on allelic frequencies calculated from queens.

Locus	Allelic size variants detected (sizes given in base pairs)	$H_e \pm$ s.e.
Camp4	205, 207, 209, 211	0.60 ± 0.10
Camp8	121, 123	0.11 ± 0.50
Ccon12	165, 171, 173, 175, 177, 179, 181	0.82 ± 0.07
Ccon20	289, 291, 293, 295, 297, 299	0.78 ± 0.07
Ccon42	262, 264, 266	0.62 ± 0.10
Ccon70	159, 161, 163, 165	0.56 ± 0.16
Ccon79	356, 358, 364, 370, 381, 385, 389, 395, 412	0.86 ± 0.05
Average across all loci		0.62 ± 0.10

Composition of nests

In total, nine queen genotypes were obtained, corresponding to the three foundress queens, the two queens found in the queenright nests, and the four queens inferred to have produced the workers in all four worker nests (Table.6.6).

Table 6.6. Breakdown of nest composition.

Nest	No. of workers	Queen present?	Queen present mother of workers?
1	26	Yes	No
2	30	Yes	Maybe
3	55	No	n.a.
4	52	No	n.a.
5	0	Yes	n.a.
6	0	Yes	n.a.
7	0	Yes	n.a.

Queen-worker incompatibility

The multilocus genotype of the queen present in Nest 1 was found to be incompatible with her being the mother of the workers found in the nest (Figure 6.1).

	Loci						
	Camp4	Camp8	Ccon12	Ccon20	Ccon42	Ccon70	Ccon79
Queen Genotype:	207/211	123/123	171/181	293/295	262/264	161/165	358/389
Worker Genotypes:	209/209	123/123	173/177	293/297	262/264	161/163	385/385
	207/209		173/179		262/262		356/385

Figure 6.1 Multilocus genotypes of the resident queen and workers in Nest 1 demonstrating queen-worker incompatibility. Solidly boxed loci are those at which the queen and workers have no alleles in common, with queen and one genotypic class of workers sharing one allele at the loci indicated by dashed boxes.

The queen-worker genotypic incompatibility in Nest 1 was reflected by the negative queen-worker relatedness estimate for Nest 1 (R_{QW} = -0.235 ± 0.227 s.e.) which was significantly different from the expected value of 0.5 for queen-daughter relatedness (Table 6.1) (t_6 = 3.2, P = 0.02), but not significantly different from zero (t_6 = 1.0, P = 0.36).

Only three loci successfully amplified in the queen found in Nest 2, namely Camp4, Camp8 and Ccon12. At one of these loci (Ccon12), the queen had an genotype consistent with her being the sister, but not the mother of the workers. However, the other two loci for which this queen was typed (Camp4 and Camp8) did not exclude her as the queen. As no genotypic data was obtained from the remaining four loci despite numerous amplification attempts, it cannot be stated with certainty that this queen was not the mother of the sampled offspring. This is due to the fact that the observed incongruence at one locus could be due to a germline mutational event, or an amplification artefact.

Linkage Disequilibrium and Hardy Weinberg

No significant linkage disequilibrium between any pairwise combination of the seven loci was detected across all genotyped individuals after strict Bonferroni correction for multiple applications of the same test ($P > 0.5$) (Rice, 1989). A similar results was obtained when only queens were included. This indicates that within this sample of *C. klugii*, these microsatellite loci segregate independently and can be treated as independent, unlinked markers.

No significant deviation from Hardy-Weinberg equilibrium was detected at the population level at any of the seven microsatellite loci, resulting in a non-significant result at the population level for all loci combined ($\chi^2_{[8]} = 7.8$, $P = 0.45$).

Distribution of genotypes across colonies

Workers from all four nests could be differentiated from each other on the basis of their multilocus genotypes. The Ccon loci were particularly informative in this regard (Table 6.7).

Table 6.7. Distribution of worker multilocus genotypes across nests, with alleles labelled according to their size in base pairs.

Nest	Locus						
	Camp4	Camp8	Ccon12	Ccon20	Ccon42	Ccon70	Ccon79
1	207/209 209/209	123/123	173/177 173/179	293/297	262/262 262/264	161/163	385/385 356/385
2	207/209 207/207	123/123	165/181 171/181	293/299 291/293	262/262	161/161 161/165	358/370 358/364
3	205/205 209/209	121/121 121/123	173/175 173/177	293/295 291/293	262/266	161/161	395/412 389/395
4	207/209 209/209	123/123	165/173 165/179	291/293 289/291	262/264 262/266	161/161 161/163	358/370 358/381

Observed genotypic differences between nests were highly significant. Simultaneous comparisons of genotypic distributions at all seven loci across all four nests revealed significant genotypic differentiation (*G*-like test; $P < 0.0001$ for each locus). Of the 42 pair-wise comparisons of genotypic differentiation for all pairs of nests at all loci, 39 were highly significant after Bonferroni correction for multiple tests ($P < 0.01$). The remaining three comparisons were meaningless due to identical genotypes observed in those pairs.

An unusual genotype distribution was observed in Nest 3 for Camp4, where workers were fixed for one of two homozygous genotypes – 205/205bp and 209/209bp with no 205/209bp heteroyzgotes observed (Table 6.7). One possible explanation for this observed pattern is that the mother of the workers was inseminated by a male carrying a null allele at this locus i.e. the male possessed a mutation in the primer binding regions flanking the microsatellite, preventing primer binding and amplification. If the queen was heterozygous at this locus for 205/209bp, and assuming Mendelian segregation of alleles, two types of gametes, one carrying the allele 205bp and the other carrying the allele 209bp would be expected to be produced in a 1:1 ratio. Workers carrying one of the two queen alleles and the male null allele would be mistakenly genotyped as 205/205bp or 209/209bp homozygotes due to the non-amplification of the male allele. Consistent with this theory, the ratio of 209/209bp homozygote workers to 205/205bp homozygote workers in Nest 3 (31:24) was not significantly different from the expected ratio of 1:1. This conclusion is based on the results from a True Basic program (Appendix IX) written by E. Harley, where a ratio of 31:24 or higher was observed in 41% of 10 000 random simulations. Another possible explanation for the observed genotypes at this locus in Nest 3 could be that two queens produced the worker offspring, with each queen homozygous for the 205bp or 209bp allele. Each queen would have to have been inseminated by a male with the same allele at Camp4 as she carried at this locus so as to generate the two observed classes of homozygotes. However, this scenario is highly unlikely, as no evidence for more than one queen is provided by the other six loci. A maximum of three alleles were observed at any one locus,

and the genotypic classes at all these loci were consistent with a single, once-mated queen producing all the observed workers.

In order not to confound estimates of relatedness and queen mating frequency, workers from Nest 3 were scored as 'missing' for this locus in all subsequent analyses.

Queen mating frequency

All four nests containing workers were classified as monogynous. The maximum number of alleles observed at any one locus for all workers from a particular colony never exceeded three. This is consistent with the queen of each nest only having mated once. Furthermore, with the exception of the Camp4 locus in Nest 3 workers as discussed above, worker genotypes within each nest at each locus were consistent with one of the two types of genotypic distributions expected for Mendelian markers in the progeny of a single, once-mated queen. They are (i) only one genotypic class of females present or (ii) two genotypic classes present, of which one is heterozygous (Shoemaker *et al.*, 1992).

For each of the four nests, the observed distribution of worker genotypes fitted well with the expected distribution under the hypothesis of a single queen mated with one haploid male (Table 6.8). No significant deviations from the expected genotypic distributions under the hypothesis of monogyny and monandry were detected in any of the four nests, for any possible combination of informative loci.

Table 6.8. Goodness of fit *G*-test results for each nest under the hypotheses of monogyny and monandry.

Nest	*n*	No. loci analysed	Genotype combinations	G-test		
				G	d.f.*	*P*
1	26	4				
			Camp4,Ccon12,Ccon79	4.87	7	0.68
			Camp4,Ccon12, Ccon42	3.03	7	0.88
			Camp4,Ccon42, Ccon79	2.13	7	0.95
			Ccon12,Ccon42, Ccon79	5.18	7	0.64
2	30	5				
			Camp4, Ccon12, Ccon20	9.44	7	0.22
			Camp4, Ccon12, Ccon70	7.02	7	0.43
			Camp4, Ccon12, Ccon79	6.12	7	0.53
			Camp4, Ccon20, Ccon70	3.43	7	0.84
			Camp4, Ccon20, Ccon79	5.49	7	0.60
			Camp4, Ccon70, Ccon79	13.12	7	0.069
			Ccon12, Ccon20, Ccon70	6.27	7	0.51
			Ccon12, Ccon20, Ccon79	6.79	7	0.45
3	55	4				
			Camp8, Ccon12, Ccon20, Ccon79	14.57	15	0.48
4	52	6				
			Camp4, Ccon12, Ccon20, Ccon42	13.98	15	0.53
			Camp4, Ccon12, Ccon20, Ccon70	11.27	15	0.73
			Camp4, Ccon12, Ccon20, Ccon79	14.22	15	0.51
			Camp4, Ccon20, Ccon42, Ccon79	21.56	15	0.12
			Camp4, Ccon42, Ccon70, Ccon79	18.98	15	0.21
			Camp4, Ccon12, Ccon70, Ccon79	12.31	15	0.66

n = number of workers sampled.
d.f. calculated as no. categories –1.
Note: Only those genotypic combinations for which *G* could be calculated (i.e. those combinations containing no categories with observed counts equal to zero) are presented in this table.

Non-detection error estimates, both at the population level and at the level of the individual colony, were extremely small due to the combination of high allelic variability at some loci and the large number of loci used in this study (Table 6.9). This indicates that it is highly unlikely that multiple mating events were missed due to the sharing of alleles at all seven microsatellite loci by inseminating males. The probability of not observing a particular paternal genotype due to limited sampling was also negligible due to the high number of workers sampled per colony, even when assuming a particular patriline is represented by only 10% of all workers (Table 6.9).

Table 6.9. Non-detection and non-sampling error estimates for not detecting additional patrilines.

		Non-detection error	Non-sampling error	
Nest	*n*	$d_{p(nest)}$	Patriline represented by 25% of workers	Patriline represented by 10% of workers
1	26	9×10^{-4}	5×10^{-4}	0.065
2	30	6×10^{-4}	2×10^{-4}	0.042
3	55	6×10^{-4}	1×10^{-7}	0.003
4	52	$*9 \times 10^{-4}$	3×10^{-7}	0.004
Population		3×10^{-4}		

n = number of workers sampled per nest.
* this estimate is a conservative estimate of the non-detection error for this nest as the putative male contained a novel allele at Ccon79 not present in any of the queen genotypes on which the population allele frequencies were based.

Relatedness

Intra-nest relatedness estimates are consistent with observations based on genotypic data, indicating that all sampled workers from each nest are the progeny of a single, once-mated queen (Table 6.10). Nests 1, 2 and 3 all had R_{ww} values above 0.70, close to the theoretical expectation of 0.75. This was reflected in the non-significant *P*-values obtained when comparing these

observed relatedness estimates to the expected value of 0.75. Nest 4 is characterised by a markedly lower average worker-worker relatedness estimate compared to average worker-worker relatedness within the other three colonies (Table 6.10). However, the R_{ww} for Nest 4 of 0.483 ± 0.111 is not significantly different from the expected value of 0.75, due to the high standard error associated with this estimate. The mean estimate of worker-worker relatedness within nests of *C. klugii*, R_{ww} = 0.686 ± 0.037 reflects the high levels of relatedness between workers in Nests 1, 2 and 3.

Table 6.10. Regression relatedness estimates for worker-worker relatedness in *Camponotus klugii*

			t-test (sig. different from *R* = 0.75?)		
Nest	R_{ww} ± S.E.	95% CI	*t*	d.f *.	*P*
1	0.740 ± 0.079	0.55, 0.93	0.13	6	0.90
2	0.706 ± 0.091	0.49 , 0.93	0.48	6	0.65
3	0.752 ± 0.043	0.65, 0.85	0.05	6	0.96
4	0.483 ± 0.111	0.21, 0.75	2.40	6	0.053
Overall nest *R*	0.686 ± 0.037	0.60, 0.78	1.73	6	0.13

*d.f. calculated as no. loci - 1

Regression estimates of worker-worker relatedness for all pairwise combinations of nests were significantly different from the expected value of 0.75 when workers initiate new nests by moving across to occupy additional nesting space. However, these pairwise nest-relatedness values were not significantly different from the value of zero expected when each nest is independently founded by an unrelated queen inseminated by an unrelated male (Table 6.11). This was reflected by the 95% confidence intervals of pairwise colony relatedness estimates largely overlapping zero, but not overlapping 0.75 (Figure 6.2).

Table 6.11. Pairwise estimates of worker-worker relatedness between nests.

Nest comparison	symmetrical $R \pm$ S.E.	95% CI	Sig. different from $R = 0.75$?	Sig. different from $R = 0$?
1 vs. 2	-0.093 ± 0.196	-0.57, 0.57	*	N.S.
1 vs. 3	-0.203 ± 0.237	-0.78, 0.38	*	N.S
1 vs. 4	-0.052 ± 0.152	-0.42, 0.32	*	N.S
2 vs. 3	-0.194 ± 0.180	-0.63, 0.25	*	N.S
2 vs. 4	0.095 ± 0.156	-0.29, 0.48	*	N.S
3 vs. 4	-0.341 ± 0.220	-0.88, 0.20	*	N.S

*$P < 0.05$ after Bonferroni correction.

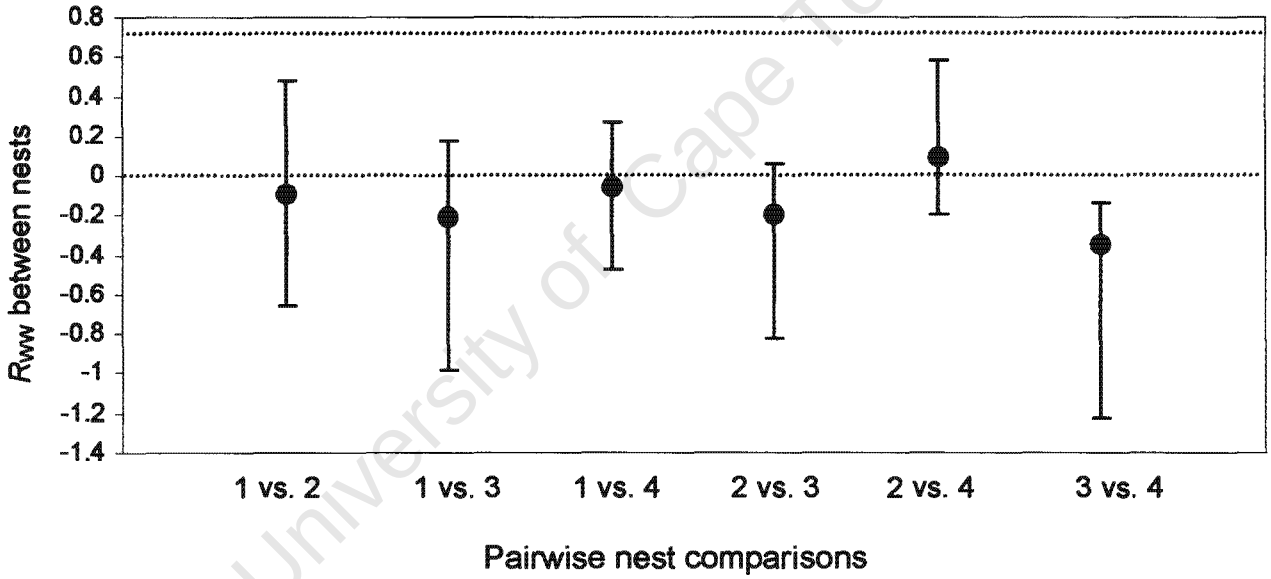


Figure 6.2. Pairwise estimates of worker-worker relatedness between nests showing 95% confidence intervals. The dashed blue line indicates the expected worker-worker relatedness value between nests if colony structure is polydomous ($R_{WW} = 0.75$). The dashed green line indicates the expected worker-worker relatedness value between nests if nests are independently founded ($R_{WW} = 0$).

Estimates of queen-mate (R_{QM} = -0.019 ± 0.17 s.e.; R_{MQ} = -0.100 ± 0.05, four queen-mate pairs) and male-male relatedness (R_{MM} = 0.074 ± 0.13 s.e., four male mates) were not significantly different from zero ($t_6 = 0.11$, $P = 0.92$; $t_6 = 2$, $P = 0.09$; $t_6 = 0.57$, $P = 0.59$), suggesting that queens mated with unrelated males. This was reflected by the inbreeding coefficient F that did not differ

significantly from zero (F = -0.034 ± 0.09 s.e., t_6 = 0.38, P = 0.71), indicating the absence of inbreeding in this local population of *C. klugii*.

Effective queen mating frequency

The effective queen mating frequency, based on the overall intracolonial worker-worker relatedness estimate, was 1.1. This value is within the range of mean effective paternity frequencies (1 to 1.48) estimated for 19 species of ants for which there was reliable genetic data (Boomsma and Ratnieks, 1996).

Discussion

Camponotus klugii is monodomous

Contrary to the suggestion by Skaife (1961) and Curtis (1985) that queenless nests of *C. klugii* may represent satellite nests of queenright colonies that had expanded, the two queenless nests examined in this study appeared to be orphaned, independently founded colonies. This conclusion is based on the observation that the average relatedness estimates of workers in either of these two nests to workers in other nests were indistinguishable from zero. Furthermore, significant genotypic differentiation was found between all nests sampled, indicative of independent colony founding by unrelated outbred queens. It is possible that the queenless nests were polydomous nest fragments of a queenright colony that had not been sampled. However, although the surrounding vegetation within a 15 m radius was thoroughly examined for *C. klugii* nests, none were found. Therefore, it seems likely that *C. klugii* colonies are monodomous rather than polydomous, with each individual nest representing an independently founded colony. The words 'colony' and 'nest' will therefore hereafter be used interchangeably.

Camponotus klugii is characterised by a simple colony kin structure

Camponotus klugii is characterised by a simple colony kin structure, with all workers the resulting offspring of a single, once-mated queen. This conclusion is based on the observed distribution of multilocus genotypes of workers within colonies, as well as regression relatedness values between workers, which are consistent with the theoretical expectation of 0.75 for full siblings within haplodiploid systems.

The low worker-worker relatedness estimate obtained for Colony 4 (R = 0.483), although not significantly different from the expected values of 0.75 for full-sibling workers, could indicate that within this colony the queen was multiply mated, or alternatively that worker offspring from more than one matriline were present. However, the distribution of worker genotypes did not differ significantly from that expected under a once-mated single queen, as

assessed by a log-likelihood *G*-test. Furthermore, no more than three alleles were observed at any one locus across all seven loci, indicating the presence of only one matriline and patriline. Therefore the relatively lower average relatedness estimate characterising full-sister relationships within this colony does not appear to reflect an important biological phenomenon, such as multiple mating or multiple queens. Rather, it can be attributed to peculiar allelic frequencies within this colony relative to the background population.

The estimated population effective queen mating frequency of 1.1 supports the conclusion of Boomsma and Ratnieks (1996) that low effective queen-mating frequency is the norm for most ant species, with high effective paternity (> 2) restricted to phylogenetically isolated genera such as *Atta*, *Acromyrmex*, and *Pogonomyrmex* (Fjerdingstad *et al.*, 1998; Bekkevold *et al.*, 1999; Boomsma *et al.*, 1999; Cole and Wiernasz, 2000). It is highly unlikely that additional patrilines were missed due to non-sampling error or non-detection error, as these error estimates were very low as a result of the large number of workers sampled per colony, and the high levels of polymorphism observed for certain loci.

Mating in *C. klugii* appears to occur between unrelated individuals. This is reflected both by queen-mate relatedness and the inbreeding coefficient *F* not differing significantly from zero. This is not an unexpected result when considering the mating behaviour of *Camponotus spp.*, which follows the male aggregation syndrome (Hölldobler and Bartz, 1985; Robertson and Zachariades, 1997). In response to heavy rainfall, winged reproductives are released more-or-less synchronously from different nests. Individual queens then fly into aerial swarms of males, who call the virgin queens by producing pheromones. Afterwards, inseminated queens typically disperse widely before excavating or finding a nest. This type of mating strategy minimises the chance of mating between related individuals, with the consequence that inbreeding is expected to be very rare (Hölldobler and Bartz, 1985; Bourke and Franks, 1995).

Potential kin conflicts in Camponotus klugii

The simple sociogenetic structure of *C. klugii* colonies implies that potential queen-worker conflict is expected to be maximal under kin selection theory, both with respect to production of male haploid eggs as well as to sex allocation. From a kin-selection perspective, the high average within-colony relatedness of *C. klugii* workers indicates that workers will be under selective pressure to produce their own sons and raise their sister's sons, rather than rear their brothers. However, the small size of *C. klugii* colonies may allow the queen to effectively police the production of worker-laid male eggs, either by pheromonal control, physical aggression or cannibalisation of the worker-laid eggs (Ratnieks and Reeve, 1992; Bourke and Franks, 1995). Sampling and genetic analysis of colonies containing a queen, workers and males are vital to ascertain whether males are of queen or worker origin. Furthermore, behavioural studies and nest manipulation experiments are required to conclusively address these hypotheses.

The full-sibling status of *C. klugii* workers within colonies also indicates that they may potentially prefer a female-biased reproductive sex ratio of 3:1. This is in contrast to the queen, who will favour equal investment in both sexes. Determination of whether sexual allocation is under queen-control or worker control in this species is contingent upon comprehensive sampling of colonies in order to determine the population sex-ratio allocation, as well as genetic analyses of these colonies to confirm their kin structure (e.g. Brown and Keller, 2000).

Queen replacement occurs in Camponotus klugii

Surprisingly, this study revealed unambiguous evidence for queen replacement by an unrelated queen in one of the sampled nests of *C. klugii*. In Nest 1, the resident queen's multilocus genotype was incompatible with her being the mother of the workers present. This must have been a fairly recent replacement event, as only one worker matriline (set of siblings derived from the same mother) was detected within this nest, even though five pupae were included in the analysis.

This finding is surprising, as it was commonly assumed that a monogynous colony was doomed to death after the death of the queen. Although it is theoretically possible for orphaned workers to raise another queen by selectively feeding some of the previous queen's larvae or female eggs, this has rarely been found to be the case (Hölldobler and Wilson, 1994; Heinze and Keller, 2000). Without a queen present, no workers can be produced to maintain the colony, which will eventually decline until the last worker dies. In recent years, however, the investigation of sociogenetic structure of social insects using sensitive molecular markers like microsatellites has resulted in the documentation of queen replacement events in numerous monogynous species (e.g. Foitzik and Heinze, 1998; Gadau *et al.*, 1998; Foitzik and Herbers, 2001), indicating that this might not be as rare a phenomenon as was previously thought.

From a worker perspective, invasion of a colony by an unrelated queen amounts to nothing less than intraspecific parasitism, as workers gain no inclusive fitness benefits from rearing the offspring of an unrelated queen. The only circumstance under which orphaned workers may gain inclusive fitness benefits from accepting an alien queen is when the sexual brood from the previous queen is present, and the presence of the alien queen increases the probability of colony survival. This is predicted to occur in species where the larval development stage is longer than the average lifespan of a worker, and has been used to explain the adoption of unrelated queens into orphaned colonies of *C. ligniperdus* (Gadau *et al.*, 1998). From a queen perspective, the advantage of invading an established colony is readily apparent: a greatly enhanced chance of successful reproduction when compared to the highly risky undertaking of independent colony founding. This is especially apparent when there are strong ecological constraints on successful independent colony founding, such as nest-site limitation, habitat patchiness or predation (Bourke and Franks, 1995; Banschbach and Herbers, 1996).

In the monogynous myrmicine ant *Leptothorax nylanderi*, new queens were observed to invade both queenright and queenless colonies when there was a scarcity of suitable nesting sites (Foitzik and Herbers, 1998). Like *C. klugii*,

this species is incapable of constructing its own nests, and likewise depends on abandoned pupal cavities of longhorn beetles, or hollow acorns, hazelnuts or grass stems for suitable nesting cavities. The paucity of suitable nest sites was shown to be a major ecological constraint on successful independent colony founding in this species. As a consequence, new foundress queens were found to seek adoption into alien colonies, rather than attempting to found their own colonies.

It is tempting to speculate that as *C klugii* also relies on empty cavities in dead fynbos plants for nests, and is unable to construct nests, nest-site limitation may be an important ecological constraint shaping the organization of *C. klugii* colonies. Furthermore, competition among various cavity-dwelling species for empty nest cavities in fynbos appears to be strong. This is based on the observation that when potential nest cavities for *C. klugii* are examined, they are inevitably found to be occupied by cockroaches, beetles or various other ant species e.g. *Crematogaster spp.* (H. Robertson, pers. com.). However, no ecological studies on nest-site limitation have been conducted for this species, therefore the relative importance of this ecological factor is unknown in *C. klugii.* Furthermore, additional sampling is required to determine how widespread the phenomenon of queen replacement is.

The possibility of queen replacement occurring should be taken into consideration in future studies on *C. klugii* when evaluating colony genetic structure, as multiple paternity may incorrectly be inferred from worker genotypes where queen replacement has occurred and two or more matrilines are present within the colony. Lower-than-expected relatedness estimates between workers or between worker and brood may also serve as an indication of queen replacement (e.g. Hammond *et al.*, 2001).

No evidence of thelytokous parthenogenesis in orphaned colonies of Camponotus klugii.

The clonal genetic signature of thelytokous parthenogenesis was not observed in either of the queenless colonies, as workers from these two colonies were not fixed for a particular multilocus genotype. Rather, the observed multilocus genotype distribution of workers within these two colonies was fully consistent with a single-mated queen origin, as previously discussed. Further laboratory-rearing experiments are required to corroborate the finding of Skaife (1961), namely that queenless colonies with all brood removed were capable of producing female offspring. If this finding is corroborated, widespread sampling would be required to determine whether the ability to reproduce by thelytokous parthenogenesis is sporadic in *C. klugii*, as is thought likely for many hymenopteran species (Crozier and Pamilo, 1996), or whether it constitutes a major part of the life cycle of this ant species.

Sampling strategies and eusocial insects

Eusocial insects represent an interesting challenge for population genetics studies in terms of sampling, particularly if kin structure within a colony is simple i.e. a once-mated single queen. This is because high relatedness due to kinship ties between individuals within colonies renders them genetically non-independent. If multiple workers from eusocial insect colonies are included as independent individuals in population genetic analyses, population genetic inferences may be biased (Goodisman *et al.*, 2001). A number of ingenious resampling techniques have been developed to avoid the problems of genetic non-independence in estimating social insect population genetic parameters of interest (e.g. Goodisman *et al.*, 2001; Goropashnaya *et al.*, 2001; Goodisman and Crozier, 2002). For a study on the hierarchical population structure of the wasp, *Vespula germanica* in Australia, Goodisman *et al.* (2001) wrote a computer program that randomly selected a single wasp's multilocus genotype from each nest. This resulted in a dataset with the number of individuals equivalent to the number of nests sampled. This procedure was repeated multiple times, and population genetic

parameters of interest were then calculated for each of these multiple datasets. The median of these values was then taken as the unbiased estimate of the parameter of interest.

In this study, the small number of colonies sampled precluded testing for Hardy-Weinberg equilibrium and linkage disequilibrium from reduced resampled data sets as described above, and necessitated the use of the independent queen genotypes to infer possible departures from these population genetic parameters. The large standard errors associated with the expected heterozygosity estimates of some loci reflect the potential effects of small sample size (Nei, 1978). Nevertheless, the major conclusions drawn from this study concerning the relatedness structure of colonies and genotypic differentiation between colonies are to a large extent independent of expected heterozygosity estimates. However, increased sampling for future studies would be desirable in order to obtain more accurate estimates of population allelic frequencies and thus classic population genetic parameters e.g. Wright's hierarchical *F*-statistics F_{ST}, F_{IS} and F_{IT} (Weir and Cockerham, 1984).

A further challenge faced when studying eusocial insects, particularly in Africa, is the dearth of ecological knowledge for the vast majority of species (H. Robertson, pers. com.). Without baseline ecological data on a species, it is very difficult to know, *a priori*, how to sample populations for genetic analyses in a way that most effectively reflects their demographic properties.

The sampling strategy implemented in this study limited the amount of information that could be extracted, and hence the conclusions that were drawn from genetic analyses. Nevertheless, the results presented in this study make a significant contribution to the baseline ecological and genetic knowledge of *C. klugii*. In summary, the major findings of Part II of this thesis are that *C. klugii* is characterised by a monodomous colony structure, with workers within colonies produced by a single queen mated with a single male. Additionally, queen replacement was documented in this study. However, no evidence was found of thelytokous parthenogenesis by female workers in orphaned colonies.

Knowledge of *C. klugii* colony structure can be applied in future studies of this species in order to facilitate both accurate ecological sampling and subsequent fine-scale colony and population genetic analyses. Furthermore, the high level of resolution provided by the suite of microsatellite markers utilised in this study, in particular the Ccon loci, indicate that they are ideal genetic markers with which to confirm and investigate the sociobiology of the endemic fynbos ant, *Camponotus klugii*.

Future research

The results of this microsatellite-based investigation of *C. klugii* sociogenetic structure raise a significant number of intriguing questions. How frequent is queen replacement? Does this phenomenon only occur in orphaned colonies, or can unrelated queens invade queenright colonies? How do the usurper queens gain access to colonies, and how do they overcome or avoid the kin recognition system of workers? Is queen replacement by an unrelated queen a function of nest-site limitation, as hypothesised in this study? Are males of this species worker-derived, as hypothesised from the simple colony structure of this species, or queen-derived? What is the stable population sex ratio? Are queens uniformly singly-mated, or is there a low incidence of double or triple mating in some colonies? Are workers of *C. klugii* capable of theylokous parthenogenesis, and if so what role does this play in the life cycle of this species?

A combination of field and laboratory observational and experimental work, as well as wide-scale sampling and genetic analyses are required to address the questions and hypotheses posed above.

PART III

CONCLUDING REMARKS

Molecular markers are one of the most important tools we, as evolutionary biologists, have at our disposal to achieve our goal of understanding the natural history and evolutionary biology of organisms. The multifaceted biological applications of molecular markers to addressing both macro- and micro-evolutionary questions are embodied in this thesis, wherein mitochondrial and microsatellite markers were used to investigate two diverse aspects of the biology of the ant genus *Camponotus*.

At a macro-evolutionary level, the utility of two mitochondrial genes for resolving species relationships within the genus *Camponotus* were investigated. *Camponotus* molecular phylogenetics is significantly underdeveloped, despite the complex taxonomy of this genus, with only four published molecular phylogenies available prior to this study (Sauer *et al.*, 2000; Brady *et al.*, 1999; Gadau *et al.*, 1999; Sameshima *et al.*, 1999). Since all these studies were based on sequence data from a single mitochondrial gene, no inferences can be made regarding the comparative utility of the genetic marker used for resolving phylogenetic relationships. This is the first study, to my knowledge, in which the comparative utility of the cytochrome *b* and cytochrome oxidase II mitochondrial genes for resolving phylogenetic relationships within *Camponotus* is explicitly assessed. As the comparative utility of a gene is highly contingent upon its molecular evolutionary dynamics, the emphasis in Part I of this thesis was on investigating and characterising differences in rates and patterns of molecular evolution between the two gene fragments. The finding that cytochrome *b* showed greater utility in phylogenetic reconstruction than cytochrome oxidase II, based on a number of considerations, will hopefully allow more informed decisions to be made regarding molecular phylogenetic sampling for future studies. Furthermore, Part I of this study provides a new perspective on the intricate taxonomy of this group, with existing morphologically based subgeneric classifications shown to be often inaccurate. In addition, some novel phylogenetic

relationships were revealed in the course of this investigation, such as the sister association between *Camponotus bifossus* and *Camponotus sp.* 12, and the paraphyletic association of subgenera *Myrmopiromis* and *Myrmopsamma*.

At a micro-evolutionary level, microsatellite markers provided unambiguous genetic data, facilitating elucidation of the colony genetic structure of *Camponotus klugii*. This is the first study, to the best of my knowledge, to investigate the social genetic structure of any *Camponotus* species in Africa, and represents one of the few molecular genetic investigations of the social structure of a species in this diverse genus world-wide. Although sampling was limited, the results of this study have revealed preliminary and novel insights into the social structure of this species. Furthermore, Part II of this thesis provides an important contribution to the sparse baseline ecological and genetic data for *Camponotus klugii*, and adds to the knowledge of life history patterns in this large genus of ants. The finding of queen replacement by an unrelated queen is of particular interest, as it is one of the few cases of intraspecific parasitism reported in ants thus far. Hitherto, intraspecific parasitism, whereby foundress queens invade unrelated orphaned colonies, had only been reported in the fire ant, *Solenopsis invicta* (Tschinkel, 1996) and *Leptothorax nylanderi* (Fotizik and Heinze, 1998). The observation of this phenomenon in the cavity-dwelling *Camponotus klugii*, where nest site limitation is likely to be an important ecological constraint on independent colony founding, indicates that the social and genetic structure of colonies of this species may be strongly shaped by ecological factors. It is hoped that knowledge of the colony kin structure of *Camponotus klugii*, contributed to by Part II of this thesis, will inform future sampling schemes, thus facilitating insights into fundamental questions such as reproductive conflicts within the colony, and sex ratio theory.

REFERENCES

Agosti, D. (1991). Revision of the oriental ant genus *Cladomyrma*, with an outline of the higher classification of the Formicinae (Hymenoptera: Formicidae). Syst. Entomol. *16*, 293-310.

Akre, R. D., Hansen, L. D., and Myhre, E. A. (1994). Colony size and polygyny in carpenter ants (Hymenoptera: Formicidae). Journal of the Kansas Entomological Society *67*, 1-9.

Arnold, G. (1924). A monograph of the Formicidae of South Africa. Part VI. Camponotinae. Ann. S. Afr. Mus. *14*, 675-766.

Asakawa, S., Kumazawa, Y., Araki, T., Himeno, H., Miura, K., and Watanabe, K. (1991). Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. J. Mol. Evol. *32*, 511-20.

Ashmead, W. H. (1905). A skeleton of a new arrangement of the families, subfamilies, tribes and genera of the ants, or the superfamily Formicoidea. Can. Entomol. *37*, 381-384.

Avise, J. C. (1994). Molecular markers, natural history and evolution. (New York: Chapman and Hall).

Avise, J. C. (2000). Phylogeography. The history and formation of species. (Cambridge: Harvard University Press).

Ayala, F. J., Wetterer, J. K., Longino, J. T., and Hartl, D. L. (1996). Molecular phylogeny of Azteca ants (Hymenoptera:Formicidae) and the colonization of *Cecropia* trees. Mol. Phylogenet. Evol. *5*, 423-8.

Baker, R. H., and DeSalle, R. (1997). Multiple sources of character information and the phylogeny of Hawaiin drosophilids. Syst. Biol. *46*, 654-673.

Baker, R. H., Wilkinson, G. S., and DeSalle, R. (2001). Phylogenetic utility of different types of molecular data used to infer evolutionary relationships among stalk-eyed flies (Diopsidae). Syst. Biol. *50*, 87-105.

Baker, R. H., Yu, X., and DeSalle, R. (1998). Assessing the relative contribution of molecular and morphological characters in simultaneous analysis trees. Mol. Phylogenet. Evol. *9*, 427-36.

Banschbach, V. S., and Herbers, J. M. (1996). Complex colony structure in social insects: I. Ecological determinants and genetic consequences. Evolution. *50*, 285-297.

Baroni Urbani, C. (2000). Rediscovery of the Baltic amber genus *Prionomyrmex* (Hymenoptera, Formicidae) and its taxonomic consequences. Eclogae geol. Helv. *93*, 471-480.

Baroni Urbani, C., Bolton, B., and Ward, P. S. (1992). The internal phylogeny of ants (Hymenoptera: Formicidae). Syst. Entomol. *17*, 310-329.

Baur, A., Buschinger, A., and Zimmermannn, F. K. (1993). Molecular cloning and sequencing of 18S rDNA gene fragments from six different ant species. Insectes Soc. *40*, 325-335.

Baur, A., Chalwatzis, N., Buschinger, A., and Zimmermann, F. K. (1995). Mitochondrial DNA sequences reveal close relationships between social parasitic ants and their host species. Curr. Genet. *28*, 242-247.

Baur, A., Sanetra, M., Chalwatzis, N., Buschinger, A., and Zimmerman, F. K. (1996). Sequence comparisons of the internal transcribed spacer region of ribosomal genes support close relationships between parasitic ants and their respective host species (Hymenoptera: Formicidae). Insectes Soc. *43*, 53-67.

Beckenbach, A. T., Wei, Y. W., and Liu, H. (1993). Relationships in the *Drosophila obscura* species group, inferred from mitochondrial cytochrome oxidase II sequences. Mol. Biol. Evol. *10*, 619-634.

Bekkevold, D., Frydenberg, J., and Boomsma, J. J. (1999). Multiple mating and facultative polygyny in the Panamanian leafcutter ant *Acromyrmex echinatior*. Behav. Ecol. Sociobiol. *46*, 103-109.

Bensasson, D., Zhang, D., Hartl, D. L., and Hewitt, G. M. (2001). Mitochondrial pseudogenes: Evolution's misplaced witnesses. Trends Ecol. Evol. *16*, 314-321.

Bensasson, D., Zhang, D. X., and Hewitt, G. M. (2000). Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. Mol. Biol. Evol. *17*, 406-415.

Bielawski, J. P., and Gold, J. R. (1996). Unequal synonymous substitution rates within and between two protein-coding mitochondrial genes. Mol. Biol. Evol. *13*, 889-892.

Bolton, B. (1994). Identification guide to the ant genera of the world. PSW Edition (Cambridge: Harvard University Press).

Bolton, B. (1995). A new general catalogue of the ants of the world. PSW Edition (London: Harvard University Press).

Boomsma, J. J. (1989). Sex investment ratios in ants: Has female bias been systematically overestimated? Am. Nat. *133*, 517-532.

Boomsma, J. J., and Ratnieks, F. L. W. (1996). Paternity in eusocial Hymenoptera. Phil. Trans. R. Soc. Lond. B. *351*, 947-975.

Bourke, A. F., Green, H. A., and Bruford, M. W. (1997). Parentage, reproductive skew and queen turnover in a multiple-queen ant analysed with microsatellites. Proc R Soc. Lond B Biol Sci. *264*, 277-283.

Bourke, A. F. G. (1988). Worker reproduction in the higher eusocial hymenoptera. Q. Rev. Biol. *63*, 291-307.

Bourke, A. F. G., and Franks, N. R. (1995). Social evolution in ants. J. R. Krebs and T. Clutton-Brock, eds. (New Jersey: Princeton University Press).

Brady, S. G., Gadau, J., and Ward, P. S. (1999). Is the ant genus *Camponotus* paraphyletic? Fourth International Hymenopterist Conference.

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

Brocchieri, L. (2001). Phylogenetic inferences from molecular sequences: Review and critique. Theoretical population biology *59*, 27-40.

Broughton, R. E., Stanley, S. E., and Durrett, R. T. (2000). Quantification of homoplasy for nucleotide transitions and transversions and a reexamination of assumptions in weighted phylogenetic analysis. Syst. Biol. *49*, 617-627.

Brown, J. M., Pellmyr, O., Thompson, J. N., and Harrison, R. G. (1994). Phylogeny of *Greya* (Lepidoptera: Prodoxidae), based on nucleotide sequence variation in mitochondrial cytochrome oxidase I and II: Congruence with morphological data. Mol. Biol. Evol. *11*, 128-141.

Brown, W. D., and Keller, L. (2000). Colony sex ratios vary with queen number but not relatedness asymmetry in the ant *Formica exsecta*. Proc R Soc. Lond B Biol Sci *267*, 1751-1757.

Brown, W. L., Jr. (1973). A comparison of the Hylean and Congo-West African rain forest ant faunas. *In* Tropical forest ecosystems in Africa and South America: A comparative review. B. J. Meggers, Ayensu, E. S., Duckworth, W. D., ed. (Washington D.C.: Smithsonian Institution Press), pp. 161-185.

Buckley, T. R., Simon, C., and Chambers, G. K. (2001b). Exploring among-site rate variation models in a maximum likelihood framework using empirical data: Effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst. Biol. *50*, 67-86.

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

Buckley, T. R., Simon, C., Shimodaira, H., and Chambers, G. K. (2001a). Evaluating hypotheses on the origin and evolution of the New Zealand alpine cicadas (*Maoricicada*) using multiple-comparison tests of tree topology. Mol. Biol. Evol. *18*, 223-234.

Bull, J. J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L., and Waddell, P. J. (1993). Partitioning and combining data in phylogenetic analysis. Syst. Biol. *42*, 384-397.

Cagniant, H. (1979). La parthenogenese thelytoque et arrhenotoque chez la fourmi *Cataglyphis cursor* Fonsc. (Hymenopteres Formicidae). Cycle biologique en elevage des colonies avec reine et des colonies sans reine. Insectes Soc. *26*, 51-60.

Cameron, S. A., and Mardulyn, P. (2001). Multiple molecular data sets suggest independent origins of highly eusocial behavior in bees (Hymenoptera: Apinae). Syst. Biol. *50*, 194-214.

Cao, Y., Adachi, J., Janke, A., Paabo, S., and Hasegawa, M. (1994). Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: Instability of a tree based on a single gene. J. Mol. Evol. *39*, 519-527.

Carlin, N. F., Reeve, H. K., and Cover, S. P. (1993). Kin discrimination and division of labour among matrilines in the polygynous carpenter ant, *Camponotus planatus*. *In* Queen number and sociality in insects, L. Keller, ed. (Oxford: Oxford University Press), pp. 362-401.

Caterino, M. S., Cho, S., and Sperling, F. A. (2000). The current state of insect molecular systematics: A thriving Tower of Babel. Ann. Rev. Entomol. *45*, 1-54.

Caterino, M. S., Reed, R. D., Kuo, M. M., and Sperling, F. A. (2001). A partitioned likelihood analysis of swallowtail butterfly phylogeny (Lepidoptera:Papilionidae) Interaction of process partitions in phylogenetic analysis: An example from the swallowtail butterfly genus *Papilio*. Syst. Biol. *50*, 106-127.

Chapuisat, M. (1996). Characterization of microsatellite loci in *Formica lugubris B* and their variability in other ant species. Mol. Ecol. *5*, 599-601.

Chapuisat, M., Goudet, J. and Keller, L. (1997). Microsatellites reveal high population viscosity and limited dispersal in the ant *Formica paralugubris*. Evolution. *51*, 475-482.

Chapuisat, M., and Crozier, R. H. (2001). Low relatedness among cooperatively breeding workers of the greenhead ant *Rhytidoponera metallica*. J. Evol. Biol. *14*, 564-573.

Chapuisat, M., Painter, J. N., and Crozier, R. H. (2000). Microsatellite markers for *Rhytidoponera metallica* and other ponerine ants. Mol. Ecol. *9*, 2219-2221.

Chenuil, A., and McKey, D. B. (1996). Molecular phylogenetic study of a myrmecophyte symbiosis: Did *Leonardoxa*/ ant associations diversify via cospeciation? Mol. Phylogenet. Evol. *6*, 270-286.

Chiotis, M., Jermiin, L. S., and Crozier, R. H. (2000). A molecular framework for the phylogeny of the ant subfamily Dolichoderinae. Mol. Phylogenet. Evol. *17*, 108-116.

Chippindale, P. T., and Wiens, J. J. (1994). Weighting, partitioning, and combining characters in phylogenetic analysis. Syst. Biol. *43*, 278-287.

Cho, S. W., Mitchell, A., Regier, J. C., Mitter, C., Poole, R. W., Freidlander, T. P., and Zhao, S. W. (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.

Clary, D. O., and Wolstenholme, D. R. (1985). The mitochondrial DNA molecular of *Drosophila yakuba*: Nucleotide sequence, gene organization, and genetic code. J. Mol. Evol. *22*, 252-271.

Cole, B. J., and Wiemasz, D. C. (2000). Colony size and reproduction in the western harvester ant, *Pogonomyrmex occidentalis*. Insectes Soc. *47*, 249-255.

Collins, T. M., Wimberger, P. H., and Naylor, G. J. P. (1994). Compositional bias, character state bias, and character-state reconstruction using parsimony. Syst. Biol. *43*, 482-496.

Collura, R. V., Auerbach, M. R., and Stewart, C. B. (1996). A quick, direct method that can differentiate expressed mitochondrial genes from their nuclear pseudogenes. Curr Biol. *6*, 1337-1339.

Collura, R. V., and Stewart, C. B. (1995). Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. Nature. *378*, 485-489.

Cowling, R. M., Richardson, D. M., and Mustart, P. J. (1997). Fynbos. *In* Vegetation of southern Africa, D. R. RM Cowling, SM Pierce, ed. (Cambridge: Cambridge University Press), pp. 99-130.

Creighton, W. S. (1950). The ants of North America. Bull. Mus. Comp. Zool. *104*, 1-585.

Crespi, B. J., Carmean, D. A., Mound, L. A., Worobey, M., and Morris, D. (1998). Phylogenetics of social behavior in Australian gall-forming thrips: Evidence from mitochondrial DNA sequence, adult morphology and behavior, and gall morphology. Mol. Phylogenet. Evol. *9*, 163-180.

Crozier, R. H., and Crozier, Y. C. (1993). The mitochondrial genome of the honeybee *Apis mellifera:* Complete sequence and genome organization. Genetics. *133*, 97-117.

Crozier, R. H., Dobric, N., Imai, H. T., Graur, D., Cornuet, J.-M., and Taylor, R. W. (1995). Mitochondrial-DNA sequence evidence on the phylogeny of Australian jack-jumper ants of the *Myrmecia pilosula* complex. Mol. Phylogenet. Evol. *4*, 20-30.

Crozier, R. H., Kaufmann, B., Carew, M. E., and Crozier, Y. C. (1999). Mutability of microsatellites developed for the ant *Camponotus consobrinus*. Mol. Ecol. *8*, 271-276.

Crozier, R. H., and Pamilo, P. (1996). Evolution of social insect colonies: Sex allocation and kin selection, R. M. May and P. H. Harvey, eds. (Oxford: Oxford University Press).

Cummings, M. P., Otto, S. P., and Wakeley, J. (1995). Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. *12*, 814-822.

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

Cunningham, C. W. (1997b). Can three incongruence tests predict when data should be combined? Mol. Biol. Evol. *14*, 733-740.

Cunningham, C. W., Zhu, H., and Hillis, D. M. (1998). Best-fit maximum likelihood models for phylogenetic inference: Empirical tests with known phylogenies. Evolution. *52*, 978 - 987.

Curtis, B. A. (1985). Observations on the natural history and behaviour of the dune ant, *Camponotus detritus* Emery, in the central Namib Desert. Madoqua *14*, 279-289.

Dalecky, A., Debout, G., Mondor, G., Rasplus, J. Y., and Estoup, A. (2002). PCR primers for polymorphic microsatellite loci in the facultatively polygynous plant-ant *Petalomyrmex phylax* (Formicidae). Mol. Ecol. Notes *2*, 404-407.

Darlu, P., and Lecointre, G. (2002). When does the incongruence length difference test fail? Mol. Biol. Evol. *19*, 432-437.

Darwin, C. (1859). Chapter VIII: Instinct. *In* The origin of the species by means of natural selection. (Chicago: William Benton).

Debout, G., Dalecky, A., Mondor, G., Estoup, A., and Rasplus, J. Y. (2002). Isolation and characterisation of polymorphic microsatellites in the tropical plant-ant *Cataulacus mckeyi* (Formicidae: Myrmicinae). Mol. Ecol. Notes *2*, 459-461.

de Queiroz, A. (1993). For consensus (sometimes). Syst. Biol. *42*, 368-372.

de Queiroz, A., Donoghue, M. J., and Kim, J. (1995). Separate versus combined analysis of phylogenetic evidence. Ann. Rev. Ecol. Syst. *26*, 657-681.

de Queiroz, A., Lawson, R., and Lemos-Espinal, J. A. (2002). Phylogenetic relationships of North American garter snakes (*Thamnophis*) based on four mitochondrial genes: How much DNA sequence is enough? Mol. Phylogenet. Evol. *22*, 315-329.

DeSalle, R., and Brower, A. V. (1997). Process partitions, congruence, and the independence of characters: Inferring relationships among closely related Hawaiian *Drosophila* from multiple gene regions. Syst. Biol. *46*, 751-764.

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

Dorow, W. H. O. (1995). Revision of the ant genus *Polyrhachis* Smith, 1857 (Hymenoptera: Formicidae: Formicinae) on subgenus level with keys, checklist of species and bibliography. Cour. Forschungsinst. Senckenb. *185*, 1-113.

Doums, C. (1999). Characterization of microsatellite loci in the queenless Ponerine ant *Diacamma cyaneiventre*. Mol. Ecol. *8*, 1957-1959.

Dowton, M., and Austin, A. D. (1997). The evolution of strand-specific compositional bias. A case study in the hymenopteran mitochondrial 16S rRNA gene. Mol. Biol. Evol. *14*, 109-112.

Dowton, M., and Austin, A. D. (1998). Phylogenetic relationships among the microgastroid wasps (Hymenoptera: Braconidae): Combined analysis of 16S and 28S rDNA genes and morphological data. Mol. Phylogenet. Evol. *10*, 354-366.

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

Dowton, M., Belshaw, R., Austin, A. D., and Quicke, D. L. (2002). Simultaneous molecular and morphological analysis of braconid relationships (Insecta: Hymenoptera: Braconidae) Indicates independent mt-tRNA gene inversions within a single wasp family. J. Mol. Evol. *54*, 210-226.

Dowton, M., and Campbell, N. J. (2001). Intramitochondrial recombination - is it why some mitochondrial genes sleep around? Trends Ecol. Evol. *16*, 269-271.

Emerson, B. C., and Wallis, G. P. (1995). Phylogenetic relationships of the *Prodontria* (Coleoptera; Scarabaeidae; subfamily Melolonthinae), derived from sequence variation in the mitochondrial cytochrome oxidase II gene. Mol. Phylogenet. Evol. *4*, 433-447.

Emery, C. (1925). Hymenoptera. Fam. Formicidae. Subfam. Formicinae. Genera Insectorum *183*, 1-302.

Emery, C. (1920). Le genre *Camponotus* Mayr. Nouvel essai de la subdivision en sous-genres. Rev. Zool. Afr. (Bruss.) *8*, 229-260.

Emery, C. (1896). Saggio di un catalogo sistematico dei generi *Camponotus*, *Polyrhachis* e affini. Mem. R. Accad. Sci. Ist. Bologna *5*, 363-382.

Estoup, A., Solignac, M., and Cornuet, J. (1994). Precise assessment of the number of patrilines and of genetic relatedness in honeybee colonies. Proc. R. Soc. Lond. B. *258*, 1-7.

Evans, J. D. (1993). Parentage analyses in ant colonies using simple sequence repeat loci. Mol. Ecol. *2*, 393-397.

Evans, J. D. (1995). Relatedness threshold for the production of female sexuals in colonies of a polygynous ant, *Myrmica tahoensis*, as revealed by microsatellite DNA analysis. Proc. Natl. Acad. Sci. U. S. A. *92*, 6514-6517.

Eyre-Walker, A. (1998). Problems with parsimony in sequences of biased base composition. J. Mol. Evol. *47*, 686-690.

Fang, Q. Q., Cho, S., Regier, J. C., Mitter, C., Matthews, M., Poole, R. W., Friedlander, T. P., and 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., and Poole, R. W. (2000). Phylogenetic utility of the nuclear gene dopa decarboxylase in noctuoid moths (Insecta: Lepidoptera: noctuoidea). Mol. Phylogenet. Evol. *15*, 473-486.

Farris, J. S. (1970). Methods for computing Wagner trees. Syst. Zool. *19*, 83-92.

Farris, J. S. (1969). A successive approximations approach to character weighting. Syst. Zool. *18*, 374-385.

Farris, J. S., Kallersjo, M., Kluge, A. G., and Bult, C. (1995). Constructing a significance test for incongruence. Syst. Biol. *44*, 570-572.

Farris, J. S., Kallersjo, M., Kluge, A. G., and Bult, C. (1994). Testing significance of incongruence. Cladistics. *10*, 315-319.

Felsenstein, J. (1978). Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. *27*, 401-410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. *17*, 368-376.

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. *266*, 418-427.

Fitch, W. M. (1971). Toward defining the course of evolution: Minimal change for a specific tree topology. Syst. Zool. *20*, 406-416.

Fjerdingstad, E. J., and Boomsma, J. J. (2000). Queen mating frequency and relatedness in young *Atta sexdens* colonies. Insectes Soc. *47*, 354-336.

Fjerdingstad, E. J., Boomsma, J. J., and Thoren, P. (1998). Multiple paternity in the leafcutter ant *Atta colombica* - a microsatellite DNA study. Heredity. *80*, 118-126.

Flook, P. K., Klee, S., and Rowell, C. H. F. (1999). Combined molecular phylogenetic analysis of the Orthoptera (Arthropoda, Insecta) and implications for their higher systematics. Syst. Biol. *48*, 233-253.

Foitzik, S., Haberl, M., Gadau, J., and Heinze, J. (1997). Mating frequency of *Leptothorax nylanderi* ant queens determined by microsatellite analysis. Insectes Soc. *44*, 219-227.

Foitzik, S., and Heinze, J. (1998). Nest site limitation and colony takeover in the ant *Leptothorax nylanderi*. Behav. Ecol. *9*, 367-375.

Foitzik, S., and Herbers, J. M. (2001). Colony structure of a slavemaking ant. I. Intracolony relatedness, worker reproduction, and polydomy. Evolution. *55*, 307-315.

Forel, A. (1912). Formicides néotropiques. Part VI. 5me sous-famille Camponotinae Forel. Mém. Soc. Entomol. Belg. *20*, 59-92.

Forel, A. (1914). Le genre *Camponotus* Mayr et les genres voisins. Rev. Suisse Zool. *22*, 257-276.

Foster, K. R., and Ratnieks, F. L. (2000). Facultative worker policing in a wasp. Nature. *407*, 692-693.

Fournier, D., Aron, S., and Milinkovitch, M. C. (2002). Investigation of the population genetic structure and mating system in the ant *Pheidole pallidula*. Mol. Ecol. *11*, 1805-1814.

Fraser, V. S., Kaufmann, B., Oldroyd, B. P., and Crozier, R. H. (2000). Genetic influence on caste in the ant *Camponotus consobrinus*. Behav. Ecol. Sociobiol. *47*, 188-194.

Frati, F., Simon, C., Sullivan, J., and Swofford, D. L. (1997). Evolution of the 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., and Fang, Q. Q. (1998). Two nuclear genes yield concordant relationships within Attacini (Lepidoptera: Saturniidae). Mol. Phylogenet. Evol. *9*, 131-140.

Funk, D. J. (1999). Molecular systematics of cytochrome oxidase I and 16S from *Neochlamisus* leaf beetles and the importance of sampling. Mol. Biol. Evol. *16*, 67-82.

Funk, D. J., Futuyma, D. J., Orti, G., and Meyer, A. (1995). Mitochondrial DNA sequences and multiple data sets: A phylogenetic study of phytophagous beetles (Chrysomelidae: Ophraella). Mol. Biol. Evol. *12*, 627-640.

Gadau, J., Brady, S., and Ward, P. (1999). Systematics, distribution, and ecology of an endemic California *Camponotus quercicola* (Hymenoptera:Formicidae). Ann. Ent.Soc. Am. *92*, 514-522.

Gadau, J., Gertsch, P., Henize, J., Pamilo, P., and Hölldobler, B. (1998). Oligogyny by unrelated queens in the carpenter ant *Camponotus ligniperdus*. Behav. Ecol. Sociobiol. *44*, 23-33.

Gadau, J., Heinze, J., Hölldobler, B., and Schmid, M. (1996). Population and colony structure of the carpenter ant *Camponotus floridanus*. Mol. Ecol. *5*, 785-792.

Gaut, B. S., and Lewis, P. O. (1995). Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. *12*, 152-162.

Gertsch, P., Pamilo, P., and Varvio, S. L. (1995). Microsatellites reveal high genetic diversity within colonies of *Camponotus* ants. Mol. Ecol. *4*, 257-260.

Giraud, T., Blatrix, R., Solignac, M., and Jaisson, P. (1999). Polymorphic microsatellite DNA markers in the ant *Gnamptogenys striatula*. Mol. Ecol. *8*, 2143-2145.

Glenn, T. C., Stephan, W., Dessauer, H. C., and Braun, M. J. (1996). Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability. Mol. Biol. Evol. *13*, 1151-1154.

Goldman, N. (1993). Statistical tests of models of DNA substitution. J. Mol. Evol. *36*, 182-198.

Goldman, N., Anderson, J. P., and Rodrigo, A. G. (2000). Likelihood-based tests of topologies in phylogenetics. Syst Biol *49*, 652-70.

Goodisman, M. A., and Crozier, R. H. (2002). Population and colony genetic structure of the primitive termite *Mastotermes darwiniensis*. Evolution. *56*, 70-83.

Goodisman, M. A. D., Matthews, R. W., Spradbery, J. P., Carew, M. E., and Crozier, R. H. (2001). Reproduction and recruitment in perennial colonies of the introduced wasp *Vespula germanica*. J. Hered. *92*, 346-349.

Gopinath, A., Gadagkar, R., and Rao, M. R. S. (2001). Identification of polymorphic microsatellite loci in the queenless ponerine ant *Diacamma ceylonense*. Mol. Ecol. Notes. *1*, 126-127.

Goropashnaya, A. V., Seppa, P., and Pamilo, P. (2001). Social and genetic characteristics of geographically isolated populations in the ant *Formica cinerea*. Mol. Ecol. *10*, 2807-2818.

Grasso, D. A., Wenseleers, T., Mori, A., Le Moli, F., and Billen, J. (2000). Thelytokous worker reproduction and lack of *Wolbachia* infection in the harvesting ant *Messor capitatus*. Ethol. Ecol. Evol. *12*, 309-314.

Gray, M. W., Burger, G., and Lang, B. F. (2001). The origin and early evolution of mitochondria. Genome Biol *2*, REVIEWS 1018.

Graybeal, A. (1994). Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. Syst. Biol. *43*, 174-193.

Graziewicz, M. A., Day, B. J., and Copeland, W. C. (2002). The mitochondrial DNA polymerase as a target of oxidative damage. Nucl. Acids. Res. *30*, 2817-2824.

Green, C. L., and Oldroyd, B. P. (2002). Queen mating frequency and maternity of males in the stingless bee *Trigona carbonaria* Smith. Insectes Soc. *49*, 196-202.

Gu, X., Fu, Y. X., and Li, W. H. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. *12*, 546-557.

Gyllenstrand, N., Gertsch, P. J., and Pamilo, P. (2002). Polymorphic microsatellite DNA markers in the ant *Formica exsecta*. Mol. Ecol. Notes. *2*, 67-69.

Hamaguchi, K., Ito, Y., and Takenaka, O. (1993). GT dinucleotide repeat polymorphisms in a polygynous ant, *Leptothorax spinosior* and their use for measurement of relatedness. Naturwissenschaften. *80*, 179-181.

Hamilton, W. D. (1972). Altruism and related phenomena, mainly in social insects. Ann. Rev. Ecol. Syst *3*, 192-232.

Hamilton, W. D. (1963). The evolution of altruistic behaviour. Am. Nat. *97*, 354 - 356.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. J Theor Biol *7*, 17-52.

Hamilton, W. D. (1970). Selfish and spiteful behaviour in an evolutionary model. Nature. *228*, 1218-1220.

Hammond, R. L., Bourke, A. F., and Bruford, M. W. (2001). Mating frequency and mating system of the polygynous ant, *Leptothorax acervorum*. Mol. Ecol. *10*, 2719-2728.

Hartl, D. L., and Clark, A. G. (1997). Principles of population genetics. 3rd Edition. (Massachusetts: Sinauer Associates, Inc).

Hasegawa, E. (1995). Parental analysis using RAPD markers in the ant *Colobopsis nipponicus*: A test of RAPD markers for estimating reproductive structure within social insect colonies. Insectes Soc. *42*, 337-346.

Hedges, S. B., and Maxson, L. R. (1996). Re: Molecules and morphology in amniote phylogeny. Mol. Phylogenet. Evol. *6*, 312-314.

Hedrick, P. W., and Parker, J. D. (1997). Evolutionary genetics and genetic variation of haplodiploids and X-linked genes. Ann. Rev. Ecol. Syst. *28*, 55-83.

Heinze, J. and Hölldobler, B. (1995). Thelytokous parthenogenesis and dominance hierarchies in the ponerine ant, *Platythyrea punctata.* Naturwissenschaften. *82*, 40-41.

Heinze, J., and Keller, L. (2000). Alternative reproductive strategies: A queen perspective in ants. Trends Ecol. Evol. *15*, 508-512.

Heinze, J., Stratz, M., Pedersen, J. S., and Haberl, M. (2000). Microsatellite analysis suggests occasional worker reproduction in the monogynous ant *Crematogaster smithi.* Insectes Soc. *47*, 299-301.

Hendy, M. D., and Penny, D. (1989). A framework for the quantitative study of evolutionary trees. Syst. Zool. *38*, 297 -309.

Herbers, J. M., and Mouser, R. L. (1998). Microsatellite DNA markers reveal details of social structure in forest ants. Mol. Ecol. *7*, 299-306.

Hillis, D. M. (1995). Approaches for assessing phylogenetic accuracy. Syst. Biol. *44*, 3-16.

Hillis, D. M. (1987). Molecular versus morphological approaches to systematics. Ann. Rev. Ecol. Syst *18*, 23-42.

Hillis, D. M., Allard, M. W., and Miyamoto, M. M. (1993). Analysis of DNA sequence data: Phylogenetic inference. Methods Enzymol. *224*, 456-487.

Hillis, D. M., and Huelsenbeck, J. P. (1992). Signal, noise, and reliability in molecular phylogenetic analyses. J. Hered. *83*, 189-195.

Hillis, D. M., Huelsenbeck, J. P., and Cunningham, C. W. (1994). Application and accuracy of molecular phylogenies. Science. *264*, 671-677.

Hölldobler, B., and Engel-Siegel, H. (1985). On the metapleural gland of ants. Psyche. (Camb.) *91*, 201-224.

Hölldobler, B., and Bartz, S. H. (1985). Sociobiology and reproduction in ants. Fortschr. Zool. *31*, 237-257.

Hölldobler, B., and Wilson, E. O. (1990). The Ants (Berlin: Springer-Verlag).

Hölldobler, B., and Wilson, E. O. (1994). Journey to the ants: A story of scientific exploration. (Cambridge: Harvard University Press).

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandanavian Journal of Statistics. *6*, 65-70.

Howell, N. (1989). Evolutionary conservation of protein regions in the protonmotive cytochrome *b* and their possible roles in redox catalysis. J. Mol. Evol. *29*, 157-169.

Huelsenbeck, J. P., Swofford, D. L., Cunningham, C. W., Bull, J. J., and Waddell, P. J. (1994). Is character weighting a panacea for the problem of data heterogeneity in phylogenetic analysis? Syst. Biol. *43*, 288-291.

Huelsenbeck, J. P., and Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. Ann. Rev. Ecol. Syst. *28*, 437-66.

Huelsenbeck, J. P. and Bollback, J. P. (2001). Application of the likelihood function in phylogenetics. *In* Handbook of Statistical Genetics, D. J. Balding, M. Bishop and C. Cannings, eds. (London: John Wiley and Sons Ltd), pp. 339-415.

Huelsenbeck, J. P. (1995a). The robustness of two phylogenetic methods: Four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol. Biol. Evol. *12*, 843-849.

Huelsenbeck, J. P. (1995b). Performance of phylogenetic methods in simulation. Syst. Biol. *44*, 17-48.

Huelsenbeck, J. P., Bull, J. J., and Cunningham, C. W. (1996). Combining data in phylogenetic analysis. Trends Ecol. Evol. *11*, 152-158.

Huelsenbeck, J. P., and Bull, J. J. (1996). A likelihood ratio test to detect conflicting phylogenetic signal. Syst. Biol. *45*, 92-98.

Huelsenbeck, J. P., and Rannala, B. (1997). Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. Science. *276*, 227-232.

Huelsenbeck, J. P., and Ronquist, F. (2001). MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics. *17*, 754-755.

Huelsenbeck, J. P., Ronquist, F., and Hall, B. (2001). Manual for MrBayes: A program for the Bayesian inference of phylogeny.

Inc, S. (1998). SPSS Statistical software for windows (Chicago: SPSS Inc.).

Ingram, K. K., and Palumbi, S. R. (2002). Characterization of microsatellite loci for the Argentine ant, *Linepithema humile*, and their potential for analysis of colony structure in invading Hawaiian populations. Mol. Ecol. Notes. *2*, 94-95.

Irwin, D. M., Kocher, T. D., and Wilson, A. C. (1991). Evolution of the cytochrome *b* gene of mammals. J. Mol. Evol. *32*, 128-144.

Jame, P., David, P., and Viard, F. (1998). Microsatellites, transposable elements and the X chromosome. Mol. Biol. Evol. *15*, 28-34.

Jamet, P., and Lagoda, P. J. L. (1996). Microsatellites, from molecules to populations and back. Trends Ecol. Evol. *11*, 424-429.

Jermiin, L. S., Graur, D., Lowe, R. M., and Crozier, R. H. (1994). Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome *b* genes. J. Mol. Evol. *39*, 160-173.

Johnson, L. A., and Soltis, D. E. (1998). Assessing congruence: Empirical examples from molecular data. *In* Molecular systematics of plants II: DNA sequencing, D. E. Soltis, P. S. Soltis and J. J. Doyle, eds. (Dordrecht: Kluwer Academic), pp. 297-343.

Jow, H., Hudelot, C., Rattray, M., and Higgs, P. G. (2002). Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol. Biol. Evol. *19*, 1591-1601.

Keller, L. (1995). Social life: The paradox of multiple-queen colonies. Trends Ecol. Evol. *10*, 355-360.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. *16*, 111-120.

Kishino, H., and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. *29*, 170-179.

Kitching, I. J., Forey, P. L., Humphries, C. J., and Williams, D. (1998). Cladistics - theory and practice of parsimony analysis. In systematics association special volumes: Oxford University Press).

Kjer, K. M., Blahnik, R. J., and Holzenthal, R. W. (2001). Phylogeny of Trichoptera (caddisflies): Characterisation of signal and noise within multiple datasets. Syst. Biol. *50*, 781-816.

Kluge, A. G., and Farris, J. S. (1969). Quantitative phyletics and the evolution of anurans. Syst. Zool. *18*, 1-32.

Kluge, A. G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. *38*, 7-25.

Krieger, M. J. B., and Keller, L. (1999). Low polymorphism at 19 microsatellite loci in a French population of Argentine ants (*Linepithema humile*). Mol. Ecol. *8*, 1075-1092.

Krieger, M. J. B., and Keller, L. (1997). Polymorphism at dinucleotide microsatellite loci in fire ant *Solenopsis invicta* populations. Mol. Ecol. *6*, 997-999.

Kronauer, D. J. C., and Gadau, J. (2002). Isolation of polymorphic microsatellite markers in the new world honey ant *Myrmecocystus mimicus*. Mol. Ecol. Notes *2*, 540-541.

Kryger, P. (2001). The *capensis* pseudo-clone, a social parasite of African honey bees. *In* European Section IUSSI meeting (Berlin: International union for the study of social insects), pp. 208.

Kuhner, M. K., and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. *11*, 459-468.

Kumar, S., Tamura, K., Jakobsen, I. B., and Nei, M. (2001). MEGA2: Molecular Evolutionary Genetics Analysis software. Bioinformatics. *17*, 1244-1245.

Lahiri, D. K., and Schnabel, B. (1993). DNA isolation by a rapid method from human blood samples: Effects of MgCl2, EDTA, storage time, and temperature on DNA yield and quality. Biochem. Genet. *31*, 321-328.

Larget, B., and Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. *16*, 750-759.

Leache, A. D., and Reeder, T. W. (2002). Molecular systematics of the Eastern Fence Lizard (*Sceloporus undulatus*): A comparison of parsimony, likelihood, and Bayesian approaches. Syst. Biol. *51*, 44-68.

Lee, M. S. (2001). Uninformative characters and apparent conflict between molecules and morphology. Mol. Biol. Evol. *18*, 676-680.

Lee, P. L., Clayton, D. H., Griffiths, R., and Page, R. D. (1996). Does behavior reflect phylogeny in swiftlets (Aves: Apodidae)? A test using cytochrome *b* mitochondrial DNA sequences. Proc. Natl. Acad. Sci. U. S. A. *93*, 7091-7096.

Legendre, P., and Legendre, L. (1998). Numerical ecology (Amsterdam: Elsevier).

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. *50*, 913-925.

Lewis, P. O. (1998). Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. *In* Molecular systematics of plants II: DNA sequencing, D. E. Soltis, P. S. Soltis and J. J. Doyle, eds. (Dordrecht: Kluwer Academic), pp. 132-163.

Lewis, P. O. (2001). Phylogenetic systematics turns over a new leaf. Trends Ecol. Evol. *16*, 30-37.

Lio, P., and Goldman, N. (1998). Models of molecular evolution and phylogeny. Genome Res. *8*, 1233-1244.

Liu, H., and Beckenbach, A. T. (1992). Evolution of the mitochondrial cytochrome oxidase II gene among 10 orders of insects. Mol. Phylogenet. Evol. *1*, 41-52.

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

Maddison, D. R., and Maddison, W. P. (2000). MacClade 4 (Sunderland, Massachusetts: Sinauer Associates).

Maddison, D. R., Ruvolo, M., and Swofford, D. L. (1992). Geographic origins of human mitochondrial DNA: Phylogenetic evidence from control region sequences. Syst. Biol. *41*, 111-124.

Maddison, D. R., Swofford, D. L., and Maddison, W. P. (1997). NEXUS: An extensible file format for systematic information. Syst. Biol. *46*, 590-621.

Maddison, W. P. (1996). Molecular approaches and the growth of phylogenetic biology. In Molecular Zoology: Advances, strategies and protocols, J. D. Ferraris and S. R. Palumbi, eds. (New York: Wiley-Liss), pp. 47-63.

Mardulyn, P., and 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.

Mardulyn, P., and Whitfield, J. B. (1999). Phylogenetic signal in the COI, 16S, and 28S genes for inferring relationships among genera of Microgastrinae (Hymenoptera; Braconidae): Evidence of a high diversification rate in this group of parasitoids. Mol. Phylogenet. Evol. *12*, 282-294.

Martin, A. P. (1995). Mitochondrial DNA sequence evolution in sharks: Rates, patterns, and phylogenetic inferences. Mol. Biol. Evol. *12*, 1114-1123.

Mason-Gamer, R., and Kellogg, E. A. (1996). Testing for phylogenetic conflict among molecular data sets in the tribe Triticeae (Gramineae). Syst. Biol. *45*, 524-545.

Mau, B., and Newton, M. (1997). Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. Journal of Computational and Graphical Statistics. *6*, 122-131.

Mau, B., Newton, M. A., and Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics. *55*, 1-12.

Maynard Smith, J. (1964). Group selection and kin selection. Nature. *201*, 1145-1147.

McCauley, D. E., and Goff, P. W. (1998). Intrademic genetic structure and natural selection in insects. *In* Genetic structure and local adaptation in natural insect populations, S. Mopper and S. Y. Strauss, eds. (New York: Chapman and Hall), pp. 181-204.

McCracken, K. G., Harshman, J., McClellan, D. A., and Afton, A. D. (1999). Data set incongruence and correlated character evolution: an example of functional convergence in the hind-limbs of stifftail diving ducks. Syst. Biol. *48*, 683-714.

Meyer, A. (1994). Shortcomings of the cytochrome *b* gene as a molecular marker. Trends Ecol. Evol. *9*, 278-280.

Michod, R. E., and Hamilton, W. D. (1980). Coefficients of relatedness in sociobiology. Nature. *288*, 694-697.

Mickevich, M. F., and Farris, J. S. (1981). The implications of congruence in *Menidia*. Syst. Zool. *30*, 351-370.

Mitchell, A., Cho, S., Regier, J. C., Mitter, C., Poole, R. W., and Matthews, M. (1997). Phylogenetic utility of elongation factor-1 alpha in noctuoidea (Insecta: Lepidoptera): the limits of synonymous substitution. Mol. Biol. Evol. *14*, 381-90.

Mitchell, A., Mitter, C., and Regier, J. C. (2000). More taxa or more characters revisited: Combining data from nuclear protein-encoding genes for phylogenetic analyses of Noctuoidea (Insect: Lepidoptera). Syst. Biol. *49*, 202-224.

Miyamoto, M. (1985). Consensus cladograms and general classifications. Cladistics. *1*, 186-189.

Miyamoto, M. M., and Fitch, W. M. (1995). Testing species phylogenies and phylogenetic methods with congruence. Syst. Biol. *44*, 64-76.

Monnin, T., and Ratnieks, F. L. W. (2001). Policing in queenless ponerine ants. Behav. Ecol. Sociobiol. *50*, 97-108.

Moore, W. S. (1995). Inferring phylogenies from mtDNA variation - mitochondrial-gene tree versus nuclear-gene trees. Evolution. *49*, 718-726.

Moritz, C., and Hillis, D. M. (1996). Molecular systematics: Context and controversies. *In* molecular systematics, C. M. a. B. K. M. D.M. Hillis, ed. (Massachussetts: Sunderland), pp. 1 -13.

Muraji, M., and Nakahara, S. (2001). Phylogenetic relationships among fruit flies, *Bactrocera* (Diptera, Tephritidae), based on mitochondrial rDNA sequences. Insect Mol.Biol. *10*, 549-559.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics *89*, 583-590.

Nei, M., and Roychoudhury, A. K. (1974). Sampling variances of heterozygosity and genetic distance. Genetics. *76*, 379-390.

Nixon, K. C., and Carpenter, J. M. (1993). On outgroups. Cladistics. *9*, 413-426.

Nyman, T., Widmer, A., and Roininen, H. (2000). Evolution of gall morphology and host-plant relationships in willow-feeding sawflies (Hymenoptera: Tenthredinidae). Evolution. *54*, 526-533.

Orti, G., and Meyer, A. (1997). The radiation of characiform fishes and the limits of resolution of mitochondrial ribosomal DNA sequences. Syst. Biol. *46*, 75-100.

Ortius-Lechner, D., Gertsch, P. J., and Boomsma, J. J. (2000). Variable microsatellite loci for the leafcutter ant *Acromyrmex echinatior* and their applicability to related species. Mol. Ecol. *9*, 114-116.

Palmer, K. A., Oldroyd, B. P., Quezada-Euan, J. J., Paxton, R. J., and May-Itza W. D. J. (2002). Paternity frequency and maternity of males in some stingless bee species. Mol. Ecol. *11*, 2107-2113.

Parker, J. D., and Rissing, S. W. (2002). Molecular evidence for the origin of workerless social parasites in the ant genus *Pogonomyrmex*. Evolution. *56*, 2017 - 2028.

Paxton, R. J., Thorén, P. A., Estoup, A., and Tengö, J. (2001). Queen–worker conflict over male production and the sex ratio in a facultatively polyandrous bumblebee, *Bombus hypnorum*: The consequences of nest usurpation. Mol. Ecol. *10*, 2489-2498.

Pederson, J. S., and Boomsma, J. J. (1999). Multiple paternity in social Hymenoptera: estimating the effective mate number in single-double mating population. Mol. Ecol. *8*, 577-587.

Pennington, R. T. (1996). Molecular and morphological data provide phylogenetic resolution at different hierarchical levels in *Andira*. Syst. Biol. *45*, 496-515.

Perna, N. T., and Kocher, T. D. (1995). Unequal base frequencies and estimation of substitution rates. Mol. Biol. Evol. *12*, 359-361.

Peters, J. M. (1997). Microsatellite loci for *Pseudomyrmex pallidus* (Hymenoptera: Formicidae). Mol. Ecol. *6*, 887-888.

Poe, S., and Wiens, J. J. (2000). Character selection and the methodology of morphological phylogenetics. *In* Phylogenetic Analysis of Morphological Data, J. J. Wiens, ed. (Washington D.C.: Smithsonian Institution Press), pp. 20-35.

Posada, D., and Crandall, K. A. (1998). MODELTEST: Testing the model of DNA substitution. Bioinformatics. *14*, 817-818.

Posada, D., and Crandall, K. A. (2001). Selecting the best-fit model of nucleotide substitution. Sys. Biol. *50*, 580-601.

Prager, E. M., and Wilson, A. C. (1988). Ancient origin of lactalbumin from lysozyme: Analysis of DNA and amino acid sequences. J. Mol. Evol. *27*, 326-335.

Prendini, L. (2000). Phylogeny and classification of the superfamily Scorpionoidea Latreille 1802 (Chelicerata, Scorpiones): An exemplar approach. Cladistics.*16*, 1-78.

Queller, D. C., and Goodnight, K. F. (1989). Estimating relatedness using genetic markers. Evolution. *43*, 258-275.

Queller, D. C., and Strassmann, J. E. (1998). Kin selection and social insects. BioScience. *48*, 165-175.

Queller, D. C., Strassmann, J. E., and Hughes, C. R. (1993). Microsatellites and kinship. Trends Ecol. Evol. *8*, 285- 288.

Queller, D. C., Zacchi, F., Cervo, R., Turillazzi, S., Henshaw, M. T., Santorelli, L. A., and Strassmann, J. E. (2000). Unrelated helpers in a social insect. Nature. *405*, 784-787.

Rannala, B., and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. *43*, 304-311.

Ratnieks, F. L. W. (1988). Reproductive harmony via mutual policing by workers in eusocial hymenoptera. Am. Nat. *132*, 217-236.

Ratnieks, F. L. W., and Visscher, P. K. (1989). Worker policing in the honeybee. Nature. *342*, 796-797.

Ratnieks, R. L. W., and Reeve, H. K. (1992). Conflict in single queen hymenopteran societies: The structure of conflict and processes that reduce conflict in advance eusocial insects. J. Theor. Biol. *158*, 33-65.

Ravary, F., and Jaisson, P. (2002). The reproductive cycle of thelytokous colonies of *Cerapachys biroi* Forel (Formicidae, Cerapachyinae). Insectes Soc. *49*, 114-119.

Raymond, M., and Rousset, F. (1995). GENEPOP: Population genetic software for exact test and ecumenism. J. Hered. *86*, 248-249.

Reed, R. D., and 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.

Remsen, J., and DeSalle, R. (1998). Character congruence of multiple data partitions and the origin of the Hawaiian Drosophilidae. Mol. Phylogenet. Evol. *9*, 225-235.

Rice, W. R. (1989). Analyzing tables of statistical test. Evolution. *43*, 223-225.

Robertson, H. M., and Zachariades, C. (1997). Revision of the *Camponotus fulvopilosus* (De Geer) species group (Hymenoptera: Formicidae). African Entomology. *5*, 1-18.

Rodrigo, A. G., Kelly-Borges, M., Berquist, P. R., and Berquist, P. L. (1993). A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. New Zealand J. Bot. *31*, 257-268.

Rokas, A., Nylander, J. A., Ronquist, F., and Stone, G. N. (2002). A maximum-likelihood analysis of eight phylogenetic markers in gallwasps (Hymenoptera: Cynipidae): Implications for insect phylogenetic studies. Mol. Phylogenet. Evol. *22*, 206-219.

Ross, K. G. (1993). The breeding system of the fire ant *Solenopsis invicta*: Effects on colony genetic structure. Am. Nat. *141*, 554-576.

Ross, K. G., Shoemaker, D. D., Krieger, M. J., DeHeer, C. J., and Keller, L. (1999). Assessing genetic structure with multiple classes of molecular markers: A case study involving the introduced fire ant Solenopsis invicta. Mol. Biol. Evol. *16*, 525-543.

Rzhetsky, A., and Nei, M. (1992). A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. *9*, 945-967.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). Molecular Cloning: A laboratory manual (New York, Cold Spring Harbor Laboratory Press).

Sameshima, S., Hasegawa, E., Kitade, O., Minaka, N., and Matsumoto, T. (1999). Phylogenetic comparison of endosymbionts with their host ants based on molecular evidence. Zoological Science. *16*, 993-1000.

Sanada, S., Satoh, T., and Obara, Y. (1997). Trophallaxis and genetic relationships among workers in colonies of the polygynous ant *Camponotus yamaokai*. Ethol. Ecol. Evol. *9*, 149-158.

Sanderson, M. J., and Donoghue, M. J. (1989). Patterns of variation in levels of homoplasy. Evolution. *43*, 1781-1790.

Sanetra, M., and Crozier, R. H. (2000). Characterization of microsatellite loci in the primitive ant *Nothomyrmecia macrops* Clark. Mol. Ecol. *9*, 2169-2170.

Sanetra, M., and Crozier, R. H. (2001). Polyandry and colony genetic structure in the primitive ant *Nothomyrmecia macrops*. J. Evol. Biol. *14*, 368-378.

Santschi, F. (1921). Retouches aux sous-genres de *Camponotus*. Ann. Soc. Entomol. Belg. *61*, 310-312.

Satoh, T., Masuko, K., and Matsumoto, T. (1997). Colony genetic structure in the mono-and polygynous sibling species of the ants *Camponotus nawai* and *Camponotus yamaokai*: DNA fingerprint analysis. Ecological Research. *12*, 71-76.

Sauer, C., Stackebrandt, E., Gadau, J., Hölldobler, B., and Gross, R. (2000). Systematic relationships and cospeciation of bacterial endosymbionts and their carpenter ant host species: Proposal of the new taxon *Candidatus blochmannia* gen. nov. Int J Syst Evol Microbiol. *50*, 1877-1886.

Sauer, P., Muller, M., and Kang, J. (1998). Quantitation of DNA. *In* QIAGEN News.

Schilder, K., Heinze, J., Gross, R., and Hölldobler, B. (1999). Microsatellites reveal clonal structure of populations of the thelytokous ant *Platythyrea punctata* (F. Smith) (Hymenoptera; Formicidae). Mol. Ecol. *8*, 1497-1507.

Seppa, P. (1994). Sociogenetic organization of the ants *Myrmica ruginodis* and *Myrmica lobicornis*: Number, relatedness and longevity of reproducing individuals. J. Evol. Biol. *7*, 71-95.

Seppa, P., and Gertsch, P. (1996). Genetic relatedness in the ant *Camponotus herculeanus*. A comparison of estimates from allozyme and DNA microsatellite data. Insectes Soc. *43*, 235-246.

Sharp, P. M., and Li, W. H. (1989). On the rate of DNA sequence evolution in *Drosophila*. J. Mol. Evol. *28*, 398-402.

Shattuck, S. O. (1999). Australian ants: Their biology and identification (Collingwood: CSIRO publishing).

Shimodaira, H., and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. *16*, 1114-1116.

Shoemaker, D. D., Costa, J. T., and Ross, K. G. (1992). Estimates of heterozygosity in two social insects using a large number of electrophoretic markers. Heredity. *69*, 573-582.

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

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H., and 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.

Skaife, S. H. (1961). The study of ants (London: Longmans).

Slowinksi, J. B., and Page, R. D. M. (1999). How should species phylogenies be inferred from sequence data? Syst. Biol. *48*, 814-825.

Sokal, R., R., and Rohlf, F. J. (1995). Biometry (New York: Freeman).

Sorenson, M. D. (1999). TreeRot version 2 (Boston: Boston University).

Sorenson, M. D., and Quinn, T. W. (1998). Numts: A challenge for avian systematics and population biology. The Auk. *115*, 214-221.

Soto-Adames, F., Robertson, H. M., and Berlocher, S. H. (1994). Phylogenetic utility of partial DNA sequences of G6P dehydrogenase at different taxonomic levels in Hexapoda with emphasis on Diptera. Ann. Entomol. Soc. Amer. *87*, 723-736.

Soucy, S. L., and Danforth, B. N. (2002). Phylogeography of the socially polymorphic sweat bee *Halictus rubicundus* (Hymenoptera: Halictidae). Evolution. *56*, 330-341.

Spicer, G. S. (1995). Phylogenetic utility of the mitochondrial cytochrome oxidase gene: Molecular evolution of the *Drosophila buzzatii* species complex. J. Mol. Evol. *41*, 749-759.

SPSS, I. (1998). SPSS Statistical software for windows (Chicago: SPSS Inc.).

Stanger-Hall, K., and Cunningham, C. W. (1998). Support for a monophyletic Lemuriformes: Overcoming incongruence between data partitions. Mol. Biol. Evol. *15*, 1572-1577.

States, D. J., Gish, W., and Altschul, S. F. (1991). Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. Methods. *3*, 66-70.

Steel, M., and Penny, D. (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. *17*, 839-850.

Stone, G. N., and Cook, J. M. (1998). The structure of cynipid oak galls: Patterns in the evolution of an extended phenotype. Proc. Roy. Soc. London Series B. *265*, 979-988.

Stone, G. N., Atkinson, R., Rokas, A., Csoka, G., and Nieves-Aldrey, J. L. (2001). Differential success in northwards range expansion between ecotypes of the marble gallwasp *Andricus kollari* I: A tale of two lifecycles. Mol. Ecol. *10*, 761-778.

Sullivan, J. (1996). Combining data with different distributions of among-site rate variation. Syst. Biol. *45*, 375-380.

Sullivan, J., Holsinger, K. E., and Simon, C. (1995). Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmodontine rodents. Mol. Biol. Evol. *12*, 988-1001.

Sullivan, J., Holsinger, K. E., and Simon, C. (1996). The effect of topology on estimates of among-site rate variation. J. Mol. Evol. *42*, 308-312.

Sullivan, J., Markert, J. A., and Kilpatrick, C. W. (1997). Phylogeography and molecular systematics of the *Peromyscus aztecus* species group (Rodentia: Muridae) inferred using parsimony and likelihood. Syst. Biol. *46*, 426-440.

Sullivan, J., and Swofford, D. L. (2001). Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. *50*, 723-729.

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

Sundstrom, L. (1994). Sex ratio bias, relatedness asymmetry and queen mating frequency in ants. Nature. *367*, 266-268.

Sundstrom, L., Chapuisat, M., and Keller, L. (1996). Conditional manipulation of sex ratios by ant workers: A test of kin selection theory. Science. *274*, 993-995.

Sunnucks, P., and Hales, D. 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.

Swofford, D. L. (1998). PAUP*: Phylogenetic analyses using parsimony (and other methods), version 4.0 (Sunderland: Sinauer Associates).

Swofford, D. L. (1991). When are phylogeny estimates from molecular and morphological data incongruent? *In* Phylogenetic analysis of DNA sequences, M. M. Miyamoto and J. Cracraft, eds. (New York: Oxford University Press).

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. *In* Molecular systematics, D.M. Hillis, C. Moritz and B. K. Maple, eds. (Massachussetts: Sunderland), pp. 407-514.

Tamura, K. (1992). The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. Mol. Biol. Evol. *9*, 814-825.

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. *10*, 512-526.

Tek Tay, W. T., and Crozier, R. H. (2000). Microsatellite analysis of gamergate relatedness of the queenless ponerine ant *Rhytidoponera* sp. 12. Insectes Soc. *47*, 188-192.

Tek Tay, W., Cook, J. M., Rowe, D. J., and Crozier, R. H. (1997). Migration between nests in the Australian arid-zone Rhytidoponera sp.12 revealed by DGGE analyses of mitochondrial DNA. Mol. Ecol. *6*, 403-411.

Templeton, A. R. (1983). Phylogenetic inference from restriction site endonuclease cleavage sites maps with particular reference to humans and apes. Evolution. *37*, 221-244.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. *25*, 4876-4882.

Thoren, P. A., Paxton, R. J., and Estoup, A. (1995). Unusually high frequency of (CT)n and (GT)n microsatellite loci in a yellowjacket wasp, *Vespula rufa* (L.) (Hymenoptera: Vespidae). Insect Mol. Biol. *4*, 141-148.

Thornton, J. W., and DeSalle, R. (2000). A new method to localize and test the significance of incongruence: detecting domain shuffling in the nuclear receptor superfamily. Syst. Biol. *49*, 183-201.

Toth, G., Gaspari, Z., and Jurka, J. (2000). Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Res. *10*, 967-981.

Trivers, R. L., and Hare, H. (1976). Haplodiploidy and the evolution of social insects. Science. *191*, 249-263.

Tschinkel, W. R. (1996). A newly-discovered mode of colony founding among fire ants. Insectes Soc. *43*, 267-276.

Tsuji, K. (1988). Obligate parthenogenesis and reproductive division of labor in the Japanese queenless ant *Pristomyrmex pungens* - comparison of intranidal and extranidal workers. Behav. Ecol. Sociobiol. *23*, 247-255.

Tsuji, K., and Yamauchi, K. (1995). Production of females by parthenogenesis in the ant, *Cerapachys biroi*. Insectes. Soc. *42*, 333-336.

Tsutsui, N. D., Suarez, A. V., Holway, D. A., and Case, T. J. (2000). Reduced genetic variation and the success of an invasive species. Proc. Natl. Acad. Sci. U. S. A. *97*, 5948-5953.

Villesen, P., Gertsch, P. J., and Boomsma, J. J. (2002). Microsatellite primers for fungus-growing ants. Mol. Ecol. Notes. *2*, 320-322.

Villesen, P., Gertsch, P. J., Frydenberg, J., Mueller, U. G., and Boomsma, J. J. (1999). Evolutionary transition from single to multiple mating in fungus-growing ants. Mol. Ecol. *8*, 1819-1825.

Voelker, G., and Edwards, S. V. (1998). Can weighting improve bushy trees? Models of cytochrome *b* evolution and the molecular systematics of pipits and wagtails (Aves: Motacilliidae). Syst. Biol. *47*, 589-603.

Volny, V. P., and Gordon, D. M. (2002). Characterization of polymorphic microsatellite loci in the red harvester ant, *Pogonomyrmex barbatus*. Mol. Ecol. Notes. *2*, 302-303.

Waddell, P. J., Cao, Y., Hauf, J., and Hasegawa, M. (1999). Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid-invariant sites-LogDet plus site stripping, to detect internal conflicts in the data, with special reference to the positions of hedgehog, armadillo, and elephant. Syst. Biol. *48*, 31-53.

Waddell, P. J., Kishino, H., and Ota, R. (2000). Rapid evaluation of the phylogenetic congruence of sequence data using likelihood ratio tests. Mol. Biol. Evol. *17*, 1988-1992.

Waits, L. P., Sullivan, J., O'Brien, S. J., and Ward, R. H. (1999). Rapid radiation events in the family Ursidae indicated by likelihood phylogenetic estimation from multiple fragments of mtDNA. Mol. Phylogenet. Evol. *13*, 82-92.

Weiblen, G. D. (2001). Phylogenetic relationships of fig wasps pollinating functionally dioecious *Ficus* based on mitochondrial DNA sequences and morphology. Syst. Biol. *50*, 243-267.

Weins, J. J., and Reeder, T. W. (1995). Combining data sets with different numbers of taxa for phylogenetic analysis. Syst. Biol. *44*, 548-558.

Weir, B. S. (1996). Genetic Data Analysis II. (Massachusetts: Sinauer Associates, Inc).

Weir, B. S., and Cockerham, C. C. (1984). Estimating *F*-statistics for the analysis of population structure. Evolution. *38*, 1358-1370.

Wendel, J. F., and Doyle, J. J. (1998). Phylogenetic incongruence: Window into genome history and molecular evolution. *In* Molecular systematics of plants II: DNA sequencing, D. E. Soltis, P. S. Soltis and J. J. Doyle, eds. (Dordrecht: Kluwer Academic), pp. 265 - 296.

Wetterer, J. K., Schultz, T. R., and Meier, R. (1998). Phylogeny of fungus-growing ants (Tribe Attini) based on mtDNA sequence and morphology. Mol. Phylogenet. Evol. *9*, 42-47.

Wheeler, W. M. (1922). Ants of the genus *Formica* in the tropics. Psyche. (Camb.) *29*, 174-177.

Wheeler, W. M. (1927). The physiognomy of insects. Q. Rev. Biol. *2*,1-36.

Whitfield, J. B., and Cameron, S. A. (1998). Hierarchical analysis of variation in the mitochondrial 16S rRNA gene among Hymenoptera. Mol. Biol. Evol. *15*, 1728-1743.

Whittingham, L. A., Slikas, B., Winkler, D. W., and Sheldon, F. H. (2002). Phylogeny of the tree swallow genus, *Tachycineta* (Aves: Hirundinidae), by Bayesian analysis of mitochondrial DNA sequences. Mol. Phylogenet. Evol. *22*, 430-441.

Wiens, J. J. (1998). Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. *47*, 625-640.

Wiens, J. J., and Hollingsworth, B. D. (2000). War of the Iguanas: Conflicting molecular and morphological phylogenies and long-branch attraction in iguanid lizards. Syst. Biol. *49*, 143-159.

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

Wilgenbusch, J., and De Queiroz, K. (2000). Phylogenetic relationships among the phrynosomatid sand lizards inferred from mitochondrial DNA sequences generated by heterogeneous evolutionary processes. Syst. Biol. *49*, 592-612.

Williams, P. L., and Fitch, W. M. (1990). Phylogeny determination using dynamically weighted parsimony method. Methods Enzymol. *183*, 615-626.

Williams, S. T., and Knowlton, N. (2001). Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp genus *Alpheus*. Mol. Biol. Evol. *18*, 1484-1493.

Wilson, E. O. (1987). Causes of ecological success: The case of the ants. J. Animal Ecol. *56*, 1-9.

Wilson, E. O. (1976). Which are the most prevalent ant genera? Stud. Entomol. *19*, 187-200.

Yang, Z. (1996a). Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. *42*, 587-596.

Yang, Z. (1996b). Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. *11*, 367-372.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. *39*, 306-314.

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

Yang, Z., Goldman, N., and Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. *11*, 316-324.

Yang, Z., and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo Method. Mol. Biol. Evol. *14*, 717-24.

Yoder, A. D., Irwin, J. A., and Payseur, B. A. (2001). Failure of the ILD to determine data combinability for Slow Loris phylogeny. Syst. Biol. *50*, 408-424.

Yoder, A. D., Vilgalys, R., and Ruvolo, M. (1996). Molecular evolutionary dynamics of cytochrome *b* in strepsirrhine primates: The phylogenetic significance of third-position transversions. Mol. Biol. Evol. *13*, 1339-1350.

Zar, J. H. (1999). Biostatistical Analysis, 4th edition Edition (New Jersey: Prentice Hall).

Zhang, D., and Hewitt, G. M. (1996). Nuclear integrations: Challenges for mitochondrial DNA markers. Trends Ecol. Evol. *11*, 247-251.

APPENDIX I

(i) Glossary of anatomical and other specialized terms used in this thesis (from Holldobler and Wilson, 1990; Bolton, 1994).

Apical: The end farthest away from the body; at or toward the tip

Carina: A ridge or raised line.

Clypeus (adj., clypeal): The foremost section of the head capsule, just back of the mandibles, demarcated posteriorly by a transverse suture.

Dorsal: Toward or at the top or upper surface of the body or structure (below).

Gaster: The globular terminal four or five segments of the abdomen, immediately posterior to the waist.

Head: The principal anterior division of the body; it bears the mouthparts and antennae.

Lateral: Toward or at the side of the body, or the side margin or edge of a structure.

Mandible (adj., mandibular): The paired, heavily sclerotized biting and chewing lateral appendage of the mouthparts between the labrum and maxilla.

Mesosoma: The thorax plus the propodeum (cf. thorax).

Metapleural gland: A large gland with an external bulla and a small orifice, opening on each side of the metathorax at its lower posterior corners.

Nodiform: rounded and knob-shaped.

Nuchal carina: A ridge situated posteriorly on the head that separates the dorsal and lateral surfaces from the occipital surface.

Occiput: The rearmost portion of the head.

Petiole: The first (and many times the only) segment of the ant waist.

Pilosity: The longer, stouter hairs which are outstanding above the shorter, finer hairs that constitute the pubescence.

Propodeum: The first abdominal segment fused to the thorax to form the central of the three main body parts.

Psammophore: a basket-like array of long, curved hairs beneath the head of some desert ants, used as an aid in carrying sand.

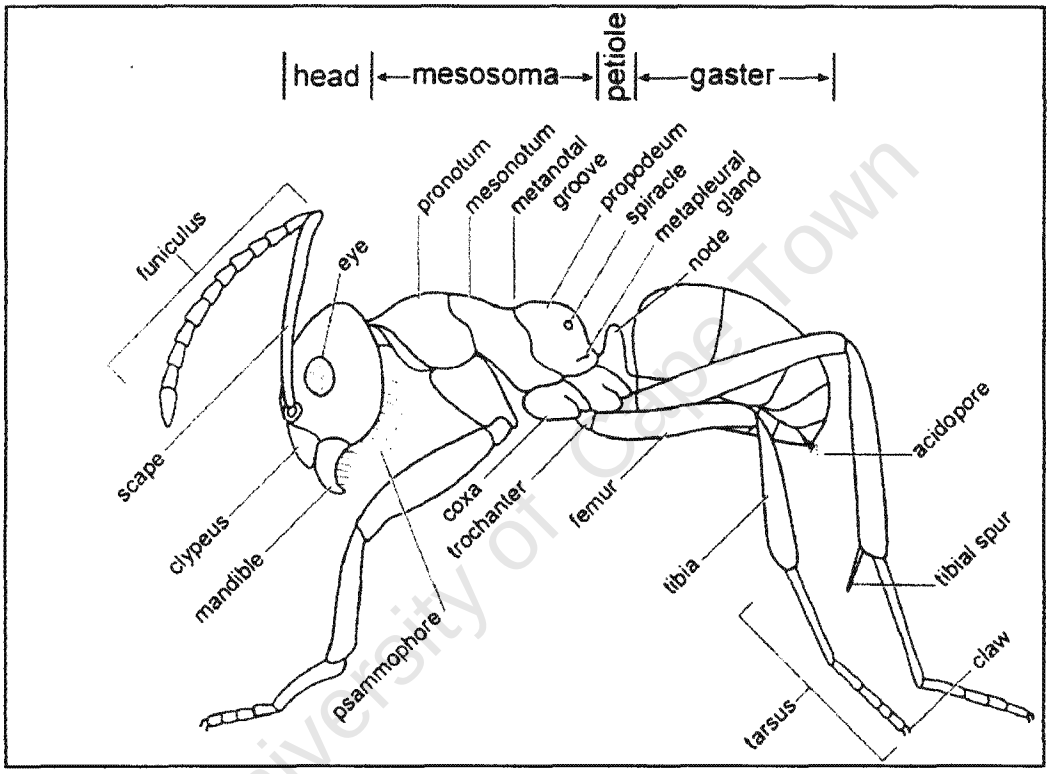
Punctate: Bearing fine, impressed points or pits.

Scape: The first elongated segment of the antenna next to the head.

Thorax: The principal middle division of the body to which the legs are attached, between the head and abdomen.

Tibia (pl., tibiae): The fourth segment of a leg, between the femur ("thigh") and the tarsus ("foot").

(ii) Diagram of a typical formicine ant in side view, with the major structures labelled (from Shattuck, 1999 pg 13).



APPENDIX II

Translation Table V for invertebrate mitochondrial protein-coding genes (sourced from Genbank: http://www.ncbi.nlm.nih.gov).

TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys
TTA	Leu	TCA	Ser	TAA	*	TGA	Trp
TTG	Leu	TCG	Ser	TAG	*	TGG	Trp
CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
CTC	Leu	CCC	Pro	CAC	His	CGC	Arg
CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
ATA	Met	ACA	Thr	AAA	Lys	AGA	Ser
ATG	Met	ACG	Thr	AAG	Lys	AGG	Ser
GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GTG	Val	GGG	Ala	GAG	Glu	GGG	Gly

* indicates a termination codon

APPENDIX III

Table showing *Camponotus spp.* for which only cytochrome *b* or cytochrome oxidase II sequence data were obtained.

Species	Cytochrome *b*	Cytochrome oxidase II
Camponotus sp. 24	x	√
Camponotus sp. 14	x	√
C. mystaceus	x	√
C. detritus	x	√
C. brevisetosus	x	√
Camponotus sp. 12	√	x
C. rufoglaucus	√	x
C. cuneiscapus	√	x
C. sericeus	√	x
C. chrysurus	√	x

√ indicates partial gene product amplified and sequenced; x indicates no amplification product was obtained.

APPENDIX IV

(i) Table showing pairwise sequence divergence estimates for cytochrome *b*. Numbers below the diagonal are Tamura-Nei corrected nucleotide distances; numbers above the diagonal represent amino acid 'p' distances.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
[1] *C.rufoglaucus*		0.105	0.132	0.127	0.155	0.159	0.109	0.128	0.123	0.118	0.136	0.173	0.164	0.134	0.136	0.095	0.132	0.293	0.22
[2] *C.sp.2*	0.243		0.141	0.132	0.150	0.150	0.114	0.142	0.136	0.132	0.145	0.173	0.159	0.144	0.141	0.114	0.127	0.268	0.244
[3] *C. cuneiscapus*	0.242	0.229		0.032	0.154	0.100	0.145	0.164	0.154	0.150	0.005	0.173	0.159	0.149	0.154	0.145	0.168	0.252	0.236
[4] *P. schistacea*	0.238	0.229	0.031		0.123	0.123	0.141	0.155	0.154	0.150	0.027	0.159	0.150	0.134	0.141	0.136	0.154	0.260	0.236
[5] *C. sp. 7*	0.254	0.248	0.276	0.207		0.164	0.127	0.160	0.159	0.159	0.159	0.145	0.118	0.119	0.123	0.145	0.145	0.292	0.244
[6] *C. sp. 10*	0.259	0.260	0.187	0.200	0.269		0.159	0.187	0.173	0.168	0.105	0.191	0.173	0.173	0.173	0.173	0.164	0.276	0.244
[7] *C. sp. 11*	0.255	0.226	0.259	0.263	0.246	0.253		0.142	0.141	0.141	0.177	0.136	0.118	0.104	0.127	0.105	0.118	0.276	0.236
[8] *C. sericeus*	0.238	0.227	0.224	0.227	0.257	0.225	0.215		0.027	0.023	0.169	0.183	0.164	0.178	0.178	0.151	0.164	0.289	0.27
[9] *C. chrysurus*	0.236	0.231	0.223	0.236	0.258	0.226	0.250	0.017		0.004	0.159	0.195	0.182	0.158	0.168	0.145	0.168	0.276	0.26
[10] *C. acvapimensis*	0.236	0.229	0.221	0.234	0.260	0.224	0.252	0.016	0.002		0.155	0.191	0.182	0.163	0.168	0.141	0.164	0.268	0.26
[11] *C. nasutus*	0.241	0.228	0.005	0.029	0.278	0.191	0.264	0.230	0.230	0.228		0.177	0.164	0.153	0.159	0.150	0.173	0.252	0.236
[12] *C. fulvopilosus*	0.263	0.226	0.246	0.262	0.247	0.240	0.240	0.216	0.214	0.216	0.253		0.064	0.114	0.127	0.173	0.182	0.268	0.228
[13] *C. storeatus*	0.233	0.215	0.233	0.250	0.239	0.211	0.207	0.188	0.210	0.210	0.235	0.137		0.094	0.105	0.173	0.159	0.268	0.221
[14] *C. sp. 12*	0.263	0.203	0.260	0.243	0.214	0.277	0.227	0.241	0.236	0.238	0.265	0.171	0.201		0.005	0.138	0.129	0.314	0.248
[15] *C. bifossus*	0.270	0.200	0.268	0.252	0.219	0.279	0.243	0.245	0.243	0.243	0.272	0.182	0.205	0.003		0.141	0.123	0.293	0.236
[16] *C. cinctellus*	0.210	0.206	0.238	0.247	0.230	0.260	0.203	0.209	0.211	0.213	0.245	0.209	0.208	0.252	0.245		0.123	0.285	0.26
[17] *C. klugii*	0.250	0.238	0.268	0.251	0.230	0.270	0.246	0.261	0.263	0.265	0.266	0.260	0.250	0.219	0.213	0.235		0.310	0.285
[18] *Oecophylla*	0.369	0.254	0.259	0.274	0.322	0.268	0.291	0.295	0.304	0.298	0.259	0.265	0.260	0.277	0.272	0.343	0.356		0.195
[19] *Formica*	0.352	0.293	0.290	0.298	0.330	0.279	0.312	0.342	0.329	0.331	0.290	0.289	0.281	0.327	0.305	0.301	0.364	0.245	

(ii) Table showing pairwise sequence divergence estimates for cytochrome oxidase II. Numbers below the diagonal are Tamura-Nei corrected nucleotide distances; numbers above the diagonal represent amino acid 'p' distances.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
[1] *C. sp.* 2		0.106	0.188	0.106	0.106	0.113	0.137	0.106	0.106	0.056	0.100	0.144	0.106	0.063	0.106	0.131	0.075	0.087	0.081	0.295	0.220
[2] *C. brevisetosus*	0.188		0.181	0.094	0.056	0.125	0.150	0.112	0.076	0.062	0.080	0.150	0.056	0.112	0.000	0.087	0.106	0.081	0.100	0.277	0.239
[3] *P. schistacea*	0.244	0.214		0.156	0.181	0.187	0.206	0.169	0.212	0.169	0.121	0.194	0.181	0.175	0.181	0.200	0.175	0.162	0.175	0.258	0.226
[4] *C. sp.* 7	0.232	0.183	0.260		0.094	0.106	0.175	0.094	0.114	0.075	0.110	0.162	0.094	0.087	0.094	0.112	0.094	0.087	0.094	0.296	0.233
[5] *C. detritus*	0.226	0.094	0.229	0.176		0.150	0.162	0.100	0.023	0.075	0.055	0.181	0.000	0.100	0.056	0.112	0.100	0.069	0.112	0.302	0.233
[6] *C. sp.* 24	0.189	0.205	0.220	0.220	0.219		0.162	0.119	0.159	0.087	0.143	0.137	0.150	0.094	0.125	0.156	0.125	0.112	0.100	0.296	0.239
[7] *C. sp.* 10	0.201	0.179	0.193	0.210	0.172	0.195		0.150	0.197	0.125	0.165	0.106	0.162	0.144	0.150	0.137	0.137	0.162	0.156	0.258	0.201
[8] *C. sp.* 11	0.203	0.192	0.251	0.224	0.180	0.221	0.214		0.129	0.081	0.110	0.137	0.100	0.087	0.112	0.100	0.087	0.094	0.106	0.290	0.214
[9] *C. sp.* 14	0.252	0.093	0.247	0.175	0.021	0.223	0.176	0.211		0.098	0.095	0.189	0.023	0.098	0.076	0.144	0.106	0.083	0.114	0.333	0.257
[10] *C. acvapimensis*	0.171	0.132	0.202	0.182	0.136	0.187	0.175	0.190	0.131		0.055	0.125	0.075	0.062	0.062	0.100	0.075	0.069	0.069	0.283	0.214
[11] *C. mystaceus*	0.214	0.106	0.176	0.179	0.037	0.221	0.178	0.164	0.056	0.114		0.187	0.055	0.110	0.080	0.154	0.121	0.088	0.110	0.244	0.166
[12] *C. nasutus*	0.223	0.233	0.238	0.213	0.268	0.197	0.180	0.225	0.251	0.209	0.253		0.181	0.125	0.150	0.144	0.125	0.162	0.131	0.233	0.214
[13] *C. fulvopilosus*	0.226	0.094	0.229	0.176	0.000	0.219	0.172	0.180	0.021	0.136	0.037	0.268		0.100	0.056	0.112	0.100	0.069	0.112	0.302	0.233
[14] *C. sp.* 19	0.182	0.177	0.202	0.191	0.161	0.199	0.202	0.184	0.150	0.159	0.124	0.192	0.161		0.112	0.131	0.081	0.062	0.075	0.277	0.207
[15] *C. storeatus*	0.188	0.000	0.214	0.183	0.094	0.205	0.179	0.192	0.093	0.132	0.106	0.233	0.094	0.177		0.087	0.106	0.081	0.100	0.277	0.239
[16] *C. sp.* 12	0.194	0.124	0.186	0.206	0.147	0.207	0.170	0.188	0.171	0.148	0.135	0.191	0.147	0.184	0.124		0.075	0.125	0.100	0.277	0.226
[17] *C. bifossus*	0.193	0.148	0.179	0.176	0.154	0.183	0.171	0.190	0.180	0.154	0.149	0.212	0.154	0.177	0.148	0.109		0.094	0.062	0.264	0.182
[18] *C. cinctellus*	0.181	0.144	0.224	0.181	0.163	0.175	0.215	0.193	0.142	0.162	0.145	0.212	0.163	0.136	0.144	0.168	0.144		0.087	0.283	0.220
[19] *C. klugii*	0.226	0.195	0.229	0.183	0.195	0.172	0.203	0.214	0.212	0.156	0.182	0.194	0.195	0.206	0.195	0.177	0.143	0.173		0.277	0.207
[20] *Formica*	0.351	0.312	0.316	0.293	0.339	0.405	0.299	0.356	0.346	0.300	0.272	0.304	0.339	0.312	0.312	0.309	0.298	0.296	0.326		0.132
[21] *Oecophylla*	0.301	0.257	0.232	0.292	0.288	0.278	0.227	0.294	0.279	0.254	0.214	0.270	0.288	0.251	0.257	0.212	0.216	0.261	0.294	0.170	

APPENDIX V

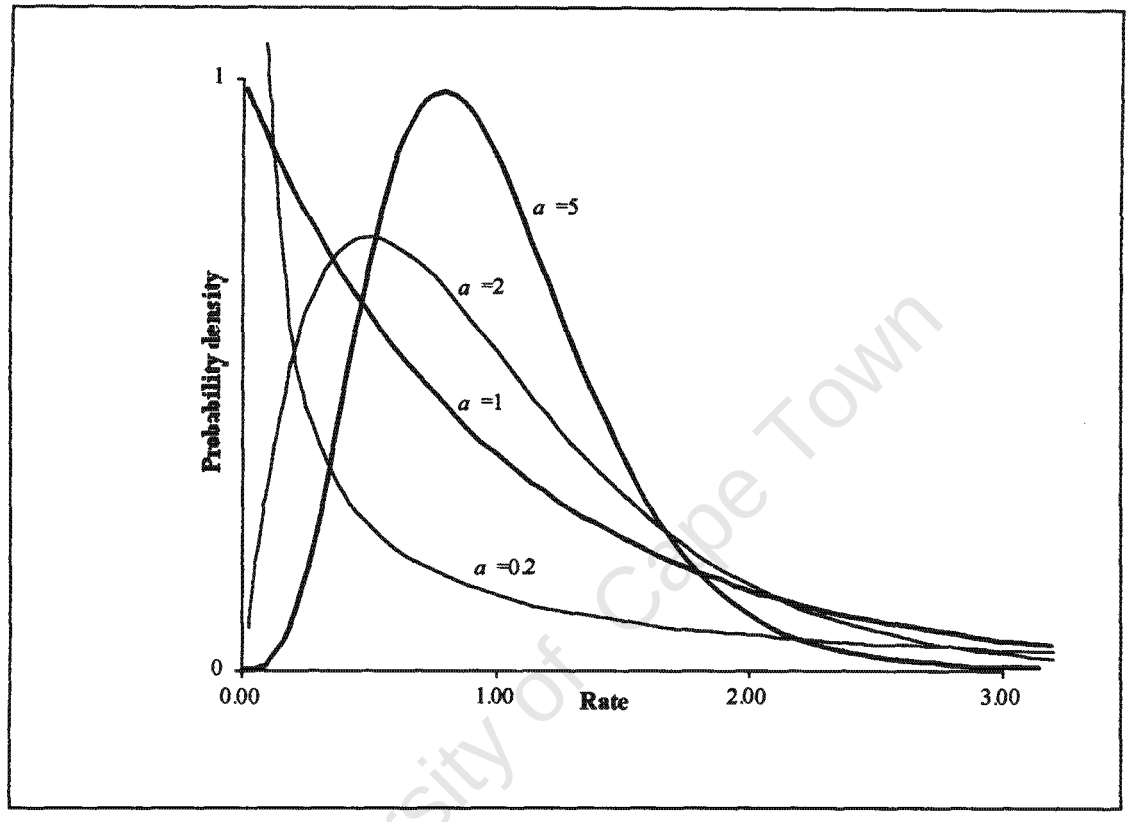


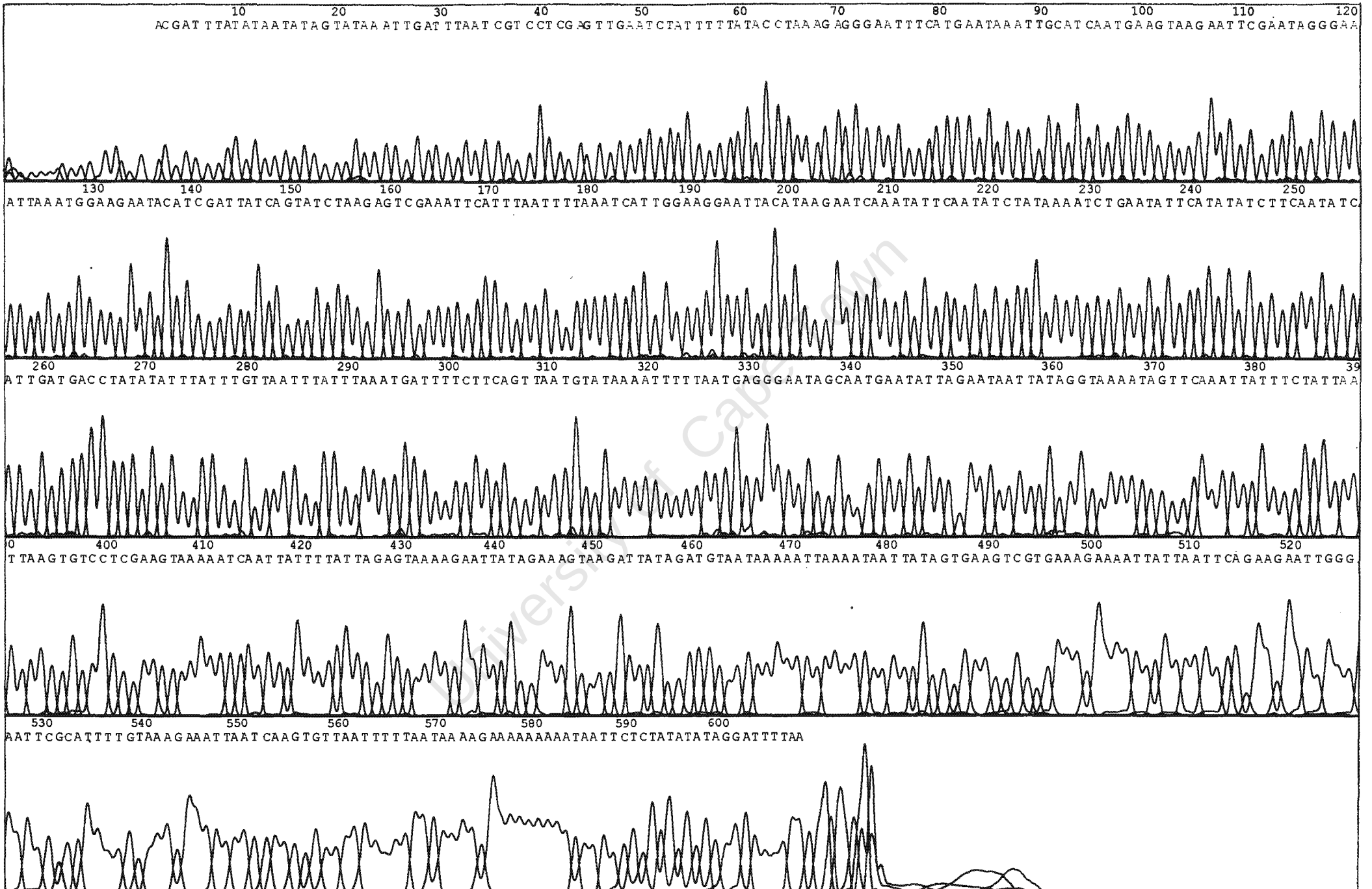
Figure showing the density function of the gamma distribution of substitution rates at sites for various values of α (after Yang, 1996b).

APPENDIX VI

Table with Genbank accession numbers of additional *Polyrhachis spp.* included in this study.

	Genbank Accession Numbers	
Species	Cytochrome *b*	Cytochrome oxidase II
P. doddi	AF191149	AF190330
P. pilosa	AF191150	AF190331
P. dives	AF191151	AF190340
P. nr. sp. R5	AF191152	AF190332
P. turneri	AF191153	AF190334
P. schellerichae	AF191154	AF190347
P. furcata	AF191156	AF190348
P. bicolor	AF191157	AF190346

APPENDIX VII



Chromatogram of cytochrome oxidase II pseudogene sequence obtained for *Camponotus rufoglaucus*. This sequence was generated using the reverse primer C2-N-3665

APPENDIX VIII

(i) Table showing inferred genotypes of queens, workers and males of *Camponotus klugii* for nest 1. Worker genotypes were predicted under the hypotheses of a single, once-mated queen.

	Camp4	Camp8	**Ccon12**	Ccon20	Ccon42	Ccon70	Ccon79
Nest 1							
Proposed mating:							
Queen	207/209	123/123	177/179	293/293 or 297/297	262/264	161/161 or 163/163	356/385
X	X	X	X	X	X	X	X
Male	209	123	173	297 or 293	262	163 or 161	385
Worker genotypes expected	207/209	123/123	173/177	293/297	262/264	161/163	385/385
	207/209	123/123	173/177	293/297	262/264	161/163	356/385
	207/209	123/123	173/177	293/297	262/262	161/163	385/385
	207/209	123/123	173/177	293/297	262/262	161/163	356/385
	207/209	123/123	173/179	293/297	262/264	161/163	385/385
	207/209	123/123	173/179	293/297	262/264	161/163	356/385
	207/209	123/123	173/179	293/297	262/262	161/163	385/385
	207/209	123/123	173/179	293/297	262/262	161/163	356/385
	209/209	123/123	173/177	293/297	262/264	161/163	385/385
	209/209	123/123	173/177	293/297	262/264	161/163	356/385
	209/209	123/123	173/177	293/297	262/262	161/163	385/385
	209/209	123/123	173/177	293/297	262/262	161/163	356/385
	209/209	123/123	173/179	293/297	262/264	161/163	385/385
	209/209	123/123	173/179	293/297	262/264	161/163	356/385
	209/209	123/123	173/179	293/297	262/262	161/163	385/385
	209/209	123/123	173/179	293/297	262/262	161/163	356/385

(ii) Table showing inferred genotypes of queens, workers and males of *Camponotus klugii* for nest 2. Worker genotypes were predicted under the hypotheses of a single, once-mated queen.

	Camp4	Camp8	**Ccon12**	Ccon20	Ccon42	Ccon70	Ccon79
Nest 2							
Proposed mating							
Queen X Male	207/209 X 207	123/123 X 123	165/171 X 181	291/299 X 293	262/262 X 262	161/165 X 161	364/370 X 358
Worker genotypes expected	207/207	123/123	165/181	293/299	262/262	161/161	358/370
	207/207	123/123	165/181	293/299	262/262	161/161	358/364
	207/207	123/123	165/181	293/299	262/262	161/165	358/370
	207/207	123/123	165/181	293/299	262/262	161/165	358/364
	207/207	123/123	165/181	291/293	262/262	161/161	358/370
	207/207	123/123	165/181	291/293	262/262	161/161	358/364
	207/207	123/123	165/181	291/293	262/262	161/165	358/370
	207/207	123/123	165/181	291/293	262/262	161/165	358/364
	207/207	123/123	171/181	293/299	262/262	161/161	358/370
	207/207	123/123	171/181	293/299	262/262	161/161	358/364
	207/207	123/123	171/181	293/299	262/262	161/165	358/370
	207/207	123/123	171/181	293/299	262/262	161/165	358/364
	207/207	123/123	171/181	291/293	262/262	161/161	358/370
	207/207	123/123	171/181	291/293	262/262	161/161	358/364
	207/207	123/123	171/181	291/293	262/262	161/165	358/370
	207/207	123/123	171/181	291/293	262/262	161/165	358/364
	207/209	123/123	165/181	293/299	262/262	161/161	358/370
	207/209	123/123	165/181	293/299	262/262	161/161	358/364
	207/209	123/123	165/181	293/299	262/262	161/165	358/370
	207/209	123/123	165/181	293/299	262/262	161/165	358/364
	207/209	123/123	165/181	291/293	262/262	161/161	358/370
	207/209	123/123	165/181	291/293	262/262	161/161	358/364
	207/209	123/123	165/181	291/293	262/262	161/165	358/370
	207/209	123/123	165/181	291/293	262/262	161/165	358/364
	207/209	123/123	171/181	293/299	262/262	161/161	358/370
	207/209	123/123	171/181	293/299	262/262	161/161	358/364
	207/209	123/123	171/181	293/299	262/262	161/165	358/370
	207/209	123/123	171/181	293/299	262/262	161/165	358/364
	207/209	123/123	171/181	291/293	262/262	161/161	358/370
	207/209	123/123	171/181	291/293	262/262	161/161	358/364
	207/209	123/123	171/181	291/293	262/262	161/165	358/370
	207/209	123/123	171/181	291/293	262/262	161/165	358/364

(iii) Table showing inferred genotypes of queens, workers and males of *Camponotus klugii* for nest 3. Worker genotypes were predicted under the hypotheses of a single, once-mated queen.

	Camp4	Camp8	**Ccon12**	Ccon20	Ccon42	Ccon70	Ccon79
Nest 3							
Proposed mating							
Queen		121/123	175/177	291/295	262/262 or 266/266	161/161	389/412
X		X	X	X	X	X	X
Male		121	173	293	266 or 262	161	395
Worker genotypes expected		121/121	173/175	293/295	262/266	161/161	395/412
		121/121	173/175	293/295	262/266	161/161	389/395
		121/121	173/175	291/293	262/266	161/161	395/412
		121/121	173/175	291/293	262/266	161/161	389/395
		121/121	173/177	293/295	262/266	161/161	395/412
		121/121	173/177	293/295	262/266	161/161	389/395
		121/121	173/177	291/293	262/266	161/161	395/412
		121/121	173/177	291/293	262/266	161/161	389/395
		121/123	173/175	293/295	262/266	161/161	395/412
		121/123	173/175	293/295	262/266	161/161	389/395
		121/123	173/175	291/293	262/266	161/161	395/412
		121/123	173/175	291/293	262/266	161/161	389/395
		121/123	173/177	293/295	262/266	161/161	395/412
		121/123	173/177	293/295	262/266	161/161	389/395
		121/123	173/177	291/293	262/266	161/161	395/412
		121/123	173/177	291/293	262/266	161/161	389/395

(iv) Table showing inferred genotypes of queens, workers and males of *Camponotus klugii* for nest 4. Worker genotypes were predicted under the hypotheses of a single, once-mated queen.

	Camp4	Camp8	Ccon12	Ccon20	Ccon42	Ccon70	Ccon79
Nest 4							
Proposed mating							
Queen	207/209	123/123	173/179	289/293	264/266	161/163	370/381
X	X	X	X	X	X	X	X
Male	209	123	165	291	262	161	358
Worker genotypes expected	207/209	123/123	165/173	291/293	262/264	161/163	358/370
	207/209	123/123	165/173	291/293	262/264	161/163	358/381
	207/209	123/123	165/173	291/293	262/264	161/161	358/370
	207/209	123/123	165/173	291/293	262/264	161/161	358/381
	207/209	123/123	165/173	291/293	262/266	161/163	358/370
	207/209	123/123	165/173	291/293	262/266	161/163	358/381
	207/209	123/123	165/173	291/293	262/266	161/161	358/370
	207/209	123/123	165/173	291/293	262/266	161/161	358/381
	207/209	123/123	165/173	289/291	262/264	161/163	358/370
	207/209	123/123	165/173	289/291	262/264	161/163	358/381
	207/209	123/123	165/173	289/291	262/264	161/161	358/370
	207/209	123/123	165/173	289/291	262/264	161/161	358/381
	207/209	123/123	165/173	289/291	262/266	161/163	358/370
	207/209	123/123	165/173	289/291	262/266	161/163	358/381
	207/209	123/123	165/173	289/291	262/266	161/161	358/370
	207/209	123/123	165/173	289/291	262/266	161/161	358/381
	207/209	123/123	165/179	291/293	262/264	161/163	358/370
	207/209	123/123	165/179	291/293	262/264	161/163	358/381
	207/209	123/123	165/179	291/293	262/264	161/161	358/370
	207/209	123/123	165/179	291/293	262/264	161/161	358/381
	207/209	123/123	165/179	291/293	262/266	161/163	358/370
	207/209	123/123	165/179	291/293	262/266	161/163	358/381
	207/209	123/123	165/179	291/293	262/266	161/161	358/370
	207/209	123/123	165/179	291/293	262/266	161/161	358/381
	207/209	123/123	165/179	289/291	262/264	161/163	358/370
	207/209	123/123	165/179	289/291	262/264	161/163	358/381
	207/209	123/123	165/179	289/291	262/264	161/161	358/370
	207/209	123/123	165/179	289/291	262/264	161/161	358/381
	207/209	123/123	165/179	289/291	262/266	161/163	358/370
	207/209	123/123	165/179	289/291	262/266	161/163	358/381
	207/209	123/123	165/179	289/291	262/266	161/161	358/370
	207/209	123/123	165/179	289/291	262/266	161/161	358/381
	209/209	123/123	165/173	291/293	262/264	161/163	358/370
	209/209	123/123	165/173	291/293	262/264	161/163	358/381
	209/209	123/123	165/173	291/293	262/264	161/161	358/370
	209/209	123/123	165/173	291/293	262/264	161/161	358/381
	209/209	123/123	165/173	291/293	262/266	161/163	358/370
	209/209	123/123	165/173	291/293	262/266	161/163	358/381
	209/209	123/123	165/173	291/293	262/266	161/161	358/370
	209/209	123/123	165/173	291/293	262/266	161/161	358/381
	209/209	123/123	165/173	289/291	262/264	161/163	358/370
	209/209	123/123	165/173	289/291	262/264	161/163	358/381
	209/209	123/123	165/173	289/291	262/264	161/161	358/370
	209/209	123/123	165/173	289/291	262/264	161/161	358/381
	209/209	123/123	165/173	289/291	262/266	161/163	358/370
	209/209	123/123	165/173	289/291	262/266	161/163	358/381
	209/209	123/123	165/173	289/291	262/266	161/161	358/370
	209/209	123/123	165/173	289/291	262/266	161/161	358/381
	209/209	123/123	165/179	291/293	262/264	161/163	358/370
	209/209	123/123	165/179	291/293	262/264	161/163	358/381
	209/209	123/123	165/179	291/293	262/264	161/161	358/370
	209/209	123/123	165/179	291/293	262/264	161/161	358/381
	209/209	123/123	165/179	291/293	262/266	161/163	358/370
	209/209	123/123	165/179	291/293	262/266	161/163	358/381
	209/209	123/123	165/179	291/293	262/266	161/161	358/370
	209/209	123/123	165/179	291/293	262/266	161/161	358/381
	209/209	123/123	165/179	289/291	262/264	161/163	358/370
	209/209	123/123	165/179	289/291	262/264	161/163	358/381
	209/209	123/123	165/179	289/291	262/264	161/161	358/370
	209/209	123/123	165/179	289/291	262/264	161/161	358/381
	209/209	123/123	165/179	289/291	262/266	161/163	358/370
	209/209	123/123	165/179	289/291	262/266	161/163	358/381
	209/209	123/123	165/179	289/291	262/266	161/161	358/370
	209/209	123/123	165/179	289/291	262/266	161/161	358/381

APPENDIX IX

True Basic code for program written by E. Harley to obtain an estimate the proportion of times in a large number of trials (10,000) that a ratio of 31:24 or higher is found.

```
RANDOMIZE
CLEAR
FOR k=1 to 10000
    LET x,y=0
    FOR j=1 to 55
        IF rnd>.5 then LET x=x+1 else LET y=y+1
    NEXT j
    IF x>=31 or y>=31 then LET C=C+1
NEXT k
PRINT "Percentage of times a 31:24 ratio (or higher) is found on
random sampling = "; C/100
END
```

sequence was numbered on the basis of the human cytochrome *b* protein sequence (PIR-PSD[1] code CBHU) which has been fully annotated with regard to predicted functional domains. The cytochrome oxidase II ant protein was numbered relative to the human cytochrome oxidase II protein sequence (PIR-PSD code OBHU2).

Mutational saturation

Nucleotide saturation plots were constructed to evaluate which, if any, of the three codon positions of the two genes appear to be saturated and therefore potentially unreliable. The proportion of observed differences (uncorrected) between pairs of species is plotted as a function of the estimated proportion of differences for the same species pairs using a model of nucleotide evolution to correct for multiple substitutions. The relationship is initially linear, but as evolutionary time between sequences increases they may become more saturated, which is reflected by the straight line approaching an asymptote; while substitutions are still occurring (estimated by the corrected sequence divergence value) they are not actually observed.

The Kimura 2-parameter model (Kimura, 1980) was used to correct for multiple substitutions in the nucleotide data sets. This distance correction allows for variable transition and transversion frequencies while assuming equal base frequencies. The Tamura-Nei distance correction, though more appropriate for sequences with biased base composition, resulted in 89 uncorrectable pairwise comparisons for the third codon positions of cytochrome *b* and 133 undefined distances for cytochrome oxidase II, indicating that these divergence values exceed the maximum calculable value of the correction method being used.

[1] The Protein Sequence Database (PSD), produced by the Protein Information Resource (PIR), National Biomedical Research Foundation, Washington, USA contains functionally annotated protein sequences and can be accessed at the following web site: http://pir.georgetown.edu.