Molecular Phylogenetics of the Superfamily Curculionoidea (Insecta: Coleoptera)

Conrad Paulus Dias Trafford Gillett

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

University of East Anglia
Norwich, Norfolk, England

March 2014

Molecular Phylogenetics of the Superfamily Curculionoidea (Insecta: Coleoptera)

Conrad Paulus Dias Trafford Gillett

March 2014

Thesis abstract

This thesis examines higher-level evolutionary history within the superfamily Curculionoidea, the most speciose family-level taxon, which includes beetles commonly known as weevils. This is achieved using a phylogenetic approach incorporating the largest datamatrix yet employed for weevil molecular systematics, and includes an investigation into the prospect of obtaining short phylogenetically informative amplicons from archival museum specimens. Newly obtained DNA sequence data is analysed from a variety of mitochondrial and nuclear loci, including 92 mitogenomes assembled through the approach of next-generation sequencing of pooled genomic DNA. The resulting trees are used to test previous morphological- and molecular-based hypotheses of weevil relationships and classification schemes.

Mitogenomic-derived trees reveal topologies that are highly congruent with previous molecular studies, but that conflict with some morphological hypotheses. Strong nodal support strengthens inferences into the relationships amongst most weevil families and suggests that the largest family, the Curculionidae, is monophyletic, if the subfamily Platypodinae is excluded. Division of the Curculionidae into two large clades is well supported and the wood-boring habit adopted by three subfamilies is shown to have arisen multiple times, contradicting most morphological analyses.

Addition of several nuclear loci to the mitogenomic data is found to provide little additional value, in terms of improving nodal bootstrap support. A suggestion is made that future efforts to enhance understanding of relationships should focus on improving taxon sampling. Statistical tests of an augmented dataset, derived from public database sequences for single mitochondrial genes, wherein multiple tribes and subfamilies within the broad-nosed weevils are constrained as monophyletic, indicate that three entimine tribes, as currently defined, are each not consistent with the hypothesis for their monophyly.

Incongruences between molecular data and classical morphological taxonomy are suggestive that the current weevil classification system is misleading if used to interpret species richness, geographic distributions or ecological traits within currently recognised lineages.

Supervisory Information

University of East Anglia

Primary supervisor:
Dr. Brent C. Emerson
Island Ecology and Evolution Research Group,
Instituto de Productos Naturales y Agrobiología (IPNA-CSIC),
C/Astrofísico Francisco Sánchez 3,
La Laguna,
Tenerife (Canary Islands, 38206),
Spain

Secondary supervisor:
Dr. David S. Richardson
School of Biological Sciences,
University of East Anglia,
Norwich Research Park,
Norwich,
Norfolk
NR4 7TJ
United Kingdom

The Natural History Museum, London

Prof. Alfried P. Vogler
Department of Life Sciences,
The Natural History Museum,
Cromwell Road,
South Kensington,
London
SW7 5BD
United Kingdom

Dr. Christopher H. Lyal
Department of Life Sciences,
The Natural History Museum,
Cromwell Road,
South Kensington,
London
SW7 5BD
United Kingdom

List of Contents

List of Tables, Figures and Appendices

Dedication

FLVCTVAT·NEC·MERGITVR

In perpetual remembrance of the fallen

1914-1918

1939-1945

Age shall not weary them

Acknowledgements

The research undertaken towards, and summarised within, this document would have been quite impossible were it not for the generosity bestowed upon me by a great number of people, throughout the entire course of four years spent engaged in pursuit of the elusive relationships within the weevils. I am especially fortunate to have worked in two prestigious institutions and therefore to have had the additional benefit of engaging directly with two groups of outstanding scholars, whose combined experiece will, without doubt, have contributed substantially to the present work.

First and foremost it is with great pleasure that I extend my most sincere gratitude to my primary supervisor, Dr. Brent Emerson, who not only offered me this superb training opportunity, but also freely shared his knowledge, experience, enthusiasm and encouragement throughout (often as a voice on skype!), without whom my struggles would have been far bleaker! Concurrently I offer my deepest thanks to Prof. Alfried Vogler and Dr. Chris Lyal of the (former) Department of Entomology at the Natural History Museum (NHM), and to Prof. David Richardson of the University of East Anglia (UEA), for their important supervisory roles lasting the duration of this project. Alfried especially shared the profound experience of his molecular laboratory group, and Chris his deep insights into weevil taxonomy, in addition to being a superb companion during our fascinating specimen collecting fieldtrip to Yunnan, China.

I am thoroughly indebted to all the individuals who so liberally dispatched on loan to me, from around the planet, the all-important weevil specimens or DNA aliquots, and in so doing, actually made this study feasible. Without their noble efforts in the field, I would have faced an insurmountable obstacle at the very first hurdle of taxon sampling. Whether their specimens were ultimately utilised in this study or not, I gratefully acknowledge all the collectors: Bengt Andersson (SWE), Max Barclay (GBR), Iain Barr (GBR), Roberto Caldara (ITA), Altanchimeg Dorjoo (MNG), Brent Emerson (ESP), Christiana Faria (BRA), Michael

Gillett (NLD), Levent Gültekin (TUR), Anna Hundsdoerfer (GER), Bjarte Jordal (NOR), James Kitson (GBR), José Luis Lencina (ESP), Ren Li (CHN), Christopher Lyal (GBR), Massimo Meregalli (ITA), Philippe Moretto (FRA), Mike Morris (GBR), Rolf Oberprieler (AUS), Charles O'Brien (USA), Pedro Oromí (ESP), Michele Pierotti (ITA), Alexandru Pintilioaie (ROU), Gordon Ramel (GRC), Clive Turner (GBR) and Robyn Waayers (USA). The curator of Coleoptera at the NHM, Max Barclay, kindly entrusted me with loans of important specimens under his care.

Several individuals deserve to be singled out for having shown exceptional patience in demonstrating and teaching me a wealth of laboratory and bioinformatics techniques. Alison Ollikainen (NHM) ensured that I persevered with the troublesome LR-PCRs until they *finally* did work, and Sujatha Sunil (NHM) offered kind assistance during my first tentative steps in sequence analysis. Alex Crampton-Platt (NHM) displayed a remarkable endurance (if not tolerance!) for my seemingly never-ending appeals for assistance during the darkest moments of NGS-assembly misery. She never failed to confer abundant and timely assistance and for that I shall always be thankful. Martijn Timmermans (NHM) also deserves particular mention for his consistent interest in my work and his eagerness to provide a helping hand with challenging analyses.

It is with a great deal of affection that I thank all my marvelous colleagues from the School of Biological Sciences (BIO), for both their joyous companionship and their unwavering support over the years, in addition to their having so well withstdood my prolonged and relentless barrage of clichéd audio-visual weevil puns, meted out at every opportunity in my talks or presentations! I could not have wished to work alongside a better bunch of friends, especially my immediate PhD cohort and associated postdocs and technicians: Christiana Faria, Gerardo Hernandez-Vera, James Kitson, Alicia Mastretta-Yanes, Edd Almond, Emma Barrett, Amanda Bretman, Kat Bruce, David Collins, Siân Diamond, Christiana Faria, Claudia Fricke, Lucy Friend, Danni Gilroy, Catalina Gonzalez, Tim Huggins, James Kitson, Phil Leftwich, Alyson Lumley, Anna Millard, Will Nash, Danielle Peruffo, Amy Romans, Damian Smith, Lewis Spurgin, Freydis Vigfusdottir, Jacob Holland, Karl Phillips, Janet Mason, and Henry Ferguson-Gow. The first four and last four scientists listed deserve special mention for having been my weevil-lab and Bambi-lab fellow members respectively and therefore have had to endure more of me than most! James Kitson especially has been a first-rate friend and colleague whose generosity is only exceeded by his humility.

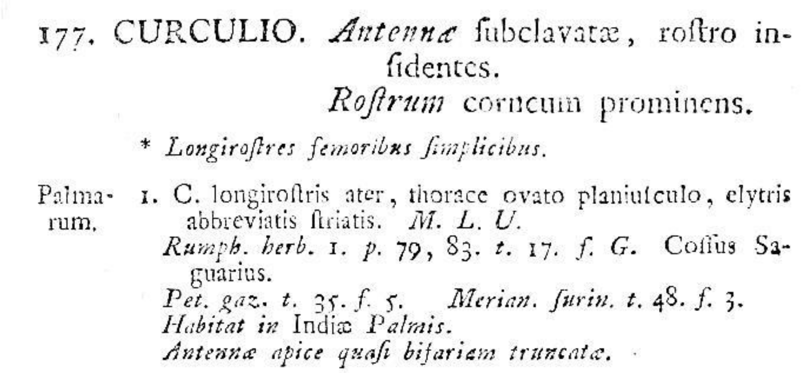
The fine City of Norwich itself merits praise for having been a fantastic home to me during the PhD. I shall treasure forever my link to the beautiful county of Norfolk and look back at my time here with abundant fondness and respect, in no small part owing to its collection of exceptional public houses, of which the White Lion certainly warrants a cheer! My two housemates, Tim Huggins and Karl Phillips, both remarkable morris dancers, also deserve special thanks; Tim for generously sharing his encyclopaedic knowledge of real ale and Karl for maintaining a steady supply of cups of tea and glasses of brandy and sherry - critical sustenance for maintaining vitality during the arduous writing-up phase!

I wish to personally extend my gratitude to the Natural Environment Research Council for having provided the financial funding for this research, and to UEA and the NHM for having provided the facilities in which to conduct it. I offer a specific thanks to the BIO administrative and finance staff, who have always been most helpful and attentive. I am also grateful to a fieldwork grant from the NHM to collect specimens in China and to two travel grants from UEA which allowed me to attend phylogenetics courses in Modena and Erice, Italy. I sincerely thank all the instructors and students who attended these courses for greatly assisting me towards an understanding of phylogenetics. I especially thank Karen Siu-Ting (NUI Maynooth) for her help with statistical hypothethis testing.

Lastly, I thank my father Dr. Michael P.T. Gillett, my mother Antoniêta D.T. Gillett, my brother James E.D.T. Gillett and my sister Dr. Caroline D.T. Gillett for having unflinchingly supported me in everything I have done; without you I would never have come this far. I especially thank my father for having single-handedly sown the seeds of passion for the natural world all those years ago in the Atlantic forests of Brazil, that I might reap the infinite rewards of its study, and thereafter nurtured it in innumerable expeditions and discussions across our planet, from desert to mountain, to tropical island and beyond.

Chapter 1

From Linnæus 1758 to next-generation sequencing 2014 - an introductory review of curculionoid higher-level phylogeny and molecular systematics



177. CURCULIO. *Antennæ* ſubclavatæ, roſtro inſidentes.
Roſtrum corneum prominens.

* *Longiroſtres femoribus ſimplicibus.*

Palmarum. 1. C. longiroſtris ater, thorace ovato planiuſculo, elytris abbreviatis ſtriatis. *M. L. U.*
Rumph. herb. 1. *p.* 79, 83. *t.* 17. *f. G.* Coſſus Saguarius.
Pet. gaz. t. 35. *f.* 5. *Merian. ſurin. t.* 48. *f.* 3.
Habitat in Indiæ *Palmis.*
Antennæ apice quaſi bifariam truncatæ.

- Carolus Linnæus, 1758

Rhynchophorus palmarum (Linnæus, 1758) (Dryophthoridae: Rhynchophorinae), the first weevil to be given a Latin binomial. Stoneiland, Suriname

Chapter 1: From Linnæus 1758 to next-generation sequencing 2014 - an introductory review of curculionoid higher-level phylogeny and molecular systematics

1.1 Weevil diversity and biology

The group of beetles (Insecta: Coleoptera) belonging to the superfamily Curculionoidea and commonly called the weevils represent the most species-rich superfamily-level taxon of living organisms (Anderson 1995), containing approximately 62,000 described species globally. However, estimates partly based upon extrapolations of regional sampling, where ratios of undescribed to described species were calculated, suggest that the actual number of extant weevil species is much higher, in the order of 220,000 (Oberprieler *et al.* 2007). Weevils are overwhelmingly phytophagous, with species recorded feeding on a vast number of plant species and on all types of living, dying and dead plant tissues across all terrestrial and subaquatic habitats, wherever there is vegetation (McKenna et al., 2009). A very few exceptions exist, and include one species that is truly entomophagous, several species that feed within the galls of Hymenoptera and other weevils, and at least two coprophagous species (Zwölfer 1969).

Approximately 9200 species of weevils (almost 15% of all those described) belong to three subfamilies (Scolytinae, Platypodinae and Cossoniae), commonly known as bark and ambrosia beetles. These are specialist wood-borers (Oberprieler *et al.* 2007), having apparently reversed the typical phytophagous life of most weevils (some other subfamilies do also contain wood-feeding species, however) to become primary decomposers of dead and dying wood (Jordal *et al.* 2011). These beetles typically construct tunnels and galleries in woody tissue and under bark, in which they feed and breed (in the case of Platypodinae, cultivating symbiotic fungi as a food source), and therefore include many species that are forestry pests. The resulting galleries can expose trees to infection from pathogenic fungi carried by the bark beetle vectors, which include the infamous Dutch elm disease (caused by fungi of the genus *Ophiostoma*) transmitted by beetles of the genus *Scolytus*. The wood-boring habit is associated with patterns of subsociality in a few lineages (Jordal et al. 2011) within which even eusociality has evolved in at least one species (Kent & Simpson 1992), ensuring that these beetles are fascinating from an evolutionary perspective.

There has been much debate about, and research into, understanding the relationships amongst the wood-boring lineages. Generally, purely morphological analyses group the three subfamilies closely (e.g. Kuschel 1995; Kuschel et al. 2000), suggesting a single origin of wood-boring behaviour, although study of larval characters does not support this (Marvaldi 1997). Purely molecular studies on the other hand, tend to suggest multiple origins for this trait (e.g. Haran et al. 2013), although poor taxon sampling has hitherto hindered robust inferences. When both data types are combined, analyses have hinted at a possible sister relationship between Scolytinae and Platypodinae (Jordal et al. 2011), leaving this aspect of weevil evolution very much still open to investigation.

Within the Curculionoidea, one family, the nominate Curculionidae, contains 4600 genera and in excess of 80% of all weevil species (Oberprieler *et al.* 2007). The evolutionary success (measured as high species diversity) of this family has been explained as resulting from multiple independent shifts in resource-use by ancestral lineages, especially shifts from the use of gymnosperm to angiosperm host-plants and the subsequent specialisation on angiosperm structures following morphological and ecological adaptations in the weevils (McKenna *et al.* 2009). That phytophagous beetle co-evolution with angiosperms led to an increase in beetle diversity was first reported using molecular data (combined with morphological characters) by Farrell (1998), who hypothesised that "repeated origins of angiosperm-feeding beetle lineages are associated with enhanced rates of beetle diversification".

An interesting peculiarity of some weevils is that, although generally very rare within the Coleoptera as a whole, parthenogenetic lineages are comparatively common within certain groups, most notably in the genus *Otiorhynchus* (Curculionidae: Entiminae), containing several apomictic species which may be triploid or tetraploid (Suomalainen 1940). Parthenogenetic species also have diploid sexual forms and it has been shown that parthenogenesis can originate multiple times from different diploid lineages of the same species and that diploid asexuals can also exist (Stenberg *et al.* 2003).

Historically, weevils have long held the fascination of naturalists, not only due to their astounding diversity but also on account of their much varied life histories. The behavior of several well-known European species was described at great length and in minute detail almost a hundred years ago (Fabre 1922) and such traits as sexual dimorphism in rostrum length and possession of horns in the males of some species was already well known to Darwin (1871) and Wallace, who observed their use in fights between rival males for access to a female (Wallace 1869). Other interesting aspects of weevil biology, such as some species' ability to stridulate, and the physical mechanism necessary to accomplish this, were noted by Wollaston (1860).

There remain many good reasons for studying the evolution of weevils, such as the fact that due to their close association with plants, the group contains many actual or potential economically important species, both harmful (pests of crops and ornamental plants, vectors of fungal diseases *etc.*) and beneficial (biological control agents, pollinators *etc.*) to humankind. Although this close association with plants is very interesting from a co-evolutionary perspective, offering opportunities to study speciation under such a system, another pertinent reason for studying weevil systematics is to establish stability in weevil classification through the discovery and confirmation of natural groups, achieved by testing their validity within a phylogenetic conceptual framework.

1.2 Weevil classification and phylogeny to date

The classification of weevils dates back to the very birth of systematics, to the publication of the tenth volume of Linnaeus' *Systema Naturae*, 256 years ago (Linnæus 1758), within which the pioneering Swedish taxonomist described and established Latin binomial names for 79 species in the genus '*Curculio*'; already then the weevils were the largest group within the Coleoptera. In this respect, nothing has changed since. From that time until now, not only have tens of thousands of additional species been described, but a continuous 'evolution' of classification schemes has taken place, which has expanded the number of families within Curculionoidea to between seven and 22 (Oberprieler *et al.* 2007).

Following Linnæus, the first truly significant step towards a higher-level organisation of the weevils was undertaken by Schoenherr (1826) who proposed a hierarchical structure of family-groups and divided the weevils into two large sections, the Orthoceri (those with straight antennae) and the Gonatoceri (those with geniculate antennae), but the Scolytinae and Platypodinae were excluded from his 'Curculionides'. The next main advance came about through Lacordaire's (Lacordaire 1863) introduction of another main division of the weevils into the Adelognatha (akin to today's 'broad-nosed weevils') and the Phanerognatha (today's 'long-nosed weevils). The broad-nosed weevils represent a very diverse group that originally consisted of several families *sensu* Alonso-Zarazaga and Lyal (1999), including Brachyceridae and Ithyceridae, as well as several subfamilies within the Curculionidae (*e.g.* Entiminae, Cyclominae, Hyperinae), but which were shown to be paraphyletic according to a study based on adult and larval morphological characters (Marvaldi 1997).

Crowson (1955) established what is generally accepted as the basis of the modern classification by including the Scolytinae and Platypodinae into the Curculionidae and recognising nine weevil families at that time, whilst also stating that "A satisfactory resolution of the Curculionidae into subfamilies and tribes is probably the largest and most important outstanding problem in the higher classification of Coleoptera".

Since then, two classification schemes that are currently widely adopted are those of Alonso-Zarazaga and Lyal, published in their important catalogues of curculionoid family-level and genus-level taxa (Alonso-Zarazaga & Lyal 1999; Alonso-Zarazaga & Lyal 2009), and that proposed by Oberprieler *et al.* (2007) which took into account more recent phylogenetic results. The two schemes are summarised in Table 1.1, and differ mainly in that Alonso-Zarazaga & Lyal 'split' the superfamily into more families (a total of 22) than Oberprieler *et al*., who retain only seven families. Recently an updated classification, heavily based upon Alonso-Zarazaga & Lyal's catalogues, has been proposed by Bouchard *et al.* (2011).

Higher-level phylogenetic studies within the Curculionoidea began with morphological analyses, of which some of the most recent and important ones were conducted by Kuschel (1995), Marvaldi and Morrone (2000) and Legalov (2006). Thompson (1992) published a study of morphological characters used in the

Alonso-Zarazaga & Lyal (1999 and 2009)		**Oberprieler *et al.* (2007)**	
Families	**Subfamilies of Curculionidae**	**Families**	**Subfamilies of Curculionidae**
Anthribidae	**Bagoinae**	Anthribidae	Baridinae
Apionidae	**Baridinae**	Attelabidae	Brachycerinae
Attelabidae	Brachyceropseinae	Belidae	Cossoninae
Belidae	**Ceutorhynchinae**	Brentidae	Curculioninae
Brachyceridae	**Conoderinae**	Caridae	Cyclominae
Brentidae	**Cossoninae**	Curculionidae →	Dryophthorinae
Cryptolaryngidae	**Cryptorhynchinae**	Nemonychidae	Entiminae
Eccoptarthridae (=**Caridae**)	**Curculioninae**		Molytinae
Curculionidae →	**Cyclominae**		Scolytinae
Dryophthoridae	**Entiminae**		Platypodinae
Eobelidae †	**Hyperinae**		
Erirhinidae	**Lixinae**		
Eurhynchidae	**Mesoptillinae**		
Ithyceridae	**Molytinae**		
Nanophyidae	**Orobitidinae**		
Nemonychidae	**Xiphaspidinae**		
Obrieniidae †	**Scolytinae***		
Oxycorynidae	**Platypodinae***		
Raymondionymidae			
Rhynchitidae			
Ulyanidae †			

Table 1.1 Summary of two widely adopted family-level classifications in the Curculionoidea. The division of the most diverse family Curculionidae into subfamilies under both schemes is shown in grey. Under the columns for the scheme of Alonso-Zarazaga and Lyal (1999) names in bold denote families and subfamilies also recognised in the more recent, though very similar, classification of Bouchard *et al.* (2011). Other family-level taxa have been reduced in rank in the latter work. † Extinct taxon

classification of weevils, with a dichotomous key to families and subfamilies included, but without attempting a phylogenetic reconstruction. The above three phylogenetic studies based on morphological characters share some fundamental results, including the repeated basal placement of the families Nemonychidae and Anthribidae (either as sister taxa or not) as sister to the remaining weevils, with the families Belidae, Attelabidae and Caridae filling the intermediate part of the tree and the two evidently most derived families Brentidae and Curculionidae forming a sister relationship at the apical clade.

In the past 15 years or so, molecular sequence data has increasingly been used either in conjunction with, or independently of, morphological characters in the reconstruction of higher-level relationships within the superfamily Curculionoidea. Both nuclear and mitochondrial marker sequences have been incorporated into analyses, although until recently most studies have focused on one or two genes, with the nuclear 18S ribosomal RNA (rRNA) being most commonly employed across studies (Table 1.2). Notable exceptions include the recent study by McKenna *et al.* (2009) which used sequence data from up to six markers to investigate the contemporaneity of weevil and angiosperm diversification, and the detailed study into the evolution of the wood-boring weevil lineages by Jordal *et al.* (2011) who incorporated five nuclear and mitochondrial markers as well as morphological characters in a comprehensively sampled phylogeny of these groups. The most recent molecular phylogenetic reconstruction of weevil relationships (Haran *et al.* 2013) used a Long-Range PCR approach to obtain long mitochondrial genomic sequences (12 protein-coding genes) and generally supported the findings of McKenna *et al.* (2009) in establishing a sound foundation of the family-level relationships, though because of limited taxon sampling, subfamilial and lower relationships could not be thoroughly tested. Molecular studies have to some extent produced congruent results to morphological analyses, regularly placing the gymnosperm feeding, primitive Nemonychidae and the Anthribidae as sister to all other Curculionoidea, and importantly, consistently placing the Brentidae sister to a monophyletic Curculionidae (*sensu* Oberprieler, 2007).

Author/year of publication	Nº. taxa	18S rRNA (nuclear)	28S rRNA (nuclear)	*rrnL* (mitochondrial)	*cox1* (mitochondrial)	EF- 1α (nuclear)	*ArgK* (nuclear)	CAD (nuclear)	Mitochondrial genomes	Morphological characters
Wink *et al.* (1997)	32			480						0
Farrell (1998)	115	~1850								212
Hunt *et al.* (2007)	222	~1927		501	724					0
Marvaldi *et al.* (2002)	100	2153								115
Marvaldi *et al.* (2009)	96	~1000	480-785							0
McKenna *et al.* (2009)	135	Y *****	Y*	Y*	Y*	Y*	Y*			0
Hundsdoerfer *et al.* (2009)	148	~2000		~500						0
Jordal *et al.* (2011)	105		778		525	836	801	675		128
Haran *et al.* (2013)	27								12 genes	0

Table 1.2 Higher-level phylogenetic studies incorporating molecular sequence data, listed by author. The markers used in each analysis (and where known the number of nucleotides incorporated in the matrix in bp) and the number of morphological characters (where used) are also listed. * A total of ~8000 bp across all six genes.

To illustrate these similarities, two recent higher-level phylogenies reconstructed from sequence data are shown in Figures 1.1 and 1.2 (Marvaldi *et al.* 2009; McKenna *et al.* 2009). Whilst largely congruent, these disagree in the exact relationships amongst the basal families, Anthribidae and Nemonychidae, which were not recovered as monophyletic in the phylogeny of McKenna *et al.* (2009). The study by Marvaldi *et al.* (2009) was based only on slow-evolving nuclear rRNA genes, which could at least partly explain its success in resolving these deeper nodes. Neither of these studies contained sufficiently dense taxon sampling within the Curculionidae to effectively test for monophyly of its constituent subfamilies.

Weevils have also been incorporated into much broader phylogenies, including both very recent purely morphological- (Lawrence *et al.* 2011) and purely molecular-based (Bocak *et al.* 2013) phylogenies of the Coleoptera as a whole. The former study analysed 516 adult and larval characters for 314 beetle families and subfamilies and recovered a monophyletic Curculionidae *sensu* Bouchard *et al.* (2011)

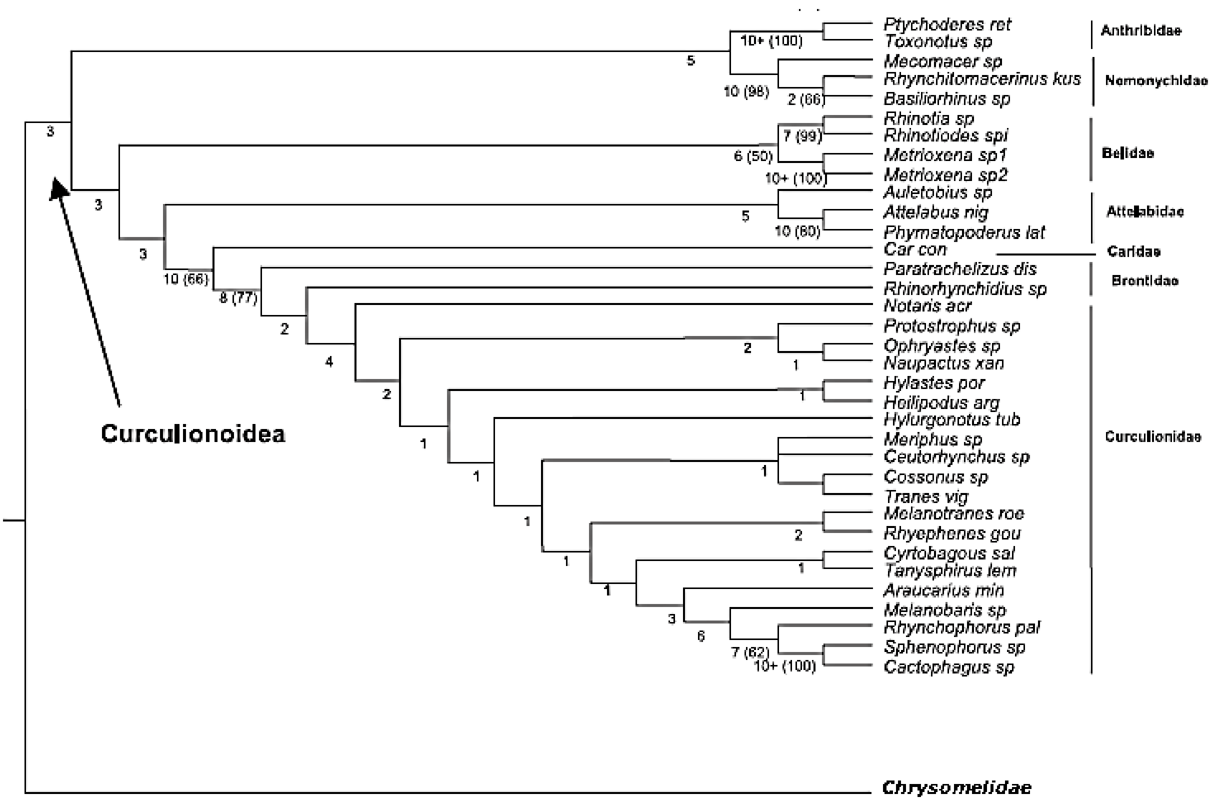


Figure 1.1 Phylogenetic relationships within the Curculionoidae reconstructed from combined 18S and 28S rRNA data. This tree is modified from a portion of a wider phylogenetic analysis of the Phytophaga (Chrysomelidae is shown as the sister group to Curculionoidea) by Marvaldi *et al.* (2009). Most weevil families (with the exception of Brentidae) are shown to be monophyletic. Taxon sampling is not dense enough to thoroughly test relationships within the family Curculionidae. Numbers below nodes are Bremer support values with Bootstrap values >50% shown in brackets.

containing a clade of Scolytinae + Platypodinae (Lawrence *et al.* 2011). In the latter study, incorporating sequences obtained from the GenBank public database from two nuclear and two mitochondrial loci for 8441 species of Coleoptera in 152 families, the Curculionidae was also retrieved as monophyletic and the inter-family relationships of Curculionoidea were similar to those recovered by McKenna *et al.* (2009) placing the ancestral gymnosperm feeding Nemonychidae and the Anthribidae at the base and the Brentidae and megadiverse Curculionidae as the most highly derived sister taxa, although differing somewhat in the ordering of family lineages compared to Haran *et al.* (2013).

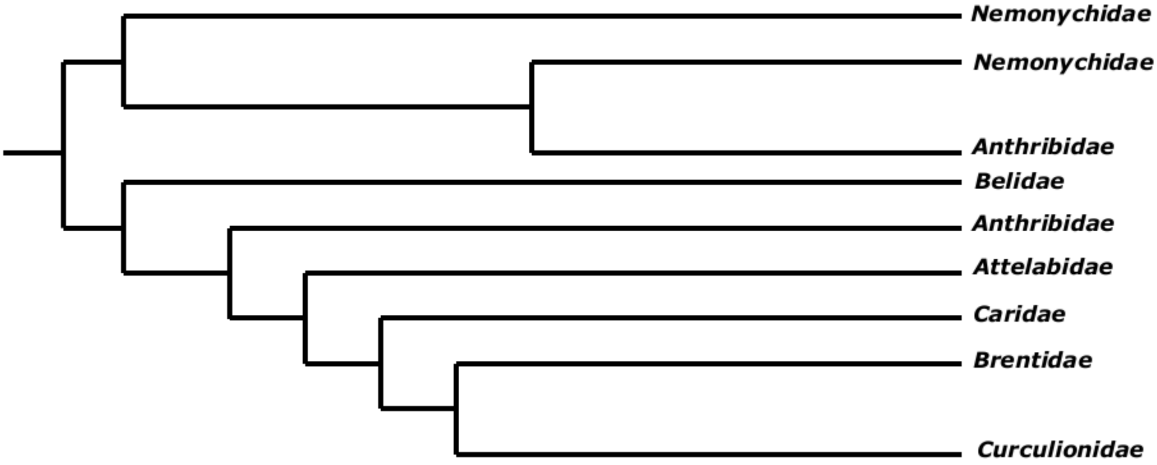


Figure 1.2 Simplified family-level Curculionoidea relationships calculated from Bayesian analysis of sequences from six gene markers by McKenna *et al.* (2009). All families except Nemonychidae and Anthribidae are retrieved as monophyletic.

That the Curculionidae has been repeatedly shown to be monophyletic does not however indicate that the relationships *within* this family have been satisfactorily resolved. In fact the inter-subfamilial and tribal phylogenetic relationships remain largely untested and unexplored, certainly from a DNA sequence data viewpoint. It is advantageous therefore, that the Curculionidae has been hypothesised to be a monophyletic group, thereby giving a sound basis for phylogenetic investigation of its constituent higher groups (subfamilies, tribes, subtribes etc.).

Some authors have argued against the undertaking of higher-level phylogeny reconstruction in the weevils, primarily citing limited taxon sampling in past studies as having been a major hurdle in being able to infer meaningful explanations for weevil evolutionary success across the entire superfamily (Franz & Engel 2010).

Narrower tribal- and generic-level studies have been advocated as alternatives to achieve this, although without incorporating a wide range of lineages themselves, thereby allowing for testing of lineage affinities, it seems clear that such reconstructions will also be prone to criticism. Regardless of these opinions, a better understanding of the phylogeny within Curculionidae will provide for a solid foundation for weevil classification, which is in a state of confusion at present.

1.3 Molecular markers employed in weevil systematics

Molecular markers used for phylogenetic studies should have certain properties making them suitable for the resolution of the particular taxonomic level or rank of investigation. Chief among these is the requirement that the nucleotide substitution or mutation rate will provide a balance between being sufficiently high to result in a suitable number of informative (variable across sequences) sites, whilst also ensuring that the rate is not so high as to lead to extensive saturation of the nucleotide sites, thereby masking phylogenetic signal. Studies looking at relatively closely related organisms (*e.g.* at the species-level within a genus) will require genes having fast substitution rates to ensure that enough informative mutations have accumulated within the relatively short time that these species have been diverging. Conversely, if looking at more distantly related organisms (*e.g.* family-level or higher), genes with slower substitution rates are required, which will minimise the effects of saturation, providing signal for these older divergences. Another vital property that is required is that the gene in question should ideally be present in a single copy or in multiple homogenous copies in the haploid genome (Cruickshank 2002) to avoid the possibility of paralogous sequences being obtained from different gene copies in different individuals. Whilst mitochondrial genes are present in many copies per cell, they usually contain the same sequence in any one individual (Cruickshank 2002). Other considerations will include how difficult sequences are to align. Protein coding sequences are easier to align than rRNA genes because gaps in their sequences will occur in multiples of three nucleotides (the amino acid codons) to prevent frameshift mutations, whereas rRNA, which is not translated, consists of both highly conserved and highly variable (both in nucleotide composition and sequence length) sections due to the three dimensional structure these molecules take up (Marvaldi *et al.* 2009). The overall proportion or base composition in a gene can also be important (*e.g.* mitochondrial DNA is generally A-T rich) and phylogenies inferred from sequences that differ greatly in base composition may group samples together based on this (Foster & Hickey 1999) rather than common descent. Some models of molecular evolution attempt to address this issue (Page & Holmes 1998).

Of lesser importance is knowing whether certain markers have been used 'successfully' to resolve relationships in previous studies on similar organisms, and that primers are already available to target the fragment of interest. Arguably using the same markers across studies will make such studies more easily comparable and can allow for sequences to be used together in future broader analyses (Cruickshank 2002).

It is unlikely that a single gene fulfilling all or even most of the above requirements exists and in general, especially for trying to establish higher-level relationships, it is expected that phylogenetic reconstructions based on data from multiple independent sources will result in more robust results.

1.3.1 Summary of loci used in weevil higher-level phylogeny reconstruction

To date seven individual mitochondrial and nuclear loci have been used to reconstruct weevil higher-level phylogeny as well as one study using mitochondrial genomes (Table 1.2). Some of these markers are briefly reviewed below.

1.3.2 Mitochondrial genes

These genes can be separated into two groups: rRNA and protein-coding genes.

1.3.2.1 Mitochondrial rRNA genes

16S rRNA (rrnL)

After the nuclear 18S rRNA gene, this has been the most widely adopted marker for reconstructing weevil phylogenies. The *rrnL* gene contains more variable sites than 18S rRNA, and this was evidently the main reason that Hundsdoerfer *et al.* (2009) decided to employ it, in conjunction with the more conserved nuclear 18S, in an attempt to obtain better resolution of the subfamilies within Curculionidae (although their taxon sampling was not dense enough to thoroughly test this) in a wider study of the Curculionoidea. Broadly their results indicated that using *rrnL* data alone recovered many of the accepted tribes and genera of the Curculionoidea as monophyletic, but there was insufficient phylogenetic information to robustly resolve higher-level relationships (inter-subfamily and above), which remained highly dependent on the alignment and reconstruction method used. Although the combination of the *rrnL* data with 18S was expected to improve resolution of higher-level clades, these relationships remained sensitive to alignment and reconstruction methodology despite this (Hundsdoerfer *et al.* 2009). In an earlier study, Wink *et al.* (1997) used *rrnL* sequence data exclusively to infer higher-level phylogeny. Their results recovered several of the more primitive families (Nemonychidae, Anthribidae *etc.*) as monophyletic, but did not place them at the base of the tree. In general there was low support for most deep nodes and the authors concluded that *rrnL* was not conservative enough to resolve these deeper branches, suggesting that *rrnL* data should be combined with 18S data, as Hundsdoerfer *et al.* (2009) did later.

Sheffield *et al.* (2008) compared *rrnL* with the other mitochondrial rRNA gene, 12S rRNA, which had hitherto not been used in curculionoid phylogeny. They concluded that *rrnL* is more variable in length than 12S, containing several stems with variable sequences and lengths. The less variable 12S is potentially a candidate for higher-level phylogenetic study in the Curculionidae.

1.3.2.2 Mitochondrial protein-coding genes

Cytochrome oxidase subunit 1 (cox1)

The gene encoding this protein has a fast substitution rate and has been traditionally used (though certainly not exclusively) in lower-level phylogenies investigating relationships amongst relatively closely related species that have recently diverged. In weevils, *cox1* sequences have been used in genus-level phylogenies, either independently, as in the phylogeny of the scolytine genus *Ips* (Cognato & Sperling 2000) or together with other genes, such as in the phylogeny of the genus *Curculio* (Hughes & Vogler 2004) which also utilised an additional mitochondrial and two nuclear genes. Used independently, *cox1* can potentially be found to be effectively too saturated to resolve internal nodes if the studied group is too divergent, leading to reduced phylogenetic support for internal nodes relative to peripheral clades (Cognato & Sperling 2000). However, in other cases *cox1* has been shown to have utility in resolving phylogenies beyond very closely related species. For example, despite its high substitution rate and high A-T bias, *cox1* performed better than 12S or 16S in resolving subgeneric-level relationships in tetranychid ticks (Navajas *et al.* 1996). An approach that is sometimes taken to reduce the effect of saturation is to exclude from the analysis the third base of each codon, which may be saturated as a result of the degenerate genetic code.

For higher-level phylogenetic studies, *cox1* is usually used in combination with other genes. Several higher-level reconstructions containing Curculioniodea have used this gene, *e.g.* Hunt *et al.* (2007) who used a combined matrix of sequences from three genes to investigate the relationships across nearly all beetle families, and McKenna *et al.* (2009) who studied the Curculionoidea exclusively.

In addition to its utility in phylogenetic analyses, *cox1* has become the gene of choice for DNA barcoding initiatives (Hebert *et al.* 2003). It is important to distinguish between barcoding, the aim of which is species identification, and phylogeny which has the aim of reconstructing evolutionary history. *Cox1* is particularly suitable to both because it is universal across eukaryotes, has a high mutation rate and lacks introns, but also because universal primers are available allowing for the comparatively easy amplification of the gene across diverse lineages. It has been shown to be both cost-effective and accurate (Hebert *et al.* 2003). More recently, *cox1* sequences have been successfully used in automated species delimitation and discovery (Pons *et al.* 2006). They are also regularly incorporated into integrative taxonomy schemes, including the so-called "turbo-taxonomy" approaches which combine *cox1* sequences, short morphological species descriptions, digital photographs and simultaneous publishing of new taxa in both traditional printed media and in closely-tied open access online databases (Riedel *et al.* 2013a). One of the first examples of such an undertaking involved the mass description of 101 new species of a diverse genus of weevils from Papua New Guinea (Riedel *et al.* 2013b).

1.3.3 Nuclear genes

Those genes so far used for weevil phylogeny reconstruction can also be separated into two groups: rRNA and protein-coding genes.

1.3.3.1 Nuclear ribosomal RNA genes

18S rRNA

This gene has been the most widely employed molecular marker in weevil higher-level phylogeny (six studies have used it, Table 1.2). As discussed above, this slowly-evolving gene was used by Hundsdoerfer *et al.* (2009) in combination with the faster-evolving *rrnL*, ultimately with results that were largely dependent on the sequence alignment and inference method. The major problem with rRNA genes is the difficulty in sequence alignment due primarily to length differences in hypervariable sites associated with maintaining the functional secondary and tertiary structure of the rRNA molecule, which is more conserved than the nucleotide sequences it is composed of. This can make it computationally impractical to align sequences. However, through a comparative approach of these difficult-to-align regions, Marvaldi *et al.* (2009) were able to obtain good alignments of 18S and 28S sequences based on secondary structure information. Their subsequent phylogenetic analysis using the combined markers yielded not only highly congruent trees using parsimony and Bayesian inferences, but also recovered relationships amongst the seven weevil families *sensu* Oberprieler *et al.* (2007) that agree considerably with current morphological hypotheses.

28S rRNA

In addition to the study by Hundsdoerfer *et al.* (2009), the 28S rRNA gene was used as a phylogenetic marker by McKenna *et al.* (2009), together with five other genes, to investigate curculionoid higher-level phylogeny and co-diversification with angiosperms. The resulting relationships at the family-level (more basal nodes) were mostly strongly supported and in agreement with current hypotheses, including a monophyletic Curculionidae, suggesting that the combined marker signal was informative for this level. However at a lower level, all but one subfamily was recovered as para- or polyphyletic. The authors suggest that increased taxon sampling within the Curculionidae subfamilies will contribute to better nodal support and resolution for these groups (McKenna *et al.* 2009).

It therefore seems clear that nuclear rRNA genes can provide important resolving power for deeper nodes at the family-level and higher within the Curculiuonidae, but this information is best accessed after careful consideration of the alignment difficulties and when the markers are used in combination with other genes.

1.3.3.2 Nuclear protein-coding genes

Recently greater emphasis has been placed on the utilisation of nuclear protein-coding genes in invertebrate systematics and the large number of such genes in the nuclear genome is a potentially rich source of phylogenetic data (Wild & Maddison 2008). Some of the positive features of such genes include their slower substitution rates and lesser susceptibility to base-composition bias when compared to mitochondrial genes. They are also generally much easier to align than ribosomal genes. However on the negative side, these genes can be present in multiple paralogous copies in the genome, decreasing their phylogenetic utility. Their sequencing and alignment can also be complicated by the presence of long introns (Wild & Maddison 2008). The substitution rates of different protein-coding genes show considerable variation, thereby potentially making them suitable for studying relationships within or between a variety of taxonomic ranks. To investigate lower-level relationships, the very variable intron sequences can potentially be exploited by designing primers within the conserved coding exons that span the introns of interest (Cruickshank 2002).

Thus far only three nuclear protein-coding genes have been utilised in weevil higher-level phylogeny (Table 1.2); two of these are briefly reviewed below.

Arginine kinase (ArgK)

This gene codes for a phosphotransferase enzyme involved with the regulation of metabolism (Wild & Maddison 2008). It is present in the majority of animal groups and is relatively conserved, with a rate of sequence divergence estimated at being six times lower than that of *cox1* (Mahon & Neigel 2008), indicating that it should be more informative in resolving deeper nodes than *cox1*. In a test of its utility in reconstructing phylogeny, Wild and Maddison (2008) used *ArgK* sequence data to reconstruct relationships among a diverse selection of beetles of known phylogeny. It was found that *ArgK* was able, with slightly more robustness, to reconstruct deeper nodes than more recent ones. The amplification of *ArgK* has been shown to be fairly straightforward and no major alignment problems arose because introns appear to be rare in the studied fragment. Additionally, no evidence of paralogous copies of the gene within Coleoptera has surfaced (Wild & Maddison 2008).

Another study explored the utility of *ArgK* to reconstruct the phylogeny of crabs (Brachyura) and reported that strongly supported monophyletic clades were found to represent known genera, but that above this level, several subfamily relationships were recovered as polyphyletic (Mahon & Neigel 2008). It was also discovered that nodes in phylogenies reconstructed from *ArgK* data were better supported than those recovered using *cox1* data. It was suggested that *ArgK* data be combined with that of other sequences having different substitution rates in order to improve phylogeny robustness.

Elongation factor-1α (EF1- α)

This gene promotes the GTP-dependent binding of aminoacyl-tRNA to ribosomes. It has been informative in a range of arthropod and insect phylogenetic studies including the multi gene analysis of Curculionoidea by McKenna *et al.* (2009). EF1- α was used to test the monophyly of a tribe of bark beetles (Scolytinae: Xyleborini), during which the gene was found to be present in two copies in these beetles, each differing by the number of introns (Jordal 2002) although the data suggested that each copy is orthologous and predates the origin of beetles. The resulting phylogeny weakly supported a monophyletic Xyleborini and was better resolved than previous studies using mitochondrial and other nuclear genes. It was also found that inclusion of sequences from variable intron regions proved informative at this level (Jordal 2002).

EF1- α has also been utilised in the much deeper-level phylogeny inference of pterygote insect orders (Simon *et al.* 2010), where it was found to be informative in resolving these older relationships. The arrangement of exons and introns witin EF1-α requires consideration because this might reflect phylogenetic relationships across insect orders (Brady & Danforth 2004). Whilst accepting that paralagous copies of the gene could present problems, these are less likely to affect phylogenies at such a high level.

Other protein-coding genes

Although only *ArgK*, EF1- α and CAD have so far been employed in weevil systematics, several other nuclear protein-coding genes have been investigated for phylogenetic utility in beetles (Wild & Maddison 2008). Of the eight other genes studied (Alpha Spectrin, CAD, Enolase, PEPCK, RNA polymerase II, Topoisomerase and *Wingless*), all were found to be informative for recent divergences within a subfamily and were also able to delineate species-level taxa. However, at deeper levels the phylogenetic utility differed considerably among them, indicating that care needs to be taken when selecting an appropriate marker for a particular level of analysis. As with other nuclear genes, problems were encountered with the presence of introns that can lead to difficulties in amplification and sequencing.

1.3.4 New approaches

In addition to the more traditional use of molecular data outlined above, two recent techniques that each take different approaches to obtaining sequence information have been recently successfully developed.

1.3.4.1 Short phylogenetically informative amplicons (SPIAs)

This technique is related to but distinct from the so-called mini-barcode approach. The latter capitalises on the fact that although full length DNA barcodes offer the best chance of species identification (the main aim of DNA barcoding), using much shorter sequences (100 bp or less) can result in correct species assignment (Meusnier *et al.* 2008). However the aim of SPIAs is to use information from these short sequences to contribute towards the reconstruction of phylogenetic relationships rather than purely for species identification. Therefore SPIAs are not restricted to the *cox1* barcode region and can be obtained from other genes and either used separately or concatenated together to provide sequence information for analysis. Through the use of this method it has been demonstrated that incorporating SPIA data into an existing robust phylogenetic 'scaffold' can enable the assignment of these taxa to particular lineages (Hernández-Vera *et al.* 2013).

Whilst longer fragments are obviously preferable, in some instances it may be difficult or impossible to obtain full length sequences from highly interesting specimens. One such case is that of specimens housed in the rich collections of natural history museums. Because DNA degrades spontaneously over time (Lindahl 1993) especially in dry material, such specimens may contain sheared DNA of low integrity (*i.e.* short fragments) prohibiting or greatly reducing the effectiveness of standard PCR success. The importance of such specimens lies in the fact that museum collections house an enormous quantity of specimens, many of which represent very rare, or perhaps even taxa already extinct in nature (Payne & Sorenson 2002). Potentially even type specimens, the actual bearers of each species name, may be sequenced and used in phylogenetic studies, thereby allowing for the unequivocal assignment of taxa. Museums hold material of species from across their historical geographic ranges, which may have since contracted dramatically, though each specimen, even if a hundred or more years old, will still contain some of its genetic information.

One important consideration is that because natural history museums exist to document diversity and to conserve and study their comprehensive collections for present and for future generations, care must be taken to limit damage to specimens under study. With the development of improved high-yielding extraction protocols, it is now quite possible to extract sufficient DNA non-destructively from museum specimens by immersing them whole or in part in extraction buffer (Gilbert *et al.* 2007) without resorting to visible physical damage. However, within Coleoptera at least, there have hitherto been few phylogenetic studies successfully incorporating sequences thus obtained (Hernández-Vera *et al.* 2013).

1.3.4.2 Whole or partial mitochondrial genome sequencing

The mitochondrial genome (mitogenome) of beetles is a circular double-stranded molecule containing approximately 16,000 nucleotides, encoding 13 proteins, 22 transfer RNA molecules and two rRNA molecules (Boore 1999). Mitogenome sequence data has successfully been utilised to reconstruct phylogenies in several animal groups (Botero-Castro *et al.* 2013), although until recently the acquisition of these sequences remained expensive and difficult, being generally achieved through many PCR amplifications of adjoining or overlapping sections (Botero-Castro *et al.* 2013). However, with the recent advent and adoption of next-generation sequencing (NGS) technologies, which allow for a much higher number of DNA molecules to be characterised compared to conventional Sanger sequencing (Grada & Weinbrecht 2013), the sequencing of mitogenomes has now become an affordable reality.

One of the challenges of using NGS for phylogenetics research is in being able to analyse multiple samples (specimens) whilst also being able to associate generated sequences with a particular individual (Timmermans *et al.* 2010). In a recent study investigating the practicalities of these techniques, Timmermans *et al.* (2010) used primers designed to be specific to conserved regions of mtDNA, to undertake Long-

Range PCR, obtaining long (up to 10000+ bp) amplification products which were pooled and sequenced using the Roche 454 platform. Sequenced reads were assembled into partially complete mitogenomic contigs and through the use of conventionally amplified PCR sequences belonging to three mitogenomic genes they were able to correctly identify to which species each contig belonged to. This technique was subsequently employed by Haran *et al.* (2013) to obtain 27 near-complete curculionoid mitogenomes, the analysis of which supported many of the family-level relationships that have been recovered in previous molecular phylogenies, thereby demonstrating that this new approach can be successfully employed to obtain multiple mitogenomic sequences and to accurately assign them to samples. Analysis of Coleoptera mitogenomes has shown that nucleotide substitution rates vary across their constituent genes, such that the effect of data-partitioning schemes (*e.g.* by gene or codon position) and model choices used in phylogeny reconstruction requires careful consideration (Pons *et al.* 2010).

1.4 Thesis aims and structure

The global aim of the research undertaken for this thesis is to obtain molecular sequence data from a wide range of loci and lineages, to use this to build phylogenetic reconstructions of the higher-level relationships in order to infer, test and define hypothesised natural evolutionary lineages within the superfamily Curculionoidea.

Higher-level phylogeny, specifically refers to the phylogenetic relationships amongst families, subfamilies and tribes within the Curculionoidea, and especially those relationships within Curculionidae *s.str.*, numerically the most important family.

A number of potentially informative nuclear and mitochondrial markers are investigated in Chapter 2 for their amplification and sequencing success across weevils. This included two strategies focusing on opposing ends of sequence length: obtaining SPIAs from old degraded museum specimens and obtaining long mitogenomic sequences from fresh material through NGS of both Long-Range PCR products and direct genomic DNA sequencing.

Chapter 3 explores an exciting new bioinformatics methodology for efficiently and economically obtaining large numbers of mtitogenomes through NGS , offering an opportunity to greatly improve upon the taxon sampling of previous studies and to investigate both the family-level relationships and subfamily affinities within the Curculionidae. This newly obtained dataset included a densely sampled assemblage of the wood-boring lineages, allowing for scrutiny of the evolution of this ecological trait.

In Chapter 4 a formal assessment is made of the benefits, if any, of supplementing mitogenomic sequence data with sequences from additional nuclear rRNA and protein-coding genes, as measured by the effect on combined nodal support across the resulting trees.

Public sequence databases are exploited in Chapter 5, through a bioinformatics sequence retrieval pipeline, used to increase taxon sampling in the broad-nosed weevils in order to investigate the relationships among their constituent subfamilies and tribes. This was achieved by exploring the ability of shorter sequences from two single mtDNA loci, incorporated into a ‘backbone’ mitogenomic dataset, to match to their closest lineages as evaluated by statistical tests of monophyly.

Finally, in Chapter 6 the main conclusions from across the studies in this thesis are drawn together and briefly discussed, whilst possible future directions for weevil molecular phylogenetics research are identified and considered.

1.5 References

Alonso-Zarazaga MA, Lyal CHC (1999) *A world catalogue of families and genera of Curculionoidea (Insecta: Coleoptera) (excepting Scolytidae and Platypodidae)* Entomopraxis, Barcelona.

Alonso-Zarazaga MA, Lyal CHC (2009) A catalogue of family and genus group names in Scolytinae and Platypodinae with nomenclatural remarks (Coleoptera: Curculionidae). *Zootaxa* **2258**, 1-134.

Anderson RS (1995) An evolutionary perspective on diversity in Curculionoidea. *Memoirs of the Entomological Society of Washington* **14**, 103-114.

Bocak L, Barton C, Crampton-Platt A*, et al.* (2013) Building the Coleoptera tree-of-life for >8000 species: composition of public DNA data and fit with Linnaean classification. *Systematic Entomology* **39**, 97-110.

Boore JL (1999) Animal mitochondrial genomes. *Nucleic Acids Research* **27**, 1767-1780.

Botero-Castro F, Tilak M-K, Justy F*, et al.* (2013) Next-generation sequencing and phylogenetic signal of complete mitochondrial genomes for resolving the evolutionary history of leaf-nosed bats (Phyllostomidae). *Molecular Phylogenetics and Evolution* **69**, 728-739.

Bouchard P, Bousquet Y, Davies AE*, et al.* (2011) Family-group names in Coleoptera (Insecta). *Zookeys* **88**, 1-972.

Brady SG, Danforth BN (2004) Recent intron gain in elongation factor-1 alpha of colletid bees (Hymenoptera : Colletidae). *Molecular Biology and Evolution* **21**, 691-696.

Cognato AI, Sperling FAH (2000) Phylogeny of Ips DeGeer species (Coleoptera : Scolytidae) inferred from mitochondrial cytochrome oxidase I DNA sequence. *Molecular Phylogenetics and Evolution* **14**, 445-460.

Crowson RA (1955) *The natural classification of Coleoptera* Nathaniel Lloyd & Co., London.

Cruickshank RH (2002) Molecular markers for the phylogenetics of mites and ticks. *Applied Aracology* **7**, 3-14.

Darwin CR (1871) *The descent of man, and selection in relation to sex* John Murray, London.

Fabre JH (1922) *The life of the weevil* Dodd, Mead and Co., New York.

Farrell BD (1998) "Inordinate fondness" explained: Why are there so many beetles? . *Science* **281**, 553-557.

Foster PG, Hickey DA (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution* **48**, 284-290.

Franz NM, Engel MS (2010) Can higher-level phylogenies of weevils explain their evolutionary success? A critical review. *Systematic Entomology* **35**, 597-606.

Gilbert MTP, Moore W, Melchior L, Worobey M (2007) DNA Extraction from Dry Museum Beetles without Conferring External Morphological Damage. *PloS one* **2**: 10.1371/journal.pone.0000272.

Grada A, Weinbrecht K (2013) Next-Generation Sequencing: Methodology and Application. *Journal of Investigative Dermatology* **133**: 10.1038/jid.2013.248.

Haran J, Timmermans MJTN, Vogler AP (2013) Mitogenome sequences stabilize the phylogenetics of weevils (Curculionoidea) and establish the monophyly of larval ectophagy. *Molecular Phylogenetics and Evolution* **67**, 156-166.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B-Biological Sciences* **270**, 313-321.

Hernández-Vera G, Caldara R, Tosevski I, Emerson BC (2013) Molecular phylogenetic analysis of archival tissue reveals the origin of a disjunct southern African-Palaearctic weevil radiation. *Journal of Biogeography* **40**, 1348-1359.

Hughes J, Vogler AP (2004) The phylogeny of acorn weevils (genus Curculio) from mitochondrial and nuclear DNA sequences: the problem of incomplete data. *Molecular Phylogenetics and Evolution* **32**, 601-615.

Hundsdoerfer AK, Rheinheimer J, Wink M (2009) Towards the phylogeny of the Curculionoidea (Coleoptera): Reconstructions from mitochondrial and nuclear ribosomal DNA sequences. *Zoologischer Anzeiger* **248**, 9-31.

Hunt T, Bergsten J, Levkanicova Z*, et al.* (2007) A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science* **318**, 1913-1916.

Jordal BH (2002) Elongation Factor 1 α resolves the monophyly of the haplodiploid ambrosia beetles Xyleborini (Coleoptera: Curculionidae). *Insect Molecular Biology* **11**, 453-465.

Jordal BH, Sequeira AS, Cognato AI (2011) The age and phylogeny of wood boring weevils and the origin of subsociality. *Molecular Phylogenetics and Evolution* **59**, 708-724.

Kent DS, Simpson JA (1992) Eusociality in the beetle *Austroplatypus incomptus* (Coleoptera: Platypodidae). *Naturwissenschaften* **79**, 86-87.

Kuschel G (1995) A phylogenetic classification of Curculionoidea to families and subfamilies. *Memoirs of the Entomological Society of Washington* **14**, 5-33.

Kuschel G, Leschen RAB, Zimmerman EC (2000) Platypodidae under scrutiny. *Invertebrate Taxonomy* **14**, 771-805.

Lacordaire T (1863) *Histoire naturelle des insectes. Genera des Coléoptères* Roret, Paris.

Lawrence JF, Ślipiński A, Seago AE*, et al.* (2011) Phylogeny of the Coleoptera based on morphological characters of adults and larvae. *Annales Zoologici (Warszawa)* **61**, 1-217.

Legalov AA (2006) Phylogenetic reconstruction of weevil superfamily Curculionoidea (Coleoptera) using the SYNAP method. *Biology Bulletin* **33**, 127-134.

Lindahl T (1993) Instability and decay of the primary structure of DNA. *Nature* **362**, 709-715.

Linnæus C (1758) *Systema Naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Editio Decima, reformata.* Holmiae, Laurentii Salvii.

Mahon BC, Neigel JE (2008) Utility of arginine kinase for resolution of phylogenetic relationships among brachyuran genera and families. *Molecular Phylogenetics and Evolution* **48**, 718-727.

Marvaldi AE (1997) Higher level phylogeny of Curculionidae (Coleoptera : Curculionoidea) based mainly on larval characters, with special reference to broad-nosed weevils. *Cladistics-the International Journal of the Willi Hennig Society* **13**, 285-312.

Marvaldi AE, Duckett CN, Kjer KM, Gillespie JJ (2009) Structural alignment of 18S and 28S rDNA sequences provides insights into phylogeny of Phytophaga (Coleoptera: Curculionoidea and Chrysomeloidea). *Zoologica Scripta* **38**, 63-77.

Marvaldi AE, Morrone JJ (2000) Phylogenetic systematics of weevils (Coleoptera : Curculionoidea): A reappraisal based on larval and adult morphology. *Insect Systematics & Evolution* **31**, 43-58.

Marvaldi AE, Sequeira AS, O'Brien CW, Farrell BD (2002) Molecular and morphological phylogenetics of weevils (Coleoptera, Curculionoidea): Do niche shifts accompany diversification? *Systematic Biology* **51**, 761-785.

McKenna DD, Sequeira AS, Marvaldi AE, Farrell BD (2009) Temporal lags and overlap in the diversification of weevils and flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7083-7088.

Meusnier I, Singer GAC, Landry J-F*, et al.* (2008) A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics* **9**: 10.1186/1471-2164-9-214.

Navajas M, Gutierrez J, Lagnel J, Boursot P (1996) Mitochondrial cytochrome oxidase I in tetranychid mites: A comparison between molecular phylogeny and changes of morphological and life history traits. *Bulletin of Entomological Research* **86**, 407-417.

Oberprieler RG, Marvaldi AE, Anderson RS (2007) Weevils, weevils, weevils everywhere. *Zootaxa* **1668**, 491-520.

Page RDM, Holmes EC (1998) *Molecular Evolution: A phylogenetic approach* Blackwell, Oxford.

Payne RB, Sorenson MD (2002) Museum collections as sources of genetic data. *Bonner zoologische Beiträge* **51**, 97-104.

Pons J, Barraclough TG, Gomez-Zurita J*, et al.* (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* **55**, 595-609.

Pons J, Ribera I, Bertranpetit J, Balke M (2010) Nucleotide substitution rates for the full set of mitochondrial protein-coding genes in Coleoptera. *Molecular Phylogenetics and Evolution* **56**, 796-807.

Riedel A, Sagata K, Suhardjono YR, Taenzler R, Balke M (2013a) Integrative taxonomy on the fast track - towards more sustainability in biodiversity research. *Frontiers in Zoology* **10**: 10.1186/1742-9994-10-15.

Riedel A, Sagata K, Surbakti S, Taenzler R, Balke M (2013b) One hundred and one new species of Trigonopterus weevils from New Guinea. *Zookeys* **280**, 1-150.

Schoenherr CJ (1826) *Curculionidum dispositio methodica*, Lipsiae.

Sheffield NC, Song H, Cameron L, Whiting MF (2008) A Comparative Analysis of Mitochondrial Genomes in Coleoptera (Arthropoda: Insecta) and Genome Descriptions of Six New Beetles. *Molecular Biology and Evolution* **25**, 2499-2509.

Simon S, Schierwater B, Hadrys H (2010) On the value of Elongation factor-1a for reconstructing pterygote insect phylogeny. *Molecular Phylogenetics and Evolution* **54**, 651-656.

Stenberg P, Lundmark M, Knutelski S, Saura A (2003) Evolution of clonality and polyploidy in a weevil system. *Molecular Biology and Evolution* **20**, 1626-1632.

Suomalainen E (1940) Polyploidy in parthenogenic Curculionidae. *Hereditas* **26**, 51-64.

Thompson RT (1992) Observations on the morphology and classification of weevils (Coleoptera, Curculionoidea) with a key to major groups. *Journal of Natural History* **26**, 835-891.

Timmermans MJTN, Dodsworth S, Culverwell CL*, et al.* (2010) Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research* **38**: 10.1093/nar/gkq807.

Wallace AR (1869) *The Malay Archipelago, the land of the orang-utan and the bird of paradise* Macmillan, London.

Wild AL, Maddison DR (2008) Evaluating nuclear protein-coding genes for phylogenetic utility in beetles. *Molecular Phylogenetics and Evolution* **48**, 877-891.

Wink M, Mikes Z, Rheinheimer J (1997) Phylogenetic relationships in weevils (Coleoptera: Curculionoidea) inferred from nucleotide sequences of mitochondrial 16S rDNA. *Naturwissenschaften* **84**, 318-321.

Wollaston TV (1860) On certain musical Curculionidae; with descriptions of two new Plithini. *Annals and Magazine of Natural History* **6**, 14-19.

Zwölfer H (1969) Rüsselkäfer mit ungewöhnlicher Lebensweise. Koprophagie, Butparasitismus und Entomophagie in der Familie der Curculionidae. *Bulletin de la Société entomologique suisse* **42**, 185-196.

Chapter 2

From mini-amplicons to mitochondrial genomes - developing molecular techniques for reconstructing curculionoid phylogeny

"What is the use of this beak, this snout, this caricature of a nose? Where did the insect find the model for it? Nowhere. The Weevil invented it and retains the monopoly. Outside his family, no beetle indulges in these nasal eccentricities."

- Jean-Henri Fabre, 1922

Larinus sp. (Curculionidae: Lixinae: Lixini), Macedonia, Greece

Chapter 2: From mini-amplicons to mitochondrial genomes – developing molecular techniques for reconstructing curculionoid phylogeny

2.1 Abstract

The robustness of molecular phylogenetic reconstruction can be improved by both reducing stochastic error through the addition of more sequences or genes, and through reducing systematic errors by the addition of more taxa. Advances in the versatility of commonly applied laboratory techniques, such as PCR amplification of degraded DNA, and novel applications of existing technologies, such as next-generation sequencing have, in recent years, opened up new possibilities to achieve both. This chapter is an exploratory test of both more traditional and more recently developed practical approaches to obtain sequence data from preserved specimens of Curculionoidea, including approaches focusing at two opposite ends of the size spectrum: mini amplicons of less than 100 bp and mitochondrial genomes of greater than 9 kbp in length. The primary objective is therefore to expand, through experimentation, the available sources of data, in an attempt to increase the number of markers and taxa that can be incorporated in subsequent in-depth phylogenetic analyses.

2.2 Introduction

The first chapter in this thesis reviewed a variety of molecular markers that have been employed in the reconstruction of phylogenies for diverse animal groups including weevils. The aim of the current chapter is to develop new techniques for the purpose of obtaining DNA sequences, from a variety of tissue sources, for use in reconstructing the higher-level phylogeny of the megadiverse beetle superfamily Curculionoidea.

Expanding taxon sampling has been shown to be an important factor in improving phylogenetic accuracy (Zwickl & Hillis 2002), whilst increasing sequence length by concatenating different genes has also been shown to lead to enhancements in support, resolution and accuracy of phylogenetic analyses (e.g. Wortley *et al.* 2005). To attain the goal of a robust phylogeny, a 'two-step protocol', to maximise both taxon sampling and sequence data, was envisioned. For the first step, as many well preserved, DNA-ready representative specimens of family-level taxa were obtained, mostly through specimen loans and donations from a wide network of collaborating collectors, but also through original fieldwork. Particular focus was made on procuring as many tribal-level taxa as possible within the largest, and focal family, Curculionidae *s.str.* It was hoped that DNA extracted from such well-preserved specimens would allow for the standard polymerase chain reaction (PCR) amplification and sequencing of multiple gene markers.

In order to enhance taxon sampling with difficult to obtain and rare taxa, usually only available as dry-preserved, often old, museum specimens, an attempt was also made to develop short phylogenetically informative amplicons (SPIAs or mini-amplicons). Such amplicons theoretically enable short (~50-300 bp) fragments from degraded archival specimens to be amplified, sequenced and incorporated in phylogenetic reconstructions. This technique was employed by Hernández-Vera *et al.* (2013), who were able to sequence short (<100 bp) sections of 16S rRNA (*rrnL*) from archival samples of two weevil genera, greatly increasing their taxon sampling for the investigation of their biogeographic history. SPIAs are chosen to contain informative sites, yet to also be short, therefore increasing the likelihood of amplification from specimens that are genetically degraded. Importantly, this includes dry-preserved specimens in natural history museum collections, which contain rare and even extinct taxa that would otherwise never be incorporated into molecular analyses.

The second goal of maximising sequence data per taxon was addressed through two approaches. The first of these involved conventional amplification of several mitochondrial and nuclear protein-coding DNA and ribosomal RNA gene markers. Markers were selected based partly upon their having been successfully used in previous curculionid molecular phylogenetic studies, and on the likelihood that either existing primers could be used for PCR amplification, or that new primers could be designed where necessary. The following mitochondrial DNA markers were selected for development: protein-coding cytochrome c oxidase subunit I (*cox1*) and cytochrome b (*cytB*), and ribosomal *rrnL*. The following nuclear genome markers were also selected: protein-coding arginine kinase (*ArgK*), ribosomal small subunit (18S), and the large subunit (28S) ribosomal RNA.

To further increase the number of genetic markers available for analysis, the possibility of obtaining partial mitochondrial genome (mitogenome) sequences, composed of multiple protein-coding genes, was investigated. Therefore, another aim of this chapter is to obtain mitogenomic sequences through a Long-Range (LR) PCR amplification technique, as developed by Timmermans *et al.* (2010), followed by sequencing of the PCR products on two alternative NGS platforms: 454 Junior (Roche) and HiSeq (Ilumina). A newer development, of direct NGS from pooled genomic DNA, is also briefly investigated here. Partial weevil mitogenomes have recently been successfully analysed by Haran *et al.* (2013), but because of low taxon sampling (28 ingroup taxa), conclusions regarding monophyly of several curculionid subfamilies could not be confidently made.

Therefore, the initial broad aim was to build a robust phylogeny using the multiple-gene sequences obtained from well preserved specimens, and then to use sequences from SPIAs to assign archival specimens to specific lineages within the multi-gene tree. This would allow for the assessment of the monophyly of higher-taxa, as presently established under conventional morphological taxonomy, in the light of molecular data.

2.3 General materials and methods

2.3.1 Taxon sampling

Increased taxon sampling has been empirically shown to improve the results of phylogenetic reconstruction, irrespective of the optimality criterion under which trees are constructed (Zwickl & Hillis 2002). It is therefore, very important to sample as broadly as possible amongst the potentially distinct lineages of interest. It was initially decided to attempt to sample species belonging to as many type genera of tribes and subtribes within Curculionidae *sensu* Bouchard *et al.* (2011) as possible. The name-bearing type of a nominal family-group taxon (tribe, family etc.) is a nominal genus called the "type genus" (ICZN 1999) and, therefore, incorporating such specimens into analyses will allow for more objective conclusions regarding the resulting taxonomic relationships. Although there was a focus on Curculionidae, efforts were also made to secure as many representatives of the other curculionoid families and to obtain other genera of particular taxonomic interest, such as *Ocladius* (Brachyceridae), of dubious family assignment (Massimo Meregalli *pers. comm.*).

2.3.1.1 Collaborating collectors

It would have been impractical and virtually impossible to specifically search out and collect the necessary specimens for this study in the field. Therefore, a large number of curculionoid researchers, museum curators, coleopterists, general entomologists and collectors across the world were contacted with requests for loans or donations of DNA-ready weevil specimens stored in ethanol, or genomic DNA (gDNA) extractions. I ultimately received samples from the institutional and/or private collections of 25 collaborators (listed in the acknowledgements) located in 16 countries, which represented the bulk of the taxa used in this study.

2.3.1.2 Fieldwork collecting

Additional weevil specimens were field collected specifically for this project in England, mostly on the UEA campus, and during a collecting field trip in April 2012 to southern Yunnan Province, China. The latter was organised together with Dr. Christopher Lyal of the Natural History Museum (NHM), in collaboration with Prof. Run-Zhi Zhang and colleagues of the Institute of Zoology, Chinese Academy of Sciences. China was selected as a potentially interesting source of specimens primarily because very few Oriental species had been previously obtained through the network of collaborating collectors.

Collecting took place at 36 sites across Xishuangbanna Prefecture, between 4-18 April, primarily in lowland and hilly tropical forested country, and predominantly involved beating foliage, sifting leaf litter, setting malaise traps and hand collecting. An informal report of the trip, in the form of three blog entries, was published contemporaneously on the NHM website as a public outreach tool (http://www.nhm.ac.uk/natureplus/blogs/beetles/2012/04/15/a-necessary-weevil-collecting-in-southern-china-april-2012). Many specimens collected during this expedition were subsequently used in analyses for this thesis.

Additional weevil specimens were collected during other collecting trips undertaken either directly preceding or during the course of this study, but not specifically designated as fieldwork. This included specimens collected in France, Greece, Zambia, Suriname, Ecuador and Saba (Dutch Caribbean). Whilst specimens collected on these occasions did not necessarily end up in the final datasets, much of the preliminary PCR-optimisation work and primer development was undertaken using them.

2.3.1.3 Specimen identification

Identifying weevils, even to the level of tribe or genus, is neither a trivial nor an intuitive task. However, through collaboration with the entomologists mentioned above, most material received on loan was identified to genus- or species-level, although this was not always possible. Therefore some samples remained identified only to tribal-, subfamily- or family-level, despite comparison with specimens housed in the NHM collections (which itself is neither comprehensive nor infallible) and referral to the limited taxonomic literature. However, although this was a limitation, these identifications are nevertheless appropriate for much of the higher-level phylogeny reconstruction concerned in this study.

In total 173 identified weevil samples were obtained for the systematic part of this study, representing six curculionoid families, 16 subfamilies and 104 tribes and subtribes within the focal family Curculionidae. All chapters within this thesis are based upon different samples from within this total (dependent upon PCR success *etc.*) and therefore the exact taxa used are listed in detail separately in each chapter, together with country of origin and source (collector) of the specimens. The identified Curculionidae taxa are summarised by subfamily and number of tribes and subtribes in Table 2.1.

2.3.2 DNA extraction for standard PCR

Experimentation with extracting DNA from various specimen tissues was undertaken to empirically select the tissue from which extractions were most consistently successful. To check for extraction success, 16µl aliquots from a range of DNA extractions carried out on individual heads, heads+prothorax, single legs and entire specimens (for small species) were run out on a 1.5% agarose gel to check for a high molecular weight DNA band. Additionally, *cox1* PCR amplifications (described below) for these trial extractions were carried out. All DNA extractions for this study were performed using DNeasy blood & tissue spin column or plate kits (Qiagen), according to the printed instructions and following overnight incubation of the samples in 180µl ATL tissue lysis buffer and 20µl proteinase K solution at 56°C. After the trials it was decided that extracting DNA from the head (or head and prothorax in very small specimens) produced the most consistently successful results. This has the added advantage that the head is relatively easy to glue back onto the body after extraction, thereby maintaining complete voucher specimens.

Table 2.1 Summary of total identified families, subfamilies, tribes and subtribes of Curculionidae obtained for this study. Each chapter analysed a unique dataset which is presented in detail within the respective chapter. In grey are highlighted the non-Curculionidae families and the number of specimens available for each.

Family	Subfamily	No. of tribes/subtribes
Curculionidae	Bagoinae	1
Curculionidae	Baridinae	1
Curculionidae	Ceutorhynchinae	4
Curculionidae	Conoderinae	4
Curculionidae	Cossoninae	4
Curculionidae	Cryptorhynchinae	5
Curculionidae	Curculioninae	15
Curculionidae	Cyclominae	4
Curculionidae	Entiminae	19
Curculionidae	Hyperinae	1
Curculionidae	Lixinae	3
Curculionidae	Mesoptillinae	2
Curculionidae	Molytinae	13
Curculionidae	Platypodinae	3
Curculionidae	Orobitidinae	1
Curculionidae	Scolytinae	24
Anthribidae		2
Attelabidae		2
Brachyceridae		4
Brentidae		4
Dryophthoridae		2

2.4 Molecular markers

2.4.1 Mitochondrial markers

2.4.1.1 Cytochrome oxidase subunit 1 (cox1)

Two regions of the *cox1* gene were selected for analysis. Each is described separately below.

cox1 5' 'Barcode' region

Primers, PCR and sequencing

The 5' end of *cox1* has been widely employed as a DNA barcode (Hebert *et al.* 2003) and universal primers for the amplification of this region exist for several taxonomic groups. Primers previously used to amplify *cox1* from a variety of metazoan invertebrates, the so called 'Folmer' primers (Folmer *et al.* 1994) were optimised for greater specificity to the Curculionoidea according to the CODEHOP primer design strategy (Rose *et al.* 2003) through the incorporation of differing bases (particularly in third codon positions) towards the 3' end of the primer to allow for matching to all degenerate codon possibilities in the amino acid target template (some of this variation was observed in alignments of reference curculionid *cox1* sequences obtained from Genbank), whilst maintaining a nondegenerate 5' end as a 'clamp'. The logic behind this optimisation is that annealing of the 3' degenerate end of the primer to the weevil template DNA is stabilised by the 5' 'clamp' in the first round of PCR; in the course of the ensuing rounds of amplification, the annealing of the primer to the PCR product is bolstered by the identical match between the incorporated primer and primers remaining in the pool at the 5' consensus clamp region (Rose *et al.* 2003).

In addition, alternative versions of these modified forward and reverse primers were designed to possess a nondegenerate, nonhomologous 'tail' attached to the 5' end, in the manner described by Regier and Shi (2005). These tails are named M13REV and M13(-21) in the forward and reverse primers respectively (Appendix 2.1 A). Tailed primers have been shown to increase the yield of PCR product relative to non-tailed primers and can also increase the readable sequence length if the tails are subsequently used alone as sequencing primers (Regier & Shi 2005). The increase in yield is hypothesised to be due to the 5′ tails enhancing amplification through incorporation of degenerate primers of greater mismatch than in the absence of 5′ tails. Following the initial PCR cycle, the nondegenerate 5′ tails increase the overall T_m of the bipartite primer in subsequent cycles and thereby allow a more diverse array of mismatched gene-specific primers to amplify (Regier & Shi 2005).

All possible combinations of the tailed and untailed primers were used in PCR trials to determine the most reliable combination for amplifying *cox1* from a diverse assembly of weevil gDNA samples. The most successful primers used to amplify the *cox1* 5' region in this study are named M13REVFOLbeetF2 (forward primer) and M13(-21)FOLbeR2 (reverse primer). The fragment they amplify is 658 bp in length. Thorough optimisation of the PCR reactions was undertaken experimentally, both by diluting template DNA and altering concentrations of all reagents in reactions, especially $MgCl_2$ and NH_4 buffer, as well as varying the annealing temperature and duration of the separate cycling steps. Primer sequences and empirically determined optimum reaction chemistries and cycling conditions are shown in Appendix 2.1 A-C. All PCR reactions, unless otherwise mentioned, were undertaken using BIOTAQ DNA polymerase (Bioline). PCR products were visualised under ultraviolet light on a 1-2% high-melting agarose gels (Fisher Biosciences) stained with either ethidium bromide or GelRed (Cambridge Bioscience). Hyperladder IV (Bioline) was used to estimate molecular weight and for relative DNA quantification.

PCR products were used as templates in Sanger-sequencing reactions using the M13 tails alone as primers and employing the BigDye v3.1 Cycle Sequencing Kit (Applied Biosystems). All sequencing reactions unless otherwise noted contained the following components per sample: 5.35µl ddH_2O, 1.5µl Sequencing buffer (5X), 0.15µl sequencing primer, 1.0µl BigDye 3.1 and 2µl of PCR product as template (total volume per reaction of 10 µl). Sequencing reactions were cycled at 96°C for 10s, 50°C for 5s and 60°C for 4 mins, repeated for 25 cycles. The results were read on a 3730XL sequencer (Applied Biosystems).

All resulting sequence traces, unless otherwise noted, were viewed and edited in Geneious 5.4 (Kearse *et al.* 2012). Sequences obtained with these reactions were subsequently used as prior information in designing primers to amplify short, phylogenetically informative *cox1* mini amplicons (see below). Optimisation of PCR reactions and sequencing of other markers followed a similar pattern to that described for *cox1* 5' and are not repeated below unless different.

***cox 1* 3' Region**

Primers, PCR and sequencing

In addition to the barcode region of *cox1*, a ~536 bp portion of the 3' region, partly overlapping with the 5' barcode region, was also amplified and sequenced using the primers Jerry (forward) and SpatR (reverse). These primers had previously been successfully used in beetle phylogenetics research (e.g. Timmermans *et al.* 2010) and proved to be reliable for amplifying weevil DNA. The primer sequences and reaction conditions are given in Appendix 2.1 A-C. Sequencing was undertaken with the same primers as for the PCR, but otherwise followed the protocol described for the *cox1* 5' region. Sequence editing was undertaken in Geneious. Figure 2.1 illustrates the relative position of the *cox1* primers and the lengths of the fragments amplified.

2.4.1.2 Cytochrome b (cytB)

Primers, PCR and sequencing

This protein-coding mitochondrial gene was reliably amplified using pre-existing primers that had been successfully used in a previously published Coleoptera phylogeny (Timmermans *et al.* 2010). Primer sequences, optimum reaction chemistries and cycling conditions for *cytB* are shown in Appendix 2.1 A-C. Sequencing was undertaken with the same primers as for the PCR, but otherwise followed the protocol described for the *cox1* 5' region.

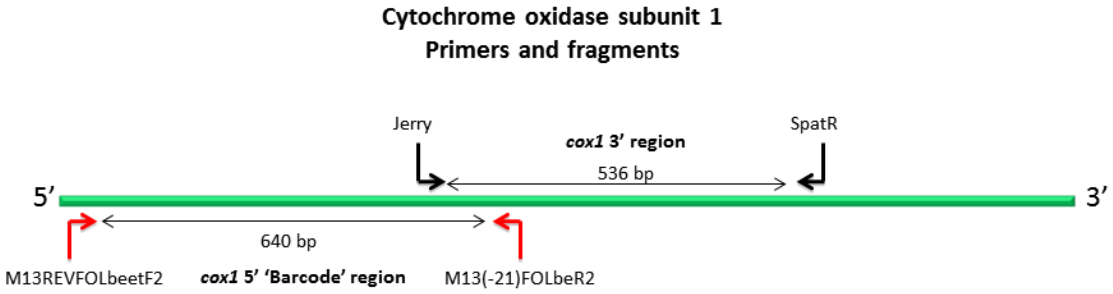


Figure 2.1 Primers used to amplify the two regions of *cox1* used in this study and the corresponding PCR fragments.

2.4.1.3 16S small subunit ribosomal RNA (rrnL)

Primers, PCR and sequencing

The *rrnL* marker was amplified without many problems using pre-existing primers that had been adopted in a previous phylogeny reconstruction of Coleoptera (Hunt *et al.* 2007). This set of primers consisted of two forward primers (LR-J-12961 and LR-J-12887) and one reverse primer (LR-N-13398). In the first instance, PCR was attempted using LR-J-12961 as the forward primer, which was empirically shown to be the most successful. Any unsuccessful reactions were then attempted again using the alternative forward primer (LR-J-12887). Primer sequences and optimum reaction chemistries and cycling conditions for *rrnL* are shown in Appendix 2.1 A-C.

Sequencing was undertaken with the same primers as for the PCR, but otherwise followed the protocol described for the *cox1* 5' region.

2.4.2 Nuclear markers

2.4.2.1 Small subunit 18S rRNA

Primers, PCR and sequencing

Because its length of approximately 1900 bp rendered it too long for amplification in a single PCR, the selected region of the 18S gene was amplified in two overlapping sections as shown in Figure 2.2, using the primers and similar methodology to that employed by Shull *et al.* (2001). The longer section (5' end) of approximately 1300 bp was sequenced using two further primers in addition to the PCR primers. The shorter section (3' end) of approximately 700 bp, was sequenced only with the two PCR primers, resulting in a total of six sequencing reactions per sample. The two sections of 18S overlapped by approximately 250 bp, allowing for unambiguous assembly of all sequences and extraction of the corresponding full-length consensus sequence (Figure 2.2), which was undertaken using Geneious 5.4 (Kearse *et al.* 2012). Primer sequences and optimum reaction chemistries and cycling conditions for 18S are shown in Appendix 2.2 A-C. Sequencing was undertaken with the same primers as for the PCR, but otherwise followed the protocol described for the *cox1* 5' region. Sequence editing was undertaken in Geneious.

2.4.2.2 Large subunit 28S rRNA

Primers, PCR and sequencing

Amplification and sequencing of 28S was straightforward, utilising the same primers which had been designed for a previous study (Monaghan *et al.* 2007). Primer sequences and optimum reaction chemistries and cycling conditions for 28S are shown in Appendix 2.2 A-C. Sequencing was undertaken with the same primers as for the PCR, but otherwise followed the protocol described for the *cox1* 5' region.

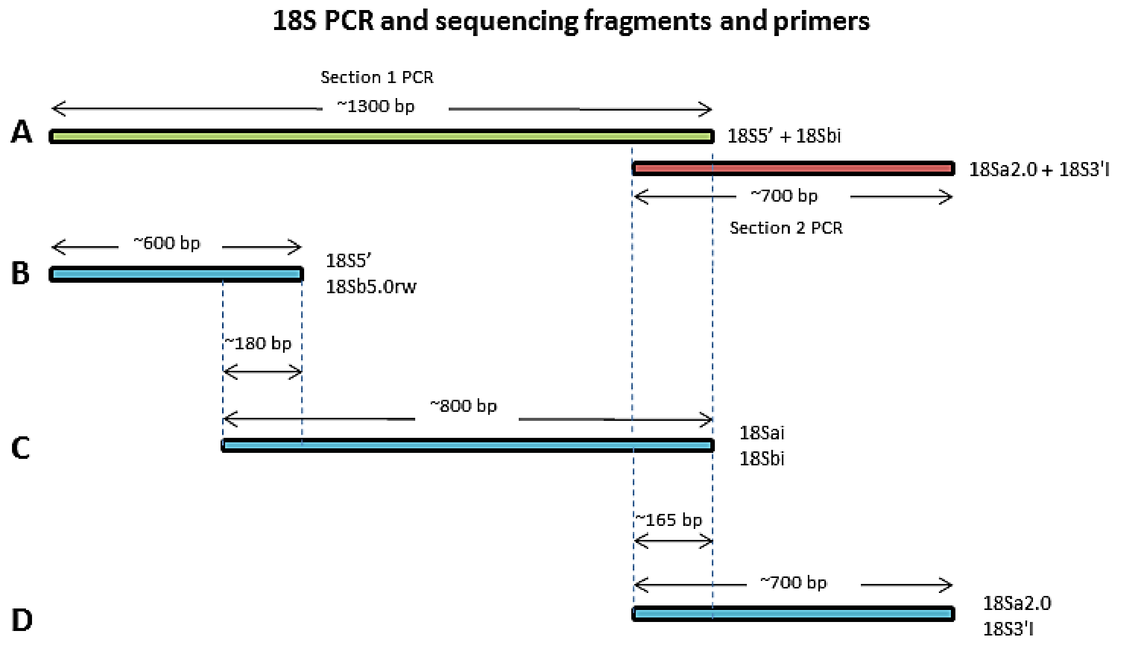


Figure 2.2 Division of the nuclear 18S marker into two sections for PCR amplification and into three section for sequencing. Primers are listed to the right of the fragments they amplify/sequence. A) PCR sections 1 and 2. B-D) the three sequenced fragments.

2.4.2.3 Arginine kinase (*ArgK*)

Primers, PCR and sequencing

Due to poor PCR success using available primers, reliable amplification and sequencing of *ArgK* was achieved using a 'nested PCR' approach, whereby the PCR product of a first amplification reaction was used as template DNA in a second-round re-amplification PCR. This reamplification was attempted using the same primers as for the initial PCR but the resulting sequences were of very poor quality; therefore, newly designed internal primers for this re-amplification step were developed. These primers were designed according to the CODEHOP principles of Rose *et al.* (2003) and as outlined above for *cox1*. To achieve this a selection of Coleoptera Polyphaga reference *ArgK* sequences from Genbank were aligned and the forward (ArgKforB2) and reverse (ArgKrevB1) *ArgK* primers used by McKenna *et al.* (2009) were mapped onto the alignment using Geneious. The sequence between the primers was searched manually for conserved regions suitable as primer-annealing sites, wherein the CODEHOP strategy was used to design primers with degenerate 3' ends. A number of potential primers were designed and experimentally trialled under varying PCR conditions, including primers with all 3rd codon positions degenerate ('fully degenerate') and primers with only 3rd codon positions in the 3' end of the primer degenerate ('semi degenerate'). After optimisation, one pair of primers proved reliable for the reamplification, composed of a semi degenerate forward primer (ArgK_F1_semidg) and a fully degenerate reverse primer (ArgK_R2_fulldg).

Figure 2.3 illustrates the relative position of the *ArgK* primers and the lengths of the fragments amplified as mapped onto the reference sequence of *Phyllotreta striolata* (Chrysomelidae; Genbank accession EU420057.1). Primer sequences and optimum reaction chemistries and cycling conditions for *ArgK* are shown in Appendix 2.2 A-C. Sequencing reactions were undertaken using the newly designed internal primers according to the standard sequencing cycling profile described above for *cox1* 5'.

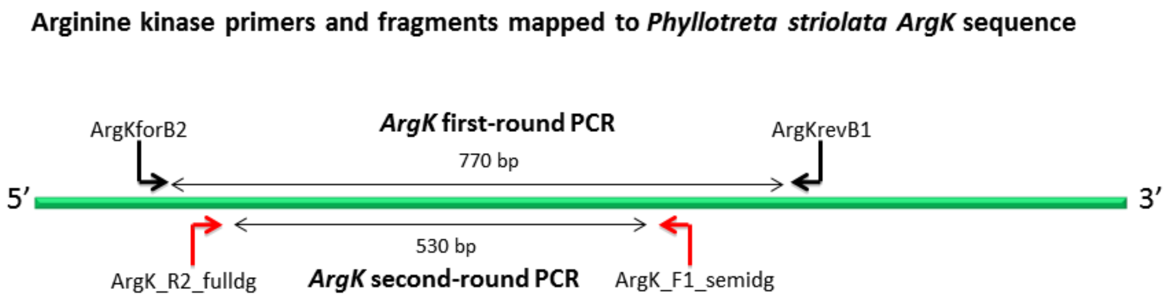


Figure 2.3 Primers used to amplify *ArgK* in a nested reamplification and the corresponding PCR fragments.

2.4.3 Short phylogenetically informative amplicons (SPIAs)

SPIAs, or mini amplicons, are similar to mini barcodes, but the important distinction is that whereas the latter are used for species identification (Hajibabaei *et al.* 2006) , the former provide sequence information to be incorporated in phylogenetic analyses. In addition, whereas the typical DNA barcode is composed of a universally agreed 650 bp section of the *cox1* gene (Hebert *et al.* 2003) or smaller fragments thereof, SPIAs are much shorter (~50-300 bp) (Hernández-Vera *et al.* 2013) and can theoretically be obtained for any gene and concatenated together if necessary to produce longer sequences composed of data from several loci.

2.4.3.1 Taxon sampling

For the development of SPIAs, 37 dry-preserved, mounted specimens were selected from the NHM entomology collection. These specimens belonged to 4 widely distributed western Palaearctic species in the genus *Curculio* (*C. glandium, C. nucum, C. pellitus* and *C. venosus*). These specimens (CU001-CU0037) are listed with their collecting data in Appendix 2.3. Because these taxa are not especially rare or irreplaceable, approval was granted from the curator of Coleoptera (M. Barclay *pers. comm.*) for initial trials to be carried out on them as a proof-of-concept. It was envisaged that upon successful optimisation of the SPIAs, more unusual and rarer taxa, important to the overall phylogeny and not available as DNA-ready specimens, would be selected to work with. The genus *Curculio* was chosen primarily because a molecular phylogeny for the genus exists (Hughes & Vogler 2004) for comparison and because of the widely distributed species available, offering potential to investigate genetic divergence between populations . The main aims of the SPIA approach were therefore, to:

1) allow for specimens to be ‘matched’ onto an existing phylogeny (e.g. Hughes & Vogler 2004)

2) identify taxa to higher level or better

3) assess the utility of SPIAs to characterise geographic patterns of genetic diversity within a species

2.4.3.2 DNA extraction

Permission was obtained for using a single leg from each specimen as a source of tissue for DNA extraction. Each leg was removed from the mounted specimen with fine forceps and placed whole in the ATL buffer and proteinase K solution for incubation, otherwise the protocol was identical to that for extracting fresh specimens as described in the DNeasy extraction kit instructions, except that these samples were not vortexed but carefully mixed to avoid damage to the fragile legs. The legs remained intact following extraction, allowing them to be re-associated with the voucher specimen afterwards.

2.4.3.3 *cox1* SPIAs

Primer design

The *cox1* 5' 'barcode' region was selected as the marker for targeting design of SPIAs, primarily due to the fact that more sequences exist for the barcode region than most other markers, increasing the amount of data available for phylogeny reconstruction and the chances of matching taxa to their closest relatives.

Primer design followed the CODEHOP and 'tailed' primers approaches described above (Regier & Shi 2005; Rose *et al.* 2003). Reference weevil barcode sequences were downloaded from Genbank, together with several other sequences of beetles belonging to the superfamily Chrysomeloidea, which is widely considered to be the sister clade to Curculionoidea (e.g. Hunt *et al.* 2007). These sequences were used as prior information for primer design and aligned in BioEdit (Hall 1999). Suitable conserved regions at the 5' end of the gene were manually searched for and sites conserved for the 1st and 2nd codon positions and with degenerate 3rd codon positions were selected. For each primer designed, two variants were developed; one was 'semi-degenerate' with only half of the 3rd codon sites degenerate (those at the 3' end of the primer), the other 'fully-degenerate' with all 3rd codon sites fully degenerate for each amino acid in the selected motif. The same forward primer as for the normal *cox1* 'barcode' amplification (FOLbeetF2) was used and two downstream reverse primers (CO1_miniR_semidg and CO1_mini_R1_semidg) were newly designed in order to obtain two differently sized fragments of 129 bp and 234bp respectively (not including primer sequences). Therefore the possibility of obtaining a 'long' fragment of 234 bp (ideally) or two 'short' fragments of 129bp and 90bp (spanning the 234bp fragment) was theoretically possible. Figure 2.4 shows the newly designed primers used to amplify *cox1* SPIAs, their relative positions and the resulting fragment lengths. Primer sequences are given in Appendix 2.4 A. Whilst both 'tailed' and 'untailed' versions of primers were designed, all successful newly designed primers incorporated the M13 nondegenerate, nonhomologous 5′ tails to increase PCR yield (Regier & Shi 2005) as for the *cox1* 5' 'barcode' primers. The reason for designing multiple primers was because in some museum specimens (particularly very old ones) the DNA could be more degraded than in others, thereby making it difficult or impossible to obtain the 'long' 234 bp fragments but hopefully possible to obtain one or the other, or both of the 'short' ones.

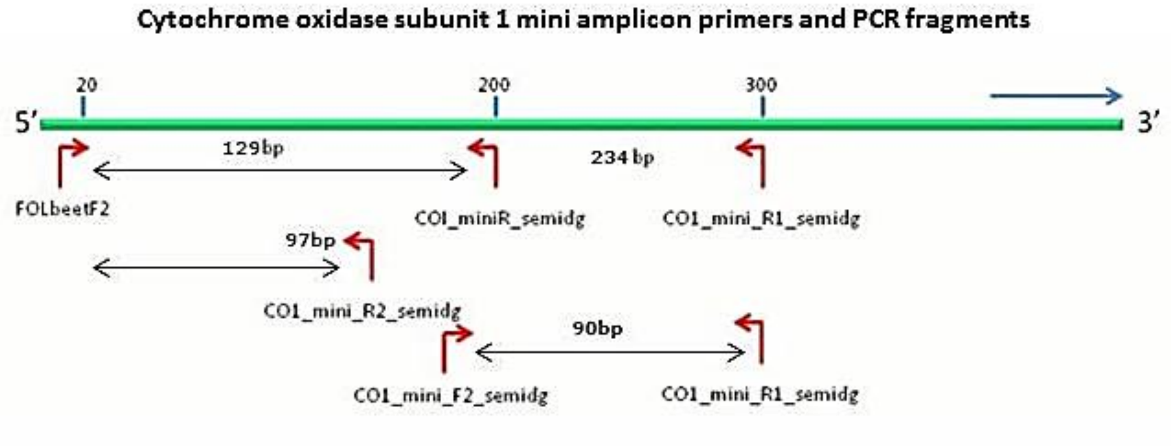


Figure 2.4 Primers used to amplify *cox1* SPIAs allowing for the amplification of a varierty of differently sized PCR fragments. All primers except FOLbeetF2 were newly designed for this study.

PCR

Initial PCR optimisation trials using eight extracted samples (CU001-CU008) tested amplification using primers for obtaining both the 234 bp and a 129 bp fragments as outlined above. Through optimisation trials, both with the museum samples and with more recent 'fresh' specimens, it was empirically determined that the optimum PCR reaction conditions included using 5µl of undiluted DNA template with a 3mM concentration of $MgCl_2$ in each reaction mix and a 45°C annealing temperature.

Optimum reaction chemistries and cycling conditions for the mini amplicon primers are shown in Appendix 2 4 B-C.

The resulting PCR reactions were encouraging, with most samples amplifying with both primer pairs, albeit weakly. Figures 2.5 and 2.6 show gel runs of the shorter and longer *cox1* SPIA fragments respectively. It should be noted that sample CU005 was amplified from DNA extracted from a 120 year old specimen. Sample CU008 consistently resulted in the most successful PCRs as evidenced by it having produced the brightest band for both fragments. Sequencing reactions were undertaken with both the forward primer (FOLbeetF2) and the M13(-21) tail of the reverse primer alone, as described for the *cox1* 5' region earlier. Sequence traces were viewed in FinchTV (Geospiza).

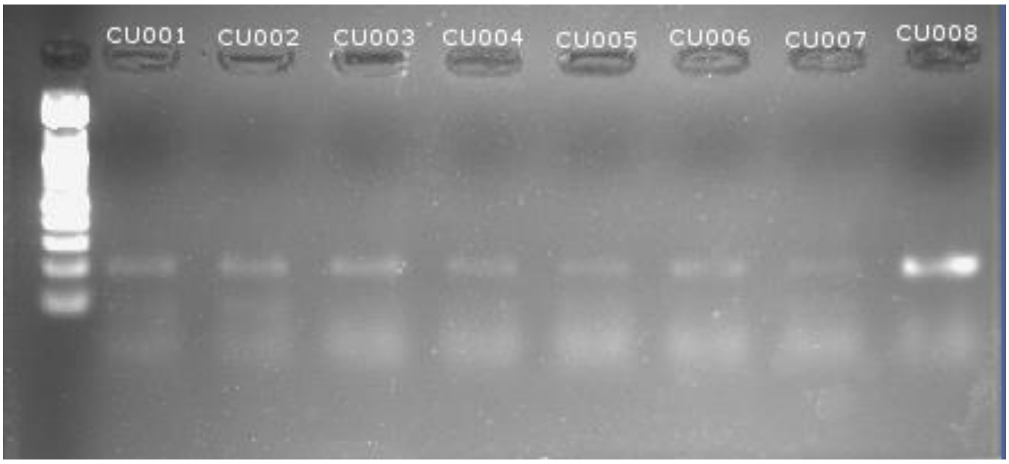


Figure 2.5 Gel run on a 2% agarose gel of PCR amplification products of the 'short' 129bp *cox1* SPIA fragment (198 bp including primers) from museum specimens. The amplified fragments have generally produced a weak but distinctive target band (sample CU008 has amplified most successfully).

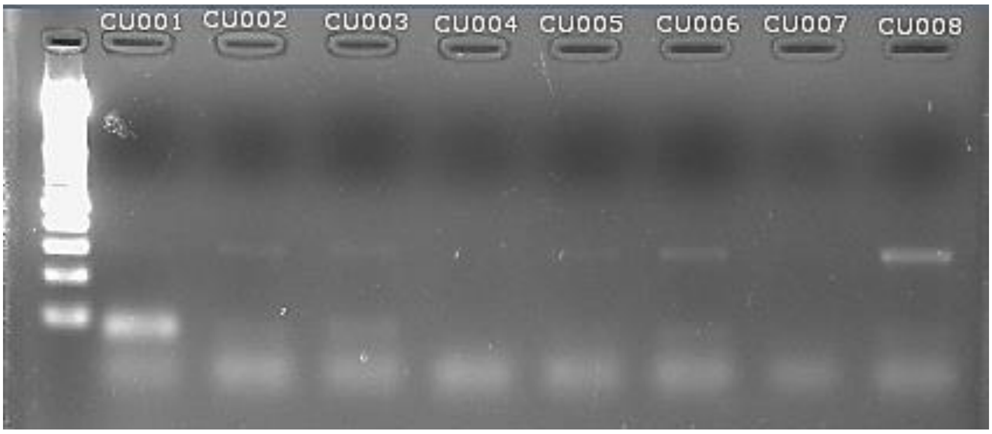


Figure 2.6 Gel run on a 2% agarose gel of PCR amplification of the 'long' 234 bp *cox1* SPIA fragment (287 bp including primers) from museum specimens. Most samples have amplified weakly but distinctly (sample CU008 has amplified most successfully).

The resulting sequences from these initial PCRs were disappointing. Only sample CU008 resulted in a readable sequence for the shorter (129 bp) fragment, all other sequences being illegible. Figure 2.7 depicts four of the resulting sequence traces (CU005-CU008) obtained with sequencing primer M13(-21) for comparison. A 123 bp length of the CU008 sequence was used as a query in a BLAST (Altschul *et al.* 1990) search of the NCBI nucleotide database (http://www.ncbi.nlm.nih.gov/) optimised for highly similar sequences (megablast), resulting in a closest-matching alignment (90% identical sites; E=5e-25) to a portion of the *cox1* sequence of *Curculio sikkimensis* (Genbank accession KC135935.1).

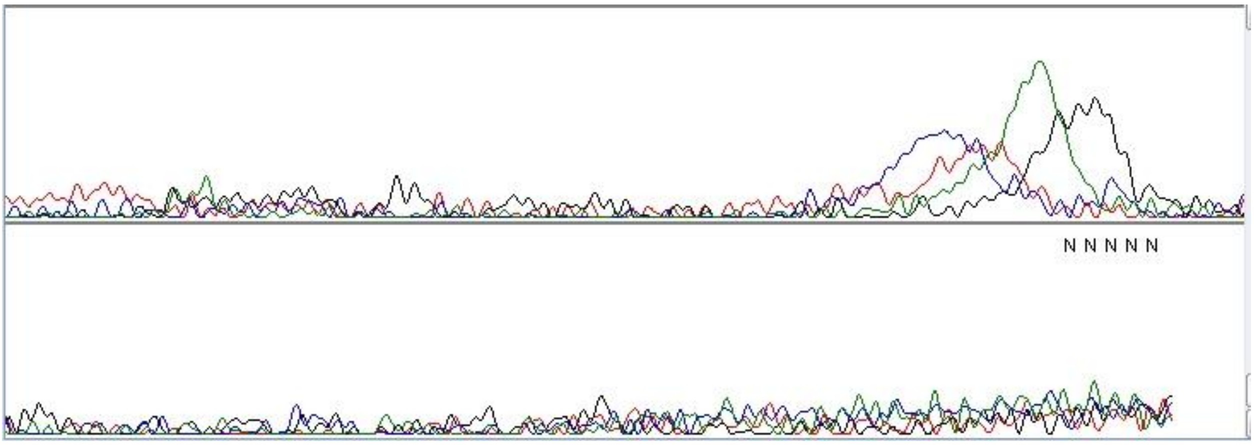
Poor quality sequences were attributed to the weak PCR bands; therefore, to test for whether inhibiting chemicals present in the genomic DNA may have been responsible for low PCR success, serial dilutions of 1:10, 1:20 and 1:50 of DNA template were used in further trials, which led to no noticeable improvement in band brightness for the 1:10 dilution and a distinct worsening for the higher dilutions.

To increase template quantity for sequencing an attempt to reamplify the original PCR products in a nested reamplification was made. This was achieved by running out 16 µl of the first-round PCR product out on a 2% agarose gel and then using a pipette tip to pick a small section of gel from the visualised band (under U.V.) containing the PCR product. This piece of gel was suspended in 30 µl of AE buffer and was used in a subsequent PCR reaction using the same primers as the first-round PCR. The increased amount of DNA template and the fact that the fragments now contained exactly matching primer sites was hoped to improve the results.

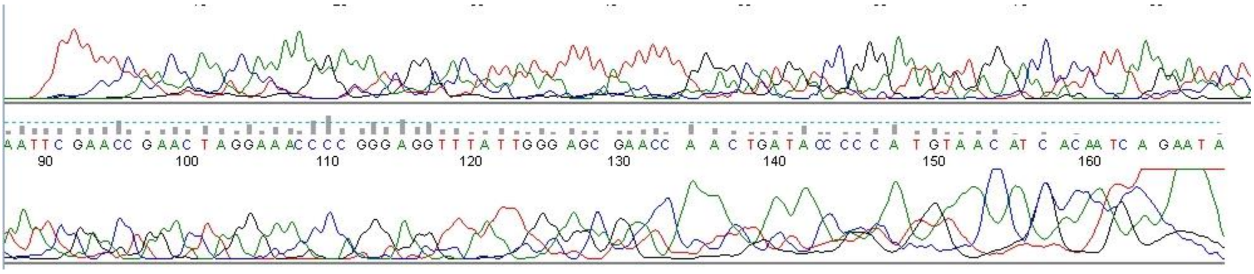
Reamplification of the SPIA resulted in substantial improvement in band intensity, with all eight trial samples amplifying well, although relatively strong primer-dimer was also evident (Figure 2.8). Reamplification of the 'longer' SPIA fragment was only achieved for sample CU008. Sequencing of the reamplifications was undertaken as previously described for SPIAs, but with the additional use of the non-tailed reverse primer (CO1minR_semdgNT) and in triplicate, using 1µl, 0.5µl and 0.2µl of PCR product. This yielded sequences that were no more legible than those obtained without reamplification, with again only sample CU008 resulting in readable sequences that most closely matched *C. sikkimensis* in BLAST searches. The best sequencing results were obtained with the reverse 'tail' primer M13(-21) and with 1µl of PCR template.

As reamplification did not result in improved sequence quality, an attempt was made to amplify alternative *cox1* SPIA regions using different primers that had been developed by Dr. Andrew Mitchell (Australian Museum, Sydney). These primers were designed to be used in nested reamplification reactions similar to those described above, to result in two partially overlapping fragments of 313 bp and 304 bp as depicted in Figure 2.9. The 'external' primers (BC1-Fm, BC3-RDm and Scar-3RDm) were published in Mitchell and Maddox (2010) and Cho *et al.* (2008) and the internal primers (Scar2RDM, Scar1Rm and miniScarFm) and two sequencing primers

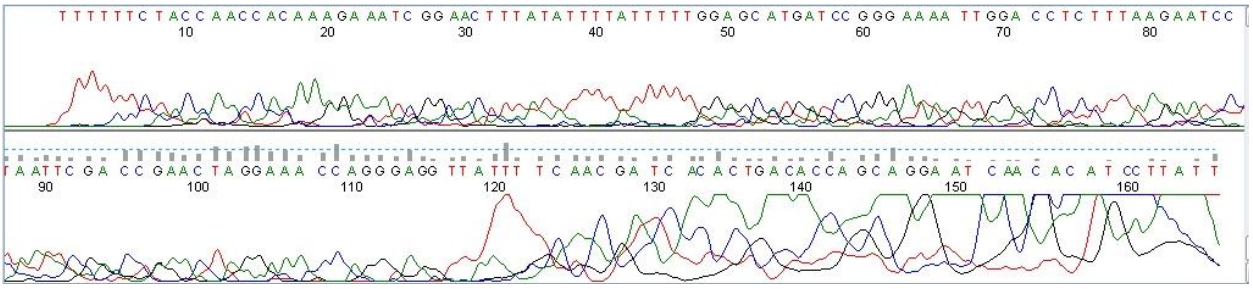
CU005



CU006



CU007



CU008

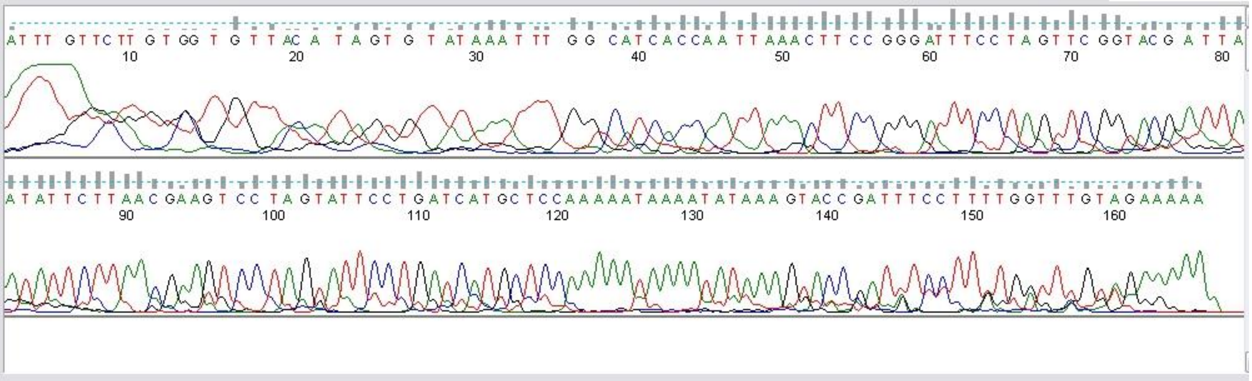


Figure 2.7 Sequence traces viewed in FinchTV (Geospiza) obtained from sequencing 2μl of PCR product of the shorter 129bp *cox1* SPIA using primer M13(-21) from samples CU005-CU008 (samples are the same as in the PCR gel shown in Figure 5).

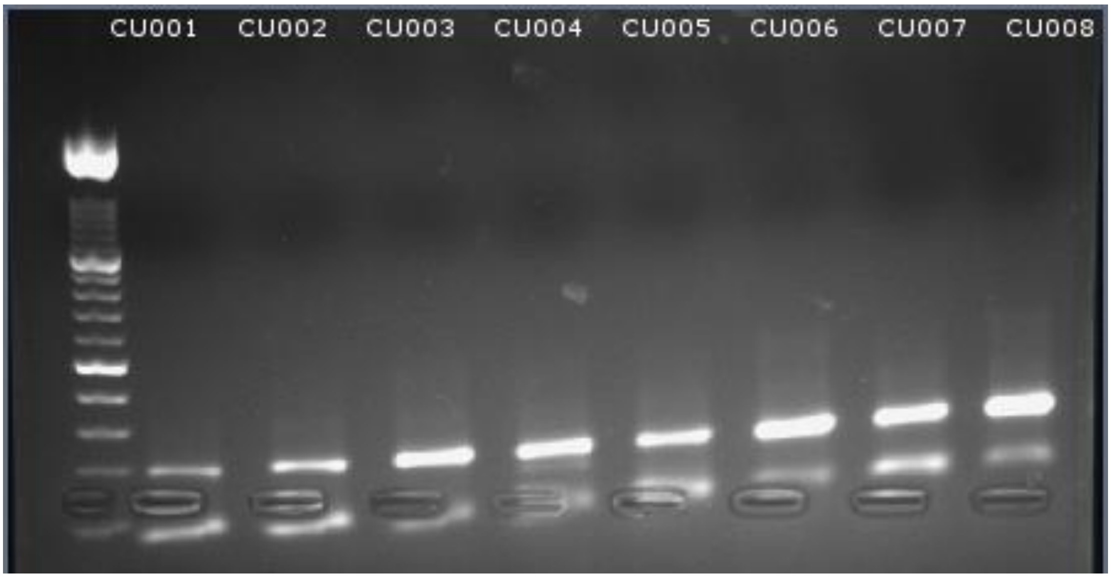


Figure 2.8 Gel run on a 2% agarose gel of PCR reamplification of products of the ‘short’ 129bp *cox1* SPIA fragment (198 bp including primers) from museum specimens, showing greatly improved amplification (*c.f.* Figs. 2.5-2.6). The samples have run at a slight angle.

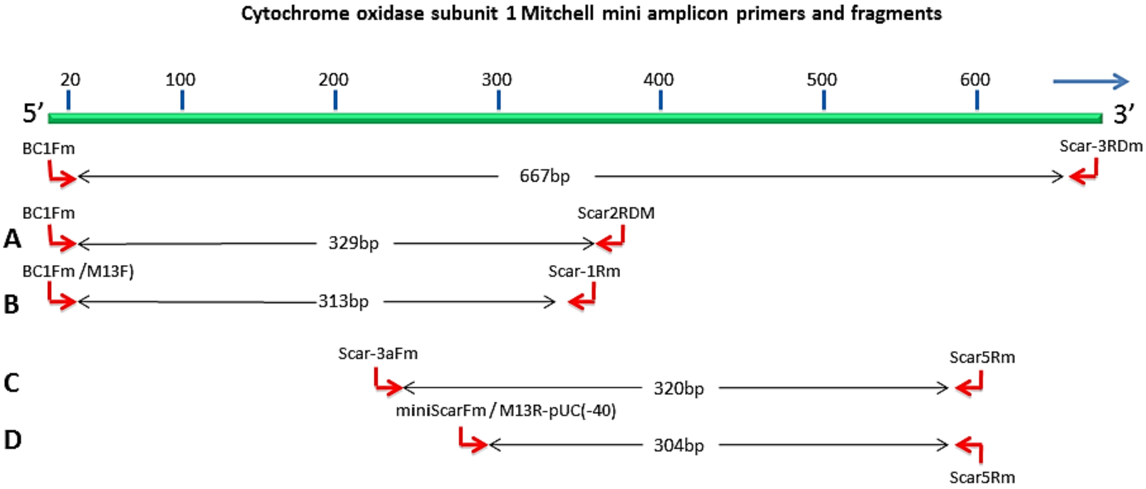


Figure 2.9 Primers designed by A. Mitchell used to amplify *cox1* SPIAs allowing for the amplification of two of differently sized PCR fragments through two nested reamplifications: (M13F and M13R-pUC(-40)) as well as nested PCR reaction conditions were communicated in confidence (A. Mitchell *pers. comm.*). A followed by B (resulting in a 313 bp fragment) and C followed by D (resulting in a 304 bp fragment).

(M13F and M13R-pUC(-40)) as well as nested PCR reaction conditions were communicated in confidence (A. Mitchell *pers. comm.*).

Initial trials on good quality 'fresh' samples confirmed that the nested PCR worked; however, when this was extended to the museum specimens (using a positive sample to confirm PCR success) initially reamplifications failed despite faint visible bands in the first-round PCR in some samples (this not being a prerequisite for success in the second-round reamplification). A subsequent reamplification trial resulted in many samples producing bright target bands but when these were sequenced and used in BLAST searches against the NCBI database, they matched *Homo sapiens* sequences, indicating that contaminant DNA had been amplified. Further work with these primers was abandoned as it was thought that amplification of 300+ bp fragments was probably impossible from the degraded museum specimens at hand.

In a study analysing the effectiveness of several polymerases, including high-fidelity polymerases (some combined with repair enzymes) to amplify various lengths of the *cox1* gene from archival specimens of varying age, it was discovered that Restorase (Sigma-Aldrich) aided PCR yield and in previous experiments allowed for the amplification of full-length barcodes from 70 year old moth specimens (Hajibabaei *et al.* 2005). A trial was therefore set up to attempt to amplify the two *cox1* SPIA regions using the newly designed primers described earlier and Restorase, which incorporates a high accuracy polymerase in a 'cocktail' with a repair enzyme. Instructions were followed allowing for an initial DNA template repair step at 37°C for 10 minutes followed by 72°C for 5 minutes, then primers were added at 65°C for a manual 'primer hot start' reaction. Several optimisation trials involving reagent concentrations, including a range of $MgCl_2$ concentrations (2.5mM to 5.5mM) and

Restorase buffer concentrations were also undertaken using both ‘fresh’ specimens and the museum samples. It was empirically determined that optimal conditions for successful PCR consisted of between 4.5-5.5mM $MgCl_2$ and2.5µl of Restorase buffer per reaction. However, whilst amplification was possible for the ‘fresh’ samples, no museum samples resulted in reaction success.

2.4.3.4 *rrnL* SPIAs

Primers

Primers designed by Hernández-Vera *et al.* (2013) to amplify two adjacent SPIAs (of 55 bp and 95 bp length) of *rrnL* from dry-mounted specimens from a private entomological collection (oldest specimen collected in 1954), and successfully used in a biogeographic study of two weevil genera, were tested for their ability to amplify the museum *Curculio* specimens together with positive samples used in the Hernández-Vera *et al.* (2013) study. Primer sequences and PCR conditions are given in Appendix 2.3 A-C.

Weak PCR target bands were obtained for only the 95 bp fragment (using primer pair 16S_7bp_FGer + 16S_7bp_RGer) amplified from museum *Curculio* samples, although positive samples produced strong bands, indicating that the PCR had been successful. No successful amplifications of *Curculio* samples resulted from PCR of the 55 bp fragment, although positive samples produced strong target bands.

Target bands from the weak 95 bp products were too weak to successfully sequence. To test whether the museum *Curculio* gDNA samples may have contained an inhibitory compound impeding PCR, six trial PCRs were run each containing 1µl of a different positive sample plus 5µl of a different *Curculio* sample ‘spiked’ into it, with a purely positive sample and a negative lacking DNA as controls.

All six trial PCRs resulted in very bright target bands indicating that inhibitory compounds in the museum *Curculio* samples were not the explanation for hindering their PCRs.

A final attempt was made to improve PCR success by using a high fidelity *Taq* enzyme, VELOCITY DNA polymerase (Bioline), but despite experimentation with both museum samples and positive samples, no improvement in PCR success could be achieved.

2.4.4 Mitochondrial genomes

The fundamental structure of typical arthropod mitogenomes comprises of 37 genes including 13 protein-coding genes, two ribosomal RNA genes and 22 transfer RNA genes encoded in a double-stranded circular extrachromosomal molecule of DNA of between 15-20kb length (Boore 1999). Mitogenomes have become a recent focus for use as markers in phylogenetic reconstruction (Botero-Castro *et al.* 2013), and have been demonstrated to provide robust resolving power up to the level of super-order in insects (Talavera & Vila 2011).

A method for obtaining multiple mitogenome sequences through NGS sequencing of long DNA fragments obtained through Long-Range PCR has recently been developed (Timmermans *et al.* 2010). This technique, which incorporates identification of subsequent mitogenomic assemblies through the use of short standard PCR-obtained ‘bait’ sequences from the same samples to match the assemblies, was subsequently used with minor modifications to obtain 27 partial curculionoid mitogenomes for phylogenetic analysis (Haran *et al.* 2013). The technique was adopted here in an attempt to obtain partial mitogenomes from recently collected specimens.

Two NGS platforms were used separately to sequence two different sets of LR-PCR products. These platforms were the 454 GS Junior (Roche) and Illumina HiSeq1000 (Illumina). Both systems have different advantages and disadvantages. The 454 platform produces longer read lengths of c. 700 bp compared to up to two 100 bp paired-reads produced by HiSeq. A 454 run is also faster to complete than a Hiseq run but its main disadvantages are its relatively low throughput (0.7Gb versus 600Gb in HiSeq) and corresponding high cost per base (Liu *et al.* 2012).

2.4.4.1 Taxon sampling

Previously obtained genomic DNA extracts from 50 of the 173 curculionoid samples obtained (see above) were used in trials of the LR-PCR technique. Samples were preferentially chosen after visualisation of 15µl of genomic DNA on a 1.5% agarose gel to select mostly those with a visible high molecular weight band denoting good DNA preservation.

2.4.4.2 'Bait' PCR

Standard PCR amplifications of the *cox1* 3' and the *cytB* markers as previously described above was undertaken using the primers, reaction chemistries and cycling conditions listed in Appendix 2.1 A-C. Standard Sanger sequencing of PCR products and sequence editing followed previous descriptions.

2.4.4.3 Long-range PCR of mitochondrial DNA

Long-range PCR (LR-PCR) reactions to obtain a single fragment of c. 9000-9200 bp length, spanning the *cox1* 3' region to *cytB* (containing sequences from 11 protein coding genes) were undertaken as described in Haran *et al.* (2013) and briefly outlined below. Primers for this reaction consisted of the previously listed forward *cox1* primer ‘Jerry’ (see description above under *cox1* 3’) and reverse *cytB* primer ‘SytB_R’ (Appendix 2.1 A). Figure 2.10 depicts the relative position of the targeted LR-PCR fragment and ‘bait’ sequence regions mapped onto a reference *Tribolium castaneum* mitogenome (Genbank accession NC_003081) indicating the genes that are partly or entirely within it, including *cox1, cox2, atp8, atp6, cox3, nad3, nad5, nad4, nad4L, nad6* and *cytB*. A specialist high-fidelity *Taq* polymerase was used in reactions, *TaKaRa LA Taq (TaKaRa)*, which combines a *Taq* DNA Polymerase with a DNA proofreading polymerase with 3’ to 5’ exonuclease activity, and is optimised for PCR amplification of very long DNA templates. The PCR chemistry and cycling profile used are shown in Appendix 2.5. PCR products were visualised on a 1.8% agarose gel using Hyperladder I (Bioline) for estimation of fragment size and band intensity.

LR-PCRs proved to be extremely unreliable despite extensive optimisation involving differing reaction component concentrations and conditions, using fresh TaKaRa buffer in reactions (old buffer was considered a possible problem; M. Timmermans *pers. comm.*) and even using DNA extractions from legs (as opposed to heads and thorax) of specimens in case mitochondrial DNA concentration was higher in these tissues. Trials using additional primer pairs outlined in Timmermans *et al.* (2010) to amplify different fragment lengths were also conducted without improved success. An attempt was also made to ‘split’ the 9 kbp fragment into two smaller fragments by designing new forward and reverse primers within the *nad5* gene, located approximately halfway between *cox1* and *cytB*, in order to amplify two smaller fragments of c. 5 kbp, which it was thought, might have been possible if poor DNA integrity, prohibiting the full 9 kbp section to be amplified, was a problem.

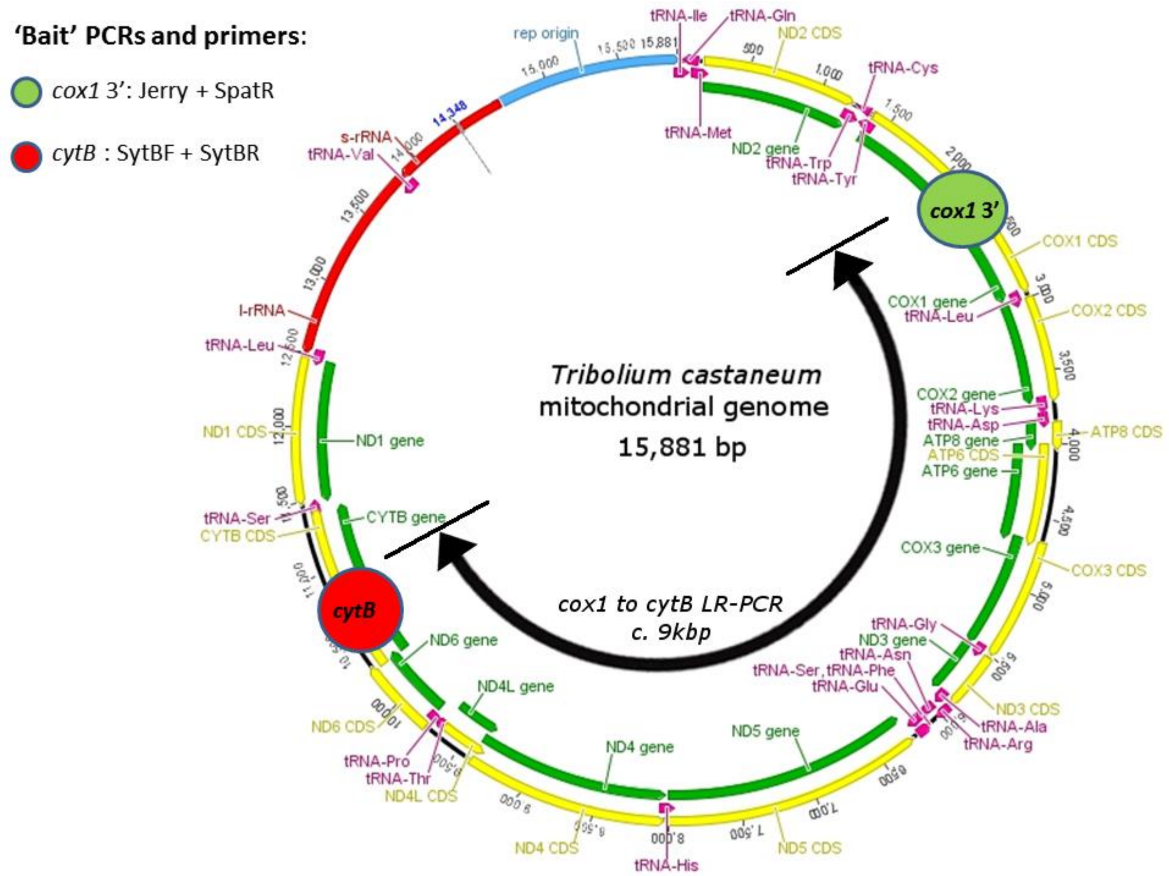


Figure 2.10 Reference *Tribolium castanaeum* mitochondrial genome showing approximate 'bait' marker sites in circles. Black arrow denotes LR-PCR fragment amplified between *cox1* and *cytB* of c. 9 kbp. Image modified from Geneious.

After comprehensive trials, 22 samples in total eventually did result in at least a weak target LR-PCR band using the Jerry and SytB_R primers. Two consistent outcomes encountered in many LR-PCRs was the presence of a very bright high molecular weight 'smear' seen on gels. In many cases it seemed that PCR product 'sat' in the wells as the wells fluoresced brightly. Both these phenomena are shown in Figure 2.11, which also depicts two successful LR-PCR reactions (samples CG093 and CG094). Both 'smeared' and 'well-sitting' samples were submitted for Sanger sequencing of *cox1* and *cytB* to determine whether the reactions had nevertheless been successful but this was not the case. Samples were run out on weaker agarose gel (down to 0.8%) in case the gel matrix was impeding movement of large LR-PCR product molecules, but no improvement was noted and these results remain unexplained.

2.4.4.4 Sample pooling, library preparation and NGS

454 GS Junior

A total of 18 successful LR-PCR products amplified from curculionoid samples were split into two pools based upon qualitative estimates of DNA quantity as viewed from target band intensity on gels. One pool of 10 samples contained those exhibiting 'bright bands' and the remaining 8 samples went into a second pool which exhibited 'faint bands'. A further 68 other LR-PCR products from unrelated studies were likewise pooled into an additional 8 pools. The 10 resulting pools were each separately purified using a QIAquick PCR purification kit (Qiagen), after which the DNA concentration of each pool was measured using a ND-1000 spectrometer (Nanodrop). Approximately equimolar quantities of each pool were combined into single final pool and this sample was used in library construction and sequencing in a single 454 GS Junior run according to standard protocols undertaken at the Department of Biochemistry, University of Cambridge.

Illumina HiSeq

A total of four successful curculionoid LR-PCR products were split into two pools based on target band intensity in the manner outlined for the 454 work. A further 81 samples from unrelated studies were also split across these pools, resulting in seven pools, which were purified, assayed and combined in approximately equimolar quantities as described above. This pool was submitted for Nextera (Illumina) library preparation followed by sequencing on a Hiseq run according to standard protocols undertaken at the Department of Biochemistry, University of Cambridge.

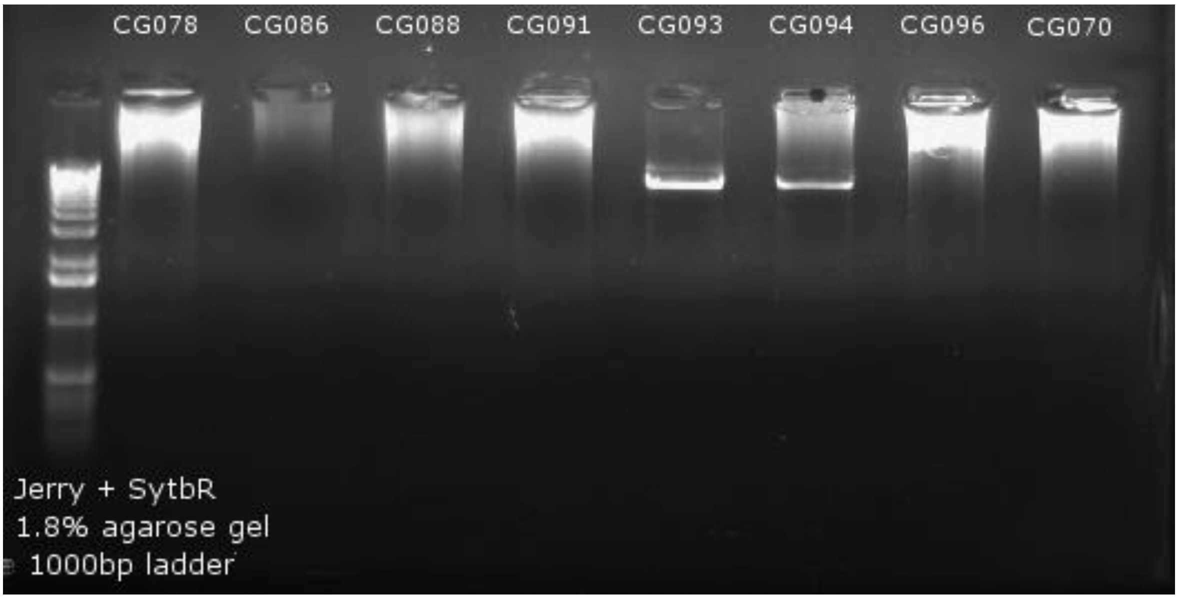


Figure 2.11 Long-Range PCR of a c. 9 kb mitogenomic section from *cox1* to *cytB* run on a 1.8% agarose gel. Samples CG093 and C094 have successfully amplified. Other samples have produced 'smears' and 'well sitters'.

2.4.4.5 Mitogenome assembly and identification

All sequence assembly steps unless otherwise specified were undertaken using the computing facilities of the NHM 'ctag' bioinformatics server. Both the 454 and Illumina reads underwent two *de novo* genome assembly steps. Illumina reads were first converted to 454 format using a custom AWK script (M. Timmermans *pers. comm.*). Initial assembly was with Newbler (Roche) using the same settings as Haran *et al.* (2013). The resulting contigs then underwent a second assembly with Phrap (http://www.phrap.org/).

To identify the resulting mitogenomic assemblies through association with the original specimens, the *cox1* and *cytB* 'bait' sequences previously obtained were used as queries against the mitogenomic assemblies in a BLAST search (blastn; E=1e-5).

2.4.4.6 Mitogenome assembly results

A total of 16 newly assembled and identified partial mitogenomes were obtained, which are listed in Appendix 2.6 with taxonomic identifications where known. Three of the assemblies resulted from samples sequenced on Illumina HiSeq and the remaining 13 from samples sequenced on 454 GS Junior. Of all successfully assembled samples, nine originated from originally 'bright' LR-PCR target bands and seven from 'faint' bands, indicating that LR-PCR success does not necessarily need to be very good for downstream assembly to be accomplished.

2.4.4. Direct NGS of pooled genomic DNA

Recent developments and novel applications of NGS have indicated that parallel *de novo* mitogenome assembly from pooled genomic DNA samples consisting of many species' DNA extracts is possible (Crampton-Platt *et al.* unpuplished data). A similar methodology was successfully employed here to sequence and assemble multiple mitogenomes from a bulk sample of curculionoid pooled genomic DNA. In combination with the 'bait' identification approach, discussed above for the LR PCR methodology, it was possible to identify assemblies to species with limited prior genome knowledge. This work is described in detail in Chapter 3.

2.5 Discussion

The various techniques explored during the course of this chapter were shown to be capable of producing sequences using the available curculionid genomic DNA templates. Following many trials and much optimisation, success varied from being generally very high for the 'standard' PCR amplifications of both mitochondrial and nuclear protein coding and rRNA markers, through substantial success for the LR-PCR amplification of partial mitogenomes, to the very low success rate for SPIAs, where only a single specimen resulted in a readable sequence.

To a large extent the quality and integrity of DNA present in the samples plays the most significant role in eventual amplification success. In particular dry-preserved, old museum specimens clearly provide a challenging prospect as a result of a variety of factors known to adversely affect DNA preservation. These include length of time since specimen collection, killing method (Dean & Ballard 2001; Lis *et al.* 2011) and exposure to pesticide fumigants (Espeland *et al.* 2010). DNA shearing into smaller fragments and cross-linking are thought to be the two most important explanations for loss in DNA integrity (Dean & Ballard 2001).

All preserved specimens will be adversely affected by time, which has been experimentally shown to very quickly reduce potential template amplification, as was demonstrated by Dean and Ballard (2001), who detected sharp declines in PCR success after only two years since collection in museum preserved *Drosophila* (Diptera) specimens. Over longer period of times much research has shown that it is generally very difficult to PCR-amplify DNA from dry-mounted insect specimens older than about 25 years, although for some groups such as Heteroptera (Lis *et al.* 2011) and Hymenoptera (Strange *et al.* 2009) older specimens have successfully amplified comparatively short loci such as microsatellites.

Killing method and exposure to chemical fumigants are more difficult to test for, because often no record exists as to how a specimen was killed or to which chemicals it may have been exposed to during the course of its existence in an insect cabinet. Nevertheless, recent interesting work has investigated these factors and determined that, for instance, in a set of *Drosophila* specimens, killing in ethanol resulted in lower DNA yields than killing in cyanide (Dean & Ballard 2001) although subsequent storage of specimens exposed to naphthalene for two years did not affect DNA yield or PCR success. Naphthalene is a common fumigant used in insect collections against potential pests, and a relatively benign one. However, other more toxic agents have been or continue to be in use in entomological collections. One such chemical is dimethyl 2,2-dichlorovinyl phosphate or Dichlorvos, more widely known under the trading name of 'Vapona' strips, which is how the chemical was used in collections, stuck onto the inside of drawers and store boxes. This hazardous substance is now known to be carcinogenic to humans (Weis *et al.* 1998) but was in widespread use at the Department of Entomology at the NHM until at least the mid to late 1970's when it was officially removed due to health and safety concerns (H. Mendel *pers. comm.*). However, remains of the 'Vapona' strips were still sometimes found as late as 2006 (*pers. obs.*), clearly indicating that specimens, as well as humans, were exposed to this chemical for some time. Lofroth (1970) was the first to indicate that Dichlorvos is able to alkylate nucleotides, which is probably linked to its biological effects and which is likely to also contribute towards or cause DNA interstrand cross-linking. Since then it has been experimentally demonstrated that DNA extractions and amplification from insect specimens exposed to Dichlorvos, even for the short time of four months, were clearly negatively affected (Espeland *et al.* 2010).

Most of the museum specimens used in this study were collected long enough ago that they could have been exposed to significant levels of Dichlorvos during the course of their time held at the NHM. However, in 2010 the NHM purchased a large private collection of weevils from a Czech collector (Oldřich Voříšek), which almost certainly had no exposure to Dichlorvos (no evidence was seen for its presence, *pers. obs.*), and from which some of the *Curculio* samples in this study were borrowed, before it was amalgamated with the rest of the NHM collection (*pers. obs.*). Significantly, the single specimen from which a readable *cox1* SPIA sequence was repeatedly obtained (CU008 in Appendix 2.3) is one of the specimens from the former Vořížek collection. This specimen was one of the eight that underwent substantial PCR trials and whilst a causal link cannot be attributed, especially because it is also a relatively recent specimen collected in 1998, the present results are consistent with the observations of Espeland *et al.* (2010). If alkylation of DNA by Dichlorvos causes cross-linking, it would be expected that PCR success would “correlate with the ability to denature the cross-linking bonds” (Dean & Ballard 2001), possibly explaining our observations.

It has been shown that silica-based DNA extraction protocols, not used in this study, can produce increased DNA yield from degraded specimens (Hajibabaei *et al.* 2005), possibly leading to improved PCR success. However, in the same study, which also tested the extraction procedure used here (DNeasy), improvements in PCR success from archival moth specimens were slight. The fact that we obtained faint PCR success for most museum specimens indicates that the extraction procedure is probably not the main problem.

Another problem encountered whilst undertaking the SPIA work was contamination with non-target DNA, as evidenced by the *Homo sapiens* sequences obtained. This is a difficult problem to avoid when undertaking nested reamplifications, of standard sized PCR fragments, let alone SPIAs. Ideally DNA extraction and PCR should be conducted in a ‘clean’, ‘ancient’ DNA lab, which unfortunately was not available at the time of this research. However, it must be stressed that all museum specimens are likely to have come into contact with ‘contaminant’ DNA through physical proximity to other specimens and through specimen handling over time, so that even if all reasonable precautions are taken, this problem is likely to persist.

Whilst the SPIA work culminated in disappointing results, newer technologies may yet enable SPIAs or indeed much larger fragments to be sequenced from degraded museum specimens. In fact, because most NGS platforms only sequence relatively short DNA fragments ('inserts'), which are usually purposely fragmented prior to sequencing (Ansorge 2009), it seems that they are inherently suitable for sequencing degraded DNA. Indeed a recent study was able to sequence and assemble an entire snail mitogenome from a museum specimen (Groenenberg *et al.* 2012) and ancient mitogenomics as a field of interest has been in existence for some time with increasing success as technology advances (Ho & Gilbert 2010).

DNA integrity is also clearly an important factor in determining success in LR-PCR reactions. Less than 25% of trialled samples resulted in mitogenomic assemblies despite extensive optimisations, during which LR-PCR reactions were observed to be highly stochastic in success, even when positive samples were used. The DNA-ready specimens used in this study were collected by a large number of collaborators across the world and naturally would not have been killed or preserved in a standard manner (although guidelines were sent to all collaborators regarding this), complicating direct comparisons. Indeed many specimens had been collected prior to the start of this project. The LR-PCR, and to a lesser extent the standard PCR results, reflect this heterogenous mixture of specimen preservation, which of course is a real world scenario rather than the ideal. Ultimately the LR-PCR technique is a feasible method for obtaining mitogenomes but the cost in both time spent optimising reactions and in terms of expensive specialist reagents weighs against it. Newer and simpler techniques for obtaining full mitogenomes from direct sequencing of pooled genomic DNA offer a better solution to this problem (Rubinstein *et al.* 2013) and are explored successfully in the following chapter.

2.6 References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Ansorge WJ (2009) Next-generation DNA sequencing techniques. *New Biotechnology* **25**, 195-203.

Boore JL (1999) Animal mitochondrial genomes. *Nucleic Acids Research* **27**, 1767-1780.

Botero-Castro F, Tilak M-K, Justy F*, et al.* (2013) Next-generation sequencing and phylogenetic signal of complete mitochondrial genomes for resolving the evolutionary history of leaf-nosed bats (Phyllostomidae). *Molecular Phylogenetics and Evolution* **69**, 728-739.

Bouchard P, Bousquet Y, Davies AE*, et al.* (2011) Family-group names in Coleoptera (Insecta). *Zookeys* **88**, 1-972.

Cho S, Mitchell A, Mitter C*, et al.* (2008) Molecular phylogenetics of heliothine moths (Lepidoptera: Noctuidae: Heliothinae), with comments on the evolution of host range and pest status. *Systematic Entomology* **33**, 581-594.

Crampton-Platt AL, Timmermans MJTN, Gimmel ML, Kutty SN, Cockerill TD, Khen CV, Vogler AP (Unpublished data) Pooled mitochondrial genome assembly for biodiversity discovery in a phylogenetic framework.

Dean MD, Ballard JWO (2001) Factors affecting mitochondrial DNA quality from museum preserved Drosophila simulans. *Entomologia Experimentalis Et Applicata* **98**, 279-283.

Espeland M, Irestedt M, Johanson KA*, et al.* (2010) Dichlorvos exposure impedes extraction and amplification of DNA from insects in museum collections. *Frontiers in Zoology* **7**: 10.1186/1742-9994-7-2.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular marine biology and biotechnology* **3**, 294-299.

Gilbert MTP, Moore W, Melchior L, Worobey M (2007) DNA Extraction from Dry Museum Beetles without Conferring External Morphological Damage. *PloS one* **2**: 10.1371/journal.pone.0000272.

Groenenberg DSJ, Pirovano W, Gittenberger E, Schilthuizen M (2012) The complete mitogenome of Cylindrus obtusus (Helicidae, Ariantinae) using Illumina next generation sequencing. *BMC Genomics* **13**, 1-10.

Hajibabaei M, DeWaard JR, Ivanova NV*, et al.* (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**, 1959-1967.

Hajibabaei M, Smith MA, Janzen DH*, et al.* (2006) A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes* **6**, 959-964.

Hall T (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *NUcleic Acids Symposium Series* **41**, 95-98.

Haran J, Timmermans MJTN, Vogler AP (2013) Mitogenome sequences stabilize the phylogenetics of weevils (Curculionoidea) and establish the monophyly of larval ectophagy. *Molecular Phylogenetics and Evolution* **67**, 156-166.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B-Biological Sciences* **270**, 313-321.

Hernández-Vera G, Caldara R, Tosevski I, Emerson BC (2013) Molecular phylogenetic analysis of archival tissue reveals the origin of a disjunct southern African-Palaearctic weevil radiation. *Journal of Biogeography* **40**, 1348-1359.

Ho SYW, Gilbert MTP (2010) Ancient mitogenomics. *Mitochondrion* **10**, 1-11.

Hughes J, Vogler AP (2004) The phylogeny of acorn weevils (genus Curculio) from mitochondrial and nuclear DNA sequences: the problem of incomplete data. *Molecular Phylogenetics and Evolution* **32**, 601-615.

Hunt T, Bergsten J, Levkanicova Z*, et al.* (2007) A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science* **318**, 1913-1916.

ICZN (1999) *International Code of Zoological Nomenclature*, Fourth edn. International Trust for Zoological Nomenclature, London.

Kearse M, Moir R, Wilson A*, et al.* (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649.

Lis JA, Ziaja DJ, Lis P (2011) Recovery of mitochondrial DNA for systematic studies of Pentatomoidea (Hemiptera: Heteroptera): successful PCR on early 20th century dry museum specimens. *Zootaxa* **2748**, 18-28.

Liu L, Li Y, Li S*, et al.* (2012) Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology* **2012**, 251364-251364.

Lofroth G (1970) Alkylation of DNA by dichlorvos. *Naturwissenschaften* **57**, 393-394.

McKenna DD, Sequeira AS, Marvaldi AE, Farrell BD (2009) Temporal lags and overlap in the diversification of weevils and flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7083-7088.

Mitchell A, Maddox C (2010) Bark beetles (Coleoptera: Curculionidae: Scolytinae) of importance to the Australian macadamia industry: an integrative taxonomic approach to species diagnostics. *Australian Journal of Entomology* **49**, 104-113.

Monaghan MT, Inward DJG, Hunt T, Vogler AP (2007) A molecular phylogenetic analysis of the Scarabaeinae (dung beetles). *Molecular Phylogenetics and Evolution* **45**, 674-692.

Posada D, Buckley T (2004) Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology* **53**, 793-808.

Rose TM, Henikoff JG, Henikoff S (2003) CODEHOP (COnsensus-DEgenerate hybrid oligonucleotide primer) PCR primer design. *Nucleic Acids Research* **31**, 3763-3766.

Rubinstein ND, Feldstein T, Shenkar N*, et al.* (2013) Deep sequencing of mixed total DNA without barcodes allows efficient assembly of highly plastic ascidian

Shull VL, Vogler AP, Baker MD, Maddison DR, Hammond PM (2001) Sequence alignment of 18S ribosomal RNA and the basal relationships of Adephagan beetles: evidence for monophyly of aquatic families and the placement of Trachypachidae. *Systematic Biology* **50**, 945-969.

Strange JP, Knoblett J, Griswold T (2009) DNA amplification from pin-mounted bumble bees (Bombus) in a museum collection: effects of fragment size and specimen age on successful PCR. *Apidologie* **40**, 134-139.

Talavera G, Vila R (2011) What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta class phylogeny. *BMC Evolutionary Biology* **11**: 10.1186/1471-2148-11-315.

Timmermans MJTN, Dodsworth S, Culverwell CL*, et al.* (2010) Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research* **38**: 10.1093/nar/gkq807.

Weis N, Stolz P, Krooss J, Meierhenrich U (1998) Dichlorvos insect strips indoors: pollution and risk assessment. *Gesundheitswesen (Bundesverband der Arzte des Offentlichen Gesundheitsdienstes (Germany))* **60**, 445-449.

Wortley AH, Rudall PJ, Harris DJ, Scotland RW (2005) How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. *Systematic Biology* **54**, 697-709.

Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology* **51**, 588-598.

2.7 Appendices

Appendix 2.1 Primers (A), PCR reaction mixes (B) and cycling conditions (C) for the mitochondrial gene amplifications: *cox1*, *cytB* and *rrnL*.

A) Primers used in the amplification and sequencing of mitochondrial gene markers with annealing temperature and length of amplified fragment.

Primer name	Marker	Dir.	Primer sequence 5'=>3'	Anneal Temp.	Length (kbp)
M13REV-FOLbeetF2*	*cox1* 5'	Fwd.	CAGGAAACAGCTATGACCTTYTCWACNAAYCAYAARGAYATYGG	52°C	0.7
M13(-21)-FOLbeR2**	*cox1* 5'	Rev.	TGTAAAACGACGGCCAGTTANACTTCWGGRTGNCCRAARAAYCA		
M13REV*	*cox1* 5' (Seq.)	Fwd.	CAGGAAACAGCTATGACC	50°C	0.7
M13(-21)**	*cox1* 5' (Seq.)	Rev.	TGTAAAACGACGGCCAGT		
Jerry	*cox1* 3'	Fwd.	CAACATTTATTTTGATTTTTTGG	53°C	0.8
SpatR	*cox1* 3'	Rev.	GCACTAWTCTGCCATATTAGA		
SytB_F	*cytb*	Fwd.	TGAGGNCAAATATCHTTYTGAGG	55°C	0.5
SytB_R	*cytb*	Rev.	GCAAATARRAARTATCATTCDGG		
LRJ-12961	*rrnL*	Fwd.	TTTAATCCAACATCGAGG	50°C	0.45
LRJ-12887	*rrnL*	Fwd.	CCGGTCTGAACTCAGATCACGT		
LRN-13398	*rrnL*	Rev.	CGCCTGTTTAACAAAAACAT		

* PCR products amplified with M13REV-FOLbeetF2 were sequenced using M13REV

** PCR products amplified with M13(-21)-FOLbeR2 were sequenced using M13(-21)

B) PCR reaction mixes for the amplification of mitochondrial gene markers

PCR Component	***cox1* 5’ X1 (µl)**	***cox13’* X1 (µl)**	***cytB* X1 (µl)**	***rrnL* X1 (µl)**
ddH$_2$O	15.925	18.15	18.15	18.8
NH$_4$ buffer X10	2.5	2.5	2.5	2.5
MgCl$_2$ (50mM)	1.5	1.0	1.0	1.0
dNTPs (10mM total/2.5mM each)	2.0	1.0	1.0	1.0
Forward primer (10µM)	1.0	0.625	0.625	0.6
Reverse primer (10µM)	1.0	0.625	0.625	0.6
Taq polymerase	0.075	0.1	0.1	0.1
DNA	1.0	1.0	1.0	1.0
Total volume	25.0	25.0	25.0	25.6

C) PCR cycling conditions for the amplification of mitochondrial gene markers

	***cox1* 5’**			***cox1* 3’**			***cytB***			***rrnL***		
	Temp.	**Duration**		**Temp.**	**Duration**		**Temp.**	**Duration**		**Temp.**	**Duration**	
Initialisation	95°C	2 mins		94°C	2 mins		94°C	5 mins		94°C	5 mins	
Denaturation	95°C	1m		94°C	30s		94°C	30s		94°C	30s	
Annealing	52°C	45s	X39	53°C	30s	X35	55°C	30s	X35	50°C	30s	X35
Extension	72°C	1m		70°C	1m		70°C	1m		72°C	30s	
Final extension	72°C	5 mins		72°C	10 mins		72°C	10 mins		72°C	7 mins	
Final hold	10°C	∞		10°C	∞		10°C	∞		10°C	∞	

Appendix 2.2 Primers (A), PCR reaction mixes (B) and cycling conditions (C) for the nuclear gene amplifications: 18S, 28S and *ArgK.*

A) Primers used in the amplification and sequencing of nuclear gene markers with annealing temperature and length of amplified fragment.

Primer name	Marker (notes)	Direction	Primer sequence 5’=>3’	Anneal Temp.	Length (kbp)
18S5’	18S (section 1)	Fwd.	GACAACCTGGTTGATCCTGCCAGT	54°C	1.3
18Sbi	18S (section 1)	Rev.	GAGTCTCGTTCGTTATCGGA	54°C	1.3
18Sb5.0-rw	18S (sequencing only)	Rev.	TAACCGCAACAACTTTAAT	58°C	0.6
18Sai	18S (sequencing only)	Fwd.	CCTGAGAAACGGCTACCACATC	58°C	0.8
18Sa2.0	18S (section 2)	Fwd.	ATGGTTGCAAAGCTGAAAC	52°C	0.7
18S3’I	18S (section 2)	Rev.	CACCTACGGAAACCTTGTTACGAC	52°C	0.7
28SFF	28S	Fwd.	TTACACACTCCTTAGCGGAT	50°C	0.8
28SDD	28S	Rev.	GGGACCCGTCTTGAAACAC	50°C	0.8
ArgKforB2	*ArgK* (1st round)	Fwd.	GAYTCCGGWATYGGWATCTAYGCTCC	50°C	0.65
ArgKrevB1	*ArgK* (1st round)	Rev.	TCNGTRAGRCCCATWCGTCTC	50°C	0.65
ArgK_F1_semidg*	*ArgK* (2nd round)	Fwd.	GATCCCATCATHGARGAYTARCA	55°C	0.6
ArgK_R2_fulldg*	*ArgK* (2nd round)	Fwd.	GTNCCYAARTTNGTNGGRCARAA	55°C	0.6

*Newly designed ‘nested re-amplification’ primers, also used for sequencing of *ArgK*

B) PCR reaction mixes for the amplification of nuclear gene markers

PCR Component	18S X1 (µl)	28S X1 (µl)	*ArgK* (1st) X1 (µl)	*ArgK* (2nd) X1 (µl)
ddH_2O	14.86	18.8	14.4	11.4
NH_4 buffer X10	2.0	2.5	2.5	2.5
$MgCl_2$ (50mM)	0.8	1.0	2.5	2.0
dNTPs (10mM total/2.5mM each)	0.8	1.0	2.0	2.0
BSA (20mg/ml)	n/a	n/a	n/a	2.0
Forward primer (10µM)	0.5	0.6	1.25	1.0
Reverse primer (10µM)	0.5	0.6	1.25	1.0
Taq polymerase	0.04	0.1	0.1	0.1
DNA	0.5	1.0	1.0	1.0*
Total volume	20.0	25.6	25.0	23.0

* 1.0 µl of PCR product from 1st round PCR

C) PCR cycling conditions for the amplification of nuclear gene markers

	18S (sections 1 and 2)			**28S**			***ArgK* (1st round)**			***ArgK* (2nd round)**		
	Temp.	**Duration**		**Temp.**	**Duration**		**Temp.**	**Duration**		**Temp.**	**Duration**	
Initialisation	94°C	5 mins		94°C	5 mins		95°C	15 secs		94°C	5 mins	
Denaturation	94°C	45s	X35	94°C	30s	X30	95°C	1m	X35	94°C	30s	X35
Annealing	54/52°C*	45s		53°C	30s		50°C	1m		50°C	30s	
Extension	72°C	2m		72°C	30s		72°C	2m		72°C	30s	
Final extension	72°C	10 mins		72°C	7 mins		72°C	2 mins		72°C	7 mins	
Final hold	10°C	∞		10°C	∞		10°C	∞		10°C	∞	

* 54°C for 18S section 1 / 52°C for section 2

Appendix 2.3 Dry-preserved, mounted museum specimens of *Curculio* from which DNA was extracted from a single leg during SPIA development.

Specimen	Species	Label data	Leg extracted
CU001	*C. glandium*	Greece, Merlin Coll., 96-275	Right hind
CU002	*C. glandium*	Europe	Left hind
CU003	*C. glandium*	Morocco: Col du Zad, 13.vii.1969, M. Vazquez., BMNH(E) 2005-108	Right middle
CU004	*C. glandium*	Shirley, Surrey., G.C.C., B.C. Champion coll., BM 1927-409	Right middle
CU005	*C. nucum*	Vernet Pyr. Or., June 1891, D.S. Sharp coll., BM 1927-409	Right middle
CU006	*C. nucum*	Slov. Mer., Domica, coll. Vorisek, BMNH(E) 2010-26, O. Vorisek	Left middle
CU007	*C. pellitus*	Hungary, J. Hughes, 274, Pic	Right middle
CU008	*C. venosus*	TR – prov. Afyon, 3km W of Basoren, 26.v.1998, J. Varisek lgt. BMNH(E) 2010-26, O. Vorisek	
CU009	*C. glandium*	St. Malo, Pascoe Coll 93-60	Middle, right
CU010	*C. glandium*	Berlin. Ruthe Coll., 58-134	Back, right
CU011	*C. glandium*	PERSIA, Astrabad, 5.99., Hauser Coll. 1904-63	Back, left
CU012	*C. glandium*	MOROCCO: Jebel Aoua, 22.vi.1969, M. Vazquez, BMNH(E) 2005 - 108	Middle, right
CU013	*C. glandium*	Europe	Back, left
CU014	*C. glandium*	GALLIA, BIGORRE, Sharp Coll. 1905-313	Middle, left
CU015	*C. venosus*	Slov. M., Heav farok, Vorisek, 29.v.63/ O. Vorisek, BMNH(E) 2010-26	Back, right
CU016	*C. glandium*	Slov. M. Plagivac, Vorisek, 16.v.53 / O. Vorisek, BMNH(E) 2010-26	Back, right
CU017	*C. glandium*	GRAECIA, Morea, Kerpini, leg. Steiner, 5.1967 / O. Vorisek, BMNH(E) 2010-26	Middle, left
CU018	*C. glandium*	86m, Bistrec 11.5.85, 40km S v. Karnobat, Kadlec + Vorisek lg. / O. Vorisek, BMNH(E) 2010-26	Back, left
CU019	*C. glandium*	Maroc. 23.5.67, Marrakech, Dr. Vazquez / O. Vorisek, BMNH(E) 2010-26	Back, right
CU020	*C. glandium*	Parkan, 19.5.1936, B. Stichce / O. Vorisek, BMNH(E) 2010-26	Middle, right
CU021	*C. glandium*	Orpington, Kent, 9.5.1948, E. Gowing-Scopes / E. Gowing -Scopes collection, BMNH(E) 2005-4	Middle, left
CU022	*C. glandium*	Darenth wood, Kent, G.C.C. / G.C. Champion Collection, B.M. 1964-540	Middle, right
CU023	*C. venosus*	MORAVIA 30.4.66, Straznice, Dr. A. Svozil lgt. / O. Vorisek, BMNH(E) 2010-26	Middle, left
CU024	*C. nucum*	40, 4 1, 146	Middle, left
CU025	*C. nucum*	Darenth wood, Kent, G.C.C. / G.C. Champion Collection, B.M. 1964-540	Middle, left
CU026	*C. nucum*	Orpington, Kent, 10.5.1947, E. Gowing-Scopes / E. Gowing -Scopes collection, BMNH(E) 2005-4	Middle, right
CU027	*C. nucum*	Ajdovscina, Lokavec / 16.5.1976, Slovenija / O. Vorisek, BMNH(E) 2010-26	Back, right
CU028	*C. nucum*	W. Bezdekov, Vanek vii.68 / O. Vorisek, BMNH(E) 2010-26	Middle, right
CU029	*C. pellitus*	Vernet Pyr. Or., June 1891, D.S. / Sharp coll., BM 1905-313	Back, right
CU030	*C. pellitus*	Slov. M., Sturovo, Vorisek, 28.v.63 / O. Vorisek, BMNH(E) 2010-26	Back, right
CU031	*C. pellitus*	Croatia littor., Makarska 9.8.2000, M. Bocan leg. / O. Vorisek, BMNH(E) 2010-26	Back, left
CU032	*C. venosus*	Vernet Pyr. Or., June 1891. D.S / Sharp Coll., 1905-313	Back, left
CU033	*C. venosus*	Berlin. Ruthe Coll., 58-134 / J. Hughes 135, Picture 135	Middle, left
CU034	*C. venosus*	Sierra d'Alfacar, 10.7.1879, D.S. / Sharp Coll., 1905-313	Back, right
CU035	*C. venosus*	Darenth wood, Kent, G.C.C. / G.C. Champion Collection, B.M. 1964-540	Middle, right
CU036	*C. venosus*	74-5, Power, Darenth. / J.A. Power Coll. , Pres. W.A. Power, BMNH(E) 1896-69	Back, left
CU037	*C. venosus*	Woking, Surrey, G.C.C. / G.C. Champion Collection, B.M. 1964-540	Middle, left

Appendix 2.4 Primers (A), PCR reaction mixes (B) and cycling conditions (C) for the *cox1* and *rrnL* SPIA PCR amplifications.

A) Primers used in the amplification and sequencing of SPIAs with annealing temperature and length of amplified fragment.

Primer name	**Marker (notes)**	**Dir.**	**Primer sequence 5'=>3'**	**Anneal Temp.**	**Length (bp)**
FOLbeetF2	*cox1* 5' SPIA	Fwd.	TTYTCWACNAAYCAYAARGAYATYGG	45°C	129/234
CO1_miniR_semidg	*cox1* 5' SPIA	Rev.	TGTAAAACGACGGCCAGT AAAAATTATAATAAADGCRTGDGC	45°C	129
CO1_mini_R1_semidg	*cox1* 5' SPIA	Rev.	TGTAAAACGACGGCCAGT CTTATATTRTTNANNCGNGG	45°C	234
CO1minR_semdgNT*	*cox1* 5' SPIA	Rev.	AAAAATTATAATAAADGCRTGDGC	50°C	129
M13(-21)**	*cox1* 5' (Sequencing)	Rev.	TGTAAAACGACGGCCAGT	50°C	129/234
16S_7bp_FGer	***rrnL***	**Fwd.**	GTAAAACGACGGCCAGTAATMATTAGTTTYTTAATT	45°C	**95**
16S_7bp_RGer	***rrnL***	**Rev.**	TAYAGGGTCTTCTCGTCTT	45°C	**95**
16S_48bpF1	***rrnL***	**Fwd.**	CGAGAAGACCCTATAGAGTTT	45°C	**55**
16S_48bpR1	***rrnL***	**Rev.**	TCAATCACCCCAAYYAAAT	45°C	**55**

*Non-tailed reverse primer, attempted in sequencing reactions
** PCR products amplified with CO1_miniR_semidg or CO1_mini_R1_semidg as reverse primers were sequenced using M13(-21)
All *rrnL* primers were designed by Hernández-Vera *et al.* (2013)

B) PCR reaction mixes for the amplification of SPIA markers

PCR Component	*cox1* 'short' SPIA 129 bp X1 (µl)	*cox1* 'long' SPIA 234 bp X1 (µl)	*rrnL* SPIA X1 (µl)
ddH$_2$O	11.4	11.9	11.4
NH$_4$ buffer X10	2.5	2.5	2.5
MgCl$_2$ (50mM)	2.0	1.5	2.0
dNTPs (10mM total/2.5mM each)	2.0	2.0	2.0
Forward primer (10µM)	1.0	1.0	1.0
Reverse primer (10µM)	1.0	1.0	1.0
Taq polymerase	0.1	0.1	0.1
DNA	5.0	5.0	5.0
Total volume	25.0	25.0	25.0

* 1.0 µl of PCR product from 1st round PCR

C) PCR cycling conditions for the amplification of SPIA markers

	cox1 'short' SPIA 129 bp			*cox1* 'short' SPIA 234 bp			*rrnL* SPIA 129 bp			Sequencing		
	Temp.	Duration		Temp.	Duration		Temp.	Duration		Temp .	Duration	
Initialisation	95°C	5 mins		95°C	5 mins		95°C	5 mins		96°C	1 min	
Denaturation	95°C	20s	X45	95°C	20s	X45	95°C	20s	X40	96°C	10s	X24
Annealing	45°C	20s		45°C	20s		45°C	20s		50°C	5s	
Extension	72°C	20s		72°C	30s		72°C	20s		60°C	4m	
Final extension	72°C	5 mins		72°C	5 mins		72°C	5 mins				
Final hold	10°C	∞		10°C	∞		10°C	∞		10°C	∞	

Appendix 2.5 PCR reaction mix and cycling conditions for the *cox1* to *cytB* partial mitogenome Long-Range PCR amplification. Primers (Jerry and SytBR) have previously been listed under the *cox1* 3' and *cytB* PCR tables respectively

PCR component	**Volume (µl) X1**	**Long-Range PCR (*cox1* to *cytB*)**			
ddH_2O	15.8		**Temp.**	**Duration**	
TaKaRa buffer (Mg^{2+} added)	3.0	**Initialisation**	94°C	1 min	
dNTPs (10mM total/2.5mM each)	4.0	**Denaturation**	98°C	5s	X35
Forward primer (10µM)	0.5	**Annealing**	53°C	30s	
Reverse primer (10µM)	0.5	**Extension**	60°C	15m	
TaKaRa LA *Taq* polymerase	0.2	**Final extension**	72°C	10 mins	
DNA	1.0	**Final hold**	10°C	∞	
Total volume	25.0				

Appendix 2.6 Samples for which successful mitogenomic assembly of NGS sequenced LR-PCR products was achieved. Three samples in grey were sequenced on Illumina HiSeq, all 13 others on 454 GS Junior. ? denotes unidentified taxa.

Code	Family	Subfamily	Tribe	Genus	Species	Origin	Source
CG050	Curculionidae	Ceutorhynchinae	Mononychini	*Mononychus*	*puntumalbum*	Italy	Caldara
CG051	Curculionidae	Curculioninae	Curculionini	*Curculio*	*glandium*	Italy	Caldara
CG060	?	?	?	?	?	Belize	Barclay
CG063	Curculionidae	Curculioninae	Ellescini	*Dorytomus*	*suratus*	Italy	Caldara
CG070	Curculionidae	Curculioninae	Curculionini	*Archarius*	*pyrrhoceras*	England	Gillett
CG093	?	?	?	?	?	England	Gillett
CG094	?	?	?	?	?	Ecuador	Gillett
CG114	?	?	?	?	?	Ecuador	Gillett
CG115	?	?	?	?	?	Ecuador	Gillett
CG119	?	?	?	?	?	Ecuador	Gillett
CG144	?	?	?	?	?	Ecuador	Gillett
CG150	?	?	?	?	?	Ecuador	Gillett
CG153	?	?	?	?	?	Ecuador	Gillett
CG157	?	?	?	?	?	Ecuador	Gillett
CG304	Curculionidae	Curculioninae	Anthonomini	*Anthonomus*	*pedicularius*	Sweden	Andersson
CG310	Curculionidae	Entiminae	Oosomini	*Barianus*	*uniformis*	Juan de Nova	Kitson

Chapter 3

Bulk *de novo* mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea)

An adaptation of this chapter was accepted for publication as an article in Molecular Biology and Evolution on 29 April, 2014, and first published online on 6 May 2014:
Gillett CPDT, Crampton-Platt A, Timmermans MJTN, Jordal BH, Emerson BC, Vogler AP (2014) Bulk *De Novo* mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution* **31** (8): 2223-2237.

"A satisfactory resolution of the Curculionidae into subfamilies and tribes is probably the largest and most important outstanding problem in the higher classification of Coleoptera, particularly as a great number of species are more or less injurious to economical plants."

- Roy Crowson, 1955

Mecysolobini sp. (Curculionidae: Molytinae), Yunnan, China

Chapter 3: Bulk *de novo* mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea)

3.1 Abstract

Complete mitochondrial genomes have been shown to be reliable markers for phylogeny reconstruction among diverse animal groups. However, the relative difficulty and high cost associated with obtaining *de novo* full mitogenomes has frequently led to conspicuously low taxon sampling in ensuing studies. Here, successful use of an economical and accessible method for assembling complete or near-complete mitogenomes through shot-gun next generation sequencing of a single library made from pooled total DNA extracts of numerous target species is reported. To avoid the use of separate indexed libraries for each specimen, and an associated increase in cost, standard PCR-based 'bait' sequences are incorporated to identify the assembled mitogenomes. The method was applied to study basal relationships in the weevils (Coleoptera: Curculionodea), producing 92 newly assembled mitogenomes obtained in a single Illumina MiSeq run, which were used to analyse the higher-level phylogenetic relationships of weevils. The analysis suported a separate origin of wood-boring behaviour by the subfamilies Scolytinae, Platypodinae and Cossoninae. This finding contradicts morphological hypotheses proposing a close relationship between the first two of these, but is congruent with previous molecular studies, reinforcing the utility of mitogenomes in phylogeny reconstruction. Our methodology provides a technically simple procedure for generating densely sampled trees from

whole mitogenomes, and is widely applicable to groups of animals for which bait sequences are the only required prior genome knowledge.

3.2 Introduction

With the advent of high-throughput next generation sequencing (NGS) technologies and their ability to generate large amounts of data suitable for genomic assembly, systematists are increasingly adopting such methods to reconstruct complete mitochondrial genomes (mitogenomes) to infer phylogenies across a diverse range of taxa. Such research has provided compelling insights in studies ranging from the investigation of deep-level metazoan relationships (e.g. Cnidaria; Osigus *et al.* 2013) to those within single phyla (e.g. Cnidaria; Kayal *et al.* 2013), orders (e.g. Primates; Finstermeier *et al.* 2013), families (e.g. Braconidae wasps; Wei *et al.* 2010) and genera (e.g. *Architeuthis* giant squid; Winkelmann *et al.* 2013). Mitogenomes have an intrinsic suitability for phylogenetic analysis due to their unambiguous orthology (Botero-Castro *et al.* 2013), varying nucleotide substitution rates that contribute to phylogenetic signal at diverse taxonomic ranks, and their uniparental inheritance consistent with bifurcating phylogenetic trees (Curole & Kocher 1999). In addition, mitochondrial DNA (mtDNA) is present in multiple copies per cell, facilitating its amplification and sequencing, which has undoubtedly contributed to the wide use of mitochondrial markers in phylogeny reconstruction. However, in spite of these advantages, complete mitogenome sequencing has been comparatively labour intensive and costly, resulting in often conspicuously few newly-generated mitogenomes per study (e.g. 17 bird mitogenomes in Pacheco *et al.* (2011), four complete Cnidarian mitogenomes in Kayal *et al.* (2013) and one cockroach and 13 termite mitogenomes in Cameron *et al.* (2012). Techniques have almost always included either shotgun sequencing of expensive multiple indexed-libraries (Botero-Castro *et al.* 2013) or a target-enrichment step such as primer walking using standard PCR amplification of overlapping fragments (Botero-Castro *et al.* 2013), long-range PCR followed by either sequencing-primer walking (Roos *et al.* 2007) or NGS (Timmermans *et al.* 2010), and hybrid-capture using sheared long-range PCR products as 'baits' immobilised on magnetic beads (Winkelmann *et al.* 2013). While these techniques can generate full mitochondrial genomes, each of them has limitations that generally restrain the number of taxa or samples that can be incorporated economically within a study.

The present study aims to address this sampling bottleneck by testing the possibility of parallel *de novo* mitogenome assembly from a single library of pooled genomic DNA from a bulk sample consisting of many species. This method has recently been applied to sequencing of environmental samples of arthropods from a rainforest canopy (Crampton-Platt *et al.* unpuplished data). This technique is applied here to investigate the higher-level phylogeny of an extremely diverse superfamily of insects, the weevils (Coleoptera: Curculionoidea), composed of no fewer than 62,000 described species distributed wherever terrestrial plants grow (Oberprieler *et al.* 2007). The current higher-level classification proposed by Bouchard *et al.* (2011) recognises 9 extant families, amongst which the Curculionidae *s.str.* is by far the largest, containing at least 51,000 species in 17 subfamilies and 292 tribes and subtribes. The phylogenetic classification of the weevils was recognised by the eminent beetle taxonomist Crowson (1955) as "…probably the largest and most important problem in the higher classification of Coleoptera…". Since that time there have been considerable advances in our understanding of the phylogeny of this group, with significant morphological analyses by Kuschel (1995) and Marvaldi

(1997). More recently, molecular data have contributed towards reconstructing weevil higher-level relationships, including studies by McKenna *et al.* (2009), Hundsdoerfer *et al.* (2009) and Jordal *et al.* (2011), which each incorporated between two and six gene markers. A recent analysis of 27 weevil mitogenomes using 12 protein-coding genes (Haran *et al.* 2013) supported the paraphyly of Curculionidae *s.str*. as currently defined because the subfamily Platypodinae was recovered in a distant position, in a clade with members of the families Dryophthoridae and Brachyceridae, that together were sister to all other Curculionidae. Although undertaken with limited taxon sampling within the Curculionidae *s.str.* (18 tribes), this last study also supported the division of the family into two large clades; one comprising the 'broad-nosed' weevils (subfamilies Entiminae, Cyclominae and Hyperinae) and another containing the remaining subfamilies (except for Platypodinae). In the same study a tRNA^{Ala} to tRNA^{Arg} gene order rearrangement was identified in a cluster of six tRNA genes, located between *nad3* and *nad5*, which appears to be a synapomorphy for the 'broad-nosed' weevil subfamilies, further supporting their monophyly. This topology was consistent with that proposed by McKenna *et al.* (2009), who concluded that the initial diversification of weevils occurred on gymnosperm plants during the Early to early Middle Jurassic.

The Platypodinae is one of several weevil subfamilies that are specialist wood-borers, together with the bark-beetles (Scolytinae) and the subfamily Cossoninae, although other subfamilies also contain xylophagous members (e.g. Molytinae, Cryptorhynchinae and Conoderinae). The evolution of wood-boring behaviour was investigated in detail by Jordal *et al.* (2011), whose analyses incorporated morphological characters together with molecular data, concluding that both Scolytinae and Platypodinae are derived lineages within the Curculionidae *sensu* Oberprieler *et al.* (2007). However several important head characters that underpin this relationship are likely to be homoplasious and associated with tunnelling habit (Jordal *et al.* 2011). Thompson (1992) identified distinct characters of the platypodine eighth abdominal sternite and male genitalia, which indicated a distant relationship to Scolytinae and a possible justification for their inclusion in a separate curculionoid family. Therefore, the question about the polyphyly of wood-boring lineages remains open, and the failure of previous mitogenome studies to recover the platypodine and scolytine lineages as monophyletic (Haran *et al.* 2013) may be due to limited taxon sampling. The issue therefore may only be resolved if Jordal et al.'s (2011) comprehensive taxon sampling of wood-boring lineages could be matched using mitochondrial genomes.

3.3 Materials and methods

3.3.1 Taxon sampling, DNA extraction and quantification

Throughout this study the most recent higher-level classification of Curculionoidea, proposed by Bouchard *et al.* (2011) is adhered to, whilst the assignment of genera to higher taxa follows the catalogue of Alonso-Zarazaga and Lyal (1999). A total of 173 weevil specimens identified to species or higher-level and obtained through collecting or loans were selected for sequencing, representing a wide range of weevil lineages, including 7 different families and 16 subfamilies and 104 tribes within the Curculionidae. DNA was extracted from each ethanol-preserved specimen individually using DNeasy blood and tissue extraction kits (Qiagen). Aliquots from 31 specimens had already been extracted for a previous study (Jordal *et al.* 2011). The concentration of double-stranded DNA (dsDNA) in most of the extractions (139 of

173) was assayed on a Qubit fluorometer using a dsDNA high-sensitivity kit (Invitrogen).

3.3.2 'Bait' sequence PCR

Standard PCR reactions to amplify 4 different fragments of mitochondrial DNA (*cox1* 5' 'barcode region', *cox1* 3' region, 16S and *Cytb*) were undertaken for each of the 173 samples. Primers and reaction conditions are listed in Table 2.2 A-C. PCR products were cleaned with a size-exclusion filter (Merck Millipore) then Sanger-sequenced and the resulting bait sequences were subsequently employed to identify mitogenomic assemblies in the manner described by Timmermans *et al.* (2010) and as detailed below.

3.3.3 Sample pooling and sequencing

A sample preparation strategy aimed to minimise the effects of DNA concentration on sequencing was employed to theoretically lead to more even read coverage and therefore to maximise assembly success across all the samples. Approximately equimolar quantities of genomic DNA from each of the samples was pooled together based on a calculation allowing approximately 10 ng of dsDNA per sample, resulting in a total mass of pooled DNA of approximately 1.5 µg. This calculation did not consider 31 samples which were not quantified because of limited sample volume. For each of these, a fixed volume of either 5 or 8 µl was added to the pool. The final concentration of dsDNA in the pooled sample was measured using the Qubit to be 27.6 ng/µl. Based on the findings of Crampton-Platt *et al.* (unpublished data), where longer insert size was found to result in longer mitochondrial contigs, a TruSeq library was prepared from the pool aiming for an insert size of 800 bp. Quantification of the final library indicated that the average insert size was 790 bp and this was sequenced on a single Illumina MiSeq run using the 500-cycle version 2 kit (250 bp paired-end reads).

Table 3.1 List of software used for the *de novo* assembly of mitogenomes, with their main function and source URL.

Program	**Function**	**URL**
FastQC	NGS quality assesment	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trimmomatic	Adapter trimming	http://www.usadellab.org/cms/index.php?page=trimmomatic
Celera	Genome assembly	http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Main_Page
IDBA-UD	Genome assembly	http://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/
Minimus2	Merging sequence sets	http://sourceforge.net/apps/mediawiki/amos/index.php?title=Minimus2
Prinseq	Sequence quality control	http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi
COVE	tRNA annotation	http://selab.janelia.org/software.html
FeatureExtract	Gene extraction	http://www.cbs.dtu.dk/services/FeatureExtract/
Geneious	Gene annotation/sequence editing	http://www.geneious.com/
MAFFT	Sequence alignment	http://mafft.cbrc.jp/alignment/software/
BLAST	Local alignment search	http://blast.ncbi.nlm.nih.gov/Blast.cgi
PartitionFinder	Partitioning scheme selection	http://www.robertlanfear.com/partitionfinder/
CIPRES	Phylogenetic analysis server	http://www.phylo.org/
RAxML	Maximum Likelihood phylogenetic analysis	http://sco.h-its.org/exelixis/software.html
APE package in R	Phylogenetic analysis	http://ape-package.ird.fr/
Grinder	NGS simulator	http://sourceforge.net/projects/biogrinder/

3.3.4 Mitogenomic assembly pipeline

The bioinformatics assembly pipeline used in this study was developed by Crampton-Platt *et al.* (unpublished data) and is followed here with minor modifications. A list of the software required (all available as freeware) is given in Table 3.1 and a schematic overview of the principal steps described below is presented in Figure 3.1. The raw MiSeq sequence reads (in both directions, R1 and R2) were first checked for quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) before

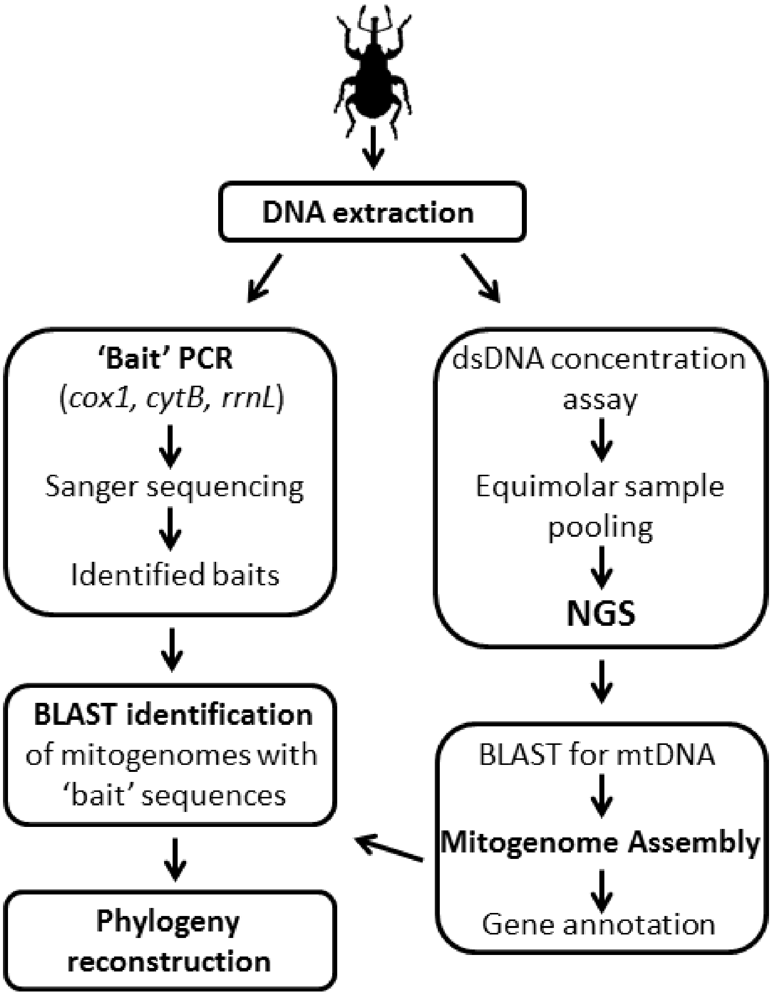


Figure 3.1 Schematic flowchart of the principal steps for the bulk *de novo* assembly of mitogenomes and identification with PCR-amplified 'bait' fragments.

the known Truseq adapters and other Illumina-specific sequences were removed with Trimmomatic (Lohse *et al.* 2012) using a combination of both palindromic and simple trimming. Once trimmed, the reads were again checked for quality using FastQC and were subsequently converted to FASTA format using a perl script (fq2fa; http://arthropods.eugenes.org/genes2/evigene/scripts/rnaseq/fq2fa.pl), then filtered (R1 and R2 separately) for mitochondrial DNA (mtDNA) through two BLAST searches (E=1e-5; no restriction in length overlap) (Altschul *et al.* 1990;

http://nebc.nerc.ac.uk/bioinformatics/documentation/blast+/user_manual.pdf) against a custom reference database of 258 Coleoptera mitogenomes (M. Timmermans, *pers. comm.*), partially obtained from GenBank (Benson *et al.* 2013). The resulting BLAST results for the R1 and R2 reads were concatenated together (using the Linux command 'cat') before the putative mtDNA paired reads were extracted from the concatenated BLAST output using a custom Perl script (FastqExtract3.pl, Appendix 3.1A). Assembly of the extracted mtDNA reads was undertaken separately using two *de novo* whole-genome shotgun DNA sequence assembler programs: Celera Assembler (Myers *et al.* 2000) and IDBA-UD (Peng *et al.* 2011); incorporating overlap-based and de Bruijn graph-based methodologies respectively. Assembly parameters for the Celera assembler (command in the .spec input file) were: doToggle=1, toggleUnitigLength=500, unitigger=bogart, createACE=1. The command for the IDBA-UD assembler contained the following parameters: --num_threads 2, --maxk 250 .

The two sets of assembled contigs, one from each assembler, were again separately filtered for mtDNA contigs using BLAST searches against the reference Coleoptera mitogenomes (E=1e-5). The resulting mtDNA matches were filtered for contigs of 1000 bp or greater length using an AWK command (Appendix 3.1B). The remaining set of contigs ≥ 1000 bp from each assembler were retrieved from the BLAST results using a custom python script (retrieve2.py, Appendix 3.1C) and were both merged using the program Minimus2 (Sommer *et al.* 2007), to combine overlapping sequences from both assemblers into longer scaffolds, reducing redundancy between them. Separate FASTA files were generated for each assembly with a custom Perl script (all2many.pl, Appendix 3.1D) and tRNA annotations were mapped onto these using COVE (Eddy & Durbin 1994), which implements covariance models of RNA secondary structure. Assemblies greater than 15 kb in length were inspected individually for circularity with reference to the *Tribolium castaneum* mitogenome (Genbank accession NC_003081). Assemblies longer than 16-17 kb were checked for duplicated regions, arising due to continued assembly beyond circularity. All resulting assemblies, containing tRNA annotations, were uploaded to the FeatureExtract 1.2 (Wernersson 2005) server, where the intervening protein and rRNA coding genes were extracted. These sequences were mapped onto the reference *Tribolium* mitogenome in Geneious for gene identification and were afterwards exported, by gene, into separate FASTA files. Short sequences of less than 1/3 the total length of each gene were discarded.

3.3.5 Identification of mitogenomic assemblies using 'bait' sequences

To identify the mitogenomic assemblies, by association with their respective originating specimen, stringent BLAST (blastn; E=1e-5) searches were conducted for each pair of bait sequence references and their corresponding combined gene sequences (separately for *cox1* 5' and 3' regions, *CytB* and 16S) extracted from the mitogenome assemblies. Only hits with 100% pairwise identity and greater than100 bp overlap were considered to have resulted in a successful identification. Where multiple bait sequences from a single specimen were available, each bait was checked to have hit the same long assembly unequivocally to test for possible chimeras. If baits from a single specimen matched multiple, non-overlapping assemblies they presumably corresponded to the same incompletely assembled mitogenome. These assemblies were combined and retained if they included eight or more genes in total.

3.3.6 tRNA gene order

Once mitogenomic assemblies were identified, the tRNA gene order in the cluster of six tRNA genes located between *nad3* and *nad5* was visually recorded for all assemblies in order to test, with our greater taxon sampling, Haran et al.'s (2013) hypothesis that a ARNSEF to RANSEF tRNA gene rearrangement in this region is a synapomorphy for the Entiminae + Cyclominae + Hyperinae clade.

3.3.7 Sequence alignment and dataset concatenation

The sequences for the genes *nad*5, *nad4*, *nad4L* and *nad1*, which are transcribed on the reverse strand of the mitochondrial genome, were reverse complemented prior to alignment. Twenty-eight additional curculionoid mitogenomic sequences were obtained from Genbank (primarily the mitogenomes generated by Haran *et al*. 2013, listed in Appendix 3.2) for inclusion in our analyses in order to maximise taxon sampling. A further two mitogenomes, that of a leaf beetle (Chrysomelidae) and a longhorn beetle (Cerambycidae) were included as outgroups. These two families belong to the superfamily Chrysomeloidea which is considered to be the sister group to the Curculionoidea (e.g. Hunt *et al.* 2007). The combined sequences from each of the separated 13 protein-coding and 2 ribosomal RNA genes were individually aligned using the MAFFT version 7 online server (Katoh *et al.* 2002), incorporating the FFT-NS-I slow iterative refinement strategy with the following parameter values: nucleotide scoring matrix 200PAM/k=2, gap open penalty = 1.53, offset value = 0.0. Alignments were thereafter checked manually in Geneious for quality and to ensure that protein-coding genes were in the correct reading frame. Genes were concatenated together to make 6 different data matrices as follows: all genes (A), only protein-coding genes (B), all genes with 3rd codon positions removed from protein coding genes (C), protein-coding genes only with 3rd codon positions removed (D), all genes with 3rd codon positions removed from protein-coding genes and first codon positions R-Y coded (E) and only protein-coding genes with 3rd codon positions removed and first codon positions R-Y coded (F).

3.3.8 Phylogenetic analyses

Each of the 6 datasets were analysed under a maximum likelihood (ML) optimality criterion using RAxML 7.6.6 (Stamatakis 2006) running on the CIPRES web-based server (Miller *et al.* 2010) to search for the best-scoring tree. To assess nodal support, a rapid bootstrap analysis (BS) with 1000 iterations was run in parallel with tree-building. The datasets were each analysed both partitioned by gene and unpartitioned separately. Additionally, three of the datasets (A, B and E) were first tested using PartitionFinder (Lanfear *et al.* 2012) in order to objectively select the best-fitting partitioning scheme and model of molecular evolution for each alignment. This was performed using the Bayesian Information Criterion from an initial partitioning of each of the 3 codon positions for each amino acid-coding sequence being separate partitions. The resulting ML trees were made ultrametric using the *chronos* function of the APE package in R (Paradis *et al.* 2004), which uses penalised likelihood to fit a chronogram to a phylogenetic tree whose branch lengths are in number of substitutions per site (Paradis 2013). In order to obtain a measure of the suitability of the mitogenomic data to robustly support relationships across different nodal ages (and putative taxonomic ranks) the pattern of distribution of nodal support across trees was investigated by calculating the branch length of each node from the base of the tree using an R script (Appendix 3.1E) and plotting this against its respective RAxML BS support. A strict consensus tree built from the 15 ML trees was also constructed to visualise the distribution of consistent nodes across all our analyses. Additional RAxML analyses were performed on datasets A and B partitioned by gene and separate codon positions for each protein-coding gene (41 and 39 partitions respectively) and various RAxML analyses on these two datasets with different combinations of partitioning schemes and topological constraints, as summarised in Table 3.2, in order to calculate the Akaike information criterion (AIC) as a means for preferred model selection (Posada & Buckley 2004).

3.3.9 **In silico** *investigations of assembly algorithms*

In order to investigate concerns regarding the possibility of inter-specific co-assembly of mitogenome sequences, via the genome assembly algorithms, two *in silico* emulations were undertaken to test for the presence of chimeric assemblies. The control region and ribosomal sequences are potentially particularly problematic to assemble due to the presence of sequence repeats in the former (Salzberg & Yorke 2005), and highly conserved regions in the latter, possibly complicating accurate assembly. These problems may be exacerbated when assembling NGS reads obtained from pooled samples because of the possibility of inter-specific chimera assembly.

The simulations were achieved using a NGS shotgun sequence simulator to create sequence libraries of known parameters from a set of known mitogenomic input sequences. The simulated sequence libraries were assembled using part of the present assembly pipeline, and the resulting assemblies were locally aligned to the original input sequences using BLAST to identify any chimeric sequences. Two sets of mitogenomic sequences were used in these emulations. The first contained the 27 partial weevil mitogenomes obtained by Haran *et al.* (2013) (Appendix 3.3A), of similar diversity and divergences to the mitogenomes newly generated in this chapter, though missing the ribosomal genes and the control region. The second contained 17 complete and near-complete delphinid (Mammalia: Cetacea) mitogenomes (one per genus) mostly sequenced and assembled by Vilstrup *et al.* (2011) (Appendix 3.3B), complete with ribosomal genes and control region sequences. Grinder (Angly *et al.* 2012) was used to simulate the Illumina MiSeq reads, with the following parameters chosen to closely match those of the actual MiSeq run undertaken for this study: insert size of 800 bp (with standard deviation of 200 bp), 250 bp pair-ended reads in the correct Illumina forward and reverse directions, and 25 x coverage (as estimated from Figure 3.5A). As described earlier, the simulated reads were then assembled using IDBA-UD, and the resulting assemblies had tRNA annotations added using COVE. The original mitogenomes from GenBank were made into a sequence database in Geneious, which all of the newly assembled simulated mitogenomes were searched against, using BLAST, in order to obtain sequence pairwise identity and query coverage values as a measure of assembly accuracy. Only assemblies of lengths greater than or equal to 1000 bp were considered.

3.4 Results

3.4.1 Mitogenomic assembly

The FastQC results indicated that the Illumina reads were of good per base, and per sequence quality. Following adapter trimming, approximately 5% of the Illumina reads resembled mitochondrial sequences after BLAST filtering (from a total of 18,341,901 paired-end reads). This search does not produce a very accurate estimate of the proportion of mitochondrial reads as it is designed to be an overestimate, ensuring that as many putatively mitochondrial reads as possible are captured. The resulting matches will almost certainly also contain bacterial reads. However, the relative proportion of reads with hits against the mitochondrial database will provide a rough comparison between datasets. The Celera and IDBA-UD assemblies resulted in 338 and 336 assemblies of >1000 bp respectively, rising to 361 assemblies when combined using Minimus2. Of these, 105 were >10 kb in length and potentially represented (largely) complete mitogenomes. The cumulative distribution of the assemblies by sequence length is shown in Figure 3.2, whilst Figure 3.3 represents the frequency distribution of assembly lengths for each of the Celera, IDBA-UD and Minimus2 assemblies. The latter produced a shift towards longer contigs, especially for the critical contig length of >15kb that corresponds to the full-length of insect mitogenomes. All subsequent analyses were conducted on the Minimus2 assemblies. It was possible to newly assemble and identify a total of 92 complete or near-complete mitogenomes comprising at least eight genes, including 75 contigs (43% of all pooled samples) containing the full complement of 15 genes, a further 15 (8.7% of pooled samples) with ≥ 12 genes, and two assemblies containing eight and nine genes respectively (Appendix 3.2). Those falling short of a full gene complement were mainly lacking the rRNA genes, in particular *rrnS*, which was the least common gene, present in only 56 of the assemblies, whilst *nad6* and *cytB* were present in all 92

The *in silico* simulation assessment of assembly accuracy resulted in 28 assembled scaffolds of lengths 2424-11269 bp from Haran *et al*'s (2013) original 27 partial mitogenomes. Twenty-six of the 28 newly simulated assemblies matched an original mitogenome with 100% pairwise identity (E = 0 in all cases) and with 100% query coverage. The three remaining assemblies matched a mitogenome with 99.9% pairwise identity (Appendix 3.3A). The discrepancy in numbers of assemblies is due to the fact that one of the original mitogenomes (JN163961) resulted in two simulated scaffolds (of lengths 2424 and 3153 bp). The original sequence of JN163961 contains a string of 52 consecutive ambiguities (Ns), approximately in the middle of the mitogenome. After mapping the two simulated scaffolds to this mitogenome in Geneious, it is clear that the two newly assembled, non-overlapping scaffolds (26 and 27) match to the original sequence with 100% pairwise identity, respectively either side ('ahead' and 'behind') of the string of Ns. The 26 simulated long assemblies,

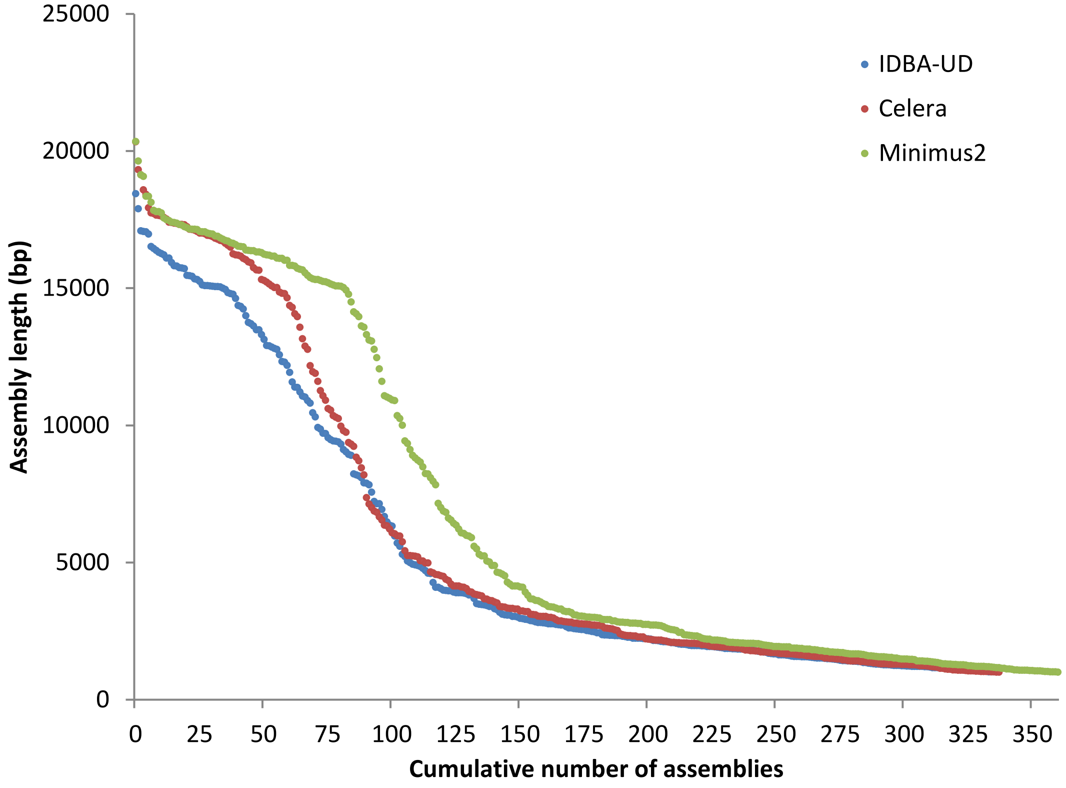


Figure 3.2 Cumulative distribution of assembly lengths from the Celera, IDBA-UD and the combined Minimus2-generated assemblies.

that each match to an original mitogenome, are consistely slightly shorter in length (by approximately 1.0%) than their respective originating mitogenome.

The simulated assembly using the 17 delphinid sequences of Vilstrup *et al.* (2011) resulted in six full mitogenomic assemblies of 16 kb or more, with 100% match to an originating mitogenome. A further seven mitogenomes were assembled in two non-overlapping sections, generally of one c. 14 kb section containing the protein genes and the control region, and a smaller section of <1.5 kb containing parts of *rrnS* and/or *rrnL*. The four remaining mitogenomes were assembled in up to 4 non-overlapping sections (Appendix 3.3B). The most problematic section for assembly appears to reside in the *rrnS*/*rrnL* ribosomal region and not in the control region, which appears to be correctly assembled in all cases.

3.4.2 Identification of mitogenomic assemblies using 'bait' sequences

From the set of 361 partial and complete contigs obtained with Minimus2, a total of 163 *cox1* (529-1560 bp), 154 *cytB* (218-1147 bp) and 162 *rrnL* (211-1340 bp) gene sequences were extracted. Sequences from each gene were grouped into libraries and used as queries in a BLAST search against each corresponding bait sequence reference library. The latter was composed of all successful PCR-based sequence from the 173 original DNA extractions and included 84 *cox1*-5', 115 *cox1*-3', 133 *cytB* and

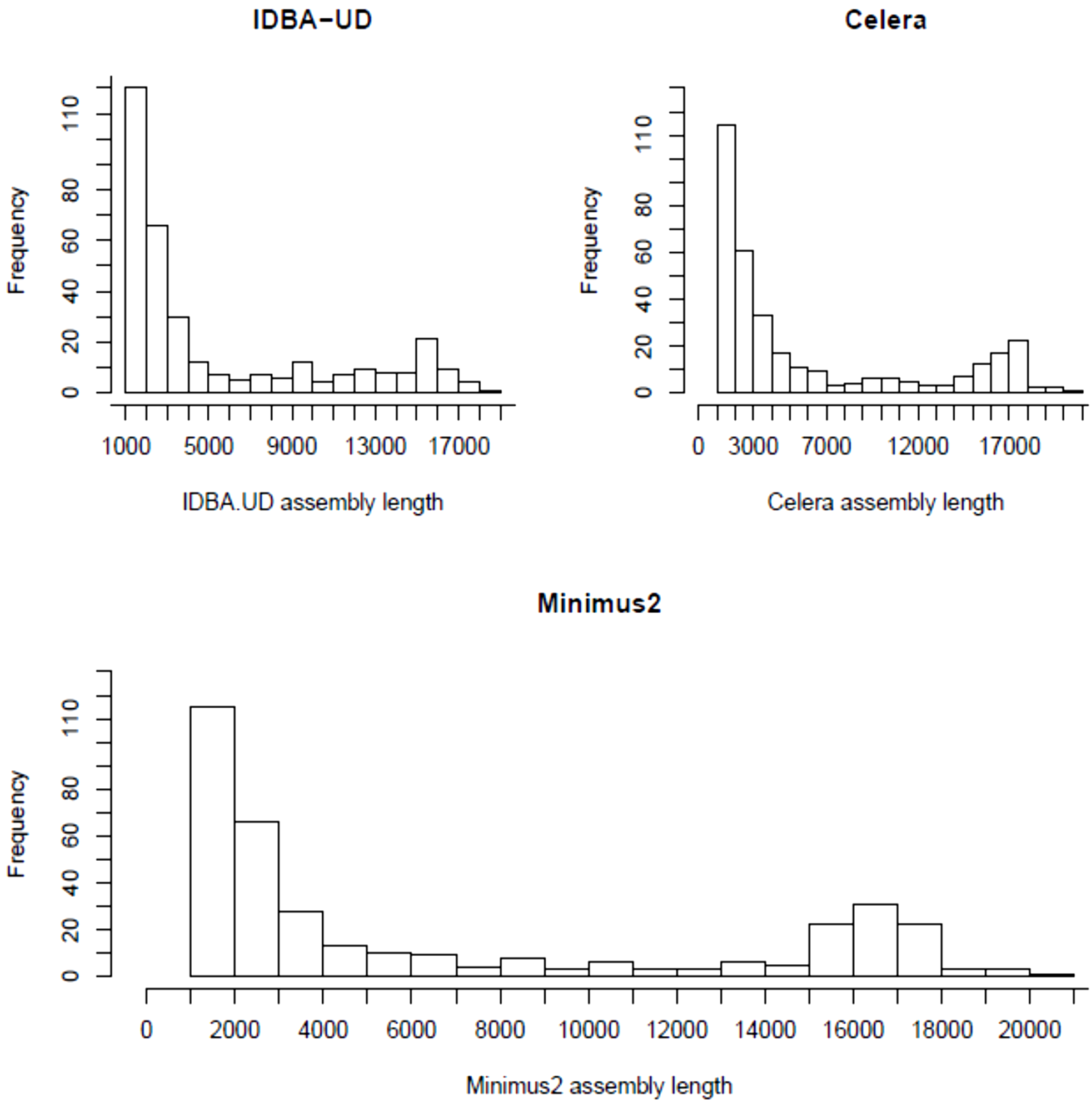


Figure 3.3 Frequency distribution of assembly lengths from the Celera, IDBA-UD and the combined Minimus2-generated assemblies.

Table 3.2 Maximum likelihood of trees under different partitioning schemes. Trees were obtained under no partitioning, under the 6-partition scheme selected by PartitionFinder, and by the maximum number of partitions tested (partitioning by gene and codon position). Each of the resulting trees was then assessed for their likelihood under the alternative models. Note the comparatively small difference in likelihood (ΔAIC) under each partitioning scheme regardless of the model used in the tree search.

Data set	Partitioning Scheme	Topological constraint	No. partitions	Substitution model	No. Parameters	LnL	AIC	ΔAIC
All genes (A)	Unpartitioned (1 partition)	None	1	GTR	8	-787773	1575562	62885
	PartitionFinder (6 partitions)	on 1 partition tree	6	GTR	48	-758061	1516219	3349
	Gene/codon-position (41 partitions)	on 1 partition tree	41	GTR	328	-756379	1513414	737
	Gene/codon-position (41 partitions)	on 6 partition tree	41	GTR	328	-756272	1513199	522
	PartitionFinder (6 partitions)	on 41 partition tree	6	GTR	48	-758010	1516116	3439
	Gene/codon-position (41 partitions)	None	41	GTR	328	-756010	1512677	n/a
	PartitionFinder (6 partitions)	on 1 partition tree	6	GTR	48	-758061	1516219	3542
Protein-coding genes (B)	Unpartitioned (1 partition)	None	1	GTR	8	-684161	1368339	34473
	Gene/codon-position (39 partitions)	on 1 partition tree	39	GTR	312	-666834	1334219	425
	PartitionFinder (5 partitions)	None	5	GTR	40	-668480	1337039	3173
	Gene/codon-position (39 partitions)	on 5 partition tree	39	GTR	312	-666678	1333981	115
	PartitionFinder (5 partitions)	on 39 partition tree	5	GTR	40	-668523	1337127	3261
	Gene/codon-position (39 partitions)	None	39	GTR	312	-666621	1333866	n/a
	PartitionFinder (5 partitions)	on 1 partition tree	5	GTR	40	-668567	1337213	3347

107 *rrnL* sequences. All samples used in the bulk sequencing were represented by at least one bait (38 samples), while 42, 57 and 36 samples were represented by two, three and four bait sequences, respectively. Matching these bait sequences to the 92 long mitogenomic assemblies, 16 assemblies showed a match to one bait, 30 assemblies matched two baits, 32 assemblies matched three baits and 14 assemblies matched all four baits. Four of the complete and near-complete mitogenomes contained sequences from two nonoverlapping assemblies that each matched at least one bait from the same specimen. Out of the remaining 81 weevil samples, there were 37 instances where baits hit a short contig that was not included in the collection of near-complete or complete mitogenome assemblies, but in 44 instances the baits did not hit any of the assembled contigs. Additionally one divergent assembly was rejected because it was found to match Coleoptera other than weevils in the reference database, possibly present in the sample due to a contamination. Appendix 3.4 summarises the bait-matching identification results, by bait, for each pooled sample, with matching contigs given by their unique number. Total number of baits available per sample, the total number of bait hits per sample and the reasons for identification failures are also listed. Overall the different baits contributed fairly equally to the final identifications, with 56% of all *cox1*-3' baits leading to a successful identification, 53% of *cytB*, 50% of *rrnL* and 45% of *cox1*-5'. Proportions of total number of baits, bait hits and hits leading to assembly identifications by gene are illustrated in Figure 3.4. A further 50 short contigs (1025 -6437 bp, mean 2472 bp) matched single baits but were not incorporated in the analyses because they contained only a maximum of four complete protein-coding or rRNA genes each. Their inclusison would have considerably increased the amount of missing data in the matrix.

The total number of reads making up each of the 92 mitogenomes (which were made up of 96 separate contigs) was used to calculate the sequencing depth (Figure

3.5). The majority of sequences showed a 10-50 x coverage that generally resulted in contigs of 15 – 20 kb. Coverage reached over 200x in a few cases but this does not appear to closely correlate with contig length. For example, two contigs of high coverage were <5kb in length and corresponded to two non-contiguous fragments from the same species (*Dryocoetes autographus*) linked by multiple baits obtained from a single specimen. In addition, read coverage was not closely correlated with the initial DNA concentration in the sequencing pool. Most samples were present at 10 ng, yet their coverage varied by more than an order of magnitude, while coverage for samples present at a concentration up to 4x lower varied over the same range (Figure 3.5).

3.4.3 Phylogenetic analyses

The 92 new assemblies were combined with existing data, for an aligned data matrix of 122 samples and 13792 positions. Of the final set of mitogenomes, 2 belonged to the family Anthribidae, 5 to Attelabidae, 3 to Brachyceridae, 4 to Brentidae, 4 to Dryophthoridae, 1 to Nemonychidae and 101 belonged to 67 identified tribes within the Curculionidae, including 19 tribes of the wood-boring Scolytinae. The optimal partitioning scheme was established using PartitionFinder, starting with a total of 39 partitions (41 partitions with the two RNA genes included) that split all 13 genes (15 in datasets A, C and E) and three codon positions in each protein-coding gene. PartitionFinder selected five partitions for the 'only protein-coding genes' dataset and six partitions for the 'all genes' dataset, whereby the two rRNA genes were grouped with the first codon positions of *nad2*, *nad*3 and *nad6* and the second codon position of *atp8* (Table 3.3). For both datasets the 1st and 3rd codon positions on forward and reverse strands were split into separate partitions, while all 2nd positions were collapsed into a single partition. Forward and reverse genes mainly differed in base frequencies, with a shift from A to T and G to C in the reverse strand partitions, and rates shifted accordingly (normalised to the time-reversible G-T changes: Figure 3.6). The dataset containing 'only protein-coding genes R-Y coded' resulted in only 2 partitions, separating 1st and 2nd codon position for both strands combined (3rd positions are removed from this dataset). The findings are in accordance with previous observations on Curculionoidea that also showed a great improvement in likelihood values when partitioning by both codon position and strand (Haran et al. 2013), reflecting the great differences in codon usage in genes coded on either strand. However, this does not extend to produce differences in variation in amino acid changes, as forward and reverse strands were consistently grouped into a single partition for the dataset using 2nd position only and for the R-Y coded matrix (eliminating 1st codon synonymous changes).

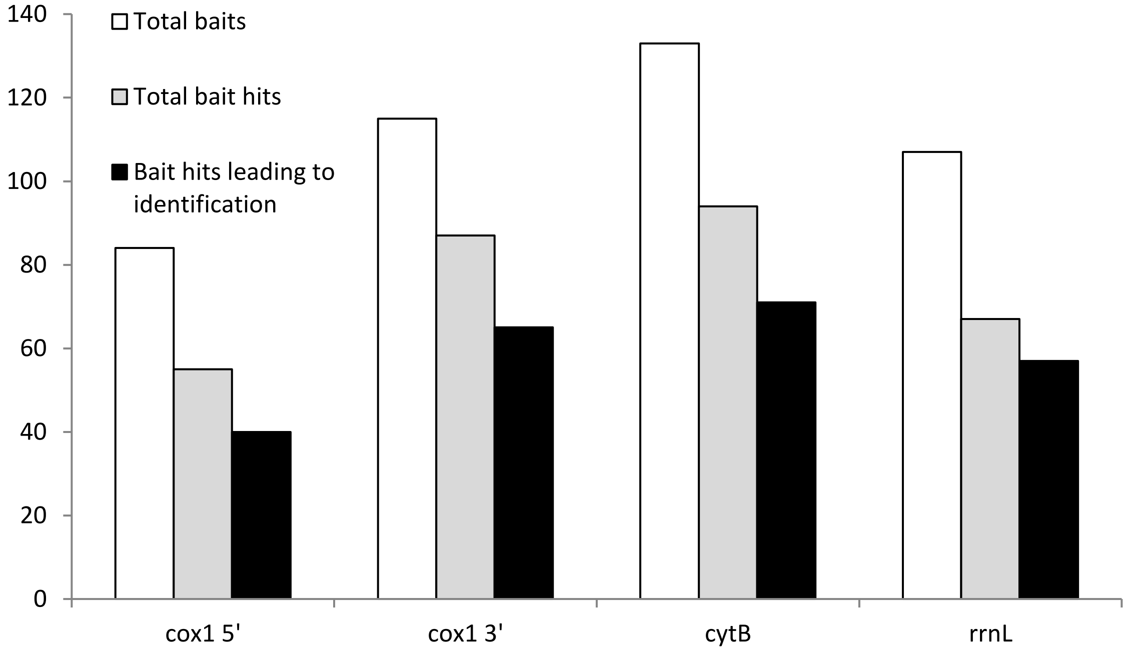


Figure 3.4 Relative proportions, by gene, of total 'bait' sequences available, 'bait' sequences with matching 'hits' to the assembled genes and matching hits that contributed to a successful mitogenome identification following a BLAST search.

The ML trees were greatly improved using six partitions over an unpartitioned analysis, but the benefit of using a model with 41 or 39 separate partitions was low, as seen from the small additional improvement in the AIC values (Table 3.2). Interestingly, the improvement in ML from using the partitioned models was very similar whether the trees were obtained directly under the partitioned model or obtained under the unpartitioned model but with the likelihood calculated under partitioning (Table 3.2). Hence, despite the greatly improved likelihood scores after partitioning, the resulting trees differ only slightly in parameters of greatest impact on the likelihood. This suggests that the topologies are little changed between the unpartitioned model, six-partition model (five-partition model without rRNA genes) and the 41 (39) partition model, given the small increase in likelihood if the simpler model is imposed on the tree obtained with the more complex model.

ML trees obtained with the various coding schemes (including or excluding rRNA genes; R-Y coding; presence of 3rd codon position: Table 3.4) also resulted in highly congruent topologies based upon strongly supported (>80% BS) nodes. Figure 3.7 depicts the best RAxML tree obtained with the 'all genes' dataset under six partitions. Indicated on this tree are nodes that are retained in the strict consensus of trees obtained from all different treatments of the data, and those nodes unresolved in the strict consensus, i.e. the nodes whose resolution is consistent with the strict consensus. Nodes with high nodal support (80-100% BS) occurred throughout the entire span of nodal ages and this pattern is found across all analyses (Figure 3.8).

3.4.4 Family-level relationships

All 15 analyses recovered the monophyletic 'ambrosia beetles', Platypodinae (100% BS) outside the other 'true weevils' (= Curculionidae *sensu* Bouchard *et al.*

2011), which would otherwise be monophyletic. In most analyses, except those including R-Y coded protein-coding genes, Platypodinae was placed in the sister clade to the rest

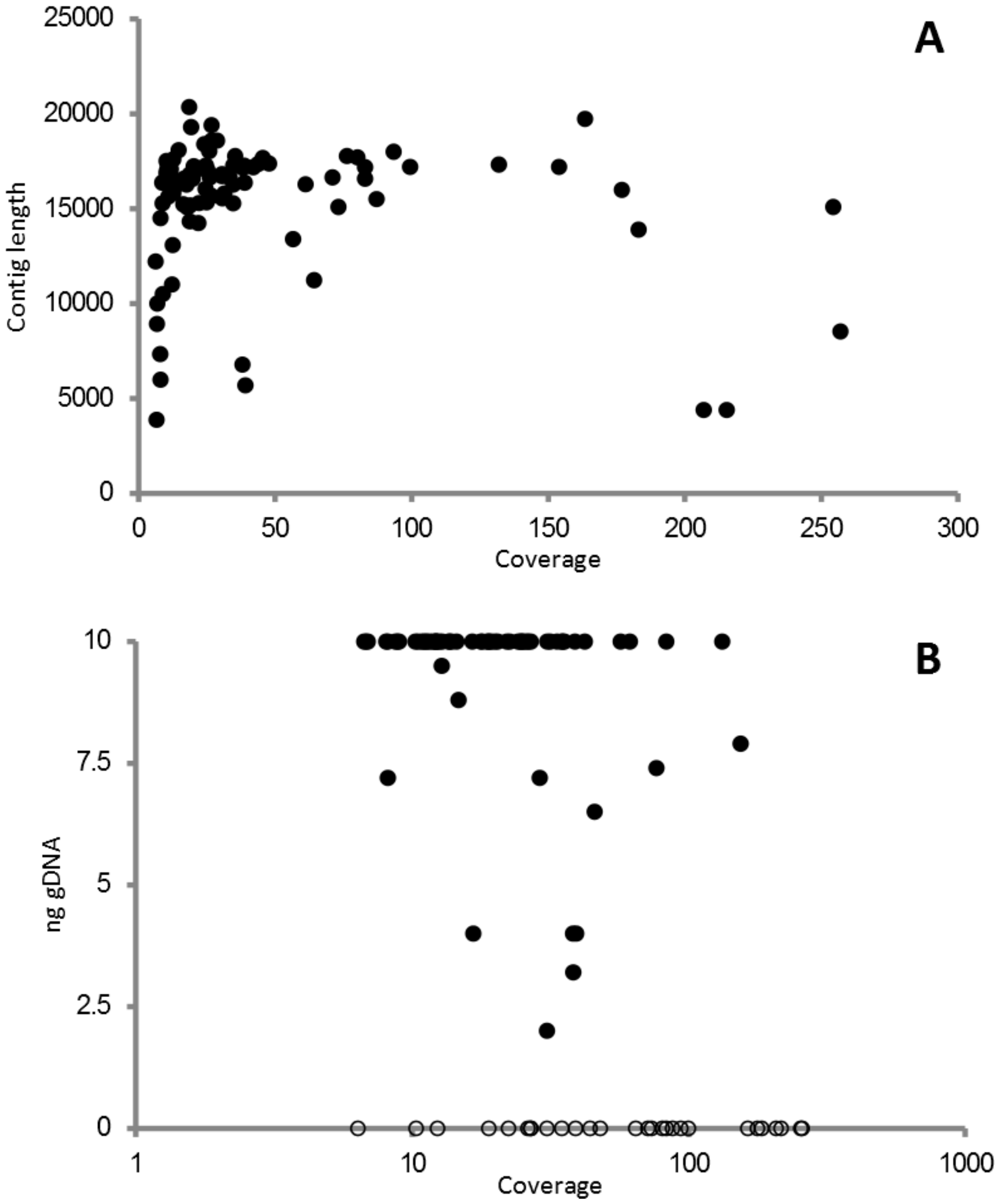


Figure 3.5 Mean sequencing coverage versus A) assembly (contig) length (bp), and B) approximate mass of genomic DNA in the sample pool (ng), for identified mitogenomic assemblies.

Table 3.3 Partitioning schemes and nucleotide substitution models selected by PartitionFinder for three datasets, according to gene and to codon position (numbered 1-3) in protein-coding genes. In yellow are the forward-strand genes, in red the reverse-strand genes and in blue the ribosomal RNA genes. Separate partitions are numbered P1 to P6 and allocated positions to each partition are coloured green. A) All genes; B) only protein-coding genes.

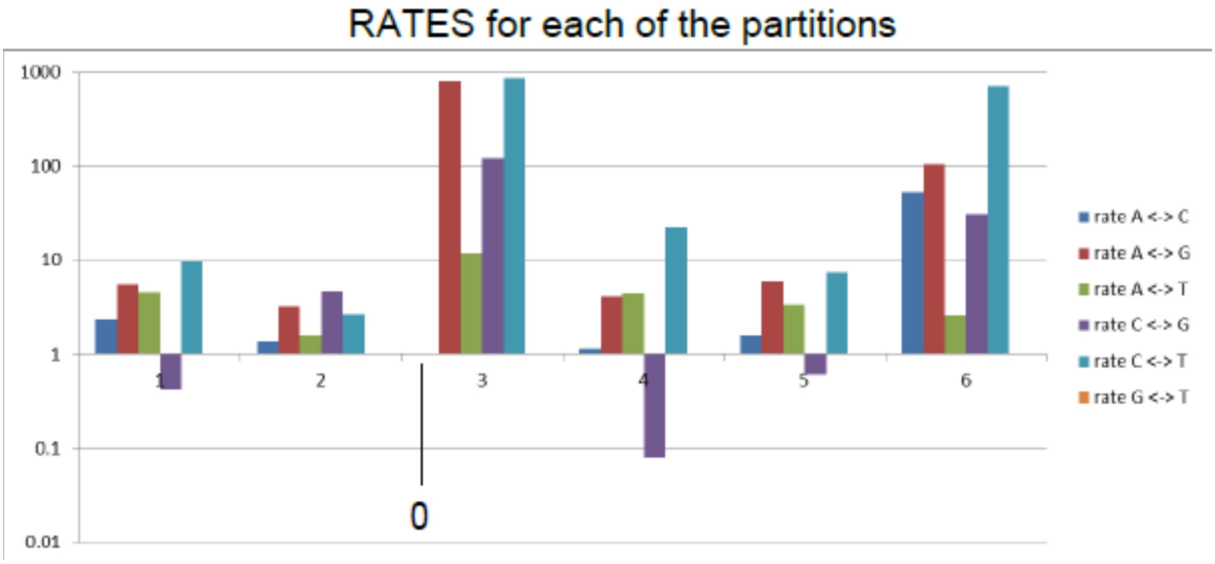
A)

Partition	*nad2*			*cox1*			*cox2*			*atp8*			*atp6*			*cox3*			*nad3*			*nad5*			*nad4*			*nad4L*			*nad6*			*cytB*			*nad1*			*rrnL*	*rrnS*
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3		
P1	🟩										🟩								🟩												🟩									🟩	🟩
P2		🟩			🟩			🟩						🟩			🟩			🟩			🟩			🟩			🟩			🟩			🟩			🟩			
P3			🟩			🟩			🟩			🟩			🟩			🟩			🟩												🟩			🟩					
P4				🟩			🟩			🟩						🟩																		🟩							
P5																						🟩			🟩			🟩									🟩				
P6																								🟩			🟩			🟩									🟩		

B)

Partition	*nad2*			*cox1*			*cox2*			*atp8*			*atp6*			*cox3*			*nad3*			*nad5*			*nad4*			*nad4L*			*nad6*			*cytB*			*nad1*		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
P1	🟩			🟩			🟩			🟩	🟩		🟩			🟩			🟩												🟩			🟩					
P2		🟩			🟩			🟩						🟩			🟩			🟩			🟩			🟩			🟩			🟩			🟩			🟩	
P3			🟩			🟩			🟩			🟩			🟩			🟩			🟩												🟩			🟩			
P4																						🟩			🟩			🟩									🟩		
P5																								🟩			🟩			🟩									🟩

A)



B)

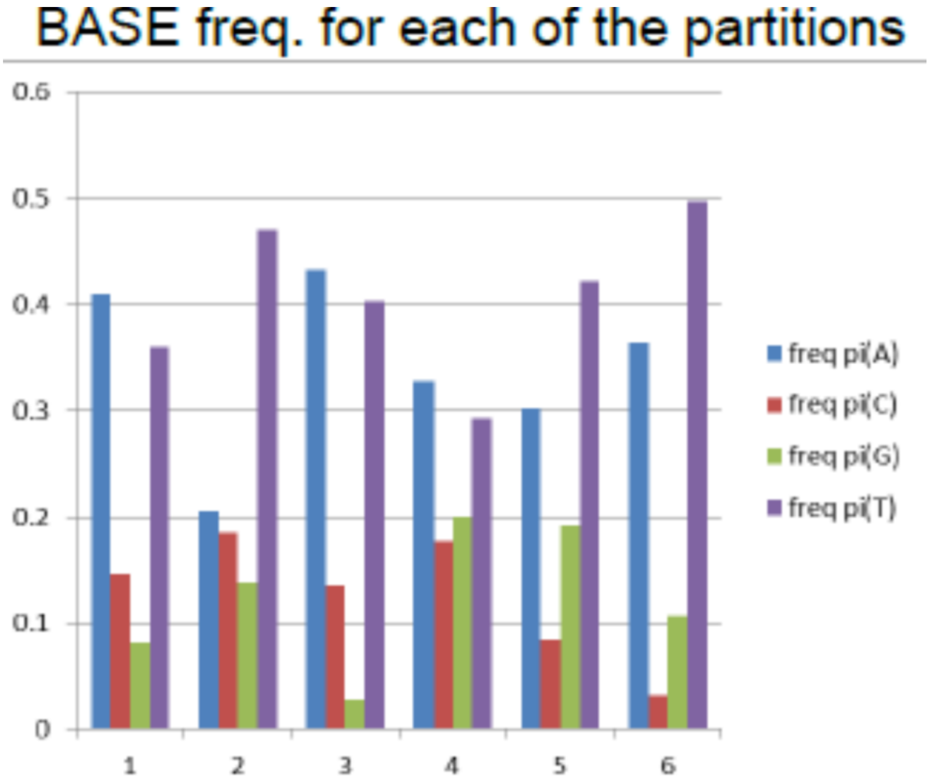


Figure 3.6 Observed nucleotide substitution rates (A) and base frequencies (B) of the six PartitionFinder-selected partitions for the 'all genes dataset'. See Table 3.3 for partition definitions.

of Curculionidae, together with the Dryophthoridae (palm weevils) and the brachycerid genus *Ocladius*, with moderate to strong support for this adelphic relationship (62-95% BS). In all analyses the monophyletic Brentidae (100% BS) were recovered as the sister taxon to a Curculionidae + Dryophthoridae + Brachyceridae clade with very strong nodal support (100% BS). The sister relationship between the monophyletic (100% BS) Attelabidae (leaf-rolling weevils) and this latter clade plus Brentidae was similarly very strongly supported (100% BS) across all analyses. The Nemonychidae was consistently recovered as sister to the clade containing Attelabidae and all other weevil families mentioned so far. Support for this relationship was very high, ranging from 98-100% BS across analyses. The two taxa belonging to the Anthribidae were always recovered as monophyletic (100% BS). Within the Attelabidae, the subfamilies Apoderinae and Rhynchitinae were recovered as monophyletic with BS support of 100% and 83-97% respectively across analyses.

3.4.5 Relationships within Curculionidae **s.str*.***

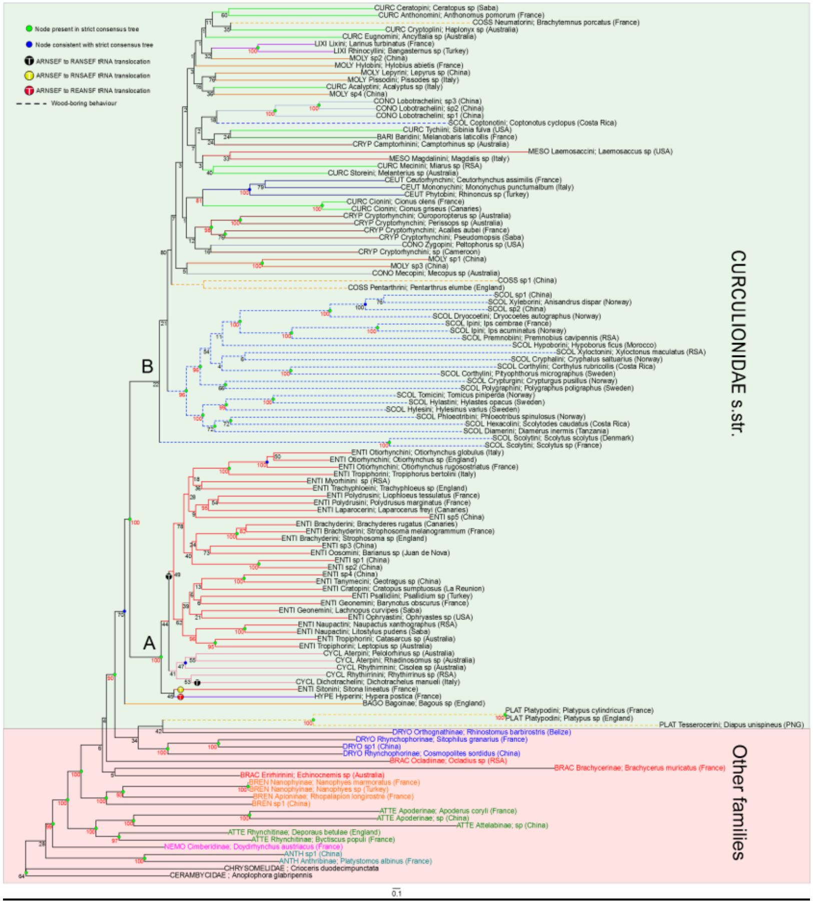
In most analyses the subfamily Bagoinae, represented only by a single *Bagous*, was recovered as the sister to all other Curculionidae (excepting Platypodinae as noted above), with BS support between 66 and 91%. Similarly, most analyses resulted in the recovery of both a monophyletic Entiminae + Cyclominae + Hyperinae clade (marked A in Figure 3.7; 100% BS) and a strongly supported sister relationship between this clade and a second clade (marked B in Figure 3.7) containing all other Curculionidae subfamilies (100% BS). Within the entimine clade, the Entiminae itself is not recovered as monophyletic because the tribe Sitonini is consistently recovered (100% BS) either as sister to the clade containing Hyperinae + Cyclominae + the rest of Entiminae, or in a sister clade also containing the Hyperinae (with generally weak nodal support for this relationship). Three entimine tribes are consistently recovered as monophyletic, with strong nodal support; the Otiorhynchini (100% BS), Brachyderini (100% BS) and the Naupactini (100% BS). The tribe Tropiphorini is apparently paraphyletic because a well-supported clade (95% BS), containing two monophyletic Australian members (*Catasarcus* and *Leptopius*), is itself sister to the Naupactini with strong support (96% BS) and is only distantly related to the other Tropiphorini species in the dataset (*Tropiphorus*), which is sister to the Otiorhynchini with strong nodal support (100% BS). All Entiminae (except *Sitona*) are marked by an ARNSEF to RANSEF rearrangement in the tRNA cluster, discovered

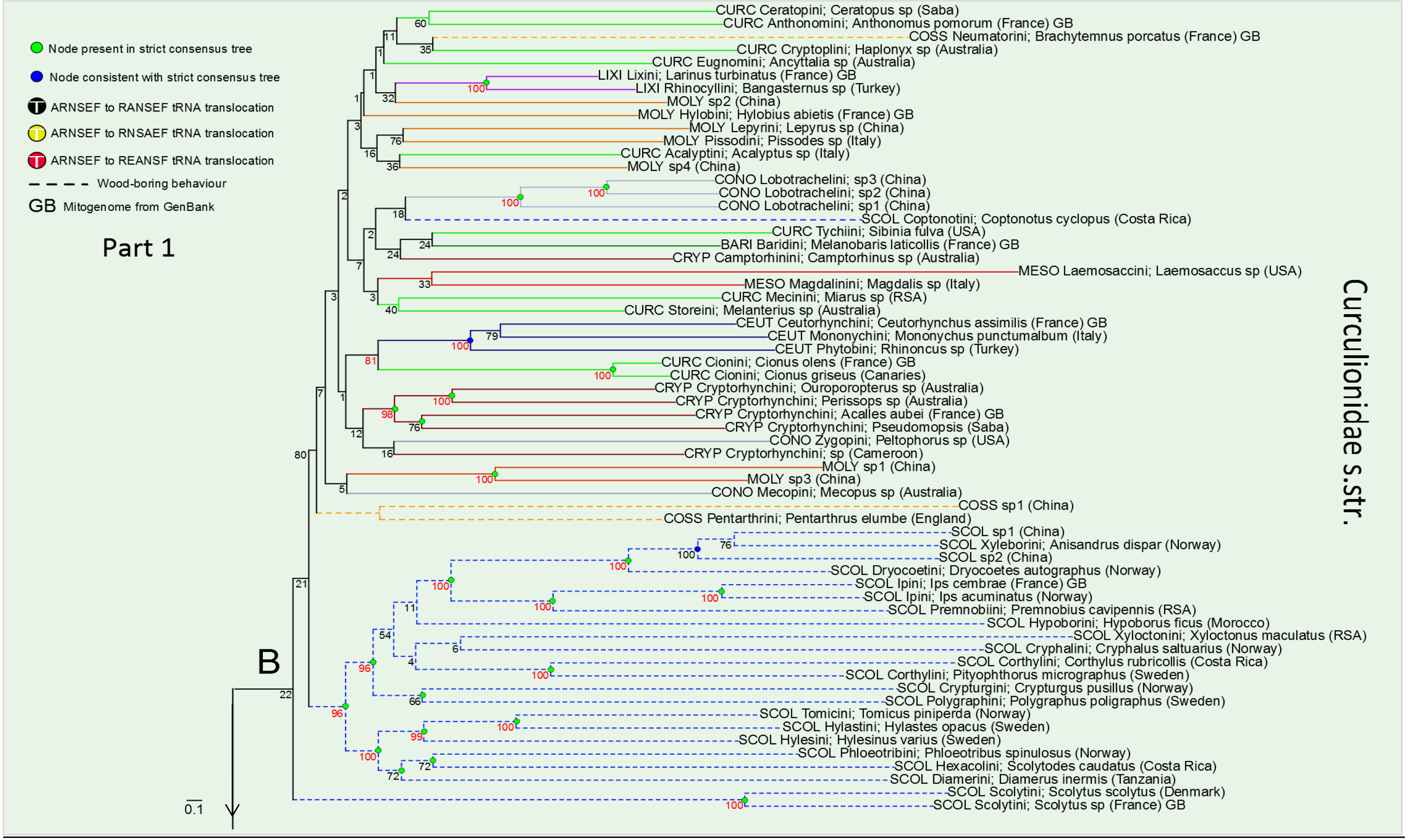
Table 3.4 Final RAxML maximum likelihood optimisation scores for the analyses of each of the 15 datasets. Analyses of datasets containing all genes are shown in grey.

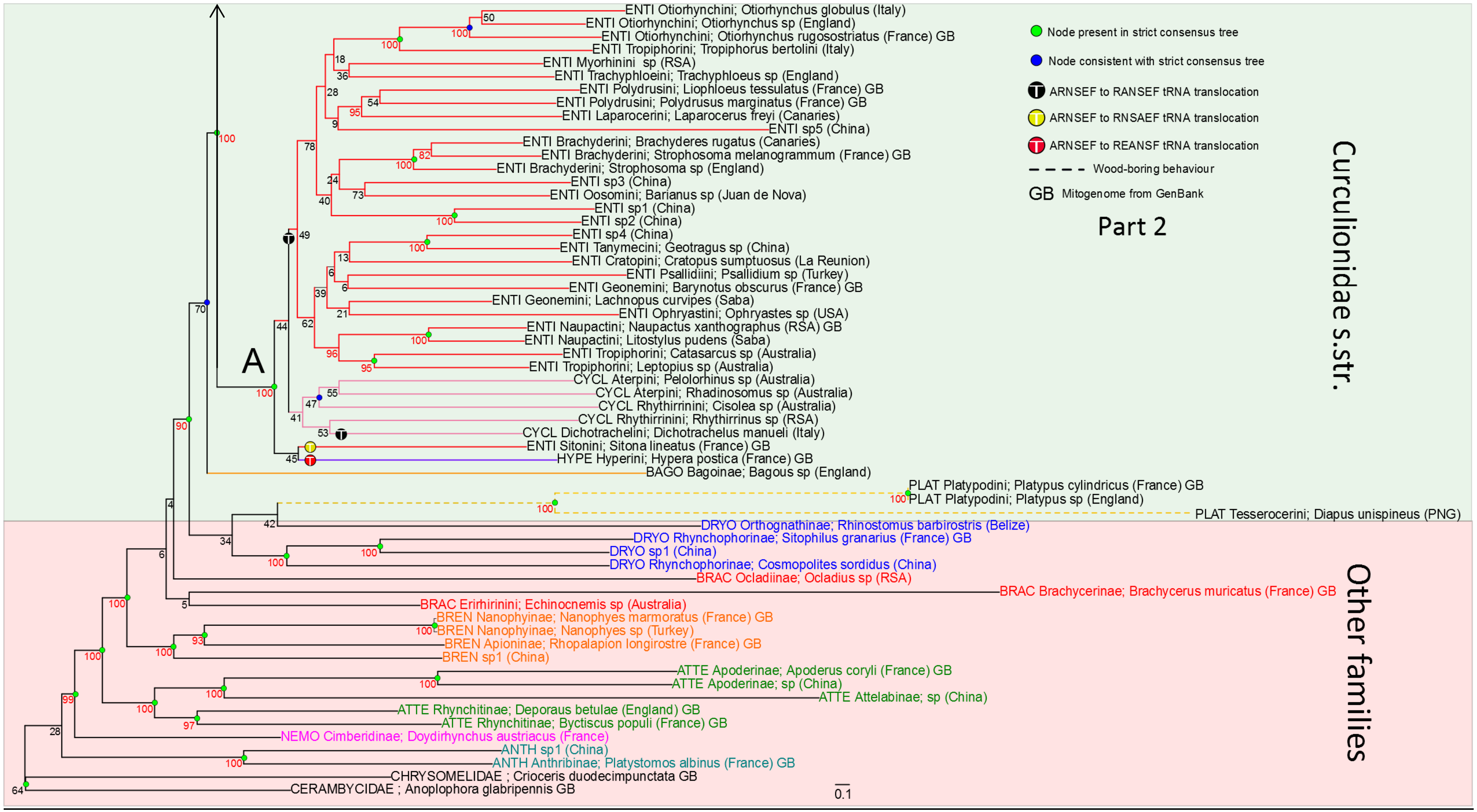
Dataset	Final ML Optimisation Likelihood
All genes partitioned by gene (A1)	-773731.4614
All genes unpartitioned (A2)	-787772.9784
All genes, PartitionFinder (A3)	-757964.9526
Only protein genes partitioned by gene (B1)	-696122.2766
Only protein genes unpartitioned (B2)	-684161.4211
Only protein genes Partition Finder (B3)	-668479.6459
All genes, protein genes without 3rd codon position, partitioned by gene (C1)	-420952.3613
All genes, protein genes without 3rd codon position, unpartitioned (C2)	-414851.3568
Only proteins genes without 3rd codon position, partitioned by gene (D1)	-328068.0482
Only proteins genes without 3rd codon position, unpartitioned (D2)	-331245.2996
All genes, proteins genes RY coded, partitioned by gene (E1)	-305075.4193
All genes, proteins genes RY coded, unpartitioned (E2)	-310588.2857
Only proteins genes RY coded, partitioned by gene (F1)	-218258.1401
Only proteins genes RY coded, unpartitioned (F2)	-219811.0759
Only protein genes RY coded PartitionFinder (E3)	-218339.2117

in earlier studies (Haran *et al.* 2013; Song *et al.* 2010) and corroborated here (Figure 3.7). One taxon, *Dichotrachelus manueli*, classified in Cyclominae by Alonso-Zarazaga and Lyal (1999), also possesses this same rearrangement, whilst the remaining Cyclominae taxa possess the common gene order, ARNSEF. *Sitona* and *Hypera* were characterised by unique RNSAEF and REANSF gene orders, respectively, observed initially by Haran *et al.* (2013) and hypothesised to constitute an initial step in the evolution of the derived gene order of the Entiminae. Here, *Hypera* + *Sitona* form a clade that is sister to all others in clade A, while the Cyclominae (minus *Dichotrachelus*), not represented in Haran et al. (2013), and exhibiting the ancestral gene order, occupy the next node as sister to the remaining Entiminae characterised by the derived gene order. This demonstrates that the gene order changes in *Hypera* and *Sitona* are independent of those in Entiminae. Within the second main curculionid clade, the scolytine taxon *Coptonotus* (Coptonotini) is never recovered together with the bulk of the scolytines, which except for Scolytini (monophyletic with 100% BS), are consistently recovered in a clade with moderate to high support values of 66-100%. The scolytine tribes Corthylini and Ipini are always recovered as monophyletic (100% BS support) within this. The following higher-level taxa from the second main Curculionidae clade are recovered as monophyletic across all analyses (BS supports follow taxon name): Ceutorhynchinae (100%), Lixinae (100%), Conoderinae Lobotrachelini (100%) and Curculioninae Cionini (100%). The Cryptorhynchini appears to be paraphyletic owing to the presence of a sample (Cryptorhynchini sp. from Cameroon) falling outside the well supported clade (98% BS) comprising all four other genera analysed.

Figure 3.7 (on following two pages) Maximum likelihood tree resulting from the analysis of the 'all genes' dataset partitioned according to PartitionFinder (see Table 3.3). Within Curculionidae s.str. (sensu Bouchard *et al.* 2011) branches are coloured according to subfamily. Other curculionoid families have their name labels coloured by family. Numbers adjacent to nodes are RAxML rapid bootstrap scores, with values >80% highlighted in red. The three principal wood-boring subfamilies are represented by dashed branches and the nodes labelled A and B indicate the two large divisions within Curculionidae referred to in the text. Nodes indicated in green correspond to nodes present in the strict consensus tree and nodes indicated in blue are consistent with it. The positions of the three tRNA rearrangements are indicated. Scale bar represents substitution rate. Family and subfamily codes precede taxa names as follows: Anthribidae (ANTH), Attelabidae (ATTE), Brachyceridae (BRAC), Brentidae (BREN), Dryophthoridae (DRYO), Nemonychidae (NEMO), Bagoinae (BAGO), Baridinae (BARI), Ceutorhynchinae (CEUT), Conoderinae (CONO), Cossoninae (COSS), Cryptorhynchinae (CRYP), Curculioninae (CURC), Lixinae (LIXI), Mesoptillinae (MESO), Molytinae (MOLY), Platypodinae (PLAT) and Scolytinae (SCOL).


Node present in strict consensus tree
Node consistent with strict consensus tree
ARNSEF to RANSEF tRNA translocation
ARNSEF to RNSAEF tRNA translocation
ARNSEF to REANSF tRNA translocation
Wood-boring behaviour
CURC Ceratopini; Ceratopus sp (Saba)
CURC Anthonomini; Anthonomus pomorum (France)
COSS Neumatorini; Brachytemnus porcatus (France)
CURC Cryptoplini; Haplonyx sp (Australia)
CURC Eugnomini; Ancyttalia sp (Australia)
LIXI Lixini; Larinus turbinatus (France)
LIXI Rhinocyllini; Bangasternus sp (Turkey)
MOLY sp2 (China)
MOLY Hylobini; Hylobius abietis (France)
MOLY Lepyrini; Lepyrus sp (China)
MOLY Pissodini; Pissodes sp (Italy)
CURC Acalyptini; Acalyptus sp (Italy)
MOLY sp4 (China)
CONO Lobotrachelini; sp3 (China)
CONO Lobotrachelini; sp2 (China)
CONO Lobotrachelini; sp1 (China)
SCOL Coptonotini; Coptonotus cyclopus (Costa Rica)
CURC Tychiini; Sibinia fulva (USA)
BARI Baridini; Melanobaris laticollis (France)
CRYP Camptorhinini; Camptorhinus sp (Australia)
MESO Laemosaccini; Laemosaccus sp (USA)
MESO Magdalini; Magdalis sp (Italy)
CURC Mecinini; Miarus sp (RSA)
CURC Storeini; Melanterius sp (Australia)
CEUT Ceutorhynchini; Ceutorhynchus assimilis (France)
CEUT Mononychini; Mononychus punctumalbum (Italy)
CEUT Phytobini; Rhinoncus sp (Turkey)
CURC Cionini; Cionus olens (France)
CURC Cionini; Cionus griseus (Canaries)
CRYP Cryptorhynchini; Ouroporopterus sp (Australia)
CRYP Cryptorhynchini; Perissops sp (Australia)
CRYP Cryptorhynchini; Acalles aubei (France)
CRYP Cryptorhynchini; Pseudomopsis (Saba)
CONO Zygopini; Peltophorus sp (USA)
CRYP Cryptorhynchini; sp (Cameroon)
MOLY sp1 (China)
MOLY sp3 (China)
CONO Mecopini; Mecopus sp (Australia)
COSS sp1 (China)
COSS Pentarthrini; Pentarthrus elumbe (England)
SCOL sp1 (China)
SCOL Xyleborini; Anisandrus dispar (Norway)
SCOL sp2 (China)
SCOL Dryocoetini; Dryocoetes autographus (Norway)
SCOL Ipini; Ips cembrae (France)
SCOL Ipini; Ips acuminatus (Norway)
SCOL Premnobiini; Premnobius cavipennis (RSA)
SCOL Hypoborini; Hypoborus ficus (Morocco)
SCOL Xyloctonini; Xyloctonus maculatus (RSA)
SCOL Cryphalini; Cryphalus saltuarius (Norway)
SCOL Corthylini; Corthylus rubricollis (Costa Rica)
SCOL Corthylini; Pityophthorus micrographus (Sweden)
SCOL Crypturgini; Crypturgus pusillus (Norway)
SCOL Polygraphini; Polygraphus poligraphus (Sweden)
SCOL Tomicini; Tomicus piniperda (Norway)
SCOL Hylastini; Hylastes opacus (Sweden)
SCOL Hylesini; Hylesinus varius (Sweden)
SCOL Phloeotribini; Phloeotribus spinulosus (Norway)
SCOL Hexacolini; Scolytodes caudatus (Costa Rica)
SCOL Diamerini; Diamerus inermis (Tanzania)
SCOL Scolytini; Scolytus scolytus (Denmark)
SCOL Scolytini; Scolytus sp (France)
ENTI Otiorhynchini; Otiorhynchus globulus (Italy)
ENTI Otiorhynchini; Otiorhynchus sp (England)
ENTI Otiorhynchini; Otiorhynchus rugosostriatus (France)
ENTI Tropiphorini; Tropiphorus bertolini (Italy)
ENTI Myorhinini sp (RSA)
ENTI Trachyphloeini; Trachyphloeus sp (England)
ENTI Polydrusini; Liophloeus tessulatus (France)
ENTI Polydrusini; Polydrusus marginatus (France)
ENTI Laparocerini; Laparocerus freyi (Canaries)
ENTI sp5 (China)
ENTI Brachyderini; Brachyderes rugatus (Canaries)
ENTI Brachyderini; Strophosoma melanogrammum (France)
ENTI Brachyderini; Strophosoma sp (England)
ENTI sp3 (China)
ENTI Oosomini; Barianus sp (Juan de Nova)
ENTI sp1 (China)
ENTI sp2 (China)
ENTI sp4 (China)
ENTI Tanymecini; Geotragus sp (China)
ENTI Cratopini; Cratopus sumptuosus (La Reunion)
ENTI Psallidiini; Psallidium sp (Turkey)
ENTI Geonemini; Barynotus obscurus (France)
ENTI Geonemini; Lachnopus curvipes (Saba)
ENTI Ophryastini; Ophryastes sp (USA)
ENTI Naupactini; Naupactus xanthographus (RSA)
ENTI Naupactini; Litostylus pudens (Saba)
ENTI Tropiphorini; Catasarcus sp (Australia)
ENTI Tropiphorini; Leptopius sp (Australia)
CYCL Aterpini; Pelolorhinus sp (Australia)
CYCL Aterpini; Rhadinosomus sp (Australia)
CYCL Rhythirrinini; Cisolea sp (Australia)
CYCL Rhythirrinini; Rhythirrinus sp (RSA)
CYCL Dichotrachelini; Dichotrachelus manueli (Italy)
ENTI Sitonini; Sitona lineatus (France)
HYPE Hyperini; Hypera postica (France)
BAGO Bagoinae; Bagous sp (England)
PLAT Platypodini; Platypus cylindricus (France)
PLAT Platypodini; Platypus sp (England)
PLAT Tesserocerini; Diapus unispineus (PNG)
DRYO Orthognathinae; Rhinostomus barbirostris (Belize)
DRYO Rhynchophorinae; Sitophilus granarius (France)
DRYO sp1 (China)
DRYO Rhynchophorinae; Cosmopolites sordidus (China)
BRAC Ocladiinae; Ocladius sp (RSA)
BRAC Brachycerinae; Brachycerus muricatus (France)
BRAC Erirhinini; Echinocnemis sp (Australia)
BREN Nanophyinae; Nanophyes marmoratus (France)
BREN Nanophyinae; Nanophyes sp (Turkey)
BREN Apioninae; Rhopalapion longirostre (France)
BREN sp1 (China)
ATTE Apoderinae; Apoderus coryli (France)
ATTE Apoderinae; sp (China)
ATTE Attelabinae; sp (China)
ATTE Rhynchitinae; Deporaus betulae (England)
ATTE Rhynchitinae; Byctiscus populi (France)
NEMO Cimberidinae; Doydirhynchus austriacus (France)
ANTH sp1 (China)
ANTH Anthribinae; Platystomos albinus (France)
CHRYSOMELIDAE ; Crioceris duodecimpunctata
CERAMBYCIDAE ; Anoplophora glabripennis
A
B
CURCULIONIDAE s.str.
Other families
0.1


Node present in strict consensus tree
Node consistent with strict consensus tree
ARNSEF to RANSEF tRNA translocation
ARNSEF to RNSAEF tRNA translocation
ARNSEF to REANSF tRNA translocation
Wood-boring behaviour
GB Mitogenome from GenBank
Part 1
B
0.1
Curculionidae s.str.
CURC Ceratopini; Ceratopus sp (Saba)
CURC Anthonomini; Anthonomus pomorum (France) GB
COSS Neumatorini; Brachytemnus porcatus (France) GB
CURC Cryptoplini; Haplonyx sp (Australia)
CURC Eugnomini; Ancyttalia sp (Australia)
LIXI Lixini; Larinus turbinatus (France) GB
LIXI Rhinocyllini; Bangasternus sp (Turkey)
MOLY sp2 (China)
MOLY Hylobini; Hylobius abietis (France) GB
MOLY Lepyrini; Lepyrus sp (China)
MOLY Pissodini; Pissodes sp (Italy)
CURC Acalyptini; Acalyptus sp (Italy)
MOLY sp4 (China)
CONO Lobotrachelini; sp3 (China)
CONO Lobotrachelini; sp2 (China)
CONO Lobotrachelini; sp1 (China)
SCOL Coptonotini; Coptonotus cyclopus (Costa Rica)
CURC Tychiini; Sibinia fulva (USA)
BARI Baridini; Melanobaris laticollis (France) GB
CRYP Camptorhinini; Camptorhinus sp (Australia)
MESO Laemosaccini; Laemosaccus sp (USA)
MESO Magdalinini; Magdalis sp (Italy)
CURC Mecinini; Miarus sp (RSA)
CURC Storeini; Melanterius sp (Australia)
CEUT Ceutorhynchini; Ceutorhynchus assimilis (France) GB
CEUT Mononychini; Mononychus punctumalbum (Italy)
CEUT Phytobini; Rhinoncus sp (Turkey)
CURC Cionini; Cionus olens (France) GB
CURC Cionini; Cionus griseus (Canaries)
CRYP Cryptorhynchini; Ouroporopterus sp (Australia)
CRYP Cryptorhynchini; Perissops sp (Australia)
CRYP Cryptorhynchini; Acalles aubei (France) GB
CRYP Cryptorhynchini; Pseudomopsis (Saba)
CONO Zygopini; Peltophorus sp (USA)
CRYP Cryptorhynchini; sp (Cameroon)
MOLY sp1 (China)
MOLY sp3 (China)
CONO Mecopini; Mecopus sp (Australia)
COSS sp1 (China)
COSS Pentarthrini; Pentarthrus elumbe (England)
SCOL sp1 (China)
SCOL Xyleborini; Anisandrus dispar (Norway)
SCOL sp2 (China)
SCOL Dryocoetini; Dryocoetes autographus (Norway)
SCOL Ipini; Ips cembrae (France) GB
SCOL Ipini; Ips acuminatus (Norway)
SCOL Premnobiini; Premnobius cavipennis (RSA)
SCOL Hypoborini; Hypoborus ficus (Morocco)
SCOL Xyloctonini; Xyloctonus maculatus (RSA)
SCOL Cryphalini; Cryphalus saltuarius (Norway)
SCOL Corthylini; Corthylus rubricollis (Costa Rica)
SCOL Corthylini; Pityophthorus micrographus (Sweden)
SCOL Crypturgini; Crypturgus pusillus (Norway)
SCOL Polygraphini; Polygraphus poligraphus (Sweden)
SCOL Tomicini; Tomicus piniperda (Norway)
SCOL Hylastini; Hylastes opacus (Sweden)
SCOL Hylesini; Hylesinus varius (Sweden)
SCOL Phloeotribini; Phloeotribus spinulosus (Norway)
SCOL Hexacolini; Scolytodes caudatus (Costa Rica)
SCOL Diamerini; Diamerus inermis (Tanzania)
SCOL Scolytini; Scolytus scolytus (Denmark)
SCOL Scolytini; Scolytus sp (France) GB


ENTI Otiorhynchini; Otiorhynchus globulus (Italy)
ENTI Otiorhynchini; Otiorhynchus sp (England)
ENTI Otiorhynchini; Otiorhynchus rugosostriatus (France) GB
ENTI Tropiphorini; Tropiphorus bertolini (Italy)
ENTI Myorhinini sp (RSA)
ENTI Trachyphloeini; Trachyphloeus sp (England)
ENTI Polydrusini; Liophloeus tessulatus (France) GB
ENTI Polydrusini; Polydrusus marginatus (France) GB
ENTI Laparocerini; Laparocerus freyi (Canaries)
ENTI sp5 (China)
ENTI Brachyderini; Brachyderes rugatus (Canaries)
ENTI Brachyderini; Strophosoma melanogrammum (France) GB
ENTI Brachyderini; Strophosoma sp (England)
ENTI sp3 (China)
ENTI Oosomini; Barianus sp (Juan de Nova)
ENTI sp1 (China)
ENTI sp2 (China)
ENTI sp4 (China)
ENTI Tanymecini; Geotragus sp (China)
ENTI Cratopini; Cratopus sumptuosus (La Reunion)
ENTI Psallidiini; Psallidium sp (Turkey)
ENTI Geonemini; Barynotus obscurus (France) GB
ENTI Geonemini; Lachnopus curvipes (Saba)
ENTI Ophryastini; Ophryastes sp (USA)
ENTI Naupactini; Naupactus xanthographus (RSA) GB
ENTI Naupactini; Litostylus pudens (Saba)
ENTI Tropiphorini; Catasarcus sp (Australia)
ENTI Tropiphorini; Leptopus sp (Australia)
CYCL Aterpini; Pelolorhinus sp (Australia)
CYCL Aterpini; Rhadinosomus sp (Australia)
CYCL Rhythirrinini; Cisolea sp (Australia)
CYCL Rhythirrinini; Rhythirrinus sp (RSA)
CYCL Dichotrachelini; Dichotrachelus manueli (Italy)
ENTI Sitonini; Sitona lineatus (France) GB
HYPE Hyperini; Hypera postica (France) GB
BAGO Bagoinae; Bagous sp (England)
PLAT Platypodini; Platypus cylindricus (France) GB
PLAT Platypodini; Platypus sp (England)
PLAT Tesserocerini; Diapus unispineus (PNG)
DRYO Orthognathinae; Rhinostomus barbirostris (Belize)
DRYO Rhynchophorinae; Sitophilus granarius (France) GB
DRYO sp1 (China)
DRYO Rhynchophorinae; Cosmopolites sordidus (China)
BRAC Ocladiinae; Ocladius sp (RSA)
BRAC Brachycerinae; Brachycerus muricatus (France) GB
BRAC Erirhinini; Echinocnemis sp (Australia)
BREN Nanophyinae; Nanophyes marmoratus (France) GB
BREN Nanophyinae; Nanophyes sp (Turkey)
BREN Apioninae; Rhopalapion longirostre (France) GB
BREN sp1 (China)
ATTE Apoderinae; Apoderus coryli (France) GB
ATTE Apoderinae; sp (China)
ATTE Attelabinae; sp (China)
ATTE Rhynchitinae; Deporaus betulae (England) GB
ATTE Rhynchitinae; Byctiscus populi (France) GB
NEMO Cimberidinae; Doydirhynchus austriacus (France)
ANTH sp1 (China)
ANTH Anthribinae; Platystomos albinus (France) GB
CHRYSOMELIDAE ; Crioceris duodecimpunctata GB
CERAMBYCIDAE ; Anoplophora glabripennis GB
50
100
100
18
36
28
54
95
9
100
78
82
100
24
73
40
100
49
100
13
6
6
39
21
62
44
100
96
95
100
55
47
41
53
45
100
70
90
4
100
42
34
100
100
6
5
100
100
93
100
100
100
100
97
99
28
100
64
A
🟢 Node present in strict consensus tree
🔵 Node consistent with strict consensus tree
T ARNSEF to RANSEF tRNA translocation
T ARNSEF to RNSAEF tRNA translocation
T ARNSEF to REANSF tRNA translocation
- - - - Wood-boring behaviour
GB Mitogenome from GenBank
Part 2
Curculionidae s.str.
Other families
0.1

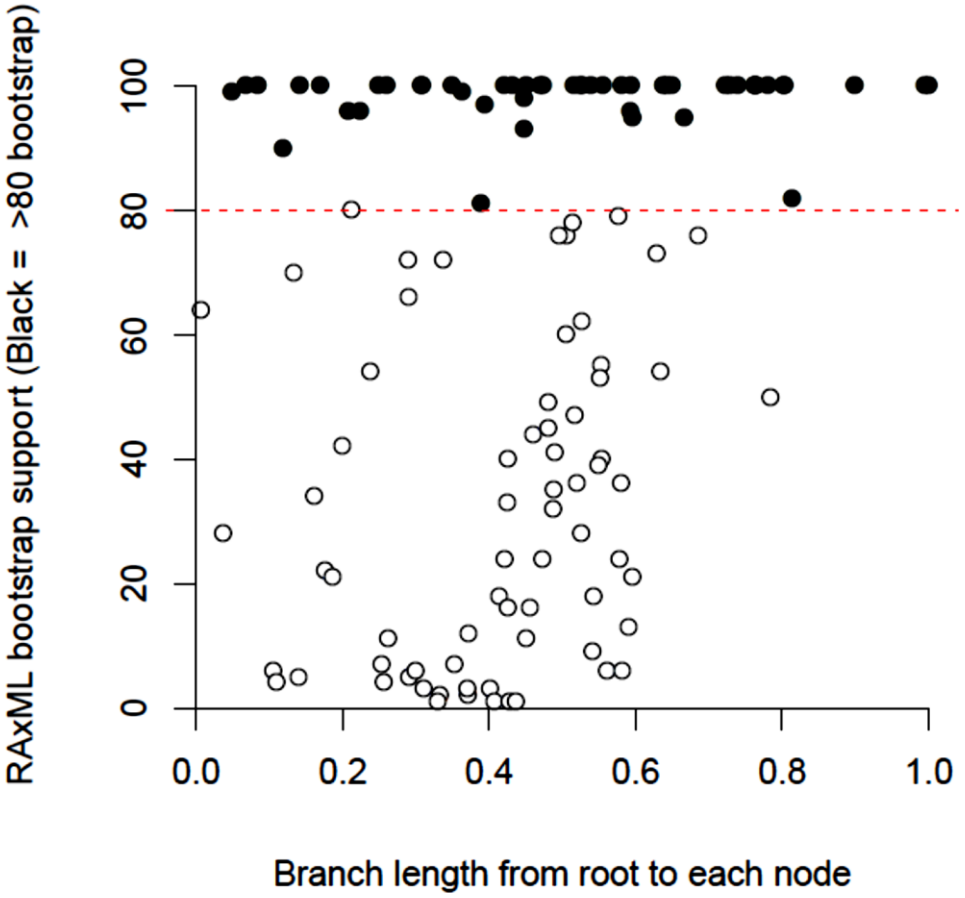


Figure 3.8. Graph of RAxML nodal bootstrap support against branch length of nodes from the root for the analysis of all 15 concatenated genes under the six partition scheme (dataset A).

3.5 Discussion

3.5.1 Contig formation from pooled total DNA sequencing

Our results provide a clear demonstration of efficient and reliable sequencing, assembly and identification of large numbers of mitogenomes from a pool of total DNA of numerous samples, without any enrichment or PCR amplification. Other recent papers attempting to generate full mitochondrial genomes from total DNA either generated a separate library for each taxon (Williams *et al.* 2014) or pooled a small number of distantly related taxa only (Rubinstein *et al.* 2013). It was possible to employ the resulting sequence data to reconstruct a higher-level phylogeny of the superfamily Curculionoidea that is highly congruent with recent molecular phylogenies and provides additional evidence for the convergent evolution of specialised wood-boring behaviour and morphology in weevils. The method has been explored previously for the analysis of bulk insect samples from a forest canopy (Crampton-Platt *et al.* unpublished data), applied to nearly 500 individuals from >200 species. They found that the assembly of mitogenomes from bulk samples is hampered by substantial differences in DNA concentration for species in the pool, due to variation in both body size and number of specimens representing a species. In addition, intra-specific variation was found to cause difficulties with assembly due to polymorphisms, mirroring the well-known problem with genome assembly from heterozygotes (e.g. Langley *et al.* 2011). The design of the current study was expected to avoid these problems by normalising the DNA concentration in the pool and by selecting a single individual per species. However, it was found that there is no close correlation of sequencing depth and assembly success (Figure 3.5), in accordance with Crampton-Platt et al. (unpublished data). Our study excludes the presence of intra-specific variation, but indicates that there is a sequencing depth at which assemblers no longer operate optimally, possibly due to the larger numbers of individual sequencing errors contributed by overlapping reads.

A concern of pooled assemblies is the formation of chimeras by the miss-assembly of different mitogenomes. The potential for this is expected to increase if closely related samples that may not differ in conserved regions of the mitogenomes are included in the pool. The prevalence of chimeras was tested using 77 taxa for which multiple baits were available. In many cases these tests involved both the *cytb* or *rrnL* and the two fragments of the *cox1* gene that map to distant positions in the mitogenome.

Not a single case of chimera formation was observed. In addition, the tree topology gave no reason to suggest chimeras, because of the monophyly of the smaller families of Curculionoidea, while chimera formation would also have produced great differences in the length of terminal branches that were not observed.

Further supporting the lack of chimeric formation using this pipeline, the *in silico* assembly test resulted in predominantly close to perfect matches of simulated assemblies to the original mitogenomes, and no inter-specific chimeras. However, because the original partial weevil mitogenomes did not contain ribosomal and control region sequences, this test may have been conservative. The single weevil mitogenome that was not completely assembled in one sequence in the simulation highlights the difficulty that the assembly algorithm has in combining scaffolds where a long string of ambiguities is present in the original sequence. However, this 'conservative' approach will also serve to prevent or reduce potential chimeric assemblies.

The results of the second simulation, using complete dolphin mitogenomes, did not indicate any difficulty in the assembly of the control region, but assembly of ribosomal genes was often incomplete and usually resulted in a separate, short scaffold containing *rrnS* and *rrnL.* Although only six original mitogenomes were reassembled in one piece, there was no evidence for chimeric assemblies in these or in any of the partial scaffolds. These results highligh the fact that only a portion of the original mitogenomes were fully recovered intact, which may contribute towards explaining why not more weevil mitogenomes were recovered in the present work.

3.5.2 Phylogenetic analysis from densely sampled mitogenomes

Together with existing mitogenome sequences, a total of 120 terminals were included in the phylogenetic analysis. As mitogenome data sets increase with the numbers of taxa needed for dense sampling, this may produce problems with tree searches and model choice. Specifically, the most complex models, such as the amino acid based CAT model used by Timmermans *et al.* (2010) that was required for resolving the deep-level relationships within the Coleoptera are not practical when the number of taxa becomes larger. This raises the question of what is the value of using complex models. Haran *et al.* (2013) have shown that likelihood trees of weevils can be substantially improved under model partitioning according to (i) codon position and (ii) forward vs. reverse strand, the latter presumably due to the well-established differences in codon usage on either strand. A formal analysis was conducted to test if this partitioning scheme by strand and codon captures the most important aspects of the nucleotide variation using the PartitionFinder software, starting from 41 potential partitions of each codon position within each gene. This could be reduced to the codon positions for all genes on either strands, similar to Haran *et al.* (2013), but maintaining a single partition for the 2nd codon position on either strand, while adding a separate partition for the rRNA genes not included in that study. The use of these six partitions over the full set of 41 partitions led only to a small reduction in likelihood, while the unpartitioned models were substantially worse (Table 3.2).

A general difficulty for comparing models is that comparisons are only possible for a single topology, but searches under different partitions favour different topologies. Therefore the optimal trees obtained under no partitioning and the six and 41-partition schemes were use to assess likelihoods of the alternative partitioning schemes on those three topologies. The likelihoods on all trees for the three models were almost identical (Table 3.2), indicating that tree topology is not a major deciding factor for the best model. Taken at face value, the 41 partition wins out over the six partition scheme in all three analyses, but the likelihood gain is minor. As likelihood values become very large with the use of numerous whole mitogenomes, AIC values may not be an appropriate approach to avoid over-parameterisation, unless they are normalised for the total likelihood values (Castoe *et al.* 2005). Therefore the six-partition scheme is believed to be fully adequate. In addition, the practicalities of tree searches on increasingly large datasets from full mitogenomes, as generated with the proposed methodology, also strongly argue for parameter reduction.

3.5.3 Implications for the systematics of weevils

The close relationship linking Platypodinae with Dryophthoridae, as sister to the Curculionidae *s.str.*, has been demonstrated multiple times (Marvaldi *et al.* 1997, McKenna *et al.* 2009 and Haran *et al.* 2013) and indicates that the family Curculionidae, as presently classified, is paraphyletic. The simplified classification system proposed by Oberprieler *et al.* (2007), recognising a broader Curculionidae also containing the presently defined Brachyceridae and Dryophthoridae as respective subfamilies (*sensu* Alonso-Zarazaga and Lyal 1999) would be consistent with our family-level results. Our results strongly support the relationships amongst the curculionoid families at the base of the tree, which are consistent with most previous molecular analyses, with the exception of the placement of Nemonychidae. This family has previously been suggested to be split off at the most basal node (e.g. McKenna *et al.* 2009), as opposed to Anthribidae in our results, but our sampling lacks two of the 'primitive' weevil families (Belidae and Caridae), prohibiting a definitive conclusion. Our results are also consistent with the previously suggested hypothesis that the Brentidae are the sister family to all the 'true weevils', Curculionidae, if Brachyceridae and Dryophthoridae are included in the latter.

A previously described deep split within the true weevils was confirmed by our substantially increased sampling. One strongly supported clade contains the Entiminae + Cyclominae + Hyperinae, and represents the monophyletic and diverse 'broad-nosed' weevils, so named because of their relatively short and blunt rostrums. Rearrangements within the cluster of six tRNA genes are restricted to this clade, even with our increased taxon coverage, further supporting its distinctiveness. The cyclomine genus *Dichotrachelus*, containing the same RANSEF rearrangement as all other Entiminae (except *Sitona*) in our analysis, has been treated as belonging to the Entiminae by some authors (Meregalli & Osella 2007) on morphological grounds. Combined with the low nodal support for its inclusion in a monophyletic Cyclominae (< 50% BS), our tRNA rearrangement data are consistent with this opinion. The second clade containing all other curculionoid subfamilies, with the exception of Bagoinae, which is placed outside of the two main clades, is much less satisfactorily resolved, with only two of its constituent subfamilies (Lixinae and Ceutorhynchinae) being monophyletic. It contains a number of very large subfamilies including the Curculioninae, Molytinae, Baridinae, Cryptorhynchinae and Conoderinae, whose relationships remain obscure due to a lack of strong nodal support. Whilst the recovery of two tribes within this group being monophyletic (Lobotrachelini and Cionini) is encouraging, in order to further investigate the confusing topology of this clade, significantly more representative taxon sampling will be required. Indeed, limitations in taxon sampling are often cited as potentially limiting factors in higher-level phylogenetics (Franz & Engel 2010) and this is certainly an important consideration in such a large group as the Curculionoidea.

An interesting finding is that strong nodal support spans the full depth of the tree and differing taxonomic ranks (families, subfamilies and tribes; Fig. 3.8). This pattern was seen in analyses of all datasets and under all partitioning models. A potential criticism of mitochondrial sequence data is that due to accelerated evolutionary rates, saturation of sites may obscure or distort phylogenetic signal at deeper nodes (Talavera & Vila 2011). It is clear from our data that at least at the intra-superfamily level in weevils, this is not necessarily the case, with phylogenetic signal being evenly distributed across the estimated 170 million year diversification history of the weevils (McKenna *et al,* 2009).

3.5.4 Evolution of wood-boring behaviour

The wood-boring weevil subfamilies are highly adapted to excavate galleries, either subcortically or in woody tissue, and feed on ligneous matter directly or cultivate symbiotic fungi in the tunnels as a food source, and for this reason many are widespread pests of forestry (Oberprieler *et al.* 2007). The taxon density of the current analysis nearly matched the extensive sampling of the wood-boring groups by Jordal *et al.* (2011), a study that is the basis for suggesting their close affinity. However, in contrast to Jordal *et al.* (2011) our results support the conclusions of Haran *et al.* (2013) and McKenna *et al.* (2009), indicating that wood-boring lineages are clearly not monophyletic, with Platypodinae consistently retrieved as closely related to the Dryophthoridae (and Brachyceridae) in a clade sister to all other Curculionidae *sensu* Bouchard *et al.* (2011). Although our analyses recovered neither the Scolytinae nor the Cossoninae as monophyletic, and they were never recovered as sister taxa or nested within the same clade, it is not possible to confidently conclude as to the relationship between them because only a series of weakly supported nodes separate the cossonine taxa and *Coptonotus* from the rest of the Scolytinae. The latter genus is interesting for consistently not being recovered in our analyses within the generally well-supported Scolytinae clade (excepting Scolytini). Based upon morphological characters, *Coptonotus* has been considered to be a transitional taxon between Platypodinae and other Curculionidae (Jordal *et al.* 2011) or alternatively as an intermediate form between Cossoninae and Scolytinae (Thompson 1992), whilst also containing morphological characters linking it with Cossoninae. Thompson (1992) has suggested a close relationship between Coptonotini and the scolytine tribe Hylastini based on structures of the aedeagus. However our results argue against this because the Hylastini sample (*Hylastesopacus*) was retrieved with strong support as the sister of Tomicini, and this clade itself was strongly supported as sister to the Hylesini, within the main Scolytinae clade.

3.5.5 Conclusions

The relative ease of obtaining a large number of mitogenome DNA sequences from a pooled mixture of DNA extracts has been demonstrated, without the need for enrichment or species specific tagging prior to genome pooling. Mitogenome sequences are confidently identified to specimen with a limited amount of prior mtDNA sequence data for each sample, and exhibit no error with regard to these bait sequences. Our mtDNA genome data yields phylogenetic relationships that are highly congruent with prior expectations, and provides phylogenetic signal with robustly supported nodes across a broad range of lineage divergence times and taxon diversity, from family-level to generic-level, which are consistent across different data partitioning schemes.

It is evident that the efficiency of our approach will be a function of the relative concentration of mitochondrial to nuclear DNA within a focal group. The average coleopteran genome size is estimated to be approximately 0.65 Gb +/- 0.05 (http://www.genomesize.com). Under the assumption that the copy number of mtDNA genomes does not differ substantially across organisms, our approach should be of broad utility within insect phylogenetics where mean nuclear genome size is estimated to be 1.22 Gb +/- 0.05. However, it may be less efficient for taxa with larger average nuclear genome sizes (e.g. crustaceans: mean nuclear genome size = approximately 4.45 Gb +/- 0.45). A further consideration for the implementation of our approach is taxon sampling and the mitogenomic assembly pipeline. Our sampling for the higher-level taxonomic relationships within the Curculionoidea provides little challenge for the pipeline, as mtDNA genomes sampled from different genera exhibit high DNA sequence divergence. Genome divergence facilitates genome reassembly from a mixed pool of genome fragments, and the pipeline efficiency will eventually be compromised as mtDNA genome relatedness increases. Our data suggests this limit lies somewhere below an uncorrected divergence of 10% for *cox*1 and *cytB* that characterises the two species of *Cionus* (*C. olens* and *C. griseus*) included in our sampling. To ascertain genome relatedness thresholds for the reassembly pipeline, simulation analyses can be employed. However, it is important to point out that as NGS technology and read lengths improve, relatedness thresholds will also become more favourable.

3.6 References

Alonso-Zarazaga MA, Lyal CHC (1999) *A world catalogue of families and genera of Curculionoidea (Insecta: Coleoptera) (excepting Scolytidae and Platypodidae)* Entomopraxis, Barcelona.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW (2012) Grinder: a versatile amplicon and shotgun sequence emulator. *Nucleic Acids Research* **40,** doi: 10.1093/nar/gks251.

Benson DA, Cavanaugh M, Clark K*, et al.* (2013) GenBank. *Nucleic Acids Research* **41**: 10.1093/nar/gks1195.

Botero-Castro F, Tilak M-K, Justy F*, et al.* (2013) Next-generation sequencing and phylogenetic signal of complete mitochondrial genomes for resolving the evolutionary history of leaf-nosed bats (Phyllostomidae). *Molecular Phylogenetics and Evolution* **69**, 728-739.

Bouchard P, Bousquet Y, Davies AE*, et al.* (2011) Family-group names in Coleoptera (Insecta). *Zookeys* **88**, 1-972.

Cameron SL, Lo N, Bourguignon T, Svenson GJ, Evans TA (2012) A mitochondrial genome phylogeny of termites (Blattodea: Termitoidae): Robust support for interfamilial relationships and molecular synapomorphies define major clades. *Molecular Phylogenetics and Evolution* **65**, 163-173.

Castoe TA, Sasa MM, Parkinson C (2005) Modeling nucleotide evolution at the mesoscale: the phylogeny of the neotropical pitvipers of the *Porthidium* group (Viperidae: Crotalinae). *Molecular Phylogenetics and Evolution* **37**, 881-898.

Crampton-Platt A, Timmermans MJTN, Gimmel ML*, et al.* (Unpublished data) Pooled mitochondrial genome assembly for biodiversity discovery in a phylogenetic framework.

Crowson RA (1955) *The natural classification of Coleoptera* Nathaniel Lloyd & Co., London.

Curole JP, Kocher TD (1999) Mitogenomics: digging deeper with complete mitochondrial genomes. *Trends in Ecology & Evolution* **14**, 394-398.

Eddy SR, Durbin R (1994) RNA sequence-analysis using covariance models. *Nucleic Acids Research* **22**, 2079-2088.

Finstermeier K, Zinner D, Brameier M*, et al.* (2013) A mitogenomic phylogeny of living primates. *PloS one* **8**: 10.1371/journal.pone.0069504.

Franz NM, Engel MS (2010) Can higher-level phylogenies of weevils explain their evolutionary success? A critical review. *Systematic Entomology* **35**, 597-606.

Haran J, Timmermans MJTN, Vogler AP (2013) Mitogenome sequences stabilize the phylogenetics of weevils (Curculionoidea) and establish the monophyly of larval ectophagy. *Molecular Phylogenetics and Evolution* **67**, 156-166.

Hundsdoerfer AK, Rheinheimer J, Wink M (2009) Towards the phylogeny of the Curculionoidea (Coleoptera): Reconstructions from mitochondrial and nuclear ribosomal DNA sequences. *Zoologischer Anzeiger* **248**, 9-31.

Hunt T, Bergsten J, Levkanicova Z*, et al.* (2007) A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science* **318**, 1913-1916.

Jordal BH, Sequeira AS, Cognato AI (2011) The age and phylogeny of wood boring weevils and the origin of subsociality. *Molecular Phylogenetics and Evolution* **59**, 708-724.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066.

Kayal E, Roure B, Philippe H, Collins AG, Lavrov DV (2013) Cnidarian phylogenetic relationships as revealed by mitogenomics. *BMC Evolutionary Biology* **13**: 10.1186/1471-2148-13-5.

Kuschel G (1995) A phylogenetic classification of Curculionoidea to families and subfamilies. *Memoirs of the Entomological Society of Washington* **14**, 5-33.

Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* **29**, 1695-1701.

Langley CH, Crepeau M, Cerdeno C, Corbett-Detig R, Stevens K (2011) Circumventing heterozygosity: sequencing the amplified genome of a single haploid *Drosophila melanogaster* embryo. *Genetics* **188**, 239-246.

Lohse M, Bolger AM, Nagel A*, et al.* (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research* **40**: 0.1093/nar/gks540.

Marvaldi AE (1997) Higher level phylogeny of Curculionidae (Coleoptera : Curculionoidea) based mainly on larval characters, with special reference to broad-nosed weevils. *Cladistics-the International Journal of the Willi Hennig Society* **13**, 285-312.

McKenna DD, Sequeira AS, Marvaldi AE, Farrell BD (2009) Temporal lags and overlap in the diversification of weevils and flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7083-7088.

Meregalli M, Osella G (2007) *Dichotrachelus kahleni* sp. n., a new weevil species from the Carnian Alps, north-eastern Italy (Coleoptera, Curculionidae, Entiminae). *Deutsche Entomologische Zeitschrift* **54**, 169-177.

Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: *Proceedings of the gateway computing environments workshop (GCE), 14 Nov. 2010*, New Orleans, LA.

Myers EW, Sutton GG, Delcher AL*, et al.* (2000) A whole-genome assembly of *Drosophila*. *Science* **287**, 2196-2204.

Oberprieler RG, Marvaldi AE, Anderson RS (2007) Weevils, weevils, weevils everywhere. *Zootaxa* **1668**, 491-520.

Osigus H-J, Eitel M, Bernt M, Donath A, Schierwater B (2013) Mitogenomics at the base of Metazoa. *Molecular Phylogenetics and Evolution* **69**, 339-351.

Pacheco MA, Battistuzzi FU, Lentino M*, et al.* (2011) Evolution of modern birds revealed by mitogenomics: timing the radiation and origin of major orders. *Molecular Biology and Evolution* **28**, 1927-1942.

Paradis E (2013) Molecular dating of phylogenies by likelihood methods: A comparison of models and a new information criterion. *Molecular Phylogenetics and Evolution* **67**, 436-444.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289-290.

Peng Y, Leung HCM, Yiu SM, Chin FYL (2011) Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* **27**, I94-I101.

Posada D, Buckley T (2004) Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology* **53**, 793-808.

Roos J, Aggarwal RK, Janke A (2007) Extended mitogenomic phylogenetic analyses yield new insight into crocodylian evolution and their survival of the Cretaceous-Tertiary boundary. *Molecular Phylogenetics and Evolution* **45**, 663-673.

Rubinstein ND, Feldstein T, Shenkar N*, et al.* (2013) Deep sequencing of mixed total DNA without barcodes allows efficient assembly of highly plastic ascidian mitochondrial genomes. *Genome Biology and Evolution* **5**, 1185-1199.

Salzberg SL, Yorke JA (2005) Beware of mis-assembled genomes. *Bioinformatics* **21**: 4320-4321.

Sommer DD, Delcher AL, Salzberg SL, Pop M (2007) Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8**: 10.1186/1471-2105-8-64.

Song HJ, Sheffield NC, Cameron SL, Miller KB, Whiting MF (2010) When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. *Systematic Entomology* **35**, 429-448.

Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690.

Talavera G, Vila R (2011) What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta class phylogeny. *BMC Evolutionary Biology* **11**: 10.1186/1471-2148-11-315.

Thompson RT (1992) Observations on the morphology and classification of weevils (Coleoptera, Curculionoidea) with a key to major groups. *Journal of Natural History* **26**, 835-891.

Timmermans MJTN, Dodsworth S, Culverwell CL, *et al.* (2010) Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research* **38**: 10.1093/nar/gkq807.

Vilstrup JT, Ho SYW, Foote AD, Morin PA, Kreb D, Krützen M, Parra GJ, Robertson KM, Stephanis R, Verborgh P, Willerslev E, Orlando L, Gilbert MTP (2011) Mitogenomic phylogenetic analyses of the Delphinidae with and emphasis on the Globicephalinae. *BMC Evolutionary Biology* **11**: doi:10.1186/1471-2148-11-65

Wei S-j, Shi M, Sharkey MJ, van Achterberg C, Chen X-x (2010) Comparative mitogenomics of Braconidae (Insecta: Hymenoptera) and the phylogenetic utility of mitochondrial genomes with special reference to Holometabolous insects. *BMC Genomics* **11**, doi:10.1186/1471-2164-1111-1371.

Wernersson R (2005) FeatureExtract - extraction of sequence annotation made easy. *Nucleic Acids Research* **33**: 10.1093/nar/gki388.

Williams S, Foster PG, Littlewood DTJ (2014) The complete mitochondrial genome of a turbinid vetigastropod from MiSeq Illumina sequencing of genomic DNA and steps towards a resolved gastropod phylogeny. *Gene* **533**, 38-47.

Winkelmann I, Campos PF, Strugnell J, *et al.* (2013) Mitochondrial genome diversity and population structure of the giant squid Architeuthis: genetics sheds new light on one of the most enigmatic marine species. *Proceedings of the Royal Society B-Biological Sciences* **280**: 10.1098/rspb.2013.0273.

3.7 Appendices

Appendix 3.1 A-E Custom bioinformatics scripts used in mitogenome assembly
Scripts developed and written by Alex Crampton-Platt and Martijn JTN Timmermans

Appendix 3.1A FastqExtract3.pl – Perl script to extract putative mtDNA paired reads from the concatenated BLAST output of the BLAST search of R1 and R2 MiSeq reads against the reference Coleoptera mitochondrial sequences.

```
print "what is the name of the blast_reader.pl output file?\n";
$blastreader = <STDIN>;
chomp $blastreader;
print "what is the name of the raw reads file?\n";
$filename = <STDIN>;
chomp $filename;
print "what is the read identifier (first 6 header characters, not inc @)?\n";
$idf = <STDIN>;
chomp $idf;

#INPUT:
$fastq = $filename;
open(FASTQ, "<$fastq") or print "could not open file $fastq";

#HEADERS:
$headers = $blastreader;
chomp $headers;
open(HEADERS, "$headers") or print "could not open file $headers";
@headers = <HEADERS>;
close HEADERS;

foreach $head(@headers){
if($head =~ m/\w/){
@head = split(" ", $head);

#add to hash
$headers{$head[0]}=1}
}
$outfile = "$fastq\.out";
open(OUTFILE, ">$outfile") or print "could not open file $outfile";

########################################

open(FASTQ, "<$fastq");
while(<FASTQ>){
$line = $_;
if($flag == 1){$flag = 2; print OUTFILE $line;}
elsif($flag == 2){$flag = 3; print OUTFILE $line;}
elsif($flag == 3){$flag = 0; print OUTFILE $line;}

        elsif($line =~ m/$idf:/){
        @line = split(" ", $line);
        $line[0] =~ s/\@//ig;


                if(exists $headers{$line[0]}){
                ++$counter; $flag = 1;
                print "$counter\:\t\t$line";
                print OUTFILE "$line";
                }

        }
}

close                                                                                    OUTFILE;
```

Appendix 3.1B

Example AWK command to filter BLAST results for contigs of ≥ 1000 bp from each assembler (ca = celera, idba = IDBA-UD)

```
    awk ‘($4>=1000){print}’ MyLibrary.ca.ctg.blastn > MyLibrary.ca.ctg.blastn.filter

    awk ‘($4>=1000){print}’ MyLibrary.idba.ctg.blastn > MyLibrary.idba.scf.blastn.filter
```

Appendix 3.1C retrieve2.py – Python script to make new FASTA files of only the mitochondrial contigs from each assembler

```
def retrieve(blastout,seqfile,outfile):
        o1=open(blastout)
        l1=o1.readlines()
        querylist=[]
        for each in l1:
                k=each.split('\t')
                if k[0]!='':
                        querylist.append(k[0])
        l1=[]
        o1.close()
        print "List of BLAST IDs built"
        o2=open(seqfile)
        l2=o2.readlines()
        seqdict={}
        poslist=[]
        for i,j in enumerate(l2):
                if ">" in j:
                        poslist.append(i)
        for i,j in enumerate(poslist):
                k1=l2[j].split('>')
                k2=k1[1].replace('\n','')
                if i!=len(poslist)-1:
                        k3=l2[j+1:poslist[i+1]]
                if i==len(poslist)-1:
                        k3=l2[j+1:]
                k4=''.join(k3)
                k5=k4.replace('\n','')
                seqdict[k2]=k5
        print "Dictionary of fasta built"
        o2.close()
        o3=open(outfile,'a')
        for each in querylist:
                o3.write('\n>'+each+'\n')
                k=seqdict.get(each)
                o3.write(k)
        print "Printed output"
        o3.close()
blastout=raw_input("Blast output filename: ")
seqfile=raw_input("Sequence fasta file: ")
outfile=raw_input("output filename: ")

retrieve(blastout,seqfile,outfile)
```

Appendix 3.1D all2many.pl – Perl script to generate an individual FASTA file for each contig/scaffold > 1000 kb

```
#!/usr/local/bin/perl
#
# Usage information updated January 21, 2005 - JDW


unless (@ARGV == 2) {
  print <<EOH;
Usage: $0  input.file  min_size_contig

    Separates a set of FASTA-format sequences in the file named
    as the first argument into individual files, each of which is
    at least as long as the  second argument. Sequences shorter
    than the second argument are ignored.  The name of each new
    file is the name of the contig.

Examples:

$0  fasta.screen.contigs  1000

creates individual files from the input file fasta.screen.contigs
for all contigs at least 1000 bases long.

$0  another.fasta.file  1

creates individual files from the input file another.fasta.file
for all contigs.
EOH
exit 0;
}
sub dump_seq {
        my($name, $seq) = @_;
        $name1 = $name;
        $name1 =~ s/>//;
         $name1 =~ s/^(\S*).*$/$1/;
        open HUNK,">$name1.fasta" or die $!;
        print HUNK "$name \n";
        $seq =~ s/\n//ig;
        print HUNK "$seq\n";
        close HUNK;
}
open BIG,$ARGV[0] or die $!;
while (<BIG>) {
  if ($_ =~ /Contig|^>/) {
        if ($len >= $ARGV[1]) {
           dump_seq($name, $seq);
         }
      chomp;
      $name = $_;
      $seq = ''; $len = 0;
  } else {
     $seq .= $_;
     $len += length($_) - 1;
  }
}
dump_seq($name, $seq);
close BIG;
```

Appendix 3.1E

R script for renaming tree terminals with taxon names and for calculating the branch length of each node from the base of the tree and plotting this against its respective RAxML BS support (Requires MrBayes to use conformat=simple and a .csv file of the OTU codes and corresponding real taxon names). Code developed by James Kitson.

```
### Clear the workspace
rm(list=ls())

### set working directory and make objects for calling the string later
setwd("C:/Working directory")
work<-as.character(getwd())
out<-paste(work,"/output_trees/",sep="")

### load the ape and phytools libraries
library(ape)
library(phytools)

### make a list of all files in the tree input directory
inputs<-list.files(paste(work,"/input_trees/",sep=""))

### subset this to include only files that are tre files
inputs<-subset(inputs,grepl('.nex$',inputs))

### make sure its ok
inputs

### extract a list of names from inputs for plotting file names
file.names<-substr(inputs,1,nchar(inputs)-nchar("_TREE.nex"))
file.names

### List control command, change this from 1-n files and rerun each time
x<-1

### Load the tree from the working directory
my.tree<-read.nexus(paste(work,"/input_trees/",inputs[[x]], sep=""))

#### make my.tree ultrametric if needed
ctrl <- chronos.control(nb.rate.cat = 1)
my.tree<-chronos(my.tree, model = "discrete", control = ctrl)
is.ultrametric(my.tree)
```

```
### rotate all the nodes so the outgroups are at the bottom
my.tree<-rotateNodes(my.tree,"all")

### make sure support values are numeric for the graph
my.tree$node.label<-as.numeric(my.tree$node.label)

### read in the list of names
name<-read.csv("Names.csv")
### read.csv turns text into factors, this gets messy later when plotting
### so make it character data
name<-data.frame(lapply(name, as.character), stringsAsFactors=FALSE)
########################################################
### replace the codes with informative names
my.tree.rename<-my.tree
### make a vector of tip colours by matching the colour column in name to the sample name
my.tip.colours<-(name$Colour[match(my.tree.rename$tip.label,name$sample)])
### the next line uses match to perform the same function as vlookup in excel
my.tree.rename$tip.label <- (name$Name[match(my.tree.rename$tip.label,name$sample)])
write.nexus(my.tree.rename,file=paste(out,file.names[[x]],"_rename.nex",sep=""))
#################################################

##### Plot node ages vs support values ##########
#################################################

### Calculate the branch depths for each node (distance from tip)
node.depths<-branching.times(my.tree)
node.depths<-as.numeric(node.depths)

### subtract all the distances from the tip to each node from the maximum depth to get node
heights
node.heights<-max(nodeHeights(my.tree))-node.depths

### extract the support values to make the plotting easier
node.support<-my.tree$node.label

###  plot the tree
pdf(paste(out,"tree_",file.names[[x]],".pdf",sep=""),30,35)
### pdf("poly_tree.pdf",30,35)
plot(my.tree.rename,
     show.node.label=FALSE,
     cex=1.5,
     tip.color=my.tip.colours)
### the node labels command below plots the node numbers
###
nodelabels(seq(from=my.tree$Nnode,to=(length(my.tree$tip.label)+my.tree$Nnode)),adj=c(1,1),
frame="none",col="red",cex=1.5)

### the node labels command below plots the support values
nodelabels(my.tree.rename$node.label,adj=c(1.1,1.3),frame="none",
          col=ifelse(my.tree$node.label>90,"red","black"),cex=1.5)
dev.off()

### handy plot for checking if my names are correct against the codes
```

```
pdf(paste(out,"cophyloplot_",file.names[[x]],".pdf",sep=""),30,35)
cophyloplot(my.tree,my.tree.rename,assoc=NULL)
dev.off()

### plot the graph of node heights against support
pdf(paste(out,"graph_",file.names[[x]],".pdf",sep=""),5,5)
### pdf("test_graph.pdf",5,5)
plot(node.support~node.heights,axes=FALSE,pch=21,col="black",
     bg=ifelse(node.support>80,"black","white"),
     xlab="Branch length from root to each node",
     ylab="RAxML bootstrap support (Black =  >80 bootstrap)")
abline(h=80,lty=2,col="red")
axis(1, pos=0)
axis(2, pos=0)
title(main=paste(file.names[[x]],sep=""))
dev.off()
```

Appendix 3.2 Taxa (organised alphabetically by family) present in the final dataset with number of genes and aligned assembly lengths. Newly assembled and identified mitogenomes are highlighted in grey, all others were obtained from Genbank.

Family	Subfamily	Tribe	Genus	Species	Origin	Source	Code	No. of genes	Total length (bp)
Anthribidae	Anthribinae	Platystomini	*Platystomos*	*albinus*	France	Haran	JN-163968	13	9460
Anthribidae				sp. 1	China	Gillett/Lyal	CG336	15	12906
Attelabidae	Apoderinae	Apoderini	*Apoderus*	*coryli*	France	Haran	JN-163966	12	8793
Attelabidae	Apoderinae			sp. 2	China	Gillett/Lyal	CG335	15	13023
Attelabidae	Attelabinae			sp. 1	China	Gillett/Lyal	CG323	15	12989
Attelabidae	Rhynchitinae	Byctiscini	*Byctiscus*	*populi*	France	Haran	JN-163965	12	8269
Attelabidae	Rhynchitinae	Deporaini	*Deporaus*	*betulae*	England	Haran	JN-163945	13	9520
Brachyceridae	Brachycerinae	Brachycerini	*Brachycerus*	*muricatus*	France	Haran	JN-163970	13	9459
Brachyceridae	Erirhininae	Erirhirinini	*Echinocnemis*	sp.	Australia	Oberprieler	CG210	15	13034
Brachyceridae	Ocladiinae	Ocladiini	*Ocladius*	sp.	RSA	Meregalli	CG288	15	13010
Brentidae	Apioninae	Apionini	*Rhopalapion*	*longirostre*	France	Haran	JN-163967	13	9460
Brentidae	Nanophyinae	Nanophyini	*Nanophyes*	*marmoratus*	France	Haran	JN-163946	13	9471
Brentidae	Nanophyinae	Nanophyini	*Nanophyes*	sp.	Turkey	Levent	CG271	14	11673
Brentidae				sp. 1	China	Gillett/Lyal	CG347	15	13021
CERAMBYCIDAE			*Anoplophora*	*glabripennis*		Genbank	NC-008222	14	11689
CHRYSOMELIDAE			*Crioceris*	*duodecimpunctata*		Genbank	NC-003372	15	13031
Curculionidae	Bagoinae		*Bagous*	sp.	England	Turner	CG220	15	13025
Curculionidae	Baridinae	Baridini	*Melanobaris*	*laticollis*	France	Haran	JN-163955	13	9453
Curculionidae	Ceutorhynchinae	Ceutorhynchini	*Ceutorhynchus*	*assimilis*	France	Haran	JN-163956	13	9495
Curculionidae	Ceutorhynchinae	Mononychini	*Mononychus*	*punctumalbum*	Italy	Caldara	CG306	12	10038
Curculionidae	Ceutorhynchinae	Phytobini	*Rhinoncus*	sp.	Turkey	Levent	CG282	15	13012
Curculionidae	Conoderinae	Lobotrachelini		sp. 1	China	Gillett/Lyal	CG321	12	9522
Curculionidae	Conoderinae	Lobotrachelini		sp. 2	China	Gillett/Lyal	CG322	13	10109
Curculionidae	Conoderinae	Lobotrachelini		sp. 3	China	Gillett/Lyal	CG328	15	13018

Curculionidae	Conoderinae	Mecopini	*Mecopus*	sp.	Australia	Oberprieler	CG248	15	13081
Curculionidae	Conoderinae	Zygopini	*Peltophorus*	sp.	USA	O'Brien	CG295	15	13022
Curculionidae	Cossoninae	Neumatorini	*Brachytemnus*	*porcatus*	France	Haran	JN-163960	13	9525
Curculionidae	Cossoninae	Pentarthrini	*Pentarthrus*	*elumbe*	England	Turner	CG222	15	13033
Curculionidae	Cossoninae			sp. 1	China	Gillett/Lyal	CG319	15	13005
Curculionidae	Cryptorhynchinae	Camptorhinini	*Camptorhinus*	sp.	Australia	Oberprieler	CG253	15	13041
Curculionidae	Cryptorhynchinae	Cryptorhynchini	*Acalles*	*aubei*	France	Haran	JN-163957	13	9505
Curculionidae	Cryptorhynchinae	Cryptorhynchini	*Ouroporopterus*	sp.	Australia	Oberprieler	CG240	15	13047
Curculionidae	Cryptorhynchinae	Cryptorhynchini	*Perissops*	sp.	Australia	Oberprieler	CG238	15	13023
Curculionidae	Cryptorhynchinae	Cryptorhynchini	*Pseudomopsis*	sp.	Saba	Gillett M	CG352	15	13037
Curculionidae	Cryptorhynchinae	Cryptorhynchini		sp.	Cameroon	Jordal	CG415	8	7924
Curculionidae	Curculioninae	Acalyptini	*Acalyptus*	sp.	Italy	Caldara	CG052	15	13029
Curculionidae	Curculioninae	Anthonomini	*Anthonomus*	*pomorum*	France	Haran	JN-163951	13	9457
Curculionidae	Curculioninae	Ceratopini	*Ceratopus*	sp.	Saba	Gillett M	CG351	12	9275
Curculionidae	Curculioninae	Cionini	*Cionus*	*griseus*	Canaries	Oromi	CG293	15	13050
Curculionidae	Curculioninae	Cionini	*Cionus*	*olens*	France	Haran	JN-163958	13	9472
Curculionidae	Curculioninae	Cryptoplini	*Haplonyx*	sp.	Australia	Oberprieler	CG235	15	13055
Curculionidae	Curculioninae	Eugnomini	*Ancyttalia*	sp.	Australia	Oberprieler	CG242	15	13026
Curculionidae	Curculioninae	Mecinini	*Miarus*	sp.	RSA	Meregalli	CG284	15	13053
Curculionidae	Curculioninae	Storeini	*Melanterius*	sp.	Australia	Oberprieler	CG257	15	13044
Curculionidae	Curculioninae	Tychiini	*Sibinia*	*fulva*	USA	O'Brien	CG298	14	11324
Curculionidae	Cyclominae	Aterpini	*Pelolorhinus*	sp.	Australia	Oberprieler	CG247	14	11049
Curculionidae	Cyclominae	Aterpini	*Rhadinosomus*	sp.	Australia	Oberprieler	CG229	14	12377
Curculionidae	Cyclominae	Dichotrachelini	*Dichotrachelus*	*manueli*	Italy	Meregalli	CG283	15	13043
Curculionidae	Cyclominae	Rhythirrinini	*Cisolea*	sp.	Australia	Oberprieler	CG226	14	11655
Curculionidae	Cyclominae	Rhythirrinini	*Rhythirrinus*	sp.	RSA	Meregalli	CG289	15	13050
Curculionidae	Entiminae	Brachyderini	*Brachyderes*	*rugatus*	Canaries	Emerson	N28	13	11245
Curculionidae	Entiminae	Brachyderini	*Strophosoma*	*melanogrammum*	France	Haran	JN-163949	13	9333
Curculionidae	Entiminae	Brachyderini	*Strophosoma*	sp.	England	Turner	CG300	15	11989
Curculionidae	Entiminae	Cratopini	*Cratopus*	*sumptuosus*	Reunion	Kitson	T-Reu3834	15	12975
Curculionidae	Entiminae	Geonemini	*Barynotus*	*obscurus*	France	Haran	JN-163950	12	8835
Curculionidae	Entiminae	Geonemini	*Lachnopus*	*curvipes*	Saba	Gillett M	CG354	15	13051
Curculionidae	Entiminae	Laparocerini	*Laparocerus*	*freyi*	Canaries	Faria	LAP007	13	10638
Curculionidae	Entiminae	Myorhinini		sp.	RSA	Meregalli	CG285	15	13046

Curculionidae	Entiminae	Naupactini	*Litostylus*	*pudens*	Saba	Gillett M	CG355	15	12398
Curculionidae	Entiminae	Naupactini	*Naupactus*	*xanthographus*	RSA	Genbank	GU-176345	15	13160
Curculionidae	Entiminae	Oosomini	*Barianus*	sp.	Juan de Nova	Kitson	CG305	15	13053
Curculionidae	Entiminae	Ophryastini	*Ophryastes*	sp.	USA	O'Brien	CG297	15	13044
Curculionidae	Entiminae	Otiorhynchini	*Otiorhynchus*	*globulus*	Italy	Caldara	CG309	13	10968
Curculionidae	Entiminae	Otiorhynchini	*Otiorhynchus*	*rugosostriatus*	France	Haran	JN-163969	13	9494
Curculionidae	Entiminae	Otiorhynchini	*Otiorhynchus*	sp.	England	Gillett	CG307	15	13046
Curculionidae	Entiminae	Polydrusini	*Liophloeus*	*tessulatus*	France	Haran	JN-163947	13	9462
Curculionidae	Entiminae	Polydrusini	*Polydrusus*	*marginatus*	France	Haran	JN-039360	12	9207
Curculionidae	Entiminae	Psallidiini	*Psallidium*	sp.	Turkey	Levent	CG272	15	13047
Curculionidae	Entiminae	Sitonini	*Sitona*	*lineatus*	France	Haran	JN-163948	13	9443
Curculionidae	Entiminae	Tanymecini	*Geotragus*	sp.	China	Li	CG311	15	12986
Curculionidae	Entiminae	Trachyphloeini	*Trachyphloeus*	sp.	England	Turner	CG301	15	13055
Curculionidae	Entiminae	Tropiphorini	*Catasarcus*	sp.	Australia	Oberprieler	CG227	9	7422
Curculionidae	Entiminae	Tropiphorini	*Leptopius*	sp.	Australia	Oberprieler	CG341	13	10485
Curculionidae	Entiminae	Tropiphorini	*Tropiphorus*	*bertolini*	Italy	Caldara	CG315	15	13050
Curculionidae	Entiminae			sp. 1	China	Gillett/Lyal	CG330	15	13041
Curculionidae	Entiminae			sp. 2	China	Gillett/Lyal	CG331	15	13042
Curculionidae	Entiminae			sp. 3	China	Gillett/Lyal	CG339	15	13036
Curculionidae	Entiminae			sp. 4	China	Gillett/Lyal	CG342	15	13050
Curculionidae	Entiminae			sp. 5	China	Gillett/Lyal	CG349	14	12103
Curculionidae	Hyperinae	Hyperini	*Hypera*	*postica*	France	Haran	JN-163953	13	9429
Curculionidae	Lixinae	Lixini	*Larinus*	*turbinatus*	France	Haran	JN-163952	12	8666
Curculionidae	Lixinae	Rhinocyllini	*Bangasternus*	sp.	Turkey	Levent	CG268	15	13034
Curculionidae	Mesoptiliinae	Laemosaccini	*Laemosaccus*	sp.	USA	O'Brien	CG296	15	13055
Curculionidae	Mesoptiliinae	Magdalinini	*Magdalis*	sp.	Italy	Caldara	CG069	15	13060
Curculionidae	Molytinae	Hylobini	*Hylobius*	*abietis*	France	Haran	JN-163954	13	9467
Curculionidae	Molytinae	Lepyrini	*Lepyrus*	sp.	China	Li	CG312	15	13070
Curculionidae	Molytinae	Pissodini	*Pissodes*	sp.	Italy	Caldara	CG055	15	13035
Curculionidae	Molytinae			sp. 1	China	Gillett/Lyal	CG317	14	11667
Curculionidae	Molytinae			sp. 2	China	Gillett/Lyal	CG332	15	13052
Curculionidae	Molytinae			sp. 3	China	Gillett/Lyal	CG340	15	12985
Curculionidae	Molytinae			sp. 4	China	Gillett/Lyal	CG350	15	13032
Curculionidae	Platypodinae	Platypodini	*Platypus*	*cylindricus*	England	Turner	CG221	15	12884

Curculionidae	Platypodinae	Platypodini	*Platypus*	*cylindricus*	France	Haran	JN-163963	13	9458
Curculionidae	Platypodinae	Tesserocerini	*Diapus*	*unispineus*	PNG	Jordal	CG419	15	12964
Curculionidae	Scolytinae	Coptonotini	*Coptonotus*	*cyclopus*	Costa Rica	Jordal	CG445	15	13016
Curculionidae	Scolytinae	Corthylini	*Corthylus*	*rubricollis*	Costa Rica	Jordal	CG435	14	11824
Curculionidae	Scolytinae	Corthylini	*Pityophthorus*	*micrographus*	Sweden	Jordal	CG436	15	13009
Curculionidae	Scolytinae	Cryphalini	*Cryphalus*	*saltuarius*	Norway	Jordal	CG437	15	12963
Curculionidae	Scolytinae	Crypturgini	*Crypturgus*	*pusillus*	Norway	Jordal	CG434	15	13001
Curculionidae	Scolytinae	Diamerini	*Diamerus*	*inermis*	Tanzania	Jordal	CG438	15	12985
Curculionidae	Scolytinae	Dryocoetini	*Dryocoetes*	*autographus*	Norway	Jordal	CG422	14	10850
Curculionidae	Scolytinae	Hexacolini	*Scolytodes*	*caudatus*	Costa Rica	Jordal	CG420	15	13002
Curculionidae	Scolytinae	Hylastini	*Hylastes*	*opacus*	Sweden	Jordal	CG423	15	13047
Curculionidae	Scolytinae	Hylesini	*Hylesinus*	*varius*	Sweden	Jordal	CG424	15	13026
Curculionidae	Scolytinae	Hypoborini	*Hypoborus*	*ficus*	Morocco	Jordal	CG439	15	12910
Curculionidae	Scolytinae	Ipini	*Ips*	*acuminatus*	Norway	Jordal	CG426	15	13009
Curculionidae	Scolytinae	Ipini	*Ips*	*cembrae*	France	Haran	JN-163961	8	4994
Curculionidae	Scolytinae	Phloeotribini	*Phloeotribus*	*spinulosus*	Norway	Jordal	CG442	15	13013
Curculionidae	Scolytinae	Polygraphini	*Polygraphus*	*poligraphus*	Sweden	Jordal	CG441	15	13013
Curculionidae	Scolytinae	Premnobiini	*Premnobius*	*cavipennis*	RSA	Jordal	CG428	15	12685
Curculionidae	Scolytinae	Scolytini	*Scolytus*	*scolytus*	Denmark	Jordal	CG429	14	11229
Curculionidae	Scolytinae	Scolytini	*Scolytus*	sp.	France	Haran	JN-163962	13	9384
Curculionidae	Scolytinae	Tomicini	*Tomicus*	*piniperda*	Norway	Jordal	CG425	15	13057
Curculionidae	Scolytinae	Xyleborini	*Anisandrus*	*dispar*	Norway	Jordal	CG431	15	12819
Curculionidae	Scolytinae	Xyloctonini	*Xyloctonus*	*maculatus*	RSA	Jordal	CG444	15	12950
Curculionidae	Scolytinae			sp. 1	China	Gillett/Lyal	CG325	15	12990
Curculionidae	Scolytinae			sp. 2	China	Gillett/Lyal	CG346	15	12988
Dryophthoridae	Orthognathinae	Rhinostomini	*Rhinostomus*	*barbirostris*	Belize	Barclay	CG074	15	13032
Dryophthoridae	Rhynchophorinae	Litosomini	*Sitophilus*	*granarius*	France	Haran	JN-163959	10	5379
Dryophthoridae	Rhynchophorinae	sp. henophorini	*Cosmopolites*	*sordidus*	China	Gillett/Lyal	CG344	15	13048
Dryophthoridae				sp. 1	China	Gillett/Lyal	CG324	15	13009
Nemonychidae	Cimberidinae	Doydirbyncbini	*Doydirhynchus*	*austriacus*	France	Haran	JN-163964	13	9515

Appendix 3.3A Results of *in silico* assembly using simulated shotgun sequence reads from 27 weevil partial mitogenomes (Haran *et al.* 2013). Newly assembled scaffold lengths, BLAST matches to originating mitogenomes, and original mitogenome sequence lengths are listed. Highlighted in grey are the three scaffolds matching to the same mitogenome as discussed in the results.

IDBA-UD Scaffold	Scaffold length (bp)	Best mitogenome BLAST match	BLAST % pairwise identity	BLAST E-value	BLAST query coverage %	Mitogenome original length (bp)
0	11269	JN163956	100	0	100	11346
1	11149	JN169357	100	0	100	11168
2	11146	JN163967	100	0	100	11152
3	10812	JN163970	100	0	100	10817
4	10757	JN163955	100	0	100	10792
5	10702	JN163948	100	0	100	10720
6	10675	JN163958	100	0	100	10702
7	10664	JN039360	100	0	100	10692
8	10652	JN163960	100	0	100	10666
9	10638	JN163953	100	0	100	10646
10	10636	JN163954	100	0	100	10673
11	10635	JN163964	100	0	100	10666
12	10626	JN163945	100	0	100	10673
13	10615	JN163969	100	0	100	10675
14	10602	JN163951	100	0	100	10606
15	10580	JN163947	100	0	100	10628
16	10555	JN163968	100	0	100	10594
17	10539	JN163963	100	0	100	10583
18	10537	JN163962	100	0	100	10567
19	10533	JN163946	100	0	100	10629

20	10482	JN163949	100	0	100	10511
21	9994	JN163950	100	0	100	10016
22	9818	JN163966	99.9	0	100	9899
23	9782	JN163952	100	0	100	9809
24	9457	JN163965	100	0	100	9468
25	6204	JN163959	99.9	0	100	6291
26	3153	JN163961	100	0	100	5670
27	2424	JN163961	99.9	0	100	5670

Appendix 3.3B Results of *in silico* assembly using simulated shotgun sequence reads from 17 dolphin mitogenomes (Vilstrup *et al.* 2011). Newly assembled scaffold lengths, BLAST matches to originating mitogenomes, and original mitogenome sequence lengths are listed. Mitogenomes that were recovered in one long scaffold are in bold type. Mitogenomes recovered in two or more non-overlapping scaffolds are highlighted in grey.

IDBA-UD Scaffold	Scaffold length (bp)	Best mitogenome BLAST match	BLAST % pairwise identity	BLAST E-value	BLAST query coverage %	Complete mitogenome original length (bp)
0	**16568**	**AJ554059 *Inia geoffrensis***	**100**	**0**	**100**	**16588**
1	**16377**	**AJ554062 *Monodon monoceros***	**100**	**0**	**100**	**16383**
2	**16362**	**AY789529 *Lipotes vexillifer***	**100**	**0**	**100**	**16392**
3	**16350**	**AJ554063 *Phocoena phoccena***	**100**	**0**	**100**	**16382**
4	**16313**	**AJ554061 *Lagenorhynchus albirostris***	**100**	**0**	**100**	**16393**
5	**16297**	**GU187186 *Orcinus orca***	**100**	**0**	**100**	**16386**
6	14035	EU557091 *Sousa chinensis*	100	0	100	16388
27	1414	EU557091 *Sousa chinensis*	99.6	0	100	16388

7	14016	HM060333 *Globicephala macrorhynchus*	100	0	100	16387
30	1298	HM060333 *Globicephala macrorhynchus*	100	0	100	16387
9	13999	HM060332 *Pseudorca crassidens*	100	0	100	16392
33	1134	HM060332 *Pseudorca crassidens*	99.3	0	100	16392
8	14005	JJF289177 *Orcaella brevirostris*	100	0	100	16383
24	2437	JJF289177 *Orcaella brevirostris*	100	0	100	16383
10	13986	JF289175 *Peponocephala electra*	100	0	100	16388
29	1340	JF289175 *Peponocephala electra*	99.9	0	99.9	16388
11	11324	EU557095 *Grampus griseus*	100	0	100	16386
19	2884	EU557095 *Grampus griseus*	100	0	100	16386
32	1170	EU557095 *Grampus griseus*	100	0	100	16386
12	11281	EU557094 *Delphinus capensis*	100	0	100	16385
22	2667	EU557094 *Delphinus capensis*	99.9	0	100	16385
13	10666	JF289171 *Feresa attenuata*	100	0	100	16387
26	1524	JF289171 *Feresa attenuata*	100	0	100	16387
34	1045	JF289171 *Feresa attenuata*	98.9	0	100	16387
14	10295	EU557096 *Stenella attenuata*	100	0	100	16386
21	2671	EU557096 *Stenella attenuata*	100	0	100	16386
25	2146	EU557096 *Stenella attenuata*	99.5	0	100	16386
31	1268	EU557096 *Stenella attenuata*	100	0	100	16386
15	8369	EU557093 *Tursiops truncatus*	100	0	100	16388
18	4732	EU557093 *Tursiops truncatus*	100	0	100	10567
16	7048	JF33998 *Steno bredanensis*	100	0	100	16385
17	4773	JF33998 *Steno bredanensis*	100	0	100	16385
20	2706	JF33998 *Steno bredanensis*	100	0	90.28	16385
23	2538	JF33998 *Steno bredanensis*	100	0	100	16385

Appendix 3.4 Results of identification of mitogenomic assemblies by BLAST searching 'bait' sequences against corresponding *cox1* 5', *cox1* 3', *cytB* and rrnL sequences from new assemblies. In grey are indicated successful assembly identifications. Numbers in the bait columns denote the unique identification numbers for the assemblies with the best 'hit' to each bait. Total number of baits available and number of successful 'hits' per sample is shown. For unsuccessful assembly identifications the reason for failure is given. * Conspecific samples CG343 and CG317 resulted in a single mitogenomic assembly.

Sample	*cox1* 5' bait	*cox1* 3' bait	*cytB* bait	*rrnL* bait	Total baits	Total bait hits	I.D. success?	Reason for I.D. failure
CG031					1	0	**n**	**No bait hits**
CG052	181	181	181	181	4	4	**y**	
CG055	104	104	104	104	4	4	**y**	
CG069	180		180	180	3	3	**y**	
CG074	112	112	112	112	4	4	**y**	
CG205					3	0	**n**	**No bait hits**
CG206	91	91	42	457514	4	4	**n**	**Short assemblies**
CG210	156	156	156	156	4	4	**y**	
CG212	197		62	74	3	3	**n**	**Short assemblies**
CG213	7				3	1	**n**	**Short assemblies**
CG215					4	0	**n**	**No bait hits**
CG220	183				1	1	**y**	
CG221	114	114			2	2	**y**	
CG222	150	150		150	3	3	**y**	

CG223					1	0	**n**	**No bait hits**
CG224					2	0	**n**	**No bait hits**
CG225	4	4			2	2	**n**	**Short assemblies**
CG226			100-27		1	1	**y**	
CG227			89		1	1	**y**	
CG229	100-31	100-31	100-31		3	3	**y**	
CG230	251327	251327			3	2	**n**	**Short assemblies**
CG231					2	0	**n**	**No bait hits**
CG232	457690	457690	59	59	4	4	**n**	**Short assemblies**
CG235	459182		459182		2	2	**y**	
CG236	41				1	1	**n**	**Short assemblies**
CG237					3	0	**n**	**No bait hits**
CG238	173	173			2	2	**y**	
CG239					3	0	**n**	**No bait hits**
CG240	177	177			2	2	**y**	
CG241					3	0	**n**	**No bait hits**
CG242		154	154	154	3	3	**y**	
CG243					1	0	**n**	**No bait hits**
CG244					4	0	**n**	**No bait hits**
CG245					2	0	**n**	**No bait hits**
CG246					4	0	**n**	**No bait hits**
CG247	43	43	79		3	3	**y**	
CG248		100-14			1	1	**y**	
CG249					4	0	**n**	**No bait hits**
CG250					3	0	**n**	**No bait hits**
CG252		51	251254		3	2	**n**	**Short assemblies**

CG253	82	82		82	3	3	**y**	
CG254		5			1	1	**n**	**Short assembly**
CG255					1	0	**n**	**No bait hits**
CG257	176	176		176	3	3	**y**	
CG258	191	191	200		4	3	**n**	**Short assemblies**
CG259					1	0	**n**	**No bait hits**
CG260	100-665				2	1	**n**	**Short assemblies**
CG261					4	0	**n**	**No bait hits**
CG263					1	0	**n**	**No bait hits**
CG264					4	0	**n**	**No bait hits**
CG265					1	0	**n**	**No bait hits**
CG266					3	0	**n**	**No bait hits**
CG267			84		2	1	**n**	**Short assemblies**
CG268	100-2	100-2	100-2	100-2	4	4	**y**	
CG269	214	214		73	4	3	**n**	**Short assemblies**
CG270	100-337	100-337	459069		4	3	**n**	**Short assemblies**
CG271		44	44	44	3	3	**y**	
CG272	96	96	96	96	4	4	**y**	
CG274					4	0	**n**	**No bait hits**
CG275					3	0	**n**	**No bait hits**
CG276					4	0	**n**	**No bait hits**
CG277	201	201			4	2	**n**	**Short assemblies**
CG278	203	203	189		3	3	**n**	**Short assemblies**
CG279		41	60	457037	4	3	**n**	**Short assemblies**
CG280		31	94		3	2	**n**	**Short assemblies**
CG281		67	14		4	2	**n**	**Short assemblies**

CG282	130	130	130	130	4	4	**y**	
CG283		127		127	2	2	**y**	
CG284	100-26	100-26	100-26		3	3	**y**	
CG285	182	182			2	2	**y**	
CG286			13		3	1	**n**	**Short assemblies**
CG287					3	0	**n**	**No bait hits**
CG288	100-6		100-6		2	2	**y**	
CG289	120	120	458630	458630	4	4	**y**	
CG290					1	0	**n**	**No bait hits**
CG291					1	0	**n**	**No bait hits**
CG293	146		146		2	2	**y**	
CG295	145	145	145		3	3	**y**	
CG296	175	175			2	2	**y**	
CG297	153	153	153		3	3	**y**	
CG298	147	147	147		3	3	**y**	
CG299	25	25			4	2	**n**	**Short assemblies**
CG300	134B	134B	134B	134B	4	4	**y**	
CG301	113	113	113	113	4	4	**y**	
CG302					3	0	**n**	**No bait hits**
CG303					3	0	**n**	**No bait hits**
CG304			195	195	4	2	**n**	**Short assemblies**
CG305		148			1	1	**y**	
CG306		458889	458889	129	3	3	**y**	
CG307		142	142	142	3	3	**y**	
CG308					1	0	**n**	**No bait hits**
CG309	108	108	108	100-231	4	4	**y**	

CG310		34	40		4	2	**n**	**Short assemblies**
CG311	95	95	95		4	3	**y**	
CG312	144	144	144	144	4	4	**y**	
CG313				147	1	1	**n**	**Ambiguous 16S hits**
CG314				147	4	1	**n**	**Ambiguous 16S hits**
CG315	126	126	126	126	4	4	**y**	
CG316		50			3	1	**n**	**Short assemblies**
CG317		81	81	81	3	3	**y**	
CG318					3	0	**n**	**No bait hits**
CG319				165	1	1	**y**	
CG320					2	0	**n**	**No bait hits**
CG321			100-94		1	1	**y**	
CG322			123		1	1	**y**	
CG323			80	80	2	2	**y**	
CG324				110	1	1	**y**	
CG325	117	117	117		3	3	**y**	
CG326					2	0	**n**	**No bait hits**
CG327		206			4	1	**n**	**Short assemblies**
CG328			163	163	2	2	**y**	
CG329			246528		1	1	**n**	**Short assemblies**
CG330			100-25	100-25	2	2	**y**	
CG331	119	119	119	119	4	4	**y**	
CG332		100-28	100-28	100-28	3	3	**y**	
CG333			36		1	1	**n**	**Short assembly**
CG334					1	0	**n**	**No bait hits**
CG335			75		1	1	**y**	

CG336		160			1	1	**y**	
CG337					1	0	**n**	**No bait hits**
CG338			10		1	1	**n**	**Short assembly**
CG339		149	149	149	3	3	**y**	
CG340		141	141	141	3	3	**y**	
CG341		152	152		2	2	**y**	
CG342		162	162	162	3	3	**y**	
CG343		81	81		2	2	**y***	
CG344			101	101	2	2	**y**	
CG346	161		161	161	3	3	**y**	
CG347		178	178		2	2	**y**	
CG348					3	0	**n**	**No bait hits**
CG349			168	168	2	2	**y**	
CG350		164	164	164	3	3	**y**	
CG351			56		2	1	**y**	
CG352			115	115	2	2	**y**	
CG353			457326		1	1	**n**	**Short assemblies**
CG354		139	139	139	3	3	**y**	
CG355		107	143	143	3	3	**y**	
CG412					1	0	**n**	**No bait hits**
CG414					3	0	**n**	**No bait hits**
CG415		102	185		2	2	**y**	
CG418					1	0	**n**	**No bait hits**
CG419			124	124	2	2	**y**	
CG420		172		172	2	2	**y**	
CG421					3	0	**n**	**No bait hits**

CG422		76	190	190	3	3	**y**	
CG423		137	137	137	3	3	**y**	
CG424		116	116		2	2	**y**	
CG425		151	151		2	2	**y**	
CG426		158	158		2	2	**y**	
CG427		210			3	1	**n**	**Short assemblies**
CG428			169	169	2	2	**y**	
CG429		118			1	1	**y**	
CG430		93			3	1	**n**	**Short assemblies**
CG431		109	109	109	3	3	**y**	
CG432			193	193	3	2	**n**	**Short assemblies**
CG434			157		1	1	**y**	
CG435		155	155		2	2	**y**	
CG436		77	77	77	3	3	**y**	
CG437		122	122	122	3	3	**y**	
CG438		136	136		2	2	**y**	
CG439			184	184	2	2	**y**	
CG440					1	0	**n**	**No bait hits**
CG441				159	1	1	**y**	
CG442		174	174		2	2	**y**	
CG443		457110	100-215	250606	3	3	**n**	**Short assemblies**
CG444				135	1	1	**y**	
CG445		170		170	2	2	**y**	
CG446					1	0	**n**	**No bait hits**
CG447	61			52	2	2	**n**	**Non-weevil**
CG448					2	0	**n**	**No bait hits**

LAP007	98	98	98		3	3	**y**
N28	134A			134A	2	2	**y**
T-Reu3834		140	140	140	3	3	**y**

Chapter 4

Augmenting mitogenomic sequence data with nuclear ribosomal and protein-coding gene sequences: an assessment of additive value using the phylogeny of weevils (Coleoptera: Curculionoidea)

"Although a large number of genera and several thousand species have been described as belonging to this group, yet we know comparatively little regarding it; it is the most anomalous and in many respects the least satisfactory of all the divisions of the Coleoptera"

- Rev. Canon W.W. Fowler, 1891

Cryptorhynchini sp. (Curculionidae), Copperbelt Province, Zambia

Chapter 4: Augmenting mitogenomic sequence data with nuclear ribosomal and protein-coding gene sequences: an assessment of additive value using the phylogeny of weevils (Coleoptera: Curculionoidea)

4.1 Abstract

Phylogenetic congruence resulting from analysis of independent datasets (such as sequences from different genes) converging on similar topologies can provide compelling support for evolutionary hypotheses. Measuring the added benefit of incorporating additional loci into an existing dataset can yield information on the utility of such a strategy. Here, this is considered by investigating whether there is any advantage, as measured by bootstrap nodal support, of supplementing a dataset of complete weevil (Coleoptera: Curculionoidea) mitogenome sequences with nuclear ribosomal and protein-coding genes. Maximum likelihood analyses of multiple concatenated datasets of differing gene composition and taxon number reveal that there is little advantage to be gained from the addition of 18S, 28S and *ArgK* gene sequence data to the mitogenomic dataset. The effect on nodal support of their inclusion is paralleled by the effect of improving taxon coverage. Faced with a choice, it is argued that enlarged taxon sampling should take priority over an increase in markers for mitogenomic phylogenetic analysis.

4.2 Introduction

Phylogenetic reconstruction employing molecular sequence data as characters frequently assumes that the gene-trees so generated are accurate approximations of the species-level phylogeny (Avise 2004; Page & Holmes 1998). This is not necessarily true because incongruences between the 'gene-tree' and the 'species-tree' can exist primarily as a result of incomplete lineage sorting, but also due to introgressive hybridisation, or the existence of homoplasious data (Sota & Vogler 2001). However, because of the stochastic mechanism by which lineage sorting, introgression and homoplasy occurs, it is extremely unlikely that two or more gene-trees, each built from an unlinked gene, will share the same topological conflicts to the species-tree, thereby enabling the possibility of identifying true congruence (Johnson & Clayton 2000).

Congruence of phylogenetic tree topologies built from independent datasets (such as morphological, molecular, behavioural) provides affirmation supporting shared relationships. For the present purposes, independent datasets are constituted of sequences from different orthologous genes, and there are contrasting views as to how best to treat such data in the context of phylogeny reconstruction. Leigh *et al.* (2011) proposed the existence of three fundamental philosophical strategies, the first being "taxonomic congruence", whereby separate phylogenetic analyses of each of the independent datasets is undertaken, which can thereafter be compared and summarised in a consensus tree. The second approach, "character congruence", argues for a 'total evidence' approach whereby all the data is combined and analysed simultaneously in one analysis and congruence of characters is assessed through statistical tests developed for partitioned data (e.g. the Incongruence Length

Difference test - ILD). The third strategy, "conditional data combination", involves initially 'testing' the data for whether it is heterogenous (did not evolve along the same tree) or not. If the data are heterogenous, then the "taxonomic congruence" approach is followed and if the data are not heterogenous, then they are combined and analysed according to the "character congruence" methodology (Leigh *et al.* 2011).

The aim of this chapter is to combine nuclear ribosomal 18S and 28S, and protein-coding arginine kinase (*ArgK*) sequences into phylogenetic analyses incorporating the mitogenomic sequence matrix generated in Chapter 3, in order to empirically evaluate the effect of the additional data on the statistical nodal support of the resulting trees in comparison to the mitogenomic sequence matrix alone. The question being addressed is whether the addition of nuclear data leads to improved nodal bootstrap support (BS) for the mitogenomic tree, particularly within the large clade of Curculionidae *s.str.*, containing the species rich subfamilies Curculioninae, Molytinae and Cryptorhynchinae, amongst others. To achieve this, a combination of the "taxonomic congruence" and "character congruence" approaches are primarily undertaken because phylogenetic relationships recovered with multiple independent datasets can arguably be considered to provide particularly strong evidence for having recovered the species-tree (Leigh *et al.* 2011) and alternatively the best hypothesis of evolution is that obtained through simultaneous analysis of the "total evidence" (DeSalle & Brower 1997). Whilst the "character congruence" and "conditional data combination" approaches alone might seem attractive due to the inclusion of statistical tests of congruence, in practice arguments have been made that such tests (e.g. the Incongruence Length Difference test) have "limited power to detect incongruence caused by differences in the evolutionary conditions or in the tree topology, except when numerous characters are present and the substitution rate is homogenous from site to site" (Darlu & Lecointre 2002: 432). There are also philosophical arguments against discriminating between character sets (e.g. genes) by making underlying *ad hoc* assumptions about the empirical data and ultimately reducing the explanatory power of a hypothesis (DeSalle & Brower 1997).

One of the most widely adopted of statistical tools for assessing the confidence intervals in phylogenies (i.e. the 'accuracy' of each clade) is non-parametric bootstrap resampling (Felsenstein 1985) which has for a long time been used to measure phylogenetic robustness (Rubinoff & Holland 2005). Whilst there have been criticisms of the technique, including claims that it is biased to be consistently too conservative, Efron *et al.* (1996), after a statistical investigation of such claims, concluded that the confidence values "obtained by Felsenstein's bootstrap method are not biased systematically downward" and that they can be thought of as "reasonable assessments of error for the estimated tree". As a pragmatic assessment of comparative nodal support across dataset analyses, we employ bootstrapping as implemented in the maximum likelihood estimating program RAxML (Stamatakis 2006).

Mitogenomes have now been shown to be reliable markers for phylogeny reconstruction across diverse taxonomic ranks (e.g. Kayal *et al.* 2013; Osigus *et al.* 2013), but partly because the wide availability of complete mitogenomic data is a relatively recent phenomenon, there have been few attempts at combining full mitogenomic sequences with nuclear markers (Janke *et al.* 2002; San Mauro *et al.* 2004), although the incorporation of single or multiple mtDNA genes together with nuclear genes has seen widespread use in studies investigating relationships across diverse taxonomic groups (e.g. Fisher-Reid & Wiens 2011), including those within the superfamily Curculionoidea (e.g. Hundsdoerfer *et al.* 2009; McKenna *et al.* 2009). Consequently there has been little research undertaken so far into the benefits of combining nuclear data with full mitogenomes, *i.e.* whether any improvement in topological resolution and nodal support can be gained with the addition of such markers over a mitogenomic-tree alone. On a pragmatic level, PCR amplification and sequencing of a further three loci (as is undertaken here), approximately doubles the sequencing costs compared to obtaining the NGS-derived mitogenomic data alone, so it is important to investigate whether the additional resources necessary for this are a worthwhile investment in terms of ultimate phylogenetic utility. Therefore, this chapter represents a practical, empirical evaluation of what, if any, gains are to be made by supplementing mitogenomic sequence data with additional nuclear loci for higher level invertebrate phylogenetics.

4.3 Materials and methods

4.3.1 Nuclear gene selection and taxon sampling

18S (small subunit) and 28S (large subunit) rRNA are both components of the eukaryotic ribosome and therefore a basic component of all eukaryotic cells. The sequences for these two nuclear genes have a long history of use in phylogeny reconstruction, and were selected for this study because they have been successfully included in recent large-scale studies within the Coleoptera (Bocak *et al.* 2013) and within the Curculionoidea (McKenna *et al.* 2009). Their widespread use in phylogeny can be attributed to their core structures containing strongly conserved regions across all life, although also possessing extremely variable regions that differ even between closely related species, and which can, as a result, complicate alignment (Marvaldi *et al.* 2009).

The other selected gene, arginine kinase (*ArgK*), codes for a phosphotransferase employed in metabolism regulation. Its nucleotide sequence has not been utilised in phylogeny reconstruction to as great an extent as rRNA genes, although it was evaluated by Wild and Maddison (2008) for phylogenegtic utility and was found to be able to more accurately reconstruct deeper nodes than more recent ones within their empirical Coleoptera test data, and did not reveal any evidence of paralogous copies. This gene has been previously used by McKenna *et al.* (2009) in curculionoid phylogeny reconstruction.

Genomic DNA aliquots from the 92 weevil species for which complete or near-complete mitogenomic sequences were generated in chapter 3 (Appendix 3.2) were selected for the PCR amplification of sections of the nuclear ribosomal 18S and 28S RNA, and the protein-coding *ArgK* genes. These samples contained taxa representing seven curculionoid families, including 13 subfamilies and 55 identified tribes within the family Curculionidae *s.str.* Taxonomy follows the most recent Coleoptera classification of Bouchard *et al.* (2011) with genera assigned to higher-level taxa according to the catalogue of Alonso-Zarazaga and Lyal (1999).

4.3.2 DNA amplification and sequencing

Standard PCR reactions were undertaken for each of the three nuclear loci and for each of the 92 samples. PCR products were used as templates in sequencing reactions employing the Big Dye v3.1 Cycle Sequencing Kit (Applied Biosystems). All sequencing reactions were cycled at 96°C for 10s, 58°C for 5s and 60°C for 4 mins, repeated for 25 cycles. The results were read on a 3730XL sequencer (Applied Biosystems).

Because its length, of approximately 1900 bp, rendered it too long for amplification in a single PCR, the selected region of the 18S gene was amplified in two overlapping sections, as shown in Figure 2.2. This was achieved using the same primers and similar methodology as those of Shull *et al.* (2001).

Amplification and sequencing of 28S was straightforward, utilising the same primers (designed by Monaghan *et al.* 2007) for both the PCR and sequencing steps, as detailed in Chapter 2 (Appendix 2.2 A-C). However, due to poor PCR success with the available primers, reliable amplification and sequencing of *ArgK* was only achieved using a 'nested PCR' methodology as described in Chapter 2 (Figure 2.3; Appendix 2.2 A-C), whereby the PCR product of a first amplification reaction is used as template DNA in a second-round PCR. Newly designed internal primers for this re-amplification step were developed, and sequencing reactions were undertaken using these same primers, according to the standard sequencing cycling profile described above. All sequences were manually edited in Geneious 5.4 prior to alignment, to remove primer sequences and poor-quality regions flanking the target regions.

4.3.3 Sequence alignment and dataset concatenation

Once all newly-generated gene sequences were edited, it was possible to cross-reference samples to those for which the mitogenomic assemblies had already been obtained in Chapter 3. Two sample-groups were thereby created; one group (A) contained all the samples sequenced for the mitogenomic, 18S and 28S loci, consisting of 79 taxa. The second, smaller sample-group (B), consisted of samples for which *ArgK* sequences were also successfully generated, and consisted of 65 taxa. All mitogenome sequences analysed in this chapter were obtained as described in Chapter 3.

Sequences for each of the three nuclear loci were individually aligned using the MAFFT 7.0 online server, under the FFT-NS-I slow iterative refinement strategy and with the following parameter values: nucleotide scoring matrix 200PAM/k=2, gap open penalty = 1.53, offset value = 0.0. (Katoh *et al.* 2002). Alignments were checked by eye for quality and to ensure that the *ArgK* sequences were consistent with the reading frame prior to analysis.

The different gene alignments within each of the two sample-groups were concatenated together using Geneious, to generate six different datasets in total, comprising mtDNA-only (MITO, 79 and 65 taxa), mtDNA + 18S and 28S (MITO+rRNA, 79 and 65 taxa), mtDNA + 18S + 28S + *ArgK* (MITO+rRNA+*ArgK*, 65 taxa), and mtDNA + *ArgK* (MITO+*ArgK*, 65 taxa), as summarised in Table 4.1.

4.3.4 Phylogenetic analyses

Each of the six datasets was analysed under a maximum likelihood (ML) optimality criterion to search for the best-scoring tree using RAxML 7.6.6 (Stamatakis 2006) running on the CIPRES web-based server (Miller *et al.* 2010). Trees were rooted with Anthribidae sp. China, the most divergent curculionoid taxon in the matrix, as ascertained in the mitogenomic ML analysis in Chapter 3, and in agreement with previous molecular studies (e.g. Haran *et al.* 2013). To assess nodal support, a rapid bootstrap analysis (BS) with 1000 iterations was conducted simultaneously with optimal tree searching. A GTRCAT model was implemented for the bootstrapping phase and a GTRGAMMA model was used for final tree inference (GTR + optimisation of substitution rates + optimisation of site-specific evolutionary rates). All datasets were partitioned by gene for analysis, whereby separate estimated models of nucleotide substitution were specified for each gene locus in the alignment. The results of analysing various partitioning schemes in Chapter 3 indicated that partitioning by gene is favoured to an unpartitioned analysis, in that the resulting ML score under such a scheme is better for a given dataset and topology (Table 3.2).

Table 4.1 The two sample groups and six datasets employed in this study, with total number of genes in the concatenations and the alignment lengths. rRNA indicates both 18S and 28S concatenated sequences. For each analysis, the mean BS nodal support is that for only the 37 nodes present or consistent in the four strict consensus trees referred to in the text and shown in the appendices.

Sample group	Dataset name	No. of genes	Alignment Length (bp)	Mean BS of consistent nodes in all 4 strict consensus trees
A	**MITO 79 taxa**	15	13792	84.40
	MITO+rRNA 79 taxa	17	17166	85.08
B	**MITO 65 taxa**	15	13792	79.49
	MITO+rRNA 65 taxa	17	17166	84.90
	MITO+rRNA+*ArgK* 65 taxa	18	17699	84.38
	MITO+*ArgK* 65 taxa	16	14325	81.70

4.3.5 Analysis of nodal bootstrap support

In order to investigate the effects on nodal support of incorporating additional gene sequences into the mitogenomic data, strict consensus trees of pairs of ML trees obtained for the MITO datasets, together with each of the following four data sets were constructed using a custom R script (Appendix 4.1) employing the APE package (Paradis *et al.* 2004): MITO+rRNA (79 and 65 taxa respectively); MITO+rRNA+*ArgK* (65 taxa); and MITO+*ArgK* (65 taxa). For each of the four resulting strict consensus trees, all nodes were numbered and the corresponding BS nodal support values in the MITO and the MITO+other genes ML trees were mapped onto the corresponding strict consensus tree. These were then used to calculate the change in mean BS support (Δ mean BS) across consistent nodes in the tree obtained with the additional genes, over the tree obtained with MITO data alone. Furthermore, all nodes that were present or consistent across all four strict consensus trees were individually coded with a letter code and mapped onto the consensus trees, and the mean BS support for all these nodes in both the originating MITO and the MITO+other_genes ML trees was also calculated. It is logical and judicious to concentrate on the BS support for the consistent nodes across analyses, as opposed to the weakly supported nodes differing amongst them, which in a statistical sense have little meaning and are of limited use for interpreting meaningful topological relationships.

To characterise the additive effect of the addition of the rRNA genes and *ArgK* for the mitogenomic data on the BS nodal support across different nodal ages (and putative taxonomic ranks), we investigated the pattern of distribution of nodal support across ML trees by calculating the branch of each node from the base of the tree using a custom R script (Appendix 3.1E) and plotting this against its respective RAxML BS support. To achieve this, the six ML trees were first made ultrametric using the *chronos* function of the APE package in R, which uses penalised likelihood to fit a chronogram to a phylogenetic tree whose branch lengths are in number of substitutions per site (Paradis 2013).

Because our MITO+rRNA data contained separate alignments with two different numbers of taxa (79 and 65), it was also possible to investigate the effect of taxon sampling on nodal support. To further explore this, an additional RAxML analysis was undertaken after reducing the 79 taxa MITO+rRNA dataset to 77 taxa through removal of two 'basal' taxa: Brentidae *Nanophyes* sp. and Attelabidae Attelabinae sp. In the reduced 65 taxa dataset, both these families remained represented.

4.4 Results

4.4.1 Taxon sampling, sequencing and dataset concatenation

Of the 92 taxa for which mitogenomic sequences were available, 79 also generated complete or near complete 18S and 28S rRNA sequence partitions. It is these latter 79 taxa, and a subset thereof, comprising 65 taxa that also generated complete or partial *ArgK* sequences, that are employed in this study (Appendix 4.3). Both the 79 and 65 taxa datasets contained members of six curculionoid families and 13 subfamilies within Curculionidae *s.str.*, whilst 49 and 43 identified tribes of Curculioninae were available for the 79 and 65 taxa datasets respectively. Thus, data consisted of two alignment matrices: 79 taxa sampled for the mitogenome, 18S and 28S loci, and 65 taxa sampled for the same loci plus *ArgK*. Alignment lengths for each locus and concatenated dataset, together with ungapped sequence lengths and number of genes per alignment are summarised in Table 4.2.

4.4.2 Phylogenetic analyses

Strict consensus trees were constructed from pairs of ML trees containing one MITO tree and a corresponding MITO+other genes tree. The four resulting strict consensus trees constructed from pairs of ML trees are shown in Appendices 4.2 – 4.5, each of which also indicates, with letter codings, the 37 shared nodes across all four strict consensus trees.

Table 4.2 Number of loci per data set with corresponding alignment and ungapped sequence lengths. Datasets used in analyses are indicated in bold type.

Loci/Dataset	No. of genes	Alignment length (bp)	Minimum ungapped sequence length (bp)	Maximum ungapped sequence length (bp)
Mitogenomes (MITO)	15	13792	13084	13588
18S rRNA	1	2530	597	1954
28S rRNA	1	844	198	732
ArgK	1	533	274	533
MITO+rRNA	17	17166	14054	16008
MITO+rRNA+*ArgK*	18	17699	14499	16425
MITO+*ArgK*	16	14325	13504	14033

All phylogenetic analyses resulted in highly congruent topologies, regardless of which additional genes were combined with mitogenomic data, and these topologies themselves were highly consistent with those obtained from the analysis of the larger mitochondrial dataset in Chapter 3 (Figure 3.7). The topology of the basal ordering of families remained identical across the four 65 taxa analyses, with the following sequence (oldest branch to most recent branch): Anthribidae, Attelabidae, Brentidae, Dryophtoridae (+ Platypodinae), Brachyceridae and Curculionidae s.str. (excluding Platypodinae). The two 79 taxa analyses resulted in alternative placements of the single Brachyceridae taxon (*Echinocnemis*), either branching off prior to Dryophthoridae + Platypodinae (with the MITO dataset) or nested within the latter clade (with MITO+rRNA), as is evidenced by the unresolved nature of these relationships in the strict consensus of these two analyses (Appendix 4.2). Similarly, the two main large clades into which Curculionidae *s.str.* is divided (Entiminae+Cyclominae, and all other Curculionidae except Bagoinae and Platypodinae), as recovered with the mitogenomic data in Chapter 3, were retrieved in all the present analyses (Appendices 4.1 – 4.4). The subfamily Bagoinae was always recovered as sister to all other Curculioninae *s.str.* (excluding Platypodinae).

4.4.3 Analysis of BS nodal support

Indicated on each node, within each strict consensus tree, are the BS nodal support values for that node in the MITO (either 79 or 65 taxa according to analysis) ML tree and the MITO+other_genes ML tree. These BS values are colour-coded to indicate whether the MITO + other gene(s) ML tree BS value for a particular shared node increased (green), decreased (red) or remained unchanged (blue) over the corresponding BS value in the MITO ML tree. The mean BS support values across all 37 shared nodes for each dataset analysed separately were also calculated, and are listed in Table 4.1 and shown graphically in Figure 4.1.

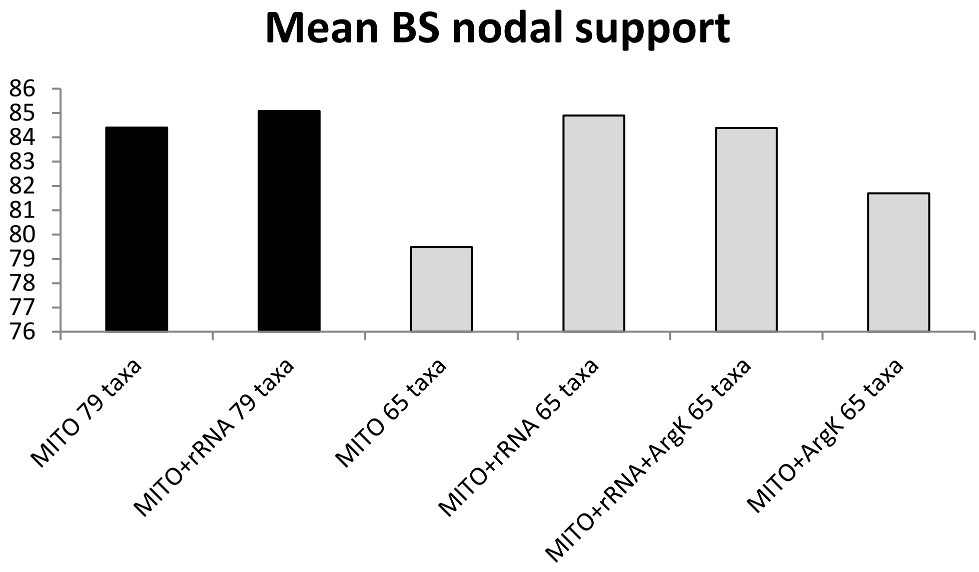


Figure 4.1 Mean BS nodal support of 37 nodes common to (or consistent with) all four strict consensus trees. In black is the 79 taxa datasets, in grey the 65 taxa datasets.

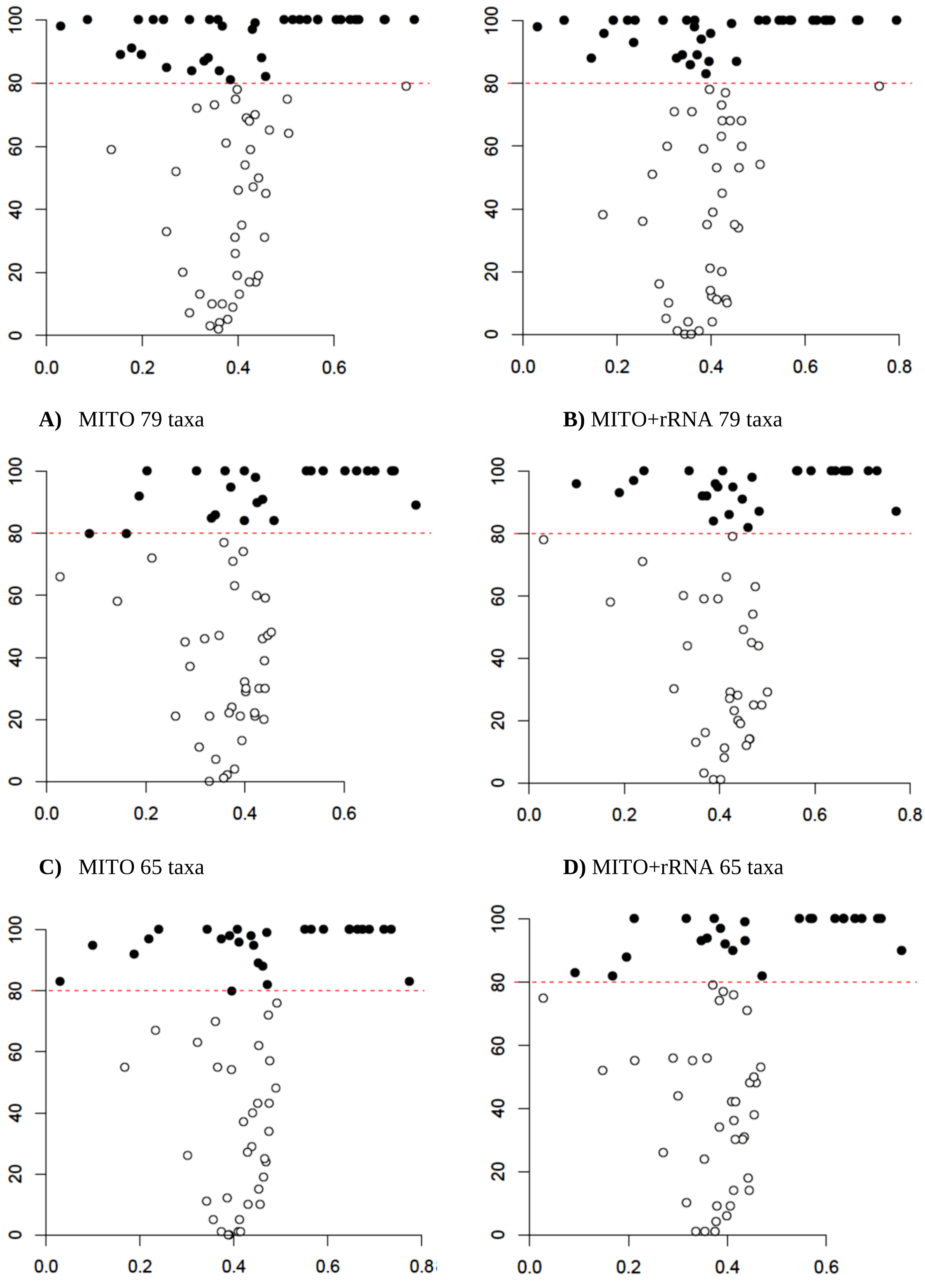
The Δ mean BS across consistent nodes in the paired analyses of MITO data alone and MITO+ other gene(s) are shown in Table 4.3, indicating that for the 79 taxa analysis, there was a marginal decrease in mean BS of <0.5%, whilst there were small increases in mean BS in all 65 taxa paired analyses (~2.5% for MITO+*Arg*K, ~3.0% for MITO+rRNA and ~5.5% for MITO+rRNA+*ArgK*). The mean BS support for only those nodes present across all four strict consensus trees are presented for each of the six datasets in Table 4.1, which reveals no appreciable gain in mean BS through the addition of rRNA sequences to the 79 taxa MITO dataset (mean BS = 85.08% for MITO+rRNA and 84.4% for MITO). Results for the 65 taxa dataset show that the greatest increase in mean BS is achieved through the addition of the rRNA sequences to the MITO sequences (mean BS = 84.9% for MITO+rRNA and 79.49% for MITO). The further addition of *ArgK* sequences to the MITO+rRNA dataset results in a slight reduction in mean BS (down to 84.38%). The addition of *ArgK* alone to the MITO data resulted in a modest increase of mean BS to 81.7%.

Reduction of the 79 taxa dataset to 77 taxa resulted in a drop in BS support for the nodes to which the two removed taxa were originally joined. Thus, nodal support for the most basal node in the trees, separating the Attelabidae as sister to the remainder of the Curculionoidea, was reduced from 98% to 68% BS. Likewise the next node along, separating Brentidae as sister to the remaining taxa dropped in BS support from 100% to 98%.

Graphs of nodal support versus nodal distance from the root of the tree are shown for each of the six ML ultrametric trees in Figure 4.2 A-F, indicating that whilst all analyses resulted in high nodal support (BS >80%) across a wide range of nodal ages, the analysis of the 79 taxa datasets resulted in noticeably greater BS support for the most basal nodes compared to the 65 taxa analyses, although with the addition of the rRNA and rRNA+*ArgK* to the latter, an observable improvement is apparent.

Table 4.3 Mean BS supports of the originating MITO and MITO + other gene(s) ML trees used in construction of the corresponding four strict consensus trees. Δ BS indicates mean BS of MITO + other gene(s) minus mean BS MITO. The 79 taxa analysis highlighted in grey.

Strict consensus tree composition	No. shared nodes	Mean BS of shared nodes in MITO tree	Mean BS of shared nodes in MITO+other gene(s) tree	Δ mean BS
MITO and MITO+rRNA 79 taxa	64	74.21875	73.75	-0.46875
MITO and MITO+rRNA 65 taxa	54	65.85185185	68.92592593	3.074074
MITO and MITO+rRNA+*ArgK* 65 taxa	48	70.75	76.27083333	5.520833
MITO and MITO+*ArgK* 65 taxa	49	70.93877551	73.40816327	2.469388



A) MITO 79 taxa

B) MITO+rRNA 79 taxa

C) MITO 65 taxa

D) MITO+rRNA 65 taxa

E) MITO+rRNA+*ArgK* 65 taxa

F) MITO+*ArgK* 65 taxa

Figure 4.2 Graph of RAxML nodal bootstrap support against branch length of nodes from the root for six datasets. BS values of 80% or greater are shown in black.

4.5 Discussion

4.5.1 Phylogenetic analyses and nodal support

Our results indicate that the addition of nuclear sequences to a dataset comprised of complete, or near-complete, mitochondrial genomes has little additive value, as measured by increase in mean BS support across shared nodes. It is also apparent that BS support is dependent upon the number of taxa in the data matrix, with generally higher mean BS values observed in the larger 79 taxa dataset (Tables 4.1 and 4.3). When comparing only nodes that are consistent across all the strict consensus trees, *i.e.* nodes that are well supported by all the datasets, it is clear that there is little benefit in adding more genes to the mitogenomic data. For these nodes, there is a marginal difference of 0.5% BS between the mean BS of the MITO 79 (84.4% BS) taxa dataset alone and that of the 'best' 65 taxa dataset – MITO+rRNA (84.9% BS) (Figure 4.1). If considering nodes shared only between pairs of MITO and MITO+ other gene(s) trees (*i.e.* some nodes not shared across all analyses), the MITO 79 taxa dataset alone resulted in a mean BS of 74.2% across nodes shared with the MITO+rRNA 79 taxa dataset; this being 8.37% higher than the MITO 65 taxa alone (65.85% BS), and is higher than all other analyses except for that of 65 taxa MITO+rRNA+*ArgK* (76.27% BS) across nodes shared with the MITO 65 taxa data (Table 4.3). This loss of BS support with decreased taxa is also supported by the analysis incorporating the reduced number of 77 taxa for the MITO+rRNA dataset, which also resulted in a loss of BS support for the basal nodes from which taxa were pruned. However further systematic tests with intermediate number of taxa between the minimum of 65 and maximum of 79 used here are necessary to investigate the extent to which this affects mean BS support.

Whilst it is difficult to isolate the effect of taxon coverage from that of additional characters on nodal support, the evidence presented here suggests that with increased taxon coverage, nodal support increases by a similar degree to how it does with the addition of more genes. This is seen in the increase in nodal support through the addition of rRNA to the MITO 65 taxa dataset, which results in very similar mean BS support to the MITO 79 taxa dataset alone. The further addition of *ArgK* to the 65 taxa data brings no further substantial improvement in BS support.

4.5.2 Practical implications for systematics

The mitogenomic dataset used here consists of 13792 aligned positions, which is considerably more data than that used by most phylogenetic analyses of the Curculionoidea to date (*e.g.* ~8000 bp in McKenna *et al.* 2009; ~2500 bp in Hundsdoerfer *et al.* 2009; ~10500 in Haran *et al.* 2013). The resulting topologies from the mitogenomic data alone contain well supported nodes across the full range of nodal ages and contain many nodes with high statistical support (BS 80-100%). That there is such good nodal support is indicative that independent data (different mitochondrial genes) are providing supporting signal and therefore that our confidence in the results can be justified.

The results presented here inevitably lead to one of the long-standing questions in systematics – should efforts be made to obtain more taxa, or more genes, for a given dataset? Which strategy is the most beneficial? In a practical sense, this trade off may be examined in the context of both options competing for limited resources.

The mean sequencing cost of obtaining a single long mitogenomic assembly through NGS for this thesis was approximately GBP £20.00 (*pers. comm.* Department of Biochemistry, University of Cambridge). Given the unequivocal bait matching success (see Chapter 3), a single Sanger-sequenced bait sequence can be used for assembly identification, yielding a per individual cost of approximately £26. It is clear from this that Sanger sequencing costs are a considerable portion of the total costs. The approximate cost of sequencing the additional three nuclear loci used here (*pers. comm.* The Natural History Museum molecular lab) raised the cost substantially, by approximately GBP £42 per sample (18S is sequenced in five separate reactions alone). The above calculations do not include PCR clean-up and relate to unidirectional sequencing. Nor do they consider possible lengthy and costly initial PCR optimisations. These additional factors could easily raise the values considerably above those indicated.

A further consideration is that often, despite considerable resources spent on PCR optimisations, some specimens/taxa prove to be very difficult or impossible to amplify with PCR (*e.g.* primer binding sites may be too divergent, or DNA may be of poor integrity), limiting the ability to generate sequences for additional markers. However, in theory, NGS of pooled genomic DNA, not being susceptible to PCR success rates, should be more reliable in being able to generate long mitogenomic sequences.

4.5.3 Conclusions

Strong arguments have been made that explicitly blame poor taxon sampling as one of the most important limiting factors to constructing meaningful higher-level relationships within the Curculionoidea (Franz & Engel 2010), precluding meaningful conclusions about the evolutionary history of this group. This view is easy to understand when one realises that most analyses to date have incorporated considerably less than 100 taxa and that there are almost 300 tribes and thousands of genera in the family Curculionidae alone, making it difficult to confidently test the monophyly of higher-level taxa.

Analyses for this chapter have broadly indicated that additional taxa result in similar gains in BS support as to those gained with additional nuclear markers. Given this, a suggestion can be made that, especially in highly diverse groups such as the Curculionoidea, where taxon sampling has been an obstacle to robust phylogenetic inferences, priority should be given to diverting limited resources towards increasing taxon coverage. Thorough taxon sampling was reported as being a very practical way to improve the accuracy of phylogeny reconstructions and, accordingly, inferences derived from them (Heath *et al.* 2008). However, Hillis *et al.* (2003) recognised that whether more taxa or characters is preferable will depend on the initial dataset and scope of the analysis, such that datasets already containing many taxa, but few characters, may benefit more from further addition of sequence data, and vice versa. An important consideration is that even if costs of increasing taxon coverage for NGS mitogenome assembly (*i.e.* additional DNA extractions and possibly fieldwork costs) are equal to the costs of sequencing additional loci, additional taxa have the advantage of benefiting nodal support, as shown here, whilst simultaneously also enriching lineage sampling.

Whilst taxon sampling itself, particular for specimens with well-preserved DNA, is nontrivial, the analyses undertaken here have indicated that greater benefit per unit cost can be gained in opting for enhanced taxon coverage over incorporating additional nuclear genes. This is likely to be generally applicable for phylogenetic studies investigating similar evolutionary depths and with similar sampling strategies. The opposing strategy of increasing sequence data with additional nuclear loci can be a time consuming and costly undertaking requiring much additional work. Technology now exists to reliably, cheaply and quickly generate mitogenomes, and analysis of these have been shown to result in highly satisfactory hypotheses of relationships across many taxa (Finstermeier *et al.* 2013; Wei *et al.* 2010). It therefore seems desirable at present, and with limited resources, that preference should be made for enhanced taxon sampling over increased marker generation for mitogenomic analyses.

4.6 References

Alonso-Zarazaga MA, Lyal CHC (1999) *A world catalogue of families and genera of Curculionoidea (Insecta: Coleoptera) (excepting Scolytidae and Platypodidae)* Entomopraxis, Barcelona.

Avise JC (2004) *Molecular markers, natural history and evolution*, Second edn. Sinauer Associates, Massachusetts.

Bocak L, Barton C, Crampton-Platt A*, et al.* (2013) Building the Coleoptera tree-of-life for >8000 species: composition of public DNA data and fit with Linnaean classification. *Systematic Entomology* **39**, 97-110.

Bouchard P, Bousquet Y, Davies AE*, et al.* (2011) Family-group names in Coleoptera (Insecta). *Zookeys* **88**, 1-972.

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

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

Efron B, Halloran E, Holmes S (1996) Bootstrap confidence levels for phylogenetic trees (vol 93, pg 7085, 1996). *Proceedings of the National Academy of Sciences of the United States of America* **93**, 13429-13434.

Felsenstein J (1985) Confidence-limits on phylogenies - an approach using the bootstrap. *Evolution* **39**, 783-791.

Finstermeier K, Zinner D, Brameier M*, et al.* (2013) A mitogenomic phylogeny of living primates. *PloS one* **8**: 10.1371/journal.pone.0069504.

Fisher-Reid MC, Wiens JJ (2011) What are the consequences of combining nuclear and mitochondrial data for phylogenetic analysis? Lessons from Plethodon salamanders and 13 other vertebrate clades. *BMC Evolutionary Biology* **11**: 10.1186/1471-2148-11-300.

Franz NM, Engel MS (2010) Can higher-level phylogenies of weevils explain their evolutionary success? A critical review. *Systematic Entomology* **35**, 597-606.

Haran J, Timmermans MJTN, Vogler AP (2013) Mitogenome sequences stabilize the phylogenetics of weevils (Curculionoidea) and establish the monophyly of larval ectophagy. *Molecular Phylogenetics and Evolution* **67**, 156-166.

Heath TA, Hedtke SM, Hillis DM (2008) Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics and Evolution* **46**, 239-257.

Hillis DM, Pollock DD, McGuire JA, Zwickl DJ (2003) Is sparse taxon sampling a problem for phylogenetic inference? *Systematic Biology* **52**, 124-126.

Hundsdoerfer AK, Rheinheimer J, Wink M (2009) Towards the phylogeny of the Curculionoidea (Coleoptera): Reconstructions from mitochondrial and nuclear ribosomal DNA sequences. *Zoologischer Anzeiger* **248**, 9-31.

Janke A, Magnell O, Wieczorek G, Westerman M, Arnason U (2002) Phylogenetic analysis of 18S rRNA and the mitochondrial genomes of the wombat, Vombatus ursinus, and the spiny anteater, Tachyglossus aculeatus: Increased support for the Marsupionta hypothesis. *Journal of Molecular Evolution* **54**, 71-80.

Johnson KP, Clayton DH (2000) Nuclear and mitochondrial genes contain similar phylogenetic signal for pigeons and doves (Aves : Columbiformes). *Molecular Phylogenetics and Evolution* **14**, 141-151.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066.

Kayal E, Roure B, Philippe H, Collins AG, Lavrov DV (2013) Cnidarian phylogenetic relationships as revealed by mitogenomics. *BMC Evolutionary Biology* **13**: 10.1186/1471-2148-13-5.

Leigh JW, Lapointe F-J, Lopez P, Bapteste E (2011) Evaluating Phylogenetic Congruence in the Post-Genomic Era. *Genome Biology and Evolution* **3**, 571-587.

Marvaldi AE, Duckett CN, Kjer KM, Gillespie JJ (2009) Structural alignment of 18S and 28S rDNA sequences provides insights into phylogeny of Phytophaga (Coleoptera: Curculionoidea and Chrysomeloidea). *Zoologica Scripta* **38**, 63-77.

McKenna DD, Sequeira AS, Marvaldi AE, Farrell BD (2009) Temporal lags and overlap in the diversification of weevils and flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7083-7088.

Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: *Proceedings of the gateway computing environments workshop (GCE), 14 Nov. 2010*, New Orleans, LA.

Monaghan MT, Inward DJG, Hunt T, Vogler AP (2007) A molecular phylogenetic analysis of the Scarabaeinae (dung beetles). *Molecular Phylogenetics and Evolution* **45**, 674-692.

Osigus H-J, Eitel M, Bernt M, Donath A, Schierwater B (2013) Mitogenomics at the base of Metazoa. *Molecular Phylogenetics and Evolution* **69**, 339-351.

Page RDM, Holmes EC (1998) *Molecular Evolution: A phylogenetic approach* Blackwell, Oxford.

Paradis E (2013) Molecular dating of phylogenies by likelihood methods: A comparison of models and a new information criterion. *Molecular Phylogenetics and Evolution* **67**, 436-444.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289-290.

Rubinoff D, Holland BS (2005) Between two extremes: Mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. *Systematic Biology* **54**, 952-961.

San Mauro D, Gower DJ, Oommen OV, Wilkinson M, Zardoya R (2004) Phylogeny of caecilian amphibians (Gymnophiona) based on complete mitochondrial genomes and nuclear RAG1. *Molecular Phylogenetics and Evolution* **33**, 413-427.

Shull VL, Vogler AP, Baker MD, Maddison DR, Hammond PM (2001) Sequence alignment of 18S ribosomal RNA and the basal relationships of Adephagan beetles: evidence for monophyly of aquatic families and the placement of Trachypachidae. *Systematic Biology* **50**, 945-969.

Sota T, Vogler AP (2001) Incongruence of mitochondrial and nuclear gene trees in the Carabid beetles Ohomopterus. *Systematic Biology* **50**, 39-59.

Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690.

Wei S-j, Shi M, Sharkey MJ, van Achterberg C, Chen X-x (2010) Comparative mitogenomics of Braconidae (Insecta: Hymenoptera) and the phylogenetic utility of mitochondrial genomes with special reference to Holometabolous insects. *BMC Genomics* **11**: 10.1186/1471-2164-1111-1371.

Wild AL, Maddison DR (2008) Evaluating nuclear protein-coding genes for phylogenetic utility in beetles. *Molecular Phylogenetics and Evolution* **48**, 877-891.

4.7 Appendices

Appendix 4.1 R script for calculating and plotting strict consensus trees from a set of input ML trees.

```
### Clear the workspace
rm(list=ls())

### set working directory
setwd("C:/Users/Conrad P.D. Gillett/Documents/Curculionidae/Chapter 4/Final analyses/Trees for
Consensus building")

### get library
library(ape)

### read in trees
trees<-read.nexus("MITO_MITO18S28S_65_2TREES.nex")

### make different consensuses
strictcons<-consensus(trees,p=1,check.labels=TRUE)
###majcons<-consensus(trees,p=0.5,check.labels=TRUE)

### root them
root(strictcons,"CG336")
###root(majcons,"CG336")

write.nexus(strictcons,file="consesnsus.nex")

### plot them
plot(strictcons)
###plot(majcons)
```

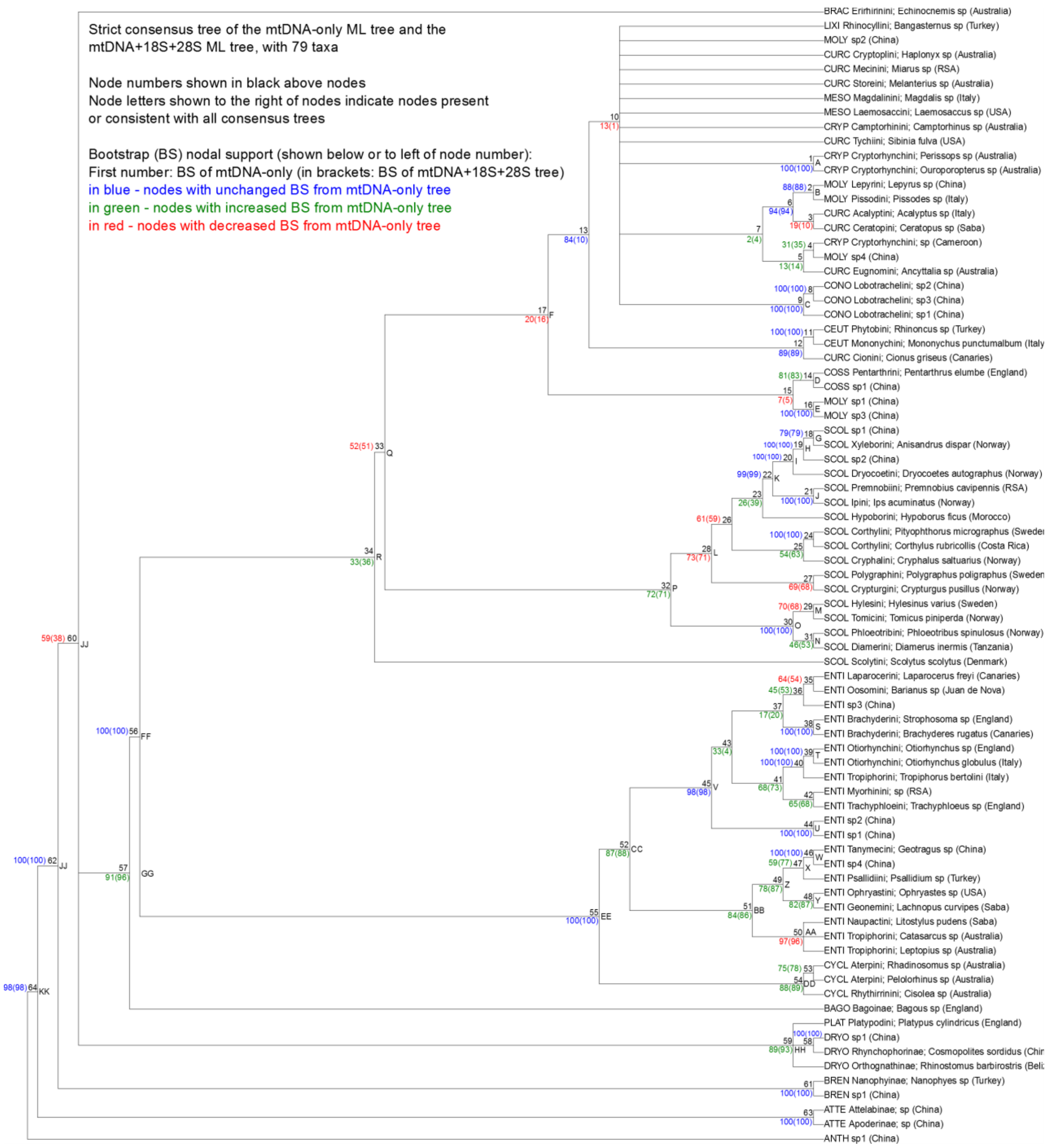
Appendix 4.2 The 79 Curculionoidea taxa identified to species or higher-level group used in this study. Highlighted in grey are 14 taxa for which only *ArgK* sequences were not generated. All other taxa alignments contained mtDNA, 18S rRNA, 28S rRNA and *ArgK* sequences

Family	Subfamily	Tribe	Genus	Species	Origin	Source	Code	No. of genes
Anthribidae				*sp. 1*	China	Gillett/Lyal	CG336	18
Attelabidae	Apoderinae			*sp. 2*	China	Gillett/Lyal	CG335	18
Attelabidae	Attelabinae			*sp. 1*	China	Gillett/Lyal	CG323	17
Brachyceridae	Erirhininae	Erirhirinini	*Echinocnemis*	*sp.*	Australia	Oberprieler	CG210	18
Brentidae	Nanophyinae	Nanophyini	*Nanophyes*	*sp.*	Turkey	Levent	CG271	16
Brentidae				*sp. 1*	China	Gillett/Lyal	CG347	18
Curculionidae	Bagoinae		*Bagous*	*sp.*	England	Turner	CG220	18
Curculionidae	Ceutorhynchinae	Mononychini	*Mononychus*	*punctumalbum*	Italy	Caldara	CG306	14
Curculionidae	Ceutorhynchinae	Phytobini	*Rhinoncus*	*sp.*	Turkey	Levent	CG282	18
Curculionidae	Conoderinae	Lobotrachelini		*sp. 1*	China	Gillett/Lyal	CG321	18
Curculionidae	Conoderinae	Lobotrachelini		*sp. 2*	China	Gillett/Lyal	CG322	18
Curculionidae	Conoderinae	Lobotrachelini		*sp. 3*	China	Gillett/Lyal	CG328	17
Curculionidae	Cossoninae	Pentarthrini	*Pentarthrus*	*elumbe*	England	Turner	CG222	18
Curculionidae	Cossoninae			*sp. 1*	China	Gillett/Lyal	CG319	18
Curculionidae	Cryptorhynchinae	Camptorhinini	*Camptorhinus*	*sp.*	Australia	Oberprieler	CG253	18
Curculionidae	Cryptorhynchinae	Cryptorhynchini	*Ouroporopterus*	*sp.*	Australia	Oberprieler	CG240	18
Curculionidae	Cryptorhynchinae	Cryptorhynchini	*Perissops*	*sp.*	Australia	Oberprieler	CG238	18
Curculionidae	Cryptorhynchinae	Cryptorhynchini		*sp.*	Cameroon	Jordal	CG415	11
Curculionidae	Curculioninae	Acalyptini	*Acalyptus*	*sp.*	Italy	Caldara	CG052	17
Curculionidae	Curculioninae	Ceratopini	*Ceratopus*	*sp.*	Saba	Gillett M	CG351	15
Curculionidae	Curculioninae	Cionini	*Cionus*	*griseus*	Canaries	Oromi	CG293	18
Curculionidae	Curculioninae	Cryptoplini	*Haplonyx*	*sp.*	Australia	Oberprieler	CG235	18
Curculionidae	Curculioninae	Eugnomini	*Ancyttalia*	*sp.*	Australia	Oberprieler	CG242	18

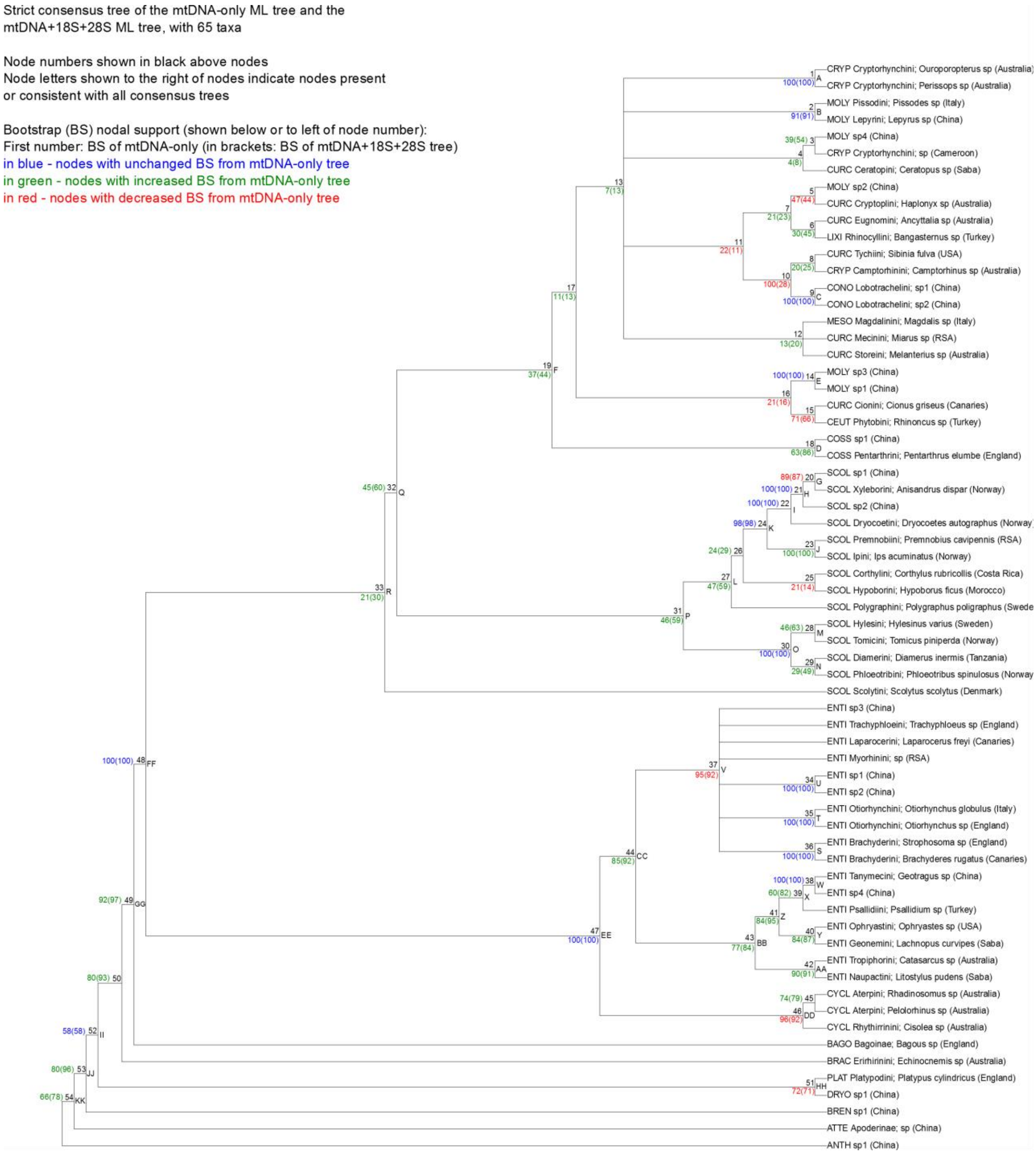
Curculionidae	Curculioninae	Mecinini	*Miarus*	*sp.*	RSA	Meregalli	CG284	18
Curculionidae	Curculioninae	Storeini	*Melanterius*	*sp.*	Australia	Oberprieler	CG257	18
Curculionidae	Curculioninae	Tychiini	*Sibinia*	*fulva*	USA	O'Brien	CG298	17
Curculionidae	Cyclominae	Aterpini	*Pelolorhinus*	*sp.*	Australia	Oberprieler	CG247	17
Curculionidae	Cyclominae	Aterpini	*Rhadinosomus*	*sp.*	Australia	Oberprieler	CG229	17
Curculionidae	Entiminae	Brachyderini	*Brachyderes*	*rugatus*	Canaries	Emerson	N28	16
Curculionidae	Entiminae	Geonemini	*Lachnopus*	*curvipes*	Saba	Gillett M	CG354	18
Curculionidae	Entiminae	Laparocerini	*Laparocerus*	*freyi*	Canaries	Faria	LAP007	16
Curculionidae	Entiminae	Myorhinini		*sp.*	RSA	Meregalli	CG285	18
Curculionidae	Entiminae	Naupactini	*Litostylus*	*pudens*	Saba	Gillett M	CG355	18
Curculionidae	Entiminae	Oosomini	*Barianus*	*sp.*	Juan de Nova	Kitson	CG305	17
Curculionidae	Entiminae	Ophryastini	*Ophryastes*	*sp.*	USA	O'Brien	CG297	18
Curculionidae	Entiminae	Otiorhynchini	*Otiorhynchus*	*globulus*	Italy	Caldara	CG309	16
Curculionidae	Entiminae	Otiorhynchini	*Otiorhynchus*	*sp.*	England	Gillett	CG307	18
Curculionidae	Entiminae	Psallidiini	*Psallidium*	*sp.*	Turkey	Levent	CG272	18
Curculionidae	Entiminae	Tanymecini	*Geotragus*	*sp.*	China	Li	CG311	18
Curculionidae	Entiminae	Trachyphloeini	*Trachyphloeus*	*sp.*	England	Turner	CG301	18
Curculionidae	Entiminae	Tropiphorini	*Catasarcus*	*sp.*	Australia	Oberprieler	CG227	12
Curculionidae	Entiminae	Tropiphorini	*Leptopius*	*sp.*	Australia	Oberprieler	CG341	16
Curculionidae	Entiminae	Tropiphorini	*Tropiphorus*	*bertolini*	Italy	Caldara	CG315	17
Curculionidae	Entiminae			*sp. 1*	China	Gillett/Lyal	CG330	18
Curculionidae	Entiminae			*sp. 2*	China	Gillett/Lyal	CG331	18
Curculionidae	Entiminae			*sp. 3*	China	Gillett/Lyal	CG339	18
Curculionidae	Entiminae			*sp. 4*	China	Gillett/Lyal	CG342	18
Curculionidae	Lixinae	Rhinocyllini	*Bangasternus*	*sp.*	Turkey	Levent	CG268	18
Curculionidae	Mesoptiliinae	Laemosaccini	*Laemosaccus*	*sp.*	USA	O'Brien	CG296	17
Curculionidae	Mesoptiliinae	Magdalinini	*Magdalis*	*sp.*	Italy	Caldara	CG069	18
Curculionidae	Molytinae	Lepyrini	*Lepyrus*	*sp.*	China	Li	CG312	18
Curculionidae	Molytinae	Pissodini	*Pissodes*	*sp.*	Italy	Caldara	CG055	18
Curculionidae	Molytinae			*sp. 1*	China	Gillett/Lyal	CG317	17
Curculionidae	Molytinae			*sp. 2*	China	Gillett/Lyal	CG332	18
Curculionidae	Molytinae			*sp. 3*	China	Gillett/Lyal	CG340	18
Curculionidae	Molytinae			*sp. 4*	China	Gillett/Lyal	CG350	18
Curculionidae	Platypodinae	Platypodini	*Platypus*	*cylindricus*	England	Turner	CG221	18

Curculionidae	Scolytinae	Corthylini	*Corthylus*	*rubricollis*	Costa Rica	Jordal	CG435	17
Curculionidae	Scolytinae	Corthylini	*Pityophthorus*	*micrographus*	Sweden	Jordal	CG436	17
Curculionidae	Scolytinae	Cryphalini	*Cryphalus*	*saltuarius*	Norway	Jordal	CG437	17
Curculionidae	Scolytinae	Crypturgini	*Crypturgus*	*pusillus*	Norway	Jordal	CG434	17
Curculionidae	Scolytinae	Hylesini	*Hylesinus*	*varius*	Sweden	Jordal	CG424	18
Curculionidae	Scolytinae	Hypoborini	*Hypoborus*	*ficus*	Morocco	Jordal	CG439	18
Curculionidae	Scolytinae	Ipini	*Ips*	*acuminatus*	Norway	Jordal	CG426	18
Curculionidae	Scolytinae	Phloeotribini	*Phloeotribus*	*sp. inulosus*	Norway	Jordal	CG442	18
Curculionidae	Scolytinae	Polygraphini	*Polygraphus*	*poligraphus*	Sweden	Jordal	CG441	18
Curculionidae	Scolytinae	Premnobiini	*Premnobius*	*cavipennis*	RSA	Jordal	CG428	18
Curculionidae	Scolytinae	Scolytini	*Scolytus*	*scolytus*	Denmark	Jordal	CG429	17
Curculionidae	Scolytinae	Tomicini	*Tomicus*	*piniperda*	Norway	Jordal	CG425	18
Curculionidae	Scolytinae	Xyleborini	*Anisandrus*	*dispar*	Norway	Jordal	CG431	18
Curculionidae	Scolytinae			*sp. 1*	China	Gillett/Lyal	CG325	18
Curculionidae	Scolytinae			*sp. 2*	China	Gillett/Lyal	CG346	18
Dryophthoridae	Orthognathinae	Rhinostomini	*Rhinostomus*	*barbirostris*	Belize	Barclay	CG074	17
Dryophthoridae	Rhynchophorinae	sp. henophorini	*Cosmopolites*	*sordidus*	China	Gillett/Lyal	CG344	17
Dryophthoridae				*sp. 1*	China	Gillett/Lyal	CG324	18

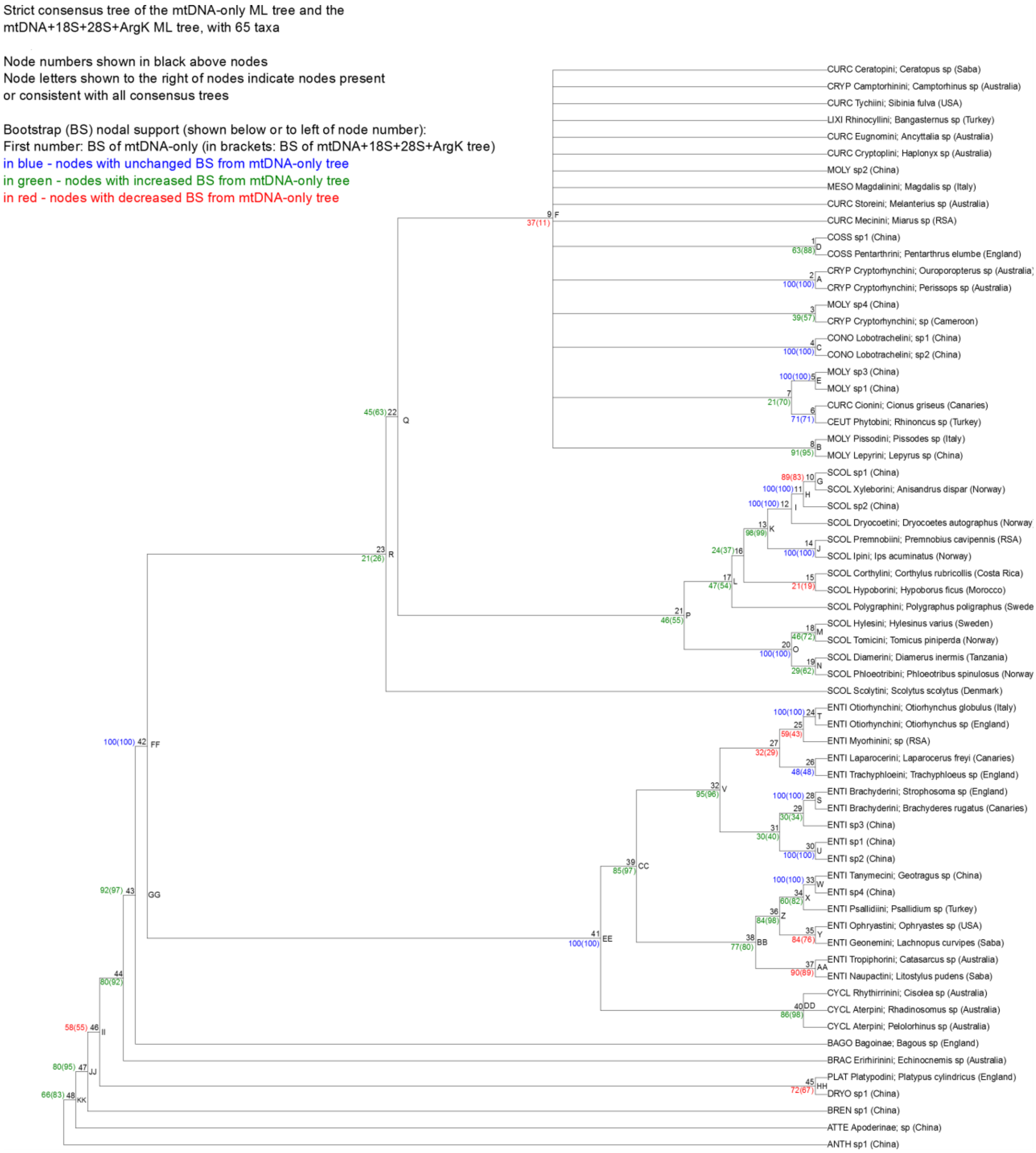
Appendix 4.3 Strict consensus tree of the MITO ML tree and the MITO+rRNA ML tree with 79 taxa. Family and subfamily codes precede taxa names as follows: Anthribidae (ANTH), Attelabidae (ATTE), Brachyceridae (BRAC), Brentidae (BREN), Dryophthoridae (DRYO), Nemonychidae (NEMO), Bagoinae (BAGO), Baridinae (BARI), Ceutorhynchinae (CEUT), Conoderinae (CONO), Cossoninae (COSS), Cryptorhynchinae (CRYP), Curculioninae (CURC), Lixinae (LIXI), Mesoptillinae (MESO), Molytinae (MOLY), Platypodinae (PLAT) and Scolytinae (SCOL).


Strict consensus tree of the mtDNA-only ML tree and the mtDNA+18S+28S ML tree, with 79 taxa
Node numbers shown in black above nodes
Node letters shown to the right of nodes indicate nodes present or consistent with all consensus trees
Bootstrap (BS) nodal support (shown below or to left of node number):
First number: BS of mtDNA-only (in brackets: BS of mtDNA+18S+28S tree)
in blue - nodes with unchanged BS from mtDNA-only tree
in green - nodes with increased BS from mtDNA-only tree
in red - nodes with decreased BS from mtDNA-only tree
BRAC Erirhinini; Echinocnemis sp (Australia)
LIXI Rhinocyllini; Bangasternus sp (Turkey)
MOLY sp2 (China)
CURC Cryptoplini; Haplonyx sp (Australia)
CURC Mecinini; Miarus sp (RSA)
CURC Storeini; Melanterius sp (Australia)
MESO Magdalinini; Magdalis sp (Italy)
MESO Laemosaccini; Laemosaccus sp (USA)
CRYP Camptorhinini; Camptorhinus sp (Australia)
CURC Tychiini; Sibinia fulva (USA)
CRYP Cryptorhynchini; Perissops sp (Australia)
CRYP Cryptorhynchini; Ouroporopterus sp (Australia)
MOLY Lepyrini; Lepyrus sp (China)
MOLY Pissodini; Pissodes sp (Italy)
CURC Acalyptini; Acalyptus sp (Italy)
CURC Ceratopini; Ceratopus sp (Saba)
CRYP Cryptorhynchini; sp (Cameroon)
MOLY sp4 (China)
CURC Eugnomini; Ancyttalia sp (Australia)
CONO Lobotrachelini; sp2 (China)
CONO Lobotrachelini; sp3 (China)
CONO Lobotrachelini; sp1 (China)
CEUT Phytobini; Rhinoncus sp (Turkey)
CEUT Mononychini; Mononychus punctumalbum (Italy
CURC Cionini; Cionus griseus (Canaries)
COSS Pentarthrini; Pentarthrus elumbe (England)
COSS sp1 (China)
MOLY sp1 (China)
MOLY sp3 (China)
SCOL sp1 (China)
SCOL Xyleborini; Anisandrus dispar (Norway)
SCOL sp2 (China)
SCOL Dryocoetini; Dryocoetes autographus (Norway)
SCOL Premnobiini; Premnobius cavipennis (RSA)
SCOL Ipini; Ips acuminatus (Norway)
SCOL Hypoborini; Hypoborus ficus (Morocco)
SCOL Corthylini; Pityophthorus micrographus (Swede
SCOL Corthylini; Corthylus rubricollis (Costa Rica)
SCOL Cryphalini; Cryphalus saltuarius (Norway)
SCOL Polygraphini; Polygraphus poligraphus (Sweden
SCOL Crypturgini; Crypturgus pusillus (Norway)
SCOL Hylesini; Hylesinus varius (Sweden)
SCOL Tomicini; Tomicus piniperda (Norway)
SCOL Phloeotribini; Phloeotribus spinulosus (Norway)
SCOL Diamerini; Diamerus inermis (Tanzania)
SCOL Scolytini; Scolytus scolytus (Denmark)
ENTI Laparocerini; Laparocerus freyi (Canaries)
ENTI Oosomini; Barianus sp (Juan de Nova)
ENTI sp3 (China)
ENTI Brachyderini; Strophosoma sp (England)
ENTI Brachyderini; Brachyderes rugatus (Canaries)
ENTI Otiorhynchini; Otiorhynchus sp (England)
ENTI Otiorhynchini; Otiorhynchus globulus (Italy)
ENTI Tropiphorini; Tropiphorus bertolini (Italy)
ENTI Myorhinini; sp (RSA)
ENTI Trachyphloeini; Trachyphloeus sp (England)
ENTI sp2 (China)
ENTI sp1 (China)
ENTI Tanymecini; Geotragus sp (China)
ENTI sp4 (China)
ENTI Psallidiini; Psallidium sp (Turkey)
ENTI Ophryastini; Ophryastes sp (USA)
ENTI Geonemini; Lachnopus curvipes (Saba)
ENTI Naupactini; Litostylus pudens (Saba)
ENTI Tropiphorini; Catasarcus sp (Australia)
ENTI Tropiphorini; Leptopius sp (Australia)
CYCL Aterpini; Rhadinosomus sp (Australia)
CYCL Aterpini; Pelolorhinus sp (Australia)
CYCL Rhythirrinini; Cisolea sp (Australia)
BAGO Bagoinae; Bagous sp (England)
PLAT Platypodini; Platypus cylindricus (England)
DRYO sp1 (China)
DRYO Rhynchophorinae; Cosmopolites sordidus (Chin
DRYO Orthognathinae; Rhinostomus barbirostris (Beli
BREN Nanophyinae; Nanophyes sp (Turkey)
BREN sp1 (China)
ATTE Attelabinae; sp (China)
ATTE Apoderinae; sp (China)
ANTH sp1 (China)
1 A 100(100)
88(88) 2 B
6 94(94)
3 19(10)
7 2(4)
31(35) 4
5 13(14)
100(100) 8
9 C 100(100)
10 13(1)
100(100) 11
12 89(89)
13 84(10)
81(83) 14 D
15 7(5)
16 E 100(100)
17 F 20(16)
79(79) 18 G
100(100) 19 H
100(100) 20 I
21 J 100(100)
99(99) 22 K
23 26(39)
100(100) 24
25 54(63)
61(59) 26
27 69(68)
28 L 73(71)
70(68) 29 M
30 O 100(100)
31 N 46(53)
32 P 72(71)
52(51) 33 Q
34 R 33(36)
64(54) 35
45(53) 36
37 17(20)
38 S 100(100)
100(100) 39 T
100(100) 40
41 68(73)
42 65(68)
43 33(4)
44 U 100(100)
45 V 98(98)
100(100) 46 W
59(77) 47 X
48 Y 82(87)
49 Z 78(87)
50 AA 97(96)
51 BB 84(86)
52 CC 87(88)
75(78) 53
54 DD 88(89)
55 EE 100(100)
100(100) 56 FF
57 GG 91(96)
100(100) 58
59 HH 89(93)
59(38) 60 JJ
61 100(100)
100(100) 62 JJ
63 100(100)
98(98) 64 KK

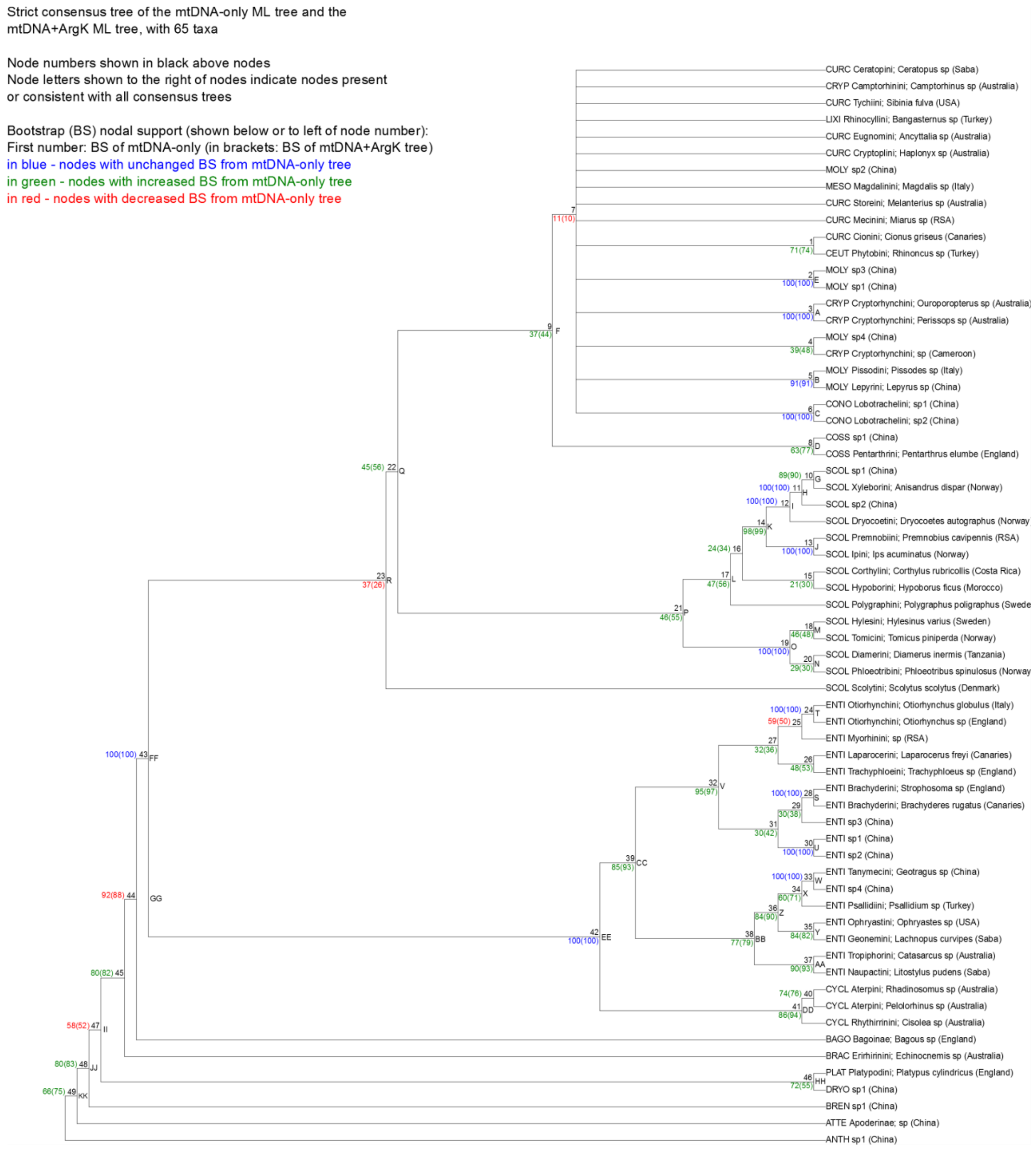
Appendix 4.4 Strict consensus tree of the MITO ML tree and the MITO+rRNA ML tree with 65 taxa. Family and subfamily codes precede taxa names as for Appendix 4.2.


Strict consensus tree of the mtDNA-only ML tree and the mtDNA+18S+28S ML tree, with 65 taxa
Node numbers shown in black above nodes
Node letters shown to the right of nodes indicate nodes present or consistent with all consensus trees
Bootstrap (BS) nodal support (shown below or to left of node number):
First number: BS of mtDNA-only (in brackets: BS of mtDNA+18S+28S tree)
in blue - nodes with unchanged BS from mtDNA-only tree
in green - nodes with increased BS from mtDNA-only tree
in red - nodes with decreased BS from mtDNA-only tree
1 100(100) A
CRYP Cryptorhynchini; Ouroporopterus sp (Australia)
CRYP Cryptorhynchini; Perissops sp (Australia)
2 91(91) B
MOLY Pissodini; Pissodes sp (Italy)
MOLY Lepyrini; Lepyrus sp (China)
39(54) 3
4 4(8)
MOLY sp4 (China)
CRYP Cryptorhynchini; sp (Cameroon)
CURC Ceratopini; Ceratopus sp (Saba)
13 7(13)
5 47(44)
7 21(23)
6 30(45)
MOLY sp2 (China)
CURC Cryptoplini; Haplonyx sp (Australia)
CURC Eugnomini; Ancyttalia sp (Australia)
LIXI Rhinocyllini; Bangasternus sp (Turkey)
11 22(11)
8 20(25)
10 100(28)
9 100(100) C
CURC Tychiini; Sibinia fulva (USA)
CRYP Camptorhinini; Camptorhinus sp (Australia)
CONO Lobotrachelini; sp1 (China)
CONO Lobotrachelini; sp2 (China)
17 11(13)
12 13(20)
MESO Magdalinini; Magdalis sp (Italy)
CURC Mecinini; Miarus sp (RSA)
CURC Storeini; Melanterius sp (Australia)
19 37(44) F
100(100) 14 E
16 21(16)
15 71(66)
MOLY sp3 (China)
MOLY sp1 (China)
CURC Cionini; Cionus griseus (Canaries)
CEUT Phytobini; Rhinoncus sp (Turkey)
18 63(86) D
COSS sp1 (China)
COSS Pentarthrini; Pentarthrus elumbe (England)
45(60) 32 Q
89(87) 20 G
100(100) 21 H
100(100) 22 I
98(98) 24 K
SCOL sp1 (China)
SCOL Xyleborini; Anisandrus dispar (Norway)
SCOL sp2 (China)
SCOL Dryocoetini; Dryocoetes autographus (Norway
24(29) 26
23 100(100) J
SCOL Premnobiini; Premnobius cavipennis (RSA)
SCOL Ipini; Ips acuminatus (Norway)
27 47(59) L
25 21(14)
SCOL Corthylini; Corthylus rubricollis (Costa Rica)
SCOL Hypoborini; Hypoborus ficus (Morocco)
33 21(30) R
SCOL Polygraphini; Polygraphus poligraphus (Swede
31 46(59) P
46(63) 28 M
30 100(100) O
29 29(49) N
SCOL Hylesini; Hylesinus varius (Sweden)
SCOL Tomicini; Tomicus piniperda (Norway)
SCOL Diamerini; Diamerus inermis (Tanzania)
SCOL Phloeotribini; Phloeotribus spinulosus (Norway
SCOL Scolytini; Scolytus scolytus (Denmark)
ENTI sp3 (China)
ENTI Trachyphloeini; Trachyphloeus sp (England)
ENTI Laparocerini; Laparocerus freyi (Canaries)
ENTI Myorhinini; sp (RSA)
100(100) 48 FF
37 95(92) V
34 100(100) U
ENTI sp1 (China)
ENTI sp2 (China)
35 100(100) T
ENTI Otiorhynchini; Otiorhynchus globulus (Italy)
ENTI Otiorhynchini; Otiorhynchus sp (England)
36 100(100) S
ENTI Brachyderini; Strophosoma sp (England)
ENTI Brachyderini; Brachyderes rugatus (Canaries)
44 85(92) CC
100(100) 38 W
60(82) 39 X
41 84(95) Z
ENTI Tanymecini; Geotragus sp (China)
ENTI sp4 (China)
ENTI Psallidiini; Psallidium sp (Turkey)
40 84(87) Y
ENTI Ophryastini; Ophryastes sp (USA)
ENTI Geonemini; Lachnopus curvipes (Saba)
92(97) 49 GG
47 100(100) EE
43 77(84) BB
42 90(91) AA
ENTI Tropiphorini; Catasarcus sp (Australia)
ENTI Naupactini; Litostylus pudens (Saba)
80(93) 50
74(79) 45
46 96(92) DD
CYCL Aterpini; Rhadinosomus sp (Australia)
CYCL Aterpini; Pelolorhinus sp (Australia)
CYCL Rhythirrinini; Cisolea sp (Australia)
58(58) 52 II
BAGO Bagoinae; Bagous sp (England)
80(96) 53 JJ
BRAC Erirhinini; Echinocnemis sp (Australia)
51 72(71) HH
66(78) 54 KK
PLAT Platypodini; Platypus cylindricus (England)
DRYO sp1 (China)
BREN sp1 (China)
ATTE Apoderinae; sp (China)
ANTH sp1 (China)

Appendix 4.5 Strict consensus tree of the MITO ML tree and the MITO+rRNA+ArgK ML tree with 65 taxa. Family and subfamily codes precede taxa names as for


Strict consensus tree of the mtDNA-only ML tree and the
mtDNA+18S+28S+ArgK ML tree, with 65 taxa
Node numbers shown in black above nodes
Node letters shown to the right of nodes indicate nodes present
or consistent with all consensus trees
Bootstrap (BS) nodal support (shown below or to left of node number):
First number: BS of mtDNA-only (in brackets: BS of mtDNA+18S+28S+ArgK tree)
in blue - nodes with unchanged BS from mtDNA-only tree
in green - nodes with increased BS from mtDNA-only tree
in red - nodes with decreased BS from mtDNA-only tree
CURC Ceratopini; Ceratopus sp (Saba)
CRYP Camptorhinini; Camptorhinus sp (Australia)
CURC Tychiini; Sibinia fulva (USA)
LIXI Rhinocyllini; Bangasternus sp (Turkey)
CURC Eugnomini; Ancyttalia sp (Australia)
CURC Cryptoplini; Haplonyx sp (Australia)
MOLY sp2 (China)
MESO Magdalinini; Magdalis sp (Italy)
CURC Storeini; Melanterius sp (Australia)
9 F
37(11)
CURC Mecinini; Miarus sp (RSA)
1 D 63(88)
COSS sp1 (China)
COSS Pentarthrini; Pentarthrus elumbe (England)
2 A 100(100)
CRYP Cryptorhynchini; Ouroporopterus sp (Australia
CRYP Cryptorhynchini; Perissops sp (Australia)
3 39(57)
MOLY sp4 (China)
CRYP Cryptorhynchini; sp (Cameroon)
4 C 100(100)
CONO Lobotrachelini; sp1 (China)
CONO Lobotrachelini; sp2 (China)
100(100) 5 E
MOLY sp3 (China)
MOLY sp1 (China)
7 21(70)
6 71(71)
CURC Cionini; Cionus griseus (Canaries)
CEUT Phytobini; Rhinoncus sp (Turkey)
45(63) 22 Q
8 B 91(95)
MOLY Pissodini; Pissodes sp (Italy)
MOLY Lepyrini; Lepyrus sp (China)
89(83) 10 G
SCOL sp1 (China)
100(100) 11 H
SCOL Xyleborini; Anisandrus dispar (Norway)
100(100) 12 I
SCOL sp2 (China)
SCOL Dryocoetini; Dryocoetes autographus (Norway
13 K 98(99)
SCOL Premnobiini; Premnobius cavipennis (RSA)
24(37) 16
14 J 100(100)
SCOL Ipini; Ips acuminatus (Norway)
23 R 21(26)
17 L 47(54)
15 21(19)
SCOL Corthylini; Corthylus rubricollis (Costa Rica)
SCOL Hypoborini; Hypoborus ficus (Morocco)
21 P 46(55)
SCOL Polygraphini; Polygraphus poligraphus (Swede
18 M 46(72)
SCOL Hylesini; Hylesinus varius (Sweden)
20 O 100(100)
SCOL Tomicini; Tomicus piniperda (Norway)
19 N 29(62)
SCOL Diamerini; Diamerus inermis (Tanzania)
SCOL Phloeotribini; Phloeotribus spinulosus (Norway
SCOL Scolytini; Scolytus scolytus (Denmark)
100(100) 24 T
ENTI Otiorhynchini; Otiorhynchus globulus (Italy)
100(100) 42 FF
25 59(43)
ENTI Otiorhynchini; Otiorhynchus sp (England)
27 32(29)
ENTI Myorhinini; sp (RSA)
26 48(48)
ENTI Laparocerini; Laparocerus freyi (Canaries)
ENTI Trachyphloeini; Trachyphloeus sp (England)
32 V 95(96)
100(100) 28 S
ENTI Brachyderini; Strophosoma sp (England)
29 30(34)
ENTI Brachyderini; Brachyderes rugatus (Canaries)
31 30(40)
ENTI sp3 (China)
30 U 100(100)
ENTI sp1 (China)
ENTI sp2 (China)
39 CC 85(97)
100(100) 33 W
ENTI Tanymecini; Geotragus sp (China)
92(97) 43 GG
34 X 60(82)
ENTI sp4 (China)
36 Z 84(98)
ENTI Psallidiini; Psallidium sp (Turkey)
35 Y 84(76)
ENTI Ophryastini; Ophryastes sp (USA)
41 EE 100(100)
38 BB 77(80)
ENTI Geonemini; Lachnopus curvipes (Saba)
37 AA 90(89)
ENTI Tropiphorini; Catasarcus sp (Australia)
ENTI Naupactini; Litostylus pudens (Saba)
44 80(92)
CYCL Rhythirrinini; Cisolea sp (Australia)
40 DD 86(98)
CYCL Aterpini; Rhadinosomus sp (Australia)
CYCL Aterpini; Pelolorhinus sp (Australia)
58(55) 46 II
BAGO Bagoinae; Bagous sp (England)
BRAC Erirhinini; Echinocnemis sp (Australia)
80(95) 47 JJ
45 HH 72(67)
PLAT Platypodini; Platypus cylindricus (England)
DRYO sp1 (China)
66(83) 48 KK
BREN sp1 (China)
ATTE Apoderinae; sp (China)
ANTH sp1 (China)

Appendix 4.6 Strict consensus tree of the MITO ML tree and the MITO+ArgK ML tree with 65 taxa. Family and subfamily codes precede taxa names as for Appendix 4.2.



Strict consensus tree of the mtDNA-only ML tree and the mtDNA+ArgK ML tree, with 65 taxa

Node numbers shown in black above nodes
Node letters shown to the right of nodes indicate nodes present or consistent with all consensus trees

Bootstrap (BS) nodal support (shown below or to left of node number):
First number: BS of mtDNA-only (in brackets: BS of mtDNA+ArgK tree)
in blue - nodes with unchanged BS from mtDNA-only tree
in green - nodes with increased BS from mtDNA-only tree
in red - nodes with decreased BS from mtDNA-only tree

CURC Ceratopini; Ceratopus sp (Saba)
CRYP Camptorhinini; Camptorhinus sp (Australia)
CURC Tychiini; Sibinia fulva (USA)
LIXI Rhinocyllini; Bangasternus sp (Turkey)
CURC Eugnomini; Ancyttalia sp (Australia)
CURC Cryptoplini; Haplonyx sp (Australia)
MOLY sp2 (China)
MESO Magdalinini; Magdalis sp (Italy)
CURC Storeini; Melanterius sp (Australia)
CURC Mecinini; Miarus sp (RSA)
CURC Cionini; Cionus griseus (Canaries)
CEUT Phytobini; Rhinoncus sp (Turkey)
MOLY sp3 (China)
MOLY sp1 (China)
CRYP Cryptorhynchini; Ouroporopterus sp (Australia)
CRYP Cryptorhynchini; Perissops sp (Australia)
MOLY sp4 (China)
CRYP Cryptorhynchini; sp (Cameroon)
MOLY Pissodini; Pissodes sp (Italy)
MOLY Lepyrini; Lepyrus sp (China)
CONO Lobotrachelini; sp1 (China)
CONO Lobotrachelini; sp2 (China)
COSS sp1 (China)
COSS Pentarthrini; Pentarthrus elumbe (England)
SCOL sp1 (China)
SCOL Xyleborini; Anisandrus dispar (Norway)
SCOL sp2 (China)
SCOL Dryocoetini; Dryocoetes autographus (Norway
SCOL Premnobiini; Premnobius cavipennis (RSA)
SCOL Ipini; Ips acuminatus (Norway)
SCOL Corthylini; Corthylus rubricollis (Costa Rica)
SCOL Hypoborini; Hypoborus ficus (Morocco)
SCOL Polygraphini; Polygraphus poligraphus (Swede
SCOL Hylesini; Hylesinus varius (Sweden)
SCOL Tomicini; Tomicus piniperda (Norway)
SCOL Diamerini; Diamerus inermis (Tanzania)
SCOL Phloeotribini; Phloeotribus spinulosus (Norway
SCOL Scolytini; Scolytus scolytus (Denmark)
ENTI Otiorhynchini; Otiorhynchus globulus (Italy)
ENTI Otiorhynchini; Otiorhynchus sp (England)
ENTI Myorhinini; sp (RSA)
ENTI Laparocerini; Laparocerus freyi (Canaries)
ENTI Trachyphloeini; Trachyphloeus sp (England)
ENTI Brachyderini; Strophosoma sp (England)
ENTI Brachyderini; Brachyderes rugatus (Canaries)
ENTI sp3 (China)
ENTI sp1 (China)
ENTI sp2 (China)
ENTI Tanymecini; Geotragus sp (China)
ENTI sp4 (China)
ENTI Psallidiini; Psallidium sp (Turkey)
ENTI Ophryastini; Ophryastes sp (USA)
ENTI Geonemini; Lachnopus curvipes (Saba)
ENTI Tropiphorini; Catasarcus sp (Australia)
ENTI Naupactini; Litostylus pudens (Saba)
CYCL Aterpini; Rhadinosomus sp (Australia)
CYCL Aterpini; Pelolorhinus sp (Australia)
CYCL Rhythirrinini; Cisolea sp (Australia)
BAGO Bagoinae; Bagous sp (England)
BRAC Erirhinini; Echinocnemis sp (Australia)
PLAT Platypodini; Platypus cylindricus (England)
DRYO sp1 (China)
BREN sp1 (China)
ATTE Apoderinae; sp (China)
ANTH sp1 (China)
7 11(10)
1 71(74)
2 100(100) E
3 100(100) A
4 39(48)
5 91(91) B
6 100(100) C
8 63(77) D
9 37(44) F
10 89(90) G
11 100(100) H
12 100(100) I
13 100(100) J
14 98(99) K
15 21(30)
16 24(34)
17 47(56) L
18 46(48) M
19 100(100) O
20 29(30) N
21 46(55) P
22 45(56) Q
23 37(26) R
24 100(100) T
25 59(50)
26 48(53)
27 32(36)
28 100(100) S
29 30(38)
30 100(100) U
31 30(42)
32 95(97) V
33 100(100) W
34 60(71) X
35 84(82) Y
36 84(90) Z
37 90(93) AA
38 77(79) BB
39 85(93) CC
40 74(76)
41 86(94) DD
42 100(100) EE
43 100(100) FF
44 92(88) GG
45 80(82)
46 72(55) HH
47 58(52) II
48 80(83) JJ
49 66(75) KK

Chapter 5

Combining whole mitogenomes with shorter sequences to evaluate tribal and subfamilial monoplyly in the broad-nosed weevils (Curculionidae: Entiminae, Cyclominae and Hyperinae)

"Classification of weevils is like a mirage in that their wonderful variety of form and the apparent distinctiveness of many major groups lead one to suppose that classifying them will be fairly straightforward but, when examined closely, the distinctions disappear in a welter of exceptions and transformation series."

- Richard Thompson, 1992

Compsus sp. (Curculionidae: Entiminae: Eustylini), Pichincha, Ecuador

Chapter 5: Combining whole mitogenomes with shorter sequences to evaluate tribal and subfamilial monoplyly in the broad-nosed weevils (Curculionidae: Entiminae, Cyclominae and Hyperinae)

5.1 Abstract

Establishing well-supported monophyletic groups is a key requirement for producing a natural classification that reflects evolutionary descent. In a phylogenetic framework this is best achieved through dense taxon sampling, ensuring a representative sampling of divergent lineages, and the analysis of a robust character dataset, combined with statistical testing of topological hypotheses. This chapter assesses the monophyly of tribes and subfamilies within the broad-nosed weevils by fulfilling these conditions. Taxon sampling is enhanced through obtaining sequence data from GenBank for regions of the mitochondrial *cox1* and *rrnL* genes, and combining these data and taxa with the mitogenomic assembly 'backbone' data obtained in Chapter 3. Phylogenetic analyses incorporating topological constraints for various higher-taxa were statistically tested using the AU, SH and KH tests and indicated that three tribes within the Entiminae are not monophyletic. Moderate and high bootstrap supports were also consistent with two entimine tribes (Peritelini and Cylydrorhinini) being retrieved as monophyletic in an unconstrained analysis. Furthermore, one genus of cyclomine weevils is recovered as belonging outside the broad-nosed weevils clade, although its taxonomic placement remains uncertain. It is apparent that this approach may be hampered in effectiveness by limited taxon

sampling in the 'backbone' dataset, rendering it difficult for divergent taxa to robustly match to their closest lineages. However, with improved taxon sampling of the mitogenomic tree, the general approach will provide a useful taxonomic tool within the weevils.

5.2 Introduction

The fundamental aim of phylogeny reconstruction is to summarise genealogically determined evolutionary relationships as phylogenetic trees, visually tracing the historical course of speciation as organised through the relative recency of common ancestry (Harrison & Langdale 2006; Wiley & Lieberman 2011). Together with other data, such as geographic distributions and ecological traits for species under consideration, phylogenies can be powerful tools for explaining observed patterns, and for testing hypothesised processes of speciation. Of central importance when inferring biological and systematic meaning from trees is the formulation of a sound basis for identifying natural groups of taxa, from which broader conclusions and predictions can be made regarding the biology of the included species. Deciphering which groups of organisms are natural (or monophyletic) is a particular prerequisite for constructing a hierarchical classification system that reflects their underlying evolutionary history.

Of paramount importance for the meaningful testing of potential monophyletic groups is a well sampled dataset, containing taxa of as many potentially separate lineages as possible in order to increase confidence in the resulting topologies and clades. However, because comprehensive taxon sampling in very diverse groups containing thousands of species is very difficult in practice, more so at present for molecular studies than for morphological analyses owing to limitations of DNA integrity and quality in older specimens, alternative sources of data other than specifically collected specimens should be investigated in order to enhance taxon coverage. Such data can be obtained from public repositories of DNA sequence data held in freely accessible online databases such as the The National Center for Biotechnology Information's GenBank (Benson *et al.* 2013), part of the International Nucleotide Sequence Database Collaboration, whose aim is to gather and publish nucleotide sequences and annotations and to allow access to data submission and retrieval tools. Other databases also exist, for example The Barcode of Life Data System (BOLD) (Ratnasingham & Hebert 2007), which specialises in the acquisition, storage and online publication of cytochrome c oxidase subunit I (*cox1*) barcode sequences only, but GenBank is by far the most comprehensive, at present holding 171123749 sequences from more than 300,000 organisms submitted by research laboratories across the world (NCBI GenBank Flat File release 200.0, 15 February 2014).

This chapter aims to test the monophyly of tribes and subfamilies within the diverse broad-nosed weevils (subfamilies Entiminae, Cyclominae and Hyperinae) using sequences obtained from GenBank to enhance the taxon coverage of these groups in the phylogeny constructed from mitogenomic sequences in Chapter 3. The approach used is analogous to the that of Hernández-Vera *et al.* (2013), who obtained short (< 100 bp) phylogenetically informative amplicons (SPIAs) of the mitochondrial 16S ribosomal large subunit gene (*rrnL*) from DNA-degraded specimens of weevils and incorporated them into a 'backbone' phylogeny built from a concatenation of longer sequences of five loci (including *rrnL*) to investigate their biogeographic history. The process tested here differs in that instead of SPIAs, longer 'whole' sequences of mitochondrial *cox1* and *rrnL* genes obtained from GenBank are added to the mitogenomic 'backbone' phylogeny, containing both those loci and 13 other genes, in order to identify the lineages to which the database sequences are most closely related under a maximum likelihood (ML) optimality criterion.

Selection of the *cox1* 5' region and *rrnL* as the short loci to be added to the mitogenomic data was made based upon the fact that a large number of sequences for these genes have been deposited on GenBank owing to their wide use in phylogenetics research, and in the case of *cox1*, its ubiquitous use as the 'barcode' region of choice for molecular-based species identifications (Hebert *et al.* 2003). The 'backbone' phylogeny of Curculionoidea constructed from mitogenomic data from 120 weevil taxa (in 7 families, including 67 tribes of Curculionidae) in Chapter 3 is highly congruent with previous molecular hypotheses of weevil relationships (e.g. Haran *et al.* 2013; McKenna *et al.* 2009) and clearly demonstrates the well supported division of the Curculionidae *s.str.* into two large clades, one of which represents the monophyletic broad-nosed weevils as defined below, recovered with 100% bootstrap (BS) support in that analysis (Figure 3.7). The broad-nosed weevils are selected for further investigation of tribal relationships because of their unambiguous monophyly and the comparatively large number of taxa represented in the mitogenomic phylogeny (33 species in 19 tribes), maximising the number of lineages and thereby increasing the probability of a close match to database sequences. Additionally, one of its component subfamilies, the Entiminae, is the most speciose subfamily-level taxon in Curculionidae, containing an estimated 12,000 described species globally (Oberprieler *et al.* 2007). Although Entiminae has generally been recovered as monophyletic (Marvaldi *et al.* 2002) or paraphyletic with respect to the other broad-nosed weevil subfamilies Hyperinae and/or Cyclominae in molecular analyses (Haran *et al.* 2013; Hundsdoerfer *et al.* 2009; McKenna *et al.* 2009), its internal tribal structure is not well understood, with as many as 55 (Alonso-Zarazaga & Lyal 1999) and as few as five (Marvaldi 1997) tribes proposed. Consequently, these relationships are in need of further investigation.

The concept of 'broad-nosed weevils' dates back to the work of Lacordaire (1863), who divided his family 'Curculionides' into two groups: the Adelognatha and the Phanerognatha. The former of these represents the broad-nosed weevils, defined morphologically by having the prementum covering the maxillae and by the possession of deciduous processes on the adult mandibles (Thompson 1992; Velazquez De Castro *et al.* 2007) as well as bearing the distinctive relatively short rostrum that gives rise to their popular name. Interpretation of precisely which taxonomic groups are characterised as broad-nosed weevils has varied according to the opinion of different authors (e.g. Kuschel 1995; Thompson 1992). One widespread definition, which was employed by Marvaldi (1997) in assessing the monophyly of broad-nosed weevils based upon larval and adult morphological characters, contained the following higher taxa *sensu* Bouchard *et. al.* (2011): Brachyceridae, Ithycerinae and Microcerinae (subfamilies of Brentidae), Gonipterini (tribe of Curculioninae), Entiminae, Cyclominae and Hyperinae (subfamilies of Curculionidae). Marvaldi (1997) concluded that broad-nosed weevils in that sense is not monophyletic, with the Ithycerinae, Microcerinae and Brachyceridae recovered as forming three stepwise basal lineages (Ithycerinae most basal) and the Entiminae + Cyclominae forming a monophyletic apical clade (Hyperinae was not analysed). This result, together with the results of the mitogenomic analysis from Chapter 3 and those based on other molecular data (e.g. McKenna *et al.* 2009) represent strong independent evidence that Brachyceridae, Ithycerinae and Microcerinae form separate paraphyletic basal lineages to those ‘broad-nosed’ weevils classified within Curculionidae *s.str.* (*i.e.* Entiminae + Cyclominae + Hyperinae). For the purposes of this study, only the latter group is defined and henceforth referred to as the ‘broad-nosed’ weevils, within which the monophyly of various groups is tested.

Statistical tests available to undertake hypothesis testing between competing ML tree topologies generally utilise the likelihood values (for each tree this is the product of all per-site likelihoods in the input alignment) for calculation of test statistics. Such tests include the Shimodaira-Hasegawa (SH) test (Shimodaira & Hasegawa 1999) and the Kishino-Hasegawa (KH) test (Kishino & Hasegawa 1989) which both compare the log-likelihoods of two trees to produce a probability statistic for each of them. In the SH test, the trees tested are selected *a posteriori*, whereas in the KH test, the trees are selected *a priori* (Schmidt 2009). Both these tests have biases and limitations, including a correlation between the SH test results with the number of trees being tested (rendering the test conservative in rejecting trees) and the inability of the KH test to control for type 1 errors (Shimodaira 2002). An alternative test that is able to correct for the tree selection bias is the approximately unbiased (AU) test of Shimodaira (2002). The AU test is based upon BS resampling of the per-site log-likelihoods of the input alignment, which allows for the alignment length to be altered and the newly bootstrapped probabilities being scaled to the original alignment length (Schmidt 2009). The AU test statistic is calculated from the change in BS probabilities for each bootstrapped set of replicates. This test is able to control for type 1 errors and is currently one of the most widely employed methods to assess topologies under the ML optimality criterion. The AU, SH and KH tests are used here for statistical tree selection.

This study is therefore both an exploration of the phylogenetic utility of incorporating shorter sections of sequence data into a longer alignment and a test of monophyly of the tribes and subfamilies for which more than one sequence is available, in a real-world scenario of combining newly generated sequences with publicly available ones.

5.3 Materials and methods

5.3.1 'Backbone' phylogeny

The mitogenomic sequences obtained for 120 curculionid taxa in Chapter 3 (plus the two Chryomelidae and Cerambycidae outgroup taxa; Appendix 3.2) were used in the phylogenetic reconstructions in this chapter, acting as a comprehensive phylogenetic framework inasmuch as they provided the 'backbone' in the resulting trees. Shorter single loci sequences for *cox1* 5' and *rrnL* obtained from GenBank as described below were added to the datamatrix for a combined analysis.

5.3.2 Bioinformatics pipeline for obtaining public database sequences

Automated extraction of sequence data from GenBank was achieved through the use of a series of Perl scripts originally developed as part of a custom-built bioinformatics pipeline for analysing public database sequence data (Hunt *et al.* 2007; Hunt & Vogler 2008). These scripts greatly facilitate the selection of both taxa and loci of interest from amongst all the sequences available, as well as greatly expediting the process of sequence retrieval. Similar scripts were recently successfully used to reconstruct a very large phylogeny of >8000 Coleoptera species from analysis of four nuclear and mitochondrial loci (Bocak *et al.* 2013) obtained from GenBank, indicating the importance of such databases as a source of freely available data. The pipeline was here used only for the selection and retrieval of sequences; subsequent sequence alignment and phylogenetic analyses were undertaken separately. All Perl scripts were run on the Natural History Museum 'ctag' Linux-based bioinformatics server and each step is briefly outlined below.

Initially, all publicly available DNA sequences labelled as belonging to Coleoptera (as of 12 October 2012) were downloaded from GenBank into a purpose-built flat file database using a custom Perl script (create_fasta_database_from_genbank_flatfiles; Appendix 5.1A). The second step involved using another custom Perl script (parse_order_from_endop_fastafile.pl; Appendix 5.1B) to automatically change the names of each sequence in the database into a short taxonomic code, based upon the first letter(s) of each hierarchical taxonomic rank from Order down to species level. This code can subsequently be used to easily identify the taxa of interest. The same script also generated a key to all ranks of the taxonomic code allowing for straightforward cross-referencing to taxonomic names. To select only those sequences belonging to the broad-nosed weevils, as defined for this study, a further custom Perl script (parse_taxa_from_fastafile.pl; Appendix 5.1C) parsed the database, selecting only those sequences having the code for Entiminae, Cyclominae, Hyperinae and all taxonomic ranks below these subfamilies. In order to further select only *cox1* 5' and *rrnL* sequences from amongst the resulting set of sequences, a small number (13) of sequences for these two loci were manually downloaded from GenBank for a wide diversity of Coleoptera and each made into a small database. These known *cox1* and *rrnL* sequences were subsequently used in turn in two separate BLAST searches (Altschul *et al.* 1990) against the broad-nosed weevil sequences database (E = 1e-5). Two custom Perl scripts were then used to select (parse_blast_output.pl; Appendix 5.1D) and to retrieve (retrieve_sequences.pl, Appendix 5.1E) only those sequences identified through the BLAST search for each locus. To avoid taxonomic redundancy in the GenBank sequences (some species may have multiple entries for the same locus), the final pipeline step used a custom Perl script (perl one_per_species.pl; Appendix 5.1F) to select only one sequence per species per locus (the longest sequence where two or more sequences differ in length).

Because several genera of broad-nosed weevils were represented by sequences from many species, the GenBank dataset was further reduced to a maximum of five species per genus following a preliminary ML analysis containing all downloaded GenBank broad-nosed weevil *cox1* and *rrnL* sequences (180 and 175 sequences respectively, representing 278 species-level taxa) combined with the mitogenomic data from Chapter 3 (122 taxa). The alignment step and analysis was otherwise identical to that described below for the unconstrained analysis. The results of this allowed for objective selection of divergent species (sometimes recovered in clearly different lineages) within each genus to ensure that no bias for closely related species was made when choosing taxa to retain for further analysis. Additionally, all taxonomic names were corrected for any mistakes and to ensure that genera had been correctly assigned to tribes and subfamilies according to the catalogue of Alonso-Zarazaga and Lyal (1999).

5.3.3 Multiple sequence alignment and dataset concatenation

Prior to alignment, the *cox1* 5' and *rrnL* GenBank sequences, obtained through the bioinformatics pipeline, were added to the corresponding mitogenomic *cox1* and *rrnL* sequences from Chapter 3 to construct the combined GenBank + whole mitogenomic dataset.

Mitogenomic sequences for the genes *nad5*, *nad4*, *nad4L* and *nad1*, which are transcribed on the reverse strand of the mitochondrial genome (mitogenome), were reverse complemented prior to alignment. Sequences for each of the 13 protein-coding and 2 ribosomal RNA genes were individually aligned using the MAFFT version 7 online server, incorporating the FFT-NS-I slow iterative refinement strategy (Katoh, et al. 2002) with the following parameter values: nucleotide scoring matrix 200PAM/k=2, gap open penalty = 1.53, offset value = 0.0. (Katoh *et al.* 2002). Alignments were thereafter checked manually in Geneious for quality and to ensure that protein-coding genes were in the correct reading frame. The resulting individual gene alignments were concatenated together in mitogenomic gene order to create the final dataset in Phylip format for phylogenetic analysis.

5.3.4 Monophyly constraints

In order to test whether monophyly of any of the subfamilies Entiminae, Cyclominae and Hyperinae, and any of the tribes within the subfamily Entiminae is consistent with the combined dataset (*i.e.* cannot be statistically rejected), a series of 20 constraint tree files in Newick format were constructed, each topologically constraining one subfamily or tribe within the broad-nosed weevils, as summarised and described in the results. Only groups with two or more species, and which were not recovered as monophyletic in the initial unconstrained ML analysis (an initial test of monophyly), were selected for constraint analysis.

5.3.5 Phylogenetic analyses

Both an unconstrained and 20 constrained (as outlined above and in the results) ML analyses were undertaken using RAxML 7.6.6 (Stamatakis 2006), run on the CIPRES web-based server (Miller *et al.* 2010). To assess nodal support, a rapid BS analysis with 1000 iterations was run simultaneously with tree-building. The dataset was analysed partitioned by gene because previous analysis of the mitogenomic dataset in Chapter 3 indicated that a partitioned analysis outperforms an unpartitioned one. Therefore separate estimated models of nucleotide substitution were specified for each gene region in the alignment. A GTRCAT model was implemented for the bootstrapping phase and a GTRGAMMA model was used for final tree inference (GTR + optimisation of substitution rates + optimisation of site-specific evolutionary rates).

All trees were visualised in Dendroscope (Huson & Scornavacca 2012) and were rooted with the branch leading to the most divergent outgroup (Chrysomelidae). Rooted trees were exported as Nexus files into R, where terminal taxon names were added using a custom R script.

5.3.6 Statistical hypothesis testing

To statistically test whether monophyly of any of the higher taxa constrained as described above could be rejected, the AU test (Shimodaira 2002) was implemented to obtain the confidence set of trees. This is achieved through resampling the per-site log-likelihood of the input alignment by changing the alignment length and drawing new BS samples from these lengths. The number of times the hypothesis is supported by the BS replicates is used to calculate the BS support for different sequence lengths; the AU test then calculates a p-value from the change in bootstrap values along the changing sequence length (Shimodaira 2002).

To undertake the AU test, the per-site log-likelihood was computed for each of the unconstrained and 20 constraint trees in RAxML using the –f g algorithm, and written to a Treepuzzle formatted file (Schmidt 2009). These values were then used in the program CONSEL (Shimodaira & Hasegawa 2001) to then perform the bootstrap resampling (100,000 replicates per tree) and to calculate the p-values for the AU, SH and KH tests.

5.4 Results

5.4.1 Bioinformatics pipeline

The GenBank-derived dataset obtained via the bioinformatics pipeline contained 107 species of Entiminae, Cyclominae and Hyperinae. Within Entiminae, 22 tribes, 62 genera and 92 species were represented. Within Cyclominae, 4 tribes, 10 genera and 13 species were represented. The Hyperinae was represented by one genus and 2 species. A total of 68 *rrnL* and 63 *cox1* sequences were obtained and 24 species were represented by sequences from both loci, with 44 species only represented by *rrnL* and 39 species only by *cox1*. Sequence lengths varied between 113-558 bp in *rrnL* and 262-748 bp in *cox1*. Appendix 5.2 summarises the GenBank-obtained sequence data matrix.

5.4.2 Phylogenetic analyses

The GenBank-obtained sequences were combined with the existing mitogenomic data from Chapter 3 to yield an aligned matrix of 229 taxa, 15 genes and 13912 positions. The final dataset contained the following broad-nosed weevil taxa: 27 tribes, 74 genera and 119 species (121 terminals) of Entiminae; 5 tribes, 14 genera and 18 species of Cyclominae; 1 genus and 3 species of Hyperinae. The following 18 tribes of Entiminae contained more than one species and therefore could be tested for monophyly, initially through the unconstrained ML analysis (as analysed by topology and BS support), and then through the individual constraint analyses: Brachyderini, Celeuthetini, Cylydrorhinini, Cyphicerini, Elytrurini, Eustylini, Geonemini, Laparocerini, Naupactini, Otiorhynchini, Peritelini, Polydrusini, Rhyncogonini, Sciaphilini, Sitonini, Tanymecini, Trachyphloeini, Tropiphorini. In addition the subfamilies Entiminae, Cyclominae and Hyperinae, and the three of them combined as the 'broad-nosed weevils', were each also tested for monophyly using constraint analyses.

The topology of weevil families and subfamilies recovered in the unconstrained ML tree (final ML optimisation likelihood: -789416.469537) shown in Appendix 5.3 is highly congruent with that of the tree generated using the mitogenomic data alone (Figure 3.7). Only the placement of *Ocladius* (Brachyceridae: Ocladiinae) differs in being placed within the Dryophthoridae + Platypodinae clade in the present analysis, and outside of it in the mitogenomic analysis. One other intriguing disparity is the sister relationship recovered between *Aphela* (Cyclominae; from GenBank) and *Bagous* (Bagoinae) in a clade sister to all other Curculionidae *s.str.* (*sensu* Bouchard *et al.* 2011) except Platypodinae. The division of Curculionidae *s.str.*

into two large clades is also recovered, although support for the dividing node is reduced to 31% BS from 100% BS in the mitogenomic tree alone (Appendix 5.3).

Relationships within the large 'long-nosed weevil' clade, *i.e.* all Curculionidae *s.str.* other than Entiminae, Cyclominae and Hyperinae (and Platypodinae) are similarly highly congruent with the previous mitogenomic analyses (Appendix 5.3), consisting of a sister relationship between the Ipini (Scolytinae) and the remaining taxa that are split into two clades, one containing the moderately well supported (70% BS) remaining Scolytinae (except *Coptonotus*) and the other containing the rest of the subfamilies with little support for the monophyly of any of them except for Lixinae (100% BS).

Within the clade of focal interest composed of the broad-nosed weevils there is generally very low nodal support for the more basal relationships although some of the more apical nodes are well supported, with 26 of them having support values of 80% BS or higher (Appendix 5.3). Two tribes of Entiminae are recovered as monophyletic with moderate nodal support in this analysis: the Peritelini (88% BS) and the Cylydrorhinini (69% BS), each represented by two genera and two species.

Because of their monophyly as evaluated through bootstrap analysis, these last two tribes are therefore not considered for further constraint analyses. The remaining 16 tribes of Entiminae were recovered as paraphyletic or polyphyletic and were consequently each constrained as monophyletic (Table 5.1) in separate RAxML analyses (identical to the unconstrained analyses other than enforcing the topological constraint). The resulting per-site log likelihoods of these trees, estimated separately in RAxML, were used to calculate the AU test statistic as detailed below.

5.4.3 Statistical hypothesis testing

Results of the statistical tests carried out in CONSEL indicate that at a significance level α = 0.05, the confidence sets are the same across the AU, SH and KH tests (Table 5.2), with only trees constraining Otiorhynchini, Brachyderini and Tropiphorini as monophyletic rejecting the null hypothesis that there is no difference between the trees (*i.e.* that all unconstrained and constrained trees are equally good explanations of the data). Consequently for these three tribes the alternative hypothesis is accepted that their likelihoods are significantly different and therefore their monophyly is rejected.

5.5 Discussion

5.5.1 Unconstrained analysis

Augmenting the mitogenomic dataset with the GenBank sequence data did not strongly affect the main topology with regards to family- and subfamily-level relationships compared to the mitogenomic data alone. This was expected because the bulk of phylogenetic signal is present in the full mitogenomic alignment and no additional taxa in the basal portion of the tree were incorporated into this analysis. The single aberrant placement of *Aphela godoti*, currently classified in the Cyclominae (Bouchard *et al.* 2011), outside the broad-nosed weevils clade together with Bagoinae was the only inconsistency. The *Aphela* + *Bagous* relationship has only weak nodal support (47% BS), but nevertheless it is striking that *Aphela,* an apparent broad-nosed weevil, is recovered outside the large Entiminae + Cyclominae + Hyperinae clade which is otherwise monophyletic in all phylogenetic analyses in this thesis.

Table 5.1 Higher-taxa constrained as monophyletic for ML analysis and the AU test of monophyly

Constrained taxon	No. of genera in constraint	No. of terminals in constraint	Final ML optimisation likelihood
Broad-nosed weevils	89	142	-789419.301277
Entiminae	74	121	-789472.134247
Cyclominae	14	18	-789458.390913
Hyperinae	1	3	-789408.960432
Brachyderini	2	6	-789643.159670
Celeuthetini	8	8	-789466.103647
Cyphicerini	1	2	-789498.841621
Elytrurini	2	3	-789423.648418
Eustylini	6	9	-789484.535716
Geonemini	5	7	-789430.817614
Laparocerini	3	9	-789411.304383
Naupactini	9	19	-789464.026787
Otiorhynchini	1	6	-789597.443422
Polydrusini	3	6	-789416.879310
Rhyncogonini	1	3	-789421.132412
Sciaphilini	4	4	-789475.627644
Sitonini	1	4	-789424.406899
Tanymecini	5	6	-789496.695133
Trachyphloeini	1	2	-789480.383429
Tropiphorini	6	9	-789872.066562
UNCONSTRAINED	147	229	-789416.469537

When *Aphela* was separately constrained within the Cyclominae and within the broad-nosed weevils, neither of the resulting ML trees was rejected by the AU, SH or KH test, prohibiting a definitive statement on its systematic placement. *Aphela* was previously classified within the Molytinae (tribe Rhythirrinini) by Alonso-Zarazaga and Lyal (1999) on morphological grounds, and clearly this fact and the present molecular findings indicate that this taxon warrants further investigation with additional sequence data (ideally a full mitogenome sequence).

Table 5.2 Results of the AU, KH and SH tests of constrained monophyly of 20 higher taxa and the unconstrained analysis, ranked by likelihood. Log likelihood difference to the best tree is shown, except for the best tree, which shows the negative distance of the second best. The *p*-values below a significance level α = 0.05 are highlighted in grey and represent the three tribes whose monophyly is rejected (Otiorhynchini, Brachycerini and Tropiphorini).

Rank (by likelihood)	**Taxon constrained in ML tree**	**Log likelihood difference to the best tree**	**AU test *p*-values**	**KH test *p*-values**	**SH test *p*-values**
1	Sitonini	-4.1	0.621	0.526	0.971
2	UNCONSTRAINED	4.1	0.605	0.474	0.948
3	Hyperinae	8.7	0.527	0.396	0.968
4	Laparocerini	11.7	0.573	0.430	0.961
5	Rhyncogonini	18.0	0.513	0.409	0.921
6	Broad-nosed weevils	21.4	0.442	0.378	0.913
7	Polydrusini	23.7	0.431	0.357	0.942
8	Cyphicerini	24.1	0.425	0.362	0.865
9	Geonemini	26.7	0.411	0.355	0.873
10	Elytrurini	29.2	0.395	0.340	0.876
11	Celeuthetini	55.6	0.202	0.213	0.719
12	Naupactini	56.4	0.206	0.185	0.726
13	Cyclominae	70.6	0.132	0.174	0.627
14	Eustylini	72.6	0.176	0.125	0.619
15	Sciaphilini	78.0	0.119	0.153	0.573
16	Entiminae	88.0	0.080	0.100	0.505
17	Trachyphloeini	94.3	0.083	0.059	0.463
18	Tanymecini	99.3	0.054	0.059	0.426
19	Otiorhynchini	204.2	2e-004	0.006	0.048
20	Brachyderini	241.0	6e-051	3e-005	0.007
21	Tropiphorini	483.0	0.001	0	0

The unconstrained analysis indicated that the tribes Peritelini and Cylydrorhinini are monophyletic in our dataset, although due to the limited taxon sampling of each, interpretation of monophyly beyond the included genera remains putative. Nevertheless, inclusion of the type genera of both these tribes (*Peritelus* and *Cylyndrorhinus* respectively) in the dataset increases objectivity and confidence in at least establishing that each of the other genera included per tribe is correctly classified at present (*Ctenochirus* in Peritelini and *Caneorhinus* in Cylydrorhini), which would not have been the case had the type genera not been analysed.

The tribe Peritelini is large, containing 76 genera with a wide distribution in the Holarctic, Afrotropical and Australian regions, with new species being continuously discovered even in the relatively well studied European fauna (e.g. Pierotti & Fink 2013; Pierotti *et al.* 2013). However, morphologically it has not been well defined, and in particular lacks apomorphies enabling a clear separation from Otiorhynchini (Pierotti *et al.* 2010). Additionaly, at least one genus, *Caenopsis,* has been recently transferred to the tribe Trachyphloeini (Pierotti *et al.* 2010), further highlighting the uncertain monophyly of the group.

In contrast, the tribe Cylydrorhinini is much smaller, containing only six genera, and of restricted distribution, occurring only in the Australian and southern Neotropical regions. It had previously been classified as a subfamily (Cylyndrorhininae) consisting of two tribes: the Cylyndrorhinini and Listroderini (Marvaldi 1998). However study of larval characters led to the conclusion that the Cylyndrorhinini (in particular the genera *Caneorhinus* and *Cylydrorhinus*, also evaluated here with molecular data) belong in the Entiminae and the Listroderini belong in the "Rhytirrhininae", *i.e.* in the current subfamily Cyclominae (Marvaldi 1998). The molecular data indicate that Listroderini is paraphyletic, consisting of three lineages, only one of which, *Germainiellus* + *Antarctobius* has low support (56% BS), with the two included *Germainiellus* species being well supported as monophletic (100% BS). Whilst the limited taxon sampling in the present study suggests that Cylydrorhinini is monophyletic, no firm conclusions can be drawn with regards to its relationship with Listroderini because of low nodal support in the intervening parts of the tree. This specific relationship was not investigated further with constraint analyses although constraining the Cyclominae as a whole did not lead to the resulting tree being rejected by the AU test statistic, suggesting that the molecular data is consistent with larval morphology and that Listroderini is distinct from Cylydrorhinini.

Although the unconstrained analysis failed to recover any of the remaining 16 tribes of Entiminae as monophyletic, some of these were recovered in two or more well supported clades. Thus within the Tropiphorini, *Tropiphorus carinatus* and *T. bertolini* form one clade (98% BS), *Malvinius* (three species) forms another (99% BS), with the remaining four genera (and species) of Tropiphorini distributed across the tree with low support. In the Celeuthetini, *Cnemidothrix, Levoecus* and *Sphaerorhinus* form a clade (90% BS), as do *Coptorhynchus* and *Heteroglymma* (99% BS). With the addition of *Samobius* and *Platysimus*, all seven aforementioned genera form a clade, but with low support (14% BS); the remaining genus of Celeutherini, *Phraotes,* is recovered away from this last clade with one moderately supported (85% BS) intervening node that groups it with members of the tribes Rhyncogonini and Elytrurini.

Whilst such clades with moderate and high nodal support appear to offer evidence for the paraphyly of several tribes, the generally low nodal supports in the intermediate nodes between such clades preclude conclusions to be drawn based on bootstrap values alone.

5.5.2 Constraint analyses and statistical tests of monophyly

In supplement to the bootstrap support results, the AU tests rejecting the three ML trees respectively containing the constrained monophyly of the tribes Otiorhynchini, Brachyderini and Tropiphorini provide further evidence for the paraphyly of these higher taxa.

Otiorhynchini is a particularly species-rich tribe containing 10 genera, of which the *Otiorhynchus* 'complex' contains about 1500 species exclusive to the Palaearctic region (except for a few introduced species in the Nearctic) which have been divided into 105 subgenera (Lachowska *et al.* 2008). No detailed phylogenetic analysis has been undertaken within this group, although a karyotype analysis of three genera was in accordance with the current classification (Lachowska *et al.* 2008). The taxa analysed in this study belong to five subgenera: *O. (Otiorhynchus) armadillo, O. (Postaremus) nodosus, O. (Dorymerus) sulcatus, O. (Nihus) globulus* and *O. (Zustalestus) rugosostriatus* (Colonnelli 2003). Four of these species (*O. globulus, O. sulcatus, O. rugosostriatus* and *Otiorhynchus* sp.) were retrieved in a monophyletic clade in the unconstrained ML analysis (69% BS), with a high support for the sister relationship between *O. sulcatus* and *Otiorhynchus* sp. (100% BS). Of the two remaining species, *O. nodosus* was retrieved with high support as sister to *Strophosoma melanogrammum,* belonging to the tribe Brachyderini (98% BS), and *O. armadillo* was weakly supported as a lineage basal to a clade containing the first group of four *Otiorhynchus* + two members of Tropiphorini (two *Tropiphorus* spp.) and one Cyclomini (*Bronchus* sp.). It is difficult to be confident about the relationships amongst these Otiorhynchini, in particular the retrieval of *O. nodosus* sister to *S. melanogrammum* is surprising. Sequences for *cox1* for these last two species were obtained from GenBank, and both originated from the same study investigating clonality and polypoidy in *Otiorhynchus* (Stenberg et al. 2003). A BLAST search against the GenBank database revealed that the *S. melanogrammum cox1* sequence very closely matches sequences from four *Otiorhynchus* species in the same study (98-99% identity over 100% of the 552 bp sequence; E=0.0) indicating a close relationship between these two genera. It is unlikely that the sample was mislabelled on Genbank, although this cannot be ruled out with certainty. *Strophosoma melanogrammum* is also represented in the present data matrix by a partial mitogenomic sequence, lacking both *cox1* and *rrnL* (Haran et al. 2013), and not recovered together with the GenBank-sequence represented *S. melanogrammum,* but in another clade containing three other Brachyderini taxa (*Brachyderes* spp.), most likely explaining the relationship with *Otiorhynchus* described above being driven by the closest-matching *cox1* sequence.

5.5.3 Conclusions

The approach used here has confirmed the utility of combining shorter sequences into a longer alignment insofar as several interesting relationships were identified, both supporting and rejecting monophyly of currently classified higher taxa. The extent to which meaningful conclusions can be made regarding how accurately shorter sequences are able to match to their correct lineages is undoubtedly a function of the depth of taxon coverage in the backbone mitogenomic alignment, from which most of the phylogenetic signal is derived. The mitogenomic dataset contained members of less than a third (19 out of 63) of the tribes within the broad-nosed weevils, so it is hardly surprising that nodal BS support for many internal nodes within this group were poorly supported with the addition of taxa represented by single mitochondrial genes from GenBank. This is a direct result of the small amount of shared comparative data for calculating BS support between taxa with long mitogenomic sequences and the taxa solely represented by short sequences.

The inability to reject several of the apparently paraphyletic clades through the constraint analyses highlights the presence of conflicting or insufficient data, and demonstrates the complex systematics of the Curculionoidea, wherein particular genera cannot confidently be ascribed to even a particular subfamily. Other limitations in this study included the use of taxa incompletely identified only to the level of subfamily, therefore not allowing for possible further scrutiny of tribal- or generic-level relationships. Additionally several sequences from the mitogenomic dataset lacked the *cox1* and *rrnL* genes, particularly those obtained from the study of Haran *et al.* (2013), confounding their utility here to act as 'backbone' sequences due to the missing data for the critical loci. Alternative or additional mitochondrial loci, such as *cytB* and *cox2* that have been used in the phylogeny of Coleoptera, could have been also incorporated in the alignment which may have increased the number of taxa available for study. Another potential limitation with this approach is that taxonomic coverage within the public databases is currently rather patchy, being dependent upon a multitude of sources such that in many cases certain higher taxa are represented by a small number of potentially highly aberrant or localised species *e.g.* most of the Cyclominae obtained from GenBank stemmed from a single study based on the fauna of the Falkland islands (Papadopoulou *et al.* 2009).

Whilst some results obtained here are cautionary in highlighting the necessity for the careful use of publicly available sequences, it has been demonstrated that it is possible to both single out interesting relationships that warrant further investigation and to test for monophyly, whilst attempting to maximising taxon sampling. One avenue of possible investigation for reconstructing supra-specific phylogenenies may involve the use or concatenation of several congeneric GenBank-obtained sequences to represent taxa rather than using single genes from individual species as used here.

5.6 References

Alonso-Zarazaga MA, Lyal CHC (1999) *A world catalogue of families and genera of Curculionoidea (Insecta: Coleoptera) (excepting Scolytidae and Platypodidae)* Entomopraxis, Barcelona.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Benson DA, Cavanaugh M, Clark K*, et al.* (2013) GenBank. *Nucleic Acids Research* **41**: 10.1093/nar/gks1195.

Bocak L, Barton C, Crampton-Platt A*, et al.* (2013) Building the Coleoptera tree-of-life for >8000 species: composition of public DNA data and fit with Linnaean classification. *Systematic Entomology* **39**, 97-110.

Bouchard P, Bousquet Y, Davies AE*, et al.* (2011) Family-group names in Coleoptera (Insecta). *Zookeys* **88**, 1-972.

Colonnelli E (2003) A revised checklist of Italian Curculionoidea (Coleoptera). *Zootaxa* **337**, 1-142.

Haran J, Timmermans MJTN, Vogler AP (2013) Mitogenome sequences stabilize the phylogenetics of weevils (Curculionoidea) and establish the monophyly of larval ectophagy. *Molecular Phylogenetics and Evolution* **67**, 156-166.

Harrison CJ, Langdale JA (2006) A step by step guide to phylogeny reconstruction. *Plant Journal* **45**, 561-572.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B-Biological Sciences* **270**, 313-321.

Hernández-Vera G, Caldara R, Tosevski I, Emerson BC (2013) Molecular phylogenetic analysis of archival tissue reveals the origin of a disjunct southern African-Palaearctic weevil radiation. *Journal of Biogeography* **40**, 1348-1359.

Hundsdoerfer AK, Rheinheimer J, Wink M (2009) Towards the phylogeny of the Curculionoidea (Coleoptera): Reconstructions from mitochondrial and nuclear ribosomal DNA sequences. *Zoologischer Anzeiger* **248**, 9-31.

Hunt T, Bergsten J, Levkanicova Z, *et al.* (2007) A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science* **318**, 1913-1916.

Hunt T, Vogler AP (2008) A protocol for large-scale rRNA sequence analysis: Towards a detailed phylogeny of Coleoptera. *Molecular Phylogenetics and Evolution* **47**, 289-301.

Huson DH, Scornavacca C (2012) Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Systematic Biology* **61**, 1061-1067.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066.

Kishino H, Hasegawa M (1989) Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA-sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution* **29**, 170-179.

Kuschel G (1995) A phylogenetic classification of Curculionoidea to families and subfamilies. *Memoirs of the Entomological Society of Washington* **14**, 5-33.

Lachowska D, Rozek M, Holecova M (2008) Cytotaxonomy and karyology of the tribe Otiorhynchini (Coleoptera : Curculionidae). *European Journal of Entomology* **105**, 175-184.

Lacordaire T (1863) *Histoire naturelle des insectes. Genera des Coléoptères* Roret, Paris.

Marvaldi AE (1997) Higher level phylogeny of Curculionidae (Coleoptera : Curculionoidea) based mainly on larval characters, with special reference to broad-nosed weevils. *Cladistics-the International Journal of the Willi Hennig Society* **13**, 285-312.

Marvaldi AE (1998) Larvae of South American Entimini (Coleoptera: Curculionidae), and phylogenetic implications of certain characters. *Revista Chilena de Entomologia* **25**, 21-44.

Marvaldi AE, Sequeira AS, O'Brien CW, Farrell BD (2002) Molecular and morphological phylogenetics of weevils (Coleoptera, Curculionoidea): Do niche shifts accompany diversification? *Systematic Biology* **51**, 761-785.

McKenna DD, Sequeira AS, Marvaldi AE, Farrell BD (2009) Temporal lags and overlap in the diversification of weevils and flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7083-7088.

Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: *Proceedings of the gateway computing environments workshop (GCE), 14 Nov. 2010*, New Orleans, LA.

Oberprieler RG, Marvaldi AE, Anderson RS (2007) Weevils, weevils, weevils everywhere. *Zootaxa* **1668**, 491-520.

Papadopoulou A, Jones AG, Hammond PM, Vogler AP (2009) DNA taxonomy and phylogeography of beetles of the Falkland Islands (Islas Malvinas). *Molecular Phylogenetics and Evolution* **53**, 935-947.

Pierotti H, Bello C, Alonso-Zarazaga MA (2010) Contribution to the systematic rearrangement of the Palaearctic Peritelini. VI. A synthesis of the Spanish Peritelini (Coleoptera: Curculionidae: Entiminae). *Zootaxa*, 1-96.

Pierotti H, Fink T (2013) New and interesting Peritelini of the Western Mediterranean fauna. XX. A novel Meira (Jacquelin du Val, 1852) species from the Ligurian Alps. *Zootaxa* **3716**, 595-598.

Pierotti H, Germann C, Braunert C (2013) New or interesting Peritelini of the West-Mediterranean fauna. XXIV. Two new Simmeiropsis Pierotti & Bello, 2013 from Portugal (Coleoptera, Curculionidae, Entiminae). *Zootaxa* **3734**, 273-280.

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes* **7**, 355-364.

Schmidt HA (2009) Testing tree topologies. In: *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing* (eds. Lemey P, Salemi M, Vandamme AM), pp. 381-404. Cambridge University Press, Cambridge, UK.

Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Systematic Biology* **51**, 492-508.

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* **16**, 1114-1116.

Shimodaira H, Hasegawa M (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246-1247.

Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690.

Stenberg P, Lundmark M, Knutelski S, Saura A (2003) Evolution of clonality and polyploidy in a weevil system. *Molecular Biology and Evolution* **20**, 1626-1632.

Thompson RT (1992) Observations on the morphology and classification of weevils (Coleoptera, Curculionoidea) with a key to major groups. *Journal of Natural History* **26**, 835-891.

Velazquez De Castro AJ, Angel Alonso-Zarazaga M, Outerelo R (2007) Systematics of Sitonini (Coleoptera : Curculionidae : Entiminae), with a hypothesis on the evolution of feeding habits. *Systematic Entomology* **32**, 312-331.

Wiley EO, Lieberman BS (2011) *Phylogenetics: Theory and Practice of Phylogenetic Systematics* Wiley-Blackwell, Hoboken, New York.

5.7 Appendices

Appendix 5.1 A-F Perl scripts used in bioinformatics pipeline for obtaining public database sequences. Scripts developed by Hunt *et al.* (2007), Hunt & Vogler (2008) and Bocak *et al.* (2013).

Appendix 5.1A create_fasta_database_from_genbank_flatfiles.pl

A file is created (*e.g.* database_fasta_Oct2012.txt) with all coleopteran sequences in GenBank.

```
#!/usr/bin/perl
# 25sept09 option to ignore the list of daily releases. these are less important if you run
this script infrequently
# 20oct2009 bugfix. all field variables reset after // is reached in flatfile.
#      bugfix: would crash if current_taxa string was found in the title but not the lineage
(eg a coevol study)
##################################################
$current_taxa = "coleoptera;";
# Hymenoptera Taxonomy ID: 7399
$download_genbank_flatfiles          = 1;          #     0==dont     download     flatfiles,
1==download. the flatfiles are updated every 3months or so
$download_daily_flatfiles            = 1;          # all taxa (not just inv). updated every
day since last full release
$download_genbank_taxonomy_database  = 1;
$ignore_daily_release                = 0;          #        this        differs        from
$download_daily_flatfiles option (which assumes these have already been downloaded, they are
still parsed).

##################################################

my $date = localtime time;
$month = $date;
$month=~ s/\w+\s+(\w+)\s+\d+\s+\S+\s+(\d+)$/$1$2/;

#              $infile = "nc0313." . $month;
#              parse_genbank_flatfile();

if($download_genbank_taxonomy_database==1)
       {
       system("rm taxdump.tar.gz");
       $command     =     "wget     ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz     -O
taxdump.tar.gz";print "command:$command\n";system($command);
       system("rm *.dmp");
       system("tar xvzf taxdump.tar.gz");
       }else{print "NOT downloading taxonomy database\n"};
#  tar  xvzf  taxdump  into  ~/hip-db/hip-db_scripts/  folder,  delete  all  but  names.dmp  and
nodes.dmp.
# copy coleoptera_fasta to ~/hip-db/ folder
# ftp://ftp.ncbi.nih.gov/blast/executables/LATEST/ for ncbi tools
# improvements:should look for holmetabola as well as endop, just in case
$outfile1_name = "database_fasta_" . $month;
system("rm $outfile1_name");
open(OUT, ">$outfile1_name") || die "cant open outfile:$outfile1_name\n";
open(OUT2,">>debugging_output") || die "cant open outfile\n";
open(OUT3, ">>database_log") || die "cant open outfile\n";
print OUT3 "\nRUNNING SCRIPT:create_fasta_database_from_genbank_flatfiles.pl\n$date\n";

##################################################################################

print "downloading list of daily update flatfiles\n" , $month , "\n";
```

```
$command = "rm flatfilelist_" . $month;print "command:$command\n";system($command);
$command = "wget ftp://ftp.ncbi.nih.gov/genbank/daily-nc/ -O " . "flatfilelist_" .
$month;print "command:$command\n";system($command);

                unless($ignore_daily_release==1)
                {
open(IN, "flatfilelist_$month") || die "cant open flatfilelist_$month\n";
while ($line= <IN>)
        {#/genbank/daily-nc/nc0219.flat.gz"
        if($line=~ /\/genbank\/daily\-nc\/(nc\d+\.)flat\.gz\"/)
                {

                if($download_daily_flatfiles == 1)
                        {
                        $command = "rm $1$month.gz";print
"command:$command\n";system($command);
                        $command = "rm $1$month";print "command:$command\n";system($command);
                        $command = "wget ftp://ftp.ncbi.nih.gov/genbank/daily-nc/" . $1 .
"flat.gz -O " . $1 . $month . ".gz";  print "command:$command\n";   system($command);
                        $command = "gunzip " . $1 . $month . ".gz";print
"command:$command\n";system($command);
                        };

                $infile = $1 . $month;
                parse_genbank_flatfile();
                system("rm $infile");

                # $infile_names[$count_inv_files] = $1 . $month;$count_inv_files++;

                }
        }
close IN;

        }
####################################################################################

print "downloading list of release flatfiles\n" , $month , "\n";
$command = "rm flatfilelist_" . $month;print "command:$command\n";system($command);
$command = "wget ftp://ftp.ncbi.nih.gov/genbank/ -O " . "flatfilelist_" . $month;print
"command:$command\n";system($command);

$debugging_counter=0;
$printed_counter=0;
$entry_counter=0;
$title_counter = 1;


$count_inv_files=0;
open(IN, "flatfilelist_$month") || die "cant open flatfilelist_$month\n";
while ($line= <IN>)
        {
        if($line=~ /(gbinv\d+\.)/)
                {
                print $line;

                if($download_genbank_flatfiles == 1)
                        {
                        $command = "rm $1$month.gz";print
"command:$command\n";system($command);
                        $command = "rm $1$month";print "command:$command\n";system($command);
                        $command = "wget ftp://ftp.ncbi.nih.gov/genbank/" . $1 . "seq.gz -O " .
$1 . $month . ".gz";  print "command:$command\n";   system($command);
                        $command = "gunzip " . $1 . $month . ".gz";print
"command:$command\n";system($command);
                        };
                $infile = $1 . $month;
                parse_genbank_flatfile();
                system("rm $infile");

                $infile_names[$count_inv_files] = $1 . $month;$count_inv_files++;

                };
        };
close(IN);
$command = "rm flatfilelist_" . $month;print "command:$command\n";system($command);
```

```
close(OUT);
close(OUT2);
close(OUT3);

print "\nEND OF SCRIPT\n";
exit;

sub parse_genbank_flatfile
{

open(IN2, $infile) || die "cant open infile:$infile\n";
print "opened $infile\n";

$in_dna = 0;
while($line = <IN2>)
        {

        $debugging_counter++;
        # if($debugging_counter<=150){print OUT2 $line};
        if($line=~/rivacindela/i ){$debugging_counter=0}; # 2586\d\d|

        $line=~ s/\n//;$line=~ s/\r//;
        if(length($current_entry_as_string)>=10000000)
                {
                # print "$current_accession $current_binomial > 10meg ";
                }else{$current_entry_as_string .= $line};
        $lineage_line_counter++;
        $line_length=length($line);
        # print "$line_length\n";
        if($line=~/^ACCESSION\s+([\w\d\.]+)/)
                {
                $entry_counter++;
                $current_accession=$1;
                $current_accession=~ s/ //g;
                $current_product="";
                };

        if($line=~/\/db_xref.+taxon.(\d+)\"/)
                {
                $current_taxid = $1;
                };

        if($line=~/\/product..(.+)\"/)
                {
                $current_product = $current_product . $1;
                };


        if($line=~/ORGANISM/)
                {
                $lineage_line_counter=0;
                $current_lineage="";
                $line=~     s/(ORGANISM\s+)cf\./$1/;$line=~     s/(ORGANISM\s+)aff\./$1/;$line=~
s/(ORGANISM\s+)nr\./$1/;

                if($line=~/ORGANISM\s+(\w+)\s+(\w+)\s*/)
                        {

                        $current_genus=$1;
                        $current_species=$2;
                        $current_binomial=$1 . "_" . $2;
                        }else{

                        if($line=~/ORGANISM\s+(\S+.+)$/)
                                {
                                $current_binomial=$1;
                                $current_binomial=~ s/ /_/g;
                                };
                        ##########print                                          "non-
standard_line:$line\ncurrent_binomial:$current_binomial\n";
                        print  OUT3  "non-standard_line:$line  current_binomial:$current_binomial
";
                        };
                };
        if($lineage_line_counter<=6)
                {
                $current_lineage=$current_lineage . $line;
                };

        if($lineage_line_counter==6)
                {       # Endopterygota
                if($current_lineage =~/$current_taxa/i){$in_endo = 1}else{$in_endo = 0};

                if($current_lineage
=~/\W(\w+ini);/){$current_tribe=$1}else{$current_tribe="unknown"};
                if($current_lineage
=~/\W(\w+inae);/){$current_subfamily=$1}else{$current_subfamily="unknown"};
                if($current_lineage
=~/\W(\w+idae);/){$current_family=$1}else{$current_family="unknown"};
                if($current_lineage
=~/\W(\w+oidea);/){$current_superfamily=$1}else{$current_superfamily="unknown"};
                if($current_lineage
=~/\W(\w+ini);/){$current_tribe=$1}else{$current_tribe="unknown"};
                };

        if($line=~/^ORIGIN/){$in_dna=1;$dna_line_counter=0;$dna_sequence=""}else{

                if($in_dna==1 && $dna_line_counter <= 10000)
                        {
                        $line=~ s/[\d ]//g;
                        $dna_sequence .= $line;
                        $dna_line_counter++;
                        };

                };

        if($line=~/^\/\//)
                {
                $in_dna = 0;$found_title = 0;
                $dna_sequence=~ s/\/\///g;
                $current_product=~ s/ /_/g;

                if(length($current_product)<=1){$current_product="no_product"};

        if($current_binomial=~/Drosophila_melanogaster|Bombyx_mori|Tribolium_castaneum|Apis_me
llifera|Anopheles_gambiae/)
                        {$genome=1}else{$genome=0};


                if($in_endo==1 && $genome==0)
                        {       # remove genome sequences, they take up inordinate space
                        $current_entry_as_string               =~               s/TITLE\s+Direct
Submission//g;while($current_entry_as_string =~ /  /){$current_entry_as_string =~ s/  / /g};

                        if($current_entry_as_string    =~    /^.+\s+TITLE\s+(\S.{20,250}\S)\s[A-
Z]{3,5}/)
                                {
                                $current_title          =          $1;$current_title          =~
s/JOURNAL.+$//;$current_title =~ s/REFERENCE.+$//;$current_title =~ s/\s/_/g;
                                # print "current_title:$current_title:end\n";

        unless(exists($title_hash{$current_title})){$title_hash{$current_title}              =
$title_counter;$title_counter++};
                                }else{

                                if($current_entry_as_string                                    =~
/^.+\s+AUTHORS\s+(\S.{6,250}\S)\s[A-Z]{3,5}/)
                                        {
                                        $current_title       =       $1;$current_title       =~
s/JOURNAL.+$//;$current_title =~ s/REFERENCE.+$//;$current_title =~ s/\s/_/g;
                                        print "current_title:$current_title:end\n";


        unless(exists($title_hash{$current_title})){$title_hash{$current_title}              =
$title_counter;$title_counter++};

                                        }else{
```

```
                                        $current_title = "UNKNOWN_TITLE";
                                        $title_hash{$current_title} = 0;
                                        }
                                # AUTHORS Ribera,I. JOURNAL


                                };

#  TITLE          Morphology  and  molecular  phylogeny  of  some  tibetan  ground  beetles
belonging to the subgenera Neoplesius and Eocechenus (coleoptera,          carabidae)

                        print   OUT   ">$current_accession   $current_taxid   $current_binomial
$current_superfamily $current_family ";
                        print OUT "$current_subfamily $current_tribe $current_product STUDY" ,
$title_hash{$current_title} , "\n$dna_sequence\n";
                        print   OUT2   "$current_accession   $current_taxid   $current_binomial
$current_product $current_superfamily $current_family";
                        print    OUT2    "    $current_subfamily    $current_tribe    "    ,
$title_hash{$current_title} , " $current_title\n";
                        $printed_counter++;

                        if ($entry_counter=~/00$/)    {
                        print   "$infile   $entry_counter   $current_accession   $current_taxid
$current_binomial " , $title_hash{$current_title} , " $current_product\n";
                        print   OUT3   "\n$entry_counter   $current_accession   $current_taxid
$current_binomial $current_product\n";
                                                        };

                        }else{$current_entry_as_string          =~           s/\sTITLE\s.+$//;
if($current_entry_as_string       =~       /$current_taxa/i       &&       $genome==0){die
"\nBUG\ncurrent_entry_as_string:$current_entry_as_string\n"}};

                $current_entry_as_string = "";
                $current_accession="";$current_taxid       =       "";$current_product       =
"";$current_genus="";$current_species="";
                $current_binomial           =           "";$current_lineage            =
"";$current_tribe="";$current_subfamily="";$current_family="";
                $current_superfamily = "";$current_tribe="";$dna_sequence = "";

                };

        };

close(IN2);
print "\n\n$infile: $printed_counter printed out of $entry_counter total\n\n";
print OUT3 "$infile: $printed_counter printed out of $entry_counter total\n";
}

#       LOCUS       FJ425915                 519 bp    DNA     linear   INV 03-FEB-2009
#       DEFINITION  Anopheles oswaldoi isolate SP22-9 5.8S ribosomal RNA gene, partial
#                   sequence; internal transcribed spacer 2, complete sequence; and 28S
#                   ribosomal RNA gene, partial sequence.
#       ACCESSION   FJ425915
#       VERSION     FJ425915.1  GI:222137854
#       KEYWORDS    .
#       SOURCE      Anopheles oswaldoi
#         ORGANISM  Anopheles oswaldoi
#                   Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
#                   Neoptera; Endopterygota; Diptera; Nematocera; Culicoidea;
#                   Culicidae; Anophelinae; Anopheles.
#       REFERENCE   1  (bases 1 to 519)
#         AUTHORS   Motoki,M.T., Santos,C.L.S. and Sallum,M.A.M.
#         TITLE     Intraspecific variation on the aedeagus of Anopheles oswaldoi
#                   (Peryassu, 1922) (Diptera: Culicidae)
#         JOURNAL   Neotrop. Entomol. (2009) In press
#       REFERENCE   2  (bases 1 to 519)
#         AUTHORS   Santos,C.L.S.
#         TITLE     Direct Submission
#         JOURNAL   Submitted (28-OCT-2008) Epidemiologia, Faculdade de Saude Publica,
#                   Universidade de Sao Paulo, Avenida Dr. Arnaldo 715 sala 200, Sao
#                   Paulo, Sao Paulo 01246904, Brazil
#       FEATURES             Location/Qualifiers
#            source          1..519
#                            /organism="Anopheles oswaldoi"
#                            /mol_type="genomic DNA"
#                           /isolate="SP22-9"
#                           /db_xref="taxon:43181"
#                           /sex="female"
#                           /tissue_type="larval and pupal exuviae"
#                           /country="Brazil: Sao Paulo state, Pariquera Acu
#                           municipality, Pariquera Mirim district"
#                           /lat_lon="24.71667 S 47.86667 W"
#                           /collection_date="15-May-2007"
#                           /collected_by="Sallum, M.A.M."
#                           /identified_by="Sallum, M.A.M."
#                           /PCR_primers="fwd_name: 5.8SF, fwd_seq:
#                           atcactcggctcgtggatcg, rev_name: 28SR, rev_seq:
#                           atgcttaaatttagggggtagtc"
#           misc_RNA        <1..>519
#                           /note="contains 5.8S ribosomal RNA, internal transcribed
#                           spacer 2, and 28S ribosomal RNA"
#      ORIGIN
#              1 cgtggatcga tgaagaccgc agctaaatgc gcgtcagaat gtgaactgca ggacacatga
#             61 acaccgacac gttgaacgca tattgcgcat tgcacgactc agtgcgatgt acacattttt
#            121 gagtgcccac attcaccgca gaaccaacta gcatagccgt cgaaagcttt gctgcgtact
#            181 gatgattggt tgaccatgtg ccaaccaagc attgaaggac tgtggcgtgg tgggtgcacc
#            241 gtgtgtgcgt cgttgcttaa tacgactcat tctctggtat cacatctgga gcgggctaac
#            301 cagtcacaat ccccagcgac atgtgcagat agccccgatg tggaggacca acatcctccc
#            361 tcaaagccag cccatgtgat acacaccaac agagagagac caaacgtacc ctgaagcaac
#            421 gtatgcgcac acgcgtgcaa ctcattgaag cgcacgatcg aaagagaacc gatcaagtgg
#            481 gcctcaaata atgtgtgact accccctaaa tttaagcat
#      //
```

Appendix 5.1B parse_order_from_endop_fastafile.pl

```
# CHANGES
# jan2010: print LOG
$database_file        = "database_fasta_Oct2012";
$key_file      = "key_Oct2012_Coleoptera";
$logfile       = "parse_order_from_endop_fastafile_LOG";

open(LOG , ">>$logfile") || die "cant open file\n";
open (IN, "$key_file") || die "cant open file\n";

while ($line= <IN>)
        {
        #    LZ1M4Mi7cal    Micropterix_calthella    41027    species    suborder:Zeugloptera
family:Micropterigidae genus:Micropterix species:Micropterix calthella

        if($line =~ /^(\S+)\s(\S+)\s(\d+)\s/)
                {
                # print
                $taxid = $3;
                $h{$taxid} = $1;
        #       print "taxid:$taxid\n";
                }else{
                #if($line =~ /^(\S+)/){die "BUG1\n$line\n"}
                }
        };

close(IN);



fasta_format_to_nexus();

exit;



sub fasta_format_to_nexus
        {
        print "sub fasta_format_to_nexus. reading $database_file. this may take a minute for
endopterygota\n";
        print LOG "sub fasta_format_to_nexus. reading $database_file.\n";

        my $current_seq_length;
        my $missing_data_character = "N";
        open(FASTA_IN,        $database_file)        ||        die        "Cant        open
input:$format_conversion_input_file.\n";
        my $file_as_string = ""; my @all_lines = ();
        while($line= <FASTA_IN>){$file_as_string .= $line};
        close(FASTA_IN);

        $file_as_string =~ s/\012\015?|\015\012?/\n/g;
        @all_lines = split />/, $file_as_string;

        print "$database_file in memory. reading seqs\n";
        for my $each_line(1 .. $#all_lines)
                {
                my $line = $all_lines[$each_line];

# >FJ041326 219766 Camponotus_femoratus Vespoidea Formicidae Formicinae unknown no_product
STUDY

                if ($line =~ /^(\S+)\s(\S+)\s.+\n/ )
                        {
                        $current_accession = $1;$taxid=$2;$line =~ s/^.+\n//;       $line     =~
s/\n//;$line =~ s/\r//;$current_seq_length = length($line);
                        }else{die "BUG2\n"};

                if(exists($h{$taxid}))
                        {
                        $key = $h{$taxid} . "_" . $current_accession;
                        $fasta_seqs{$key} = $line;
                        #print "$taxid exists\n";
                        }else{
                        # do nothing
                        #print "$taxid doesnt exist\n";
                        }


                }

        my @fasta_seqs_keys = keys %fasta_seqs;
        my $number_of_taxa = scalar @fasta_seqs_keys;
        print  "$number_of_taxa entries read into memory.\n";
        print  LOG "$number_of_taxa entries read into memory.\n";

        #############################################
        @fasta_seqs_keys = sort @fasta_seqs_keys;
############################################

        open(NEXUS_OUT    ,">$database_file.parsed")    ||    die    "cant    open    output
file:$format_conversion_output_file\n";

        for $i(0 .. $#fasta_seqs_keys)
                {
                $current_name = $fasta_seqs_keys[$i];
                $current_seq = $fasta_seqs{$current_name};
                # print "current_name:$current_name current_seq:$current_seq\n";

                if(length($current_name)<=1  ||  length($current_seq)<=1){print  "warning:  zero
length of current entry. quitting\n";die}

                print NEXUS_OUT ">$current_name\n$current_seq\n";
                };

        close(NEXUS_OUT);

        };

close(LOG);
```

Appendix 5.1C parse_taxa_from_fastafile.pl

Look for the taxon of interest in the file containing the key and then use that code to select all the sequences for that taxon; e.g. for Curculionoidea the code is CP1Cur3:

Example: perl parse_taxa_from_fastafile *database_fasta_Oct2012.txt CP1Cur3*

```
#!/usr/bin/perl

$input = $ARGV[0];
$parse_this_taxon = $ARGV[1];


$output = $input . "." . $parse_this_taxon;

print "you have chosen:\ninput file:$input
output file:$output
parse_this_taxon:$parse_this_taxon\n\n";


# globals
my $entry_counter = 0;
my $current_seq_length;
my @seqs;
my @ids;
my %seq_hash;
my %ids_r;


        # INPUT


        read_fasta();

print "end of script\n";

die;


####################################################################################
#
#
#
#
#
####################################################################################


sub read_fasta
        {
```

```
    print "subroutine to read fasta format.\nreading input file.\n";

    open(FASTA_IN, $input) || die "Cant open input:$input.\n";
    open(FASTA_OUT, ">$output") || die "Cant open output:$output.\n";



    # local variables:

    my $line;
    my $current_id;
    my $current_sequence;
    my $current_name;
    my $current_seq;
    my $number_seqs_printed =0;

    while($line = <FASTA_IN>)
        { #cb1
        $line =~ s/\n//;
        $line =~ s/\r//;

        if ($line=~/^\s{0,2}>\s{0,2}(.+)\s{0,4}$/ )
            {

            if ($entry_counter >= 1)
                {
                $current_seq_length=length($current_sequence);
                if($current_id=~/$parse_this_taxon/)
                    {
                    #       print       "cid:$current_id
ptt:$parse_this_taxon\n";
                    print                           FASTA_OUT
">$current_id\n$current_sequence\n";
                    $number_seqs_printed++;
                    };

                };

            $entry_counter  ++;#  print  "ec:$entry_counter  line:$line
cid:$current_id\n";
            $current_id = $1;
            $current_id =~ s/(\S+)\s\d{1,4}\s+bp\s*$/$1/; # remove 688 bp
from end of id if present
            $current_sequence = "";
            }else{
            $current_sequence = $current_sequence . $line;
            if ($line=~/>/){print "WARNING: unreadable id line. quitting.
CHECK YOUR INPUT FILE IS RIGHT\nline:$line\n";die};
            }

        }; #cb1

if($current_id=~/$parse_this_taxon/)
    {
    print "cid:$current_id ptt:$parse_this_taxon\n";
    print FASTA_OUT ">$current_id\n$current_sequence\n";
    $number_seqs_printed++;
    };
```

```
               close (FASTA_IN);
               close (FASTA_OUT);

               print "$number_seqs_printed printed out of $entry_counter entries.\n";

               };


        ####################################################################################
        #
        #
        #
        #
        #
        ####################################################################################
```

Appendix 5.1D parse_blast_output.pl

Identifies the position of the blasted sequences within the sequence database

Example: perl parse_blast_output.pl *cox1_blast_output.txt*

This command will generate a file with the extension .parsed that is subsequently used to retrieve the sequences

```
#!/usr/bin/perl
use lib "BioPerl-1.6.1";
use Bio::SearchIO;
use Bio::AlignIO;

$current_blast_outfile = $ARGV[0];
$current_outputfile = $current_blast_outfile . ".parsed";

print "Script name:parse_blast_output.pl\n";

$which_database = "msc_practical_database.txt";
$entry_length_cutoff = 0.2;

do_blast_searches();

print "End of script:parse_blast_output.pl\n";
die;

sub do_blast_searches {

my %start_hash = (); # these must be reset for each search
my %end_hash = ();
my %strand_hash = ();
my $query_length = "";

open(OUT3, ">$current_outputfile") || die "cant open output";



        my $in = new Bio::SearchIO      (       -format => "blast",
                                                -file =>        $current_blast_outfile
                                        );

                # use BioPerl parser for the blast output
        my $count_hits2=0;print "\nhit_no current_id start end strand\n";
        my $number_queries=0;my $query_length_sum = 0;

        while (my $result = $in->next_result)
                {
                $query_length        =        $result->query_length;$query_length_sum        +=
$query_length;$number_queries++;
                while (my $hit = $result->next_hit)
                        {
                        while (my $hsp = $hit->next_hsp)
                                {

                                my   $current_id   =   $hit->name;my   $current_strand   =   $hsp-
>strand(hit);
                                my  $current_start  =  $hsp->start(hit);my  $current_end  =  $hsp-
>end(hit);
                                $count_hits2++;if($count_hits2=~    /00$/){print    "$count_hits2
$current_id $current_start $current_end $current_strand\n"};

                                if($current_strand == 1)
                                        {
                                # some loci are submitted in opposite strand, so record strand
as well as start and end position of hit.
                                #  the  start  and  end  position  of  the  hit  to  a  given  id  are
recorded seperatly.
                                #  as  multiple  blast  searches  are  being  performed,  the  lowest
start position (of all hits to the id)
                                # and highest end position is taken.

                                        if(length($start_hash{$current_id}) >= 1)
                                                {if($start_hash{$current_id}                >=
$current_start){$start_hash{$current_id} = $current_start};
                                                }else{$start_hash{$current_id} = $current_start};

                                        if(length($end_hash{$current_id}) >= 1)
                                                {if($end_hash{$current_id}                  <=
$current_end){$end_hash{$current_id} = $current_end};
                                                }else{$end_hash{$current_id} = $current_end};
                                        $strand_hash{$current_id} = 1;
                                        }else{

                                        if(length($start_hash{$current_id}) >= 1)
                                                {if($start_hash{$current_id}                >=
$current_start){$start_hash{$current_id} = $current_start};
                                                }else{$start_hash{$current_id} = $current_start};

                                        if(length($end_hash{$current_id}) >= 1)
```

```
                                            {if($end_hash{$current_id} <=
$current_end){$end_hash{$current_id} = $current_end};
                                            }else{$end_hash{$current_id} = $current_end};

                                     $strand_hash{$current_id} = 2;
                                     };

                              };
                      };
              };
$query_length = $query_length_sum / $number_queries;$query_length = int($query_length);
print "mean_query_length:$query_length\n";
my $hit_limit = $query_length * $entry_length_cutoff;
print "results file parsed, $count_hits2 hits (total, incl repeats). \nfetching sequences with
hits > $hit_limit (query_length:$query_length)\n";



                              # now go through all the recorded hits and print if > hit_limit
my $count_hits=0; my $count_hits3=0; # start / end position and strand are taken from hash for
each id, then ncbi_fastacmd takes the sequence from
my @all_ids = keys %start_hash;       # a local blast database (use -o T option when making
this)
foreach my $current_id(@all_ids)# fastacmd prints extra bits (>lcl|sequence_id No definition
line found), 2 lines below remove these
       {
       my $current_start = $start_hash{$current_id};
       my $current_end = $end_hash{$current_id};
       my $current_length = $current_end-$current_start;
       if ($current_length >= $hit_limit)
               {
               $count_hits++;
               if($count_hits=~ /0$/){print "$count_hits $current_id $current_start
$current_end\n"};


               my $current_strand = $strand_hash{$current_id}; # -L
$current_start,$current_end
               my $for_log = $current_id . "_" . $current_start . "_" . $current_end . "_" .
$current_strand; print LOG "$for_log ";
               print OUT3 "$current_id $current_strand $current_start $current_end\n";
#              my $entry_retrieved =`fastacmd -d $which_database -s $current_id -S
$current_strand -L $current_start,$current_end`;
#              $entry_retrieved =~ s/>lcl\|/>/;
#              $entry_retrieved =~ s/\:\d+.\d+\s+No definition line found//;
#              $entry_retrieved =~ s/No definition line found//;

#              print OUT1 $entry_retrieved;

               }else{$count_hits3++;};
       };

print "found $count_hits hits longer than hit_limit($hit_limit). $count_hits3 too
short\n\n*****************************\n\n";

close(OUT3);

       }; # end of sub (do blast searches)
```

Appendix 5.1 E retrieve_sequences.pl

This script retrieves the matching sequences from the database.

Example: perl retrieve_sequences.pl *cox1_blast_output.txt.parsed*

```
#!/usr/bin/perl

        # when the command is typed into the shell
        # to run perl and this script,
        # anything that is typed after will go into
        # the $infile variable due to the following line

$infile = $ARGV[0];

        # the output file will be named by appending
        # ( . = concatenate) ".retrieved" to the infile name

$outfile = $infile . ".retrieved";

        # comments are preceeded by a hash symbol (#),
        # comments are ignored by the command interpretor,
        # they are just for the benefit of users
        # (the exeption is the very first line, which
        # looks to be 'commented-out' but isnt)

        # first open input and output files,
        # file handle (IN) is followed by name of file (blastout_parsed)

open(IN, $infile)             or die "cant open infile:$infile\n";
open(OUT1, ">$outfile")       or die "cant open outfile:$outfile\n";

$which_database = "database_fasta_Oct2012.parsed.CP1Cur3";

        # here a while loop is used.
        # everything between the curly braces {} is repeated for each loop.
        # in this case the loop is performed
        # for each line of the input file,
        # in other words we are scanning each line of the
        # input file and running ~10 commands for each line


while ($line = <IN>)
        {

        # the line contains a regular expression
        # between the (/ ... /).
        # regular expression use shorthand representations
        # for different characters we may want to look for
        # \S means a character that is not a space
        # (ie letter or number), \s means space character,
        # + means 1 or more characters.


        if($line =~ /(\S+)\s(\S+)\s(\S+)\s(\S+)\s/)
        {

        # when the condition is met (a line is scanned
        # containing non-space characters followed by space followed
        # by non-space charaters etc),
        # the brackets around the characters (\S+) place whatever is present
        # at that position into a numbered variable,
        # $1 for the first brackets, $2 for the second etc.

        $current_id    = $1;
        $current_strand = $2;
        $current_start        = $3;
        $current_end   = $4;

        # now we have the desired characters into variables,
        # we will invoke fastacmd, the result of which is placed
        # into the $entry_retrieved variable,
        # which is subsequently printed.

        $entry_retrieved =`fastacmd -d $which_database -s $current_id -S $current_strand -L
$current_start,$current_end`;

        # so for example if the line read:
        # LG1N2HDApZy3Zy4Zy5Zy7nocnoc_AJ830850 1 1 1884
        # this would be invoked:
        # fastacmd -d lepidoptera_fasta_coded -s LG1N2HDApZy3Zy4Zy5Zy7nocnoc_AJ830850 -S 1 -L
1,1884

        $entry_retrieved =~ s/>lcl\|/>/;
        $entry_retrieved =~ s/\:\d+.\d+\s+No definition line found//;
        $entry_retrieved =~ s/No definition line found//;

        print OUT1 $entry_retrieved;

        };

        };


close(IN);
close(OUT1);
```

Appendix 5.1 F perl one_per_species.pl

This script selects only one sequence (the longest available) per species

Example: perl one_per_species.pl *cox1_blast_output.parsed.retrieved*

```
#!/usr/local/bin/perl -w

#takes a file containing fasta sequences and filters it leaving only one sequence per species.
The longest
#available sequence is retained.

$infile = $ARGV[0];


open(IN, $infile) || die;
open(OUT, ">$infile.one_per_species") || die;
$/ = ">";

@allseqs = <IN>;
close IN;


shift @allseqs;
      foreach $next(@allseqs){
        $next =~ s/>//;
        $safe = $next;
#       $next =~ m/(.+)_[^_]+$/;   #matches codename of sequence
        $next =~ m/(.+)_/;   #matches codename of sequence

        $name = $1;
#       print "$name\n";

         $sequence = extract_sequence_from_fasta_data($safe);
        #print "$sequence\n";
         $len = length($sequence);


#       Take %names as a hash to store the length
#       Take %seq as a hash to store the sequence

        if (exists($names{$name})) { # have we seen this sequence already?
               if($names{$name} < $len) {    # if yes, is the length longer than what we've
already got?
                       $seq{$name} = $safe;   # if yes, store the new sequence
                       $names{$name} = $len; # and store the new length
               }
               else {                        # otherwise

                       next;                         # skip to the next one
               }
        }

        else {
               $names{$name} = $len;         # if we haven't seen the sequence before
               $seq{$name} = $safe;          # put it in the $names and $seq hash
                #print OUT ">$safe";
        }
        }
# Now we run another loop to print out the contents of %seq


        foreach $sequence(values %seq) {
               #$remove = ">$sequence";
               #$remove =~ s/>.*>//gs;
               #$printing = $remove;
               print OUT ">$sequence";
        }

# extract_sequence_from_fasta_data
#
# A subroutine to extract FASTA sequence data

sub extract_sequence_from_fasta_data {

    my(@fasta_file_data) = @_;

    use strict;
    use warnings;

    my $sequence = '';

    foreach my $line (@fasta_file_data) {

        # discard blank line
        if ($line =~ /^\s*$/) {
            next;

        # discard comment line
        } elsif($line =~ /^\s*#/) {
            next;

        # discard fasta header line
        } elsif($line =~ /^>/) {
            next;

        # keep line, add to sequence string
        } else {
            $sequence .= $line;
        }
    }

    # remove non-sequence data whitespace from $sequence string
    $sequence =~ s/\w+\s{1}\w+\s{1}//;
    $sequence =~ s/\s//g;

    return $sequence;
}

print "script fininshed\n";
```

Appendix 5.2 GenBank-obtained broad-nosed weevil taxa showing loci available and original first author of the source sequences. GenBank codes are given for all available sequences followed by the sequence length in brackets.

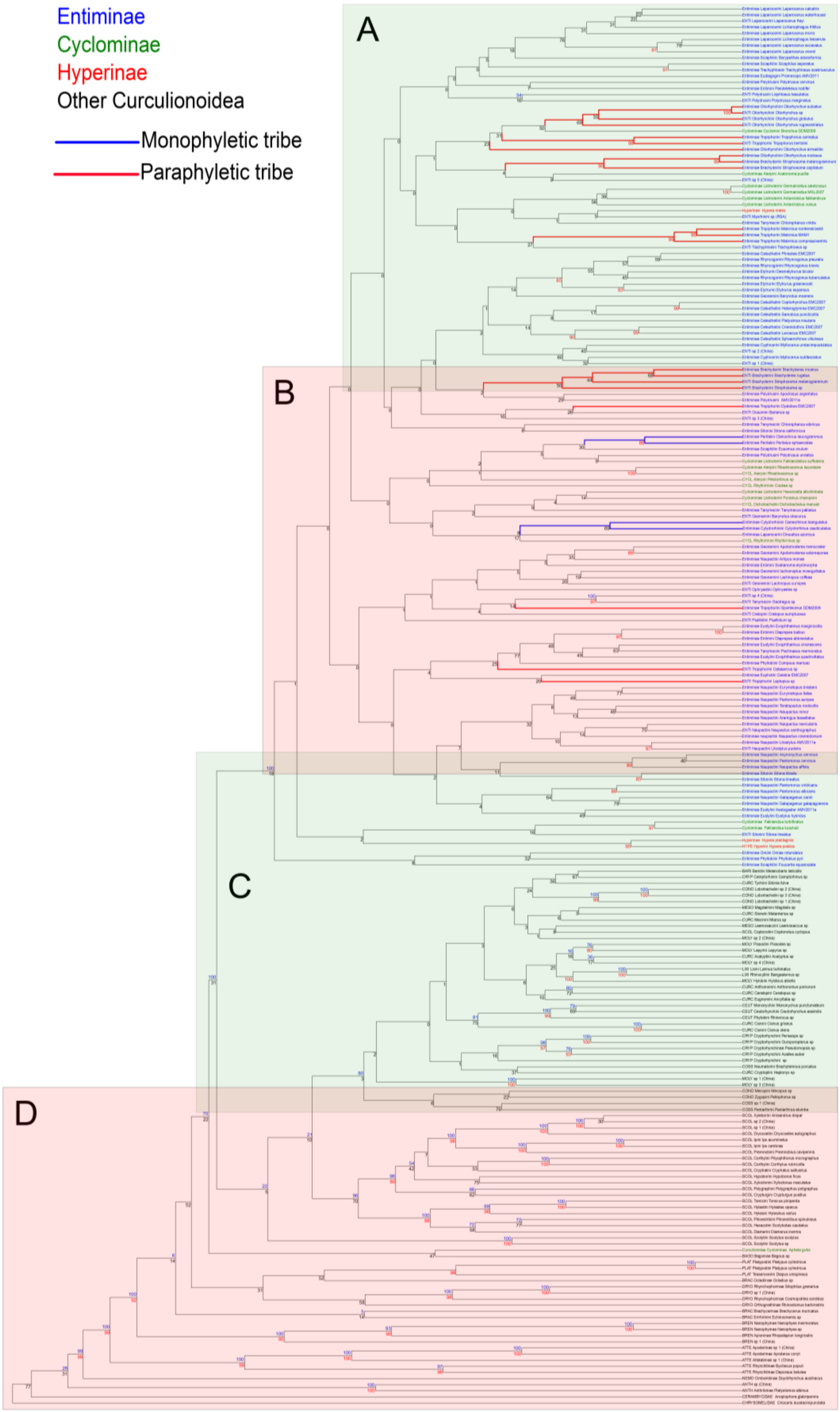
Subfamily	Tribe	Species	*cox1* 'barcode'	*rrnL*	Source author
Cyclominae	Aterpini	*Acalonoma pusilla*		AJ495588 (558 bp)	Hundsdoerfer
Cyclominae	Aterpini	*Rhadinosomus lacordairei*		AJ495587 (481 bp)	Hundsdoerfer
Cyclominae	Cyclomini	*Bronchus DDM2009*	FJ867830 (649 bp)		McKenna
Cyclominae	Listroderini	*Antarctobius falklandicus*		FM994747 (511 bp)	Papadopoulou
Cyclominae	Listroderini	*Antarctobius vulsus*		EF213957 (511bp)	Papadopoulou
Cyclominae	Listroderini	*Falklandiellus suffodens*		FM994723 (493 bp)	Papadopoulou
Cyclominae	Listroderini	*Falklandius kuscheli*		EF213968 (512 bp)	Papadopoulou
Cyclominae	Listroderini	*Falklandius turbificatus*		EF213990 (512 bp)	Papadopoulou
Cyclominae	Listroderini	*Germainiellus MSL2007*		EF213960 (513 bp)	Papadopoulou
Cyclominae	Listroderini	*Germainiellus salebrosus*		FM994694 (513 bp)	Papadopoulou
Cyclominae	Listroderini	*Haversiella albolimbata*		FM994708 (514 bp)	Papadopoulou
Cyclominae	Listroderini	*Puranius championi*		FM994818 (514 bp)	Papadopoulou
Cyclominae	Rhythirrinini	*Aphela gotoi*	AB661842 (584 bp)		Kudo
Entiminae	Brachyderini	*Brachyderes incanus*		AJ495503 (478 bp)	Hundsdoerfer
Entiminae	Brachyderini	*Strophosoma capitatum*		AJ495504 (479 bp)	Hundsdoerfer
Entiminae	Brachyderini	*Strophosoma melanogrammum*	AY196875 (552 bp)	AJ495505 (478 bp)	Stenberg/Hundsdoerfer
Entiminae	Celeuthetini	*Cnemidothrix EMC2007*	EF575506 (262 bp)	EF606992 (380 bp)	Claridge
Entiminae	Celeuthetini	*Coptorhynchus EMC2007*	EF575499 (336 bp)	EF606985 (402 bp)	Claridge
Entiminae	Celeuthetini	*Heteroglymma EMC2007*	EF575502 (401 bp)	EF606988 (401 bp)	Claridge
Entiminae	Celeuthetini	*Levoecus EMC2007*	EF575504 (414 bp)	EF606990 (403 bp)	Claridge
Entiminae	Celeuthetini	*Phraotes EMC2007*	EF575508 (335 bp)		Claridge
Entiminae	Celeuthetini	*Platysimus insularis*	EF575501 (336 bp)	EF606987 (394 bp)	Claridge
Entiminae	Celeuthetini	*Samobius puncticollis*	EF575500 (301 bp)	EF606986 (401 bp)	Claridge
Entiminae	Celeuthetini	*Sphaerorhinus villulosus*	EF575507 (414 bp)	EF606993 (382 bp)	Claridge
Entiminae	Cylydrorhinini	*Caneorhinus biangulatus*		EF213991 (524 bp)	Papadopoulou
Entiminae	Cylydrorhinini	*Cylydrorhinus caudiculatus*		EF214094 (510 bp)	Papadopoulou

Entiminae	Cyphicerini	*Myllocerus subfasciatus*	JQ280416 (613 bp)		Nagesh
Entiminae	Cyphicerini	*Myllocerus undecimpustulatus*	JX467534 (649 bp)		Geetha
Entiminae	Elytrurini	*Desmelytrurus bicolor*	EF575513 (413 bp)	EF606998 (372 bp)	Claridge
Entiminae	Elytrurini	*Elytrurus expansus*	EF575509 (414 bp)	EF606994 (401 bp)	Claridge
Entiminae	Elytrurini	*Elytrurus greenwoodi*	EF575515 (334 bp)	EF607000 (400 bp)	Claridge
Entiminae	Eudiagogini	*Promecops AMV2011*	HQ891477 (728 bp)		Mazo-Vargas
Entiminae	Eupholini	*Celebia EMC2007*	EF575490 (414 bp)	EF606976 (394 bp)	Claridge
Entiminae	Eustylini	*Compsus maricao*	HQ891431 (748 bp)		Mazo-Vargas
Entiminae	Eustylini	*Diaprepes abbreviatus*	JF302927 (748 bp)	EF042125 (436 bp)	Mazo-Vargas/Ascunce
Entiminae	Eustylini	*Diaprepes balloui*	HQ891433 (748 bp)	EF042127 (437 bp)	Mazo-Vargas/Ascunce
Entiminae	Eustylini	*Eustylus hybridus*	HQ891445 (748 bp)		Mazo-Vargas
Entiminae	Eustylini	*Exophthalmus cinerascens*	HQ891446 (748 bp)		Mazo-Vargas
Entiminae	Eustylini	*Exophthalmus marginicollis*	HQ891447 (748 bp)		Mazo-Vargas
Entiminae	Eustylini	*Exophthalmus quadrivittatus*	HQ891448 (748 bp)		Mazo-Vargas
Entiminae	Eustylini	*Scelianoma elydimorpha*	HQ891478 (748 bp)		Mazo-Vargas
Entiminae	Eustylini	*Xestogaster AMV2011a*	HQ891479 (748 bp)		Mazo-Vargas
Entiminae	Geonemini	*Apotomoderes menocrater*	HQ891426 (748 bp)		Mazo-Vargas
Entiminae	Geonemini	*Apotomoderes sotomayorae*	HQ891427 (748 bp)		Mazo-Vargas
Entiminae	Geonemini	*Barynotus moerens*		AJ495512 (479 bp)	Hundsdoerfer
Entiminae	Geonemini	*Ischionoplus niveoguttatus*	HQ891462 (748 bp)		Mazo-Vargas
Entiminae	Geonemini	*Lachnopus coffeae*	HQ891463 (748 bp)		Mazo-Vargas
Entiminae	Laparocerini	*Drouetius azoricus*		EF583417 (428 bp)	Machado
Entiminae	Laparocerini	*Laparocerus calcatrix*		EF583394 (427 bp)	Machado
Entiminae	Laparocerini	*Laparocerus excavatus*		EF583431 (427 bp)	Machado
Entiminae	Laparocerini	*Laparocerus morio*		EF583425 (427 bp)	Machado
Entiminae	Laparocerini	*Laparocerus oromii*	FJ716583 (649 bp)	FJ716536 (492 bp)	Machado
Entiminae	Laparocerini	*Laparocerus waterhousei*		EF583414 (427 bp)	Machado
Entiminae	Laparocerini	*Lichenophagus fritillus*		EF583433 (427 bp)	Machado
Entiminae	Laparocerini	*Lichenophagus tesserula*		EF583434 (427 bp)	Machado
Entiminae	Naupactini	*Aramigus tessellatus*	AY790875 (511 bp)		Scataglini
Entiminae	Naupactini	*Artipus monae*	HQ891428 (748 bp)		Mazo-Vargas
Entiminae	Naupactini	*Asynonychus cervinus*	AY790876 (541 bp)		Scataglini
Entiminae	Naupactini	*Eurymetopus birabeni*	AY790877 (480 bp)		Scataglini
Entiminae	Naupactini	*Eurymetopus fallax*	AY790878 (511 bp)		Scataglini

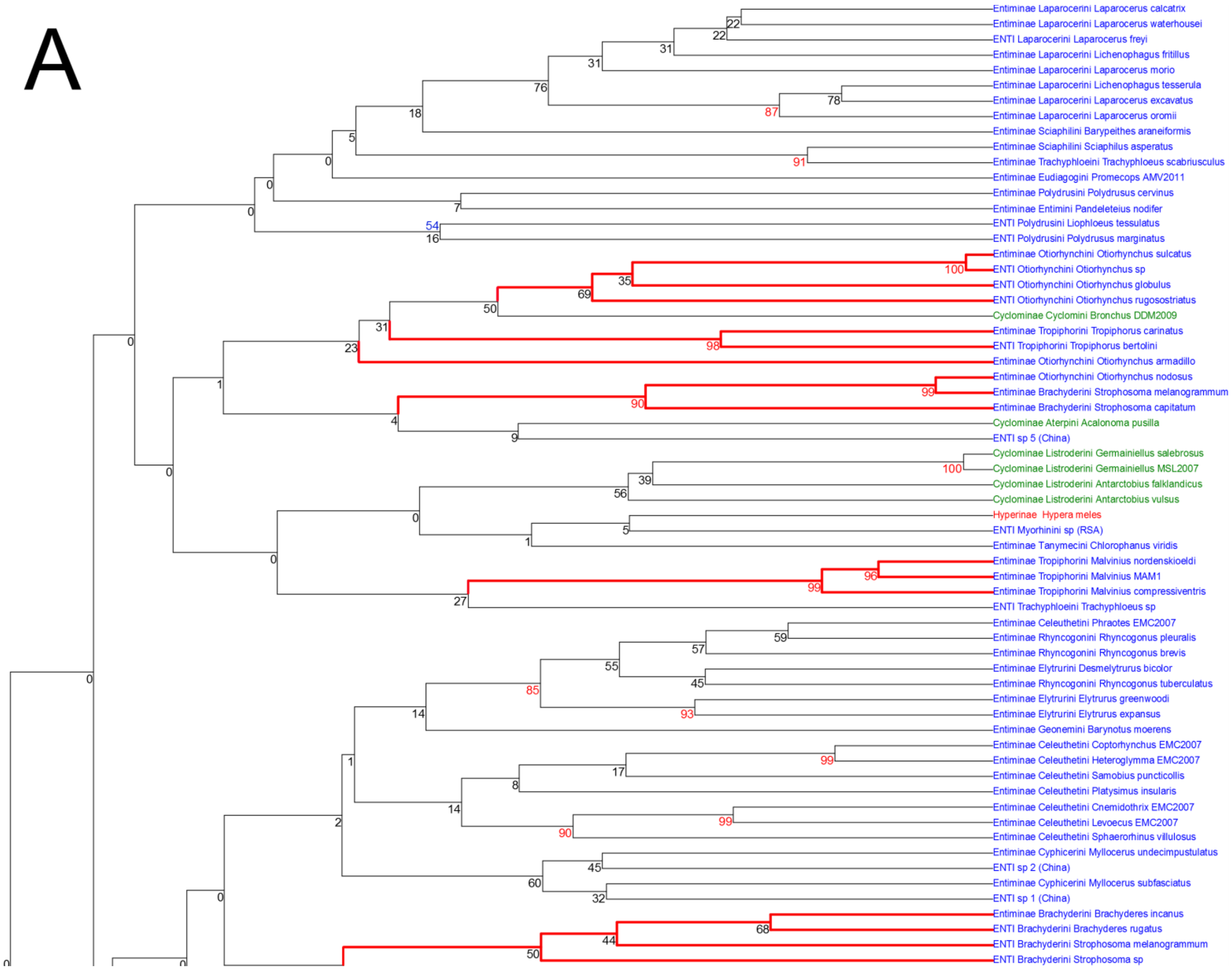
Entiminae	Naupactini	*Galapaganus caroli*	AF211486 (550 bp)	EF606979 (402 bp)	Sequeira
Entiminae	Naupactini	*Galapaganus galapagoensis*	AF015914 (561 bp)		Sequeira
Entiminae	Naupactini	*Litostylus AMV2011a*	HQ891470 (748 bp)		Mazo-Vargas
Entiminae	Naupactini	*Naupactus affinis*	GU727685 (544 bp)		Rodriguero
Entiminae	Naupactini	*Naupactus cinereidorsum*	AY770388 (541 bp)		Scataglini
Entiminae	Naupactini	*Naupactus minor*	AY790881 (511 bp)		Scataglini
Entiminae	Naupactini	*Naupactus navicularis*	AY790882 (523 bp)		Scataglini
Entiminae	Naupactini	*Pantomorus albicans*	GU565278 (558 bp)		Rosas
Entiminae	Naupactini	*Pantomorus auripes*	AY770383 (541 bp)		Scataglini
Entiminae	Naupactini	*Pantomorus cervinus*	AY790876 (541 bp)	EF606980 (401 bp)	Scataglini/Claridge
Entiminae	Naupactini	*Pantomorus viridicans*	GU565277 (559 bp)		Rosas
Entiminae	Naupactini	*Teratopactus nodicollis*	AY770387 (511 bp)		Scataglini
Entiminae	Omiini	*Omias rotundatus*		AJ495515 (479 bp)	Hundsdoerfer
Entiminae	Otiorhynchini	*Otiorhynchus armadillo*		AJ495480 (480 bp)	Hundsdoerfer
Entiminae	Otiorhynchini	*Otiorhynchus nodosus*	AY196876 (558 bp)		Stenberg
Entiminae	Otiorhynchini	*Otiorhynchus sulcatus*	EF575489 (299 bp)	AJ495482 (480 bp)	Claridge/Hundsdoerfer
Entiminae	Peritelini	*Ctenochirus leucogrammus*		AJ495484 (481 bp)	Hundsdoerfer
Entiminae	Peritelini	*Peritelus sphaeroides*		AJ495485 (482 bp)	Hundsdoerfer
Entiminae	Phyllobiini	*Phyllobius pyri*		AJ495491 (478 bp)	Hundsdoerfer
Entiminae	Polydrusini	*Apodrosus argentatus*	HQ891422 (748 bp)		Mazo-Vargas
Entiminae	Polydrusini	*Polydrusini AMV2011a*	HQ891476 (748 bp)		Mazo-Vargas
Entiminae	Polydrusini	*Polydrusus cervinus*	HQ883653 (550 bp)	AJ495494 (479 bp)	Jordal/Hundsdoerfer
Entiminae	Polydrusini	*Polydrusus undatus*		AJ495496 (479 bp)	Hundsdoerfer
Entiminae	Rhyncogonini	*Rhyncogonus brevis*	EF575535 (412 bp)	EF607019 (402 bp)	Claridge
Entiminae	Rhyncogonini	*Rhyncogonus pleuralis*	EF575525 (414 bp)	EF607009 (402 bp)	Claridge
Entiminae	Rhyncogonini	*Rhyncogonus tuberculatus*	EF575546 (407 bp)	EF607030 (402 bp)	Claridge
Entiminae	Sciaphilini	*Barypeithes araneiformis*		AJ495500 (480 bp)	Hundsdoerfer
Entiminae	Sciaphilini	*Eusomus ovulum*		AJ495499 (479 bp)	Hundsdoerfer
Entiminae	Sciaphilini	*Foucartia squamulata*		AJ495501 (479 bp)	Hundsdoerfer
Entiminae	Sciaphilini	*Sciaphilus asperatus*		AJ495502 (481 bp)	Hundsdoerfer
Entiminae	Sitonini	*Sitona californicus*	EF575488 (336 bp)	EF606974 (402 bp)	Claridge
Entiminae	Sitonini	*Sitona lineatus*		AJ495508 (484 bp)	Hundsdoerfer
Entiminae	Sitonini	*Sitona tibialis*		AJ495511 (482 bp)	Hundsdoerfer
Entiminae	Tanymecini	*Chlorophanus sibiricus*	HQ883651 (550 bp)		Jordal

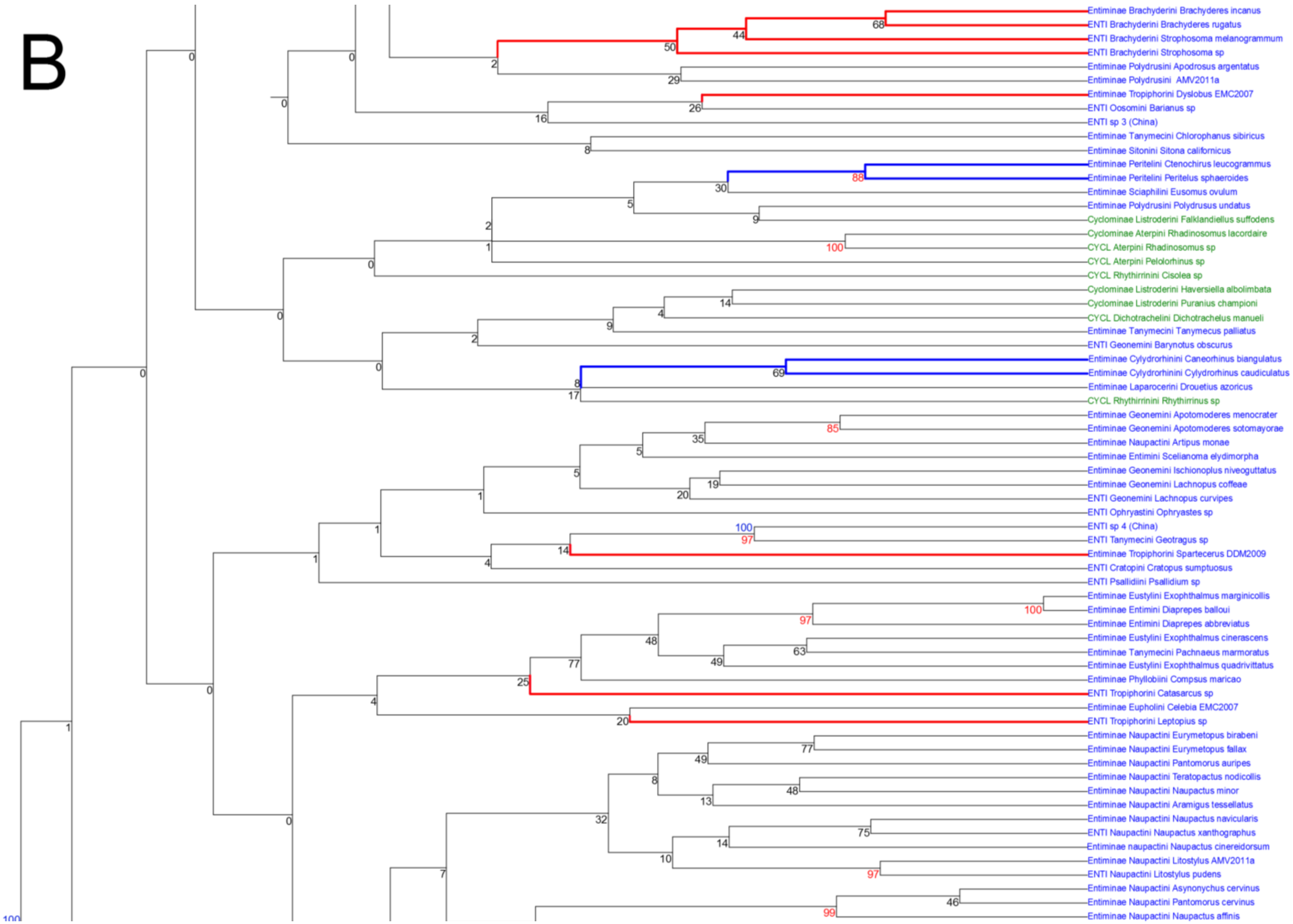
Entiminae	Tanymecini	*Chlorophanus viridis*		AJ495506 (479 bp)	Hundsdoerfer
Entiminae	Tanymecini	*Pachnaeus marmoratus*	HQ891472 (748 bp)		Mazo-Vargas
Entiminae	Tanymecini	*Pandeleteius nodifer*	HQ891474 (748 bp)		Mazo-Vargas
Entiminae	Tanymecini	*Tanymecus palliatus*		AJ495507 (481 bp)	Hundsdoerfer
Entiminae	Trachyphloeini	*Trachyphloeus scabriusculus*		AJ495516 (461 bp)	Hundsdoerfer
Entiminae	Tropiphorini	*Dyslobus EMC2007*	EF575495 (414 bp)	EF606981 (395 bp)	Claridge
Entiminae	Tropiphorini	*Malvinius compressiventris*		FM994840 (512 bp)	Papadopoulou
Entiminae	Tropiphorini	*Malvinius MAM1*		FM994730 (514 bp)	Papadopoulou
Entiminae	Tropiphorini	*Malvinius nordenskioeldi*		EF214001 (514 bp)	Papadopoulou
Entiminae	Tropiphorini	*Spartecerus DDM2009*	FJ867826 (649 bp)		McKenna
Entiminae	Tropiphorini	*Tropiphorus carinatus*		AJ495488 (477 bp)	Hundsdoerfer
Hyperinae	Hyperinae	*Hypera meles*		AJ495526 (478 bp)	Hundsdoerfer
Hyperinae	Hyperinae	*Hypera plantaginis*		JN163953 (113 bp)	Haran

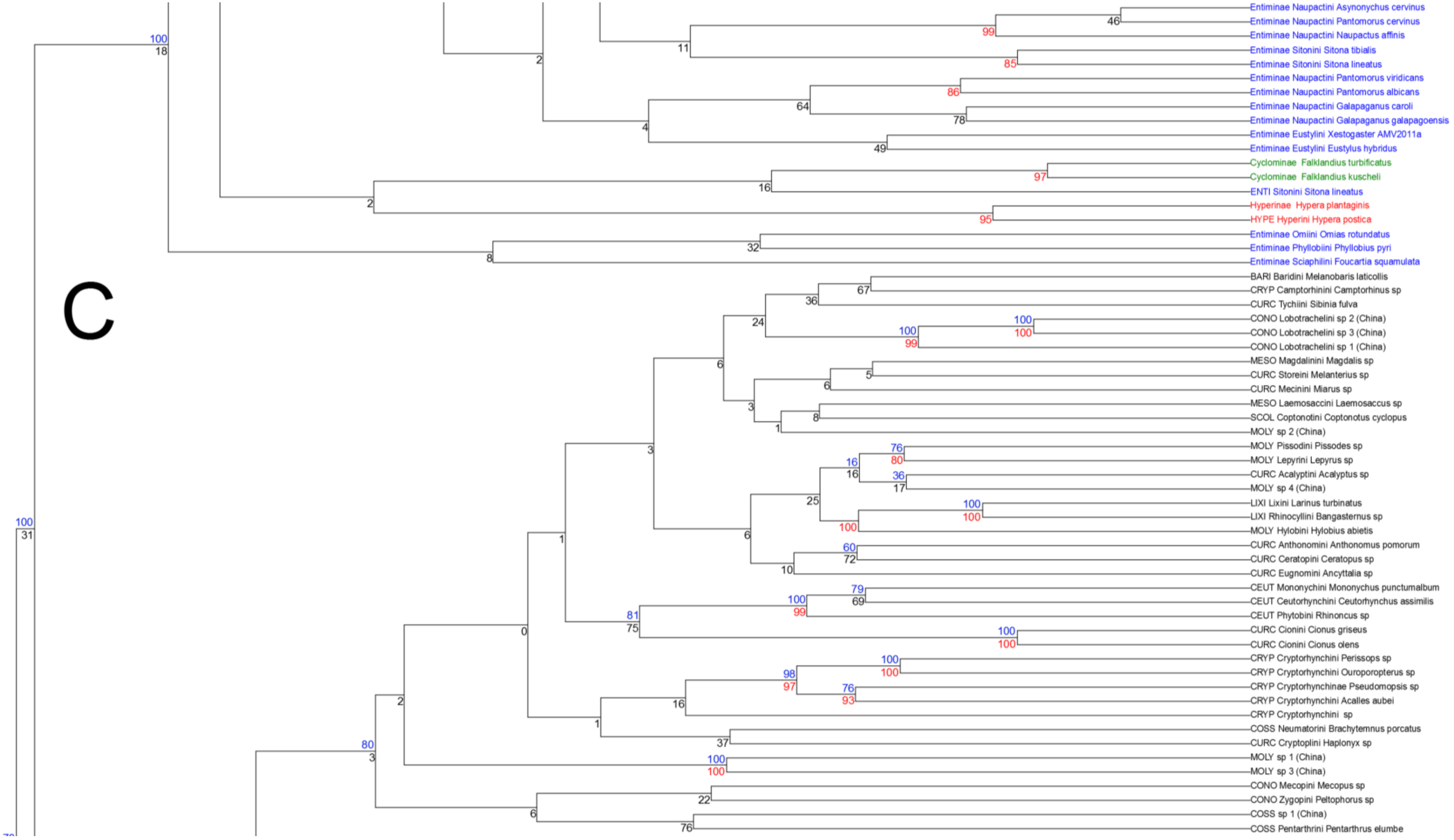
Appendix 5.3 A-D (shown in four overlapping sections on the following five pages) Unconstrained maximum likelihood tree of combined mitogenomic and GenBank sequences, rooted at Chrysomelidae. Bootstrap nodal supports are indicated below nodes, with those of 80% and higher indicated in red. Bootstrap values above nodes (in blue) are shown for consistent nodes in the mitogenomic-only data ML tree from Chapter 3. Node labelled B represents the 'long-nosed weevils (Curculionidae *s.str.* minus broad-nosed weevils and Platypodinae). In green is highlighted the aberrant position of *Aphela gotoi*, currently classified in Cyclominae. Taxa represented by mitogenomic sequences have family and subfamily codes prefixes as follows: Anthribidae (ANTH), Attelabidae (ATTE), Brachyceridae (BRAC), Brentidae (BREN), Dryophthoridae (DRYO), Nemonychidae (NEMO), Bagoinae (BAGO), Baridinae (BARI), Ceutorhynchinae (CEUT), Conoderinae (CONO), Cossoninae (COSS), Cryptorhynchinae (CRYP), Curculioninae (CURC), Lixinae (LIXI), Mesoptillinae, (MESO), Molytinae (MOLY), Platypodinae (PLAT) and Scolytinae (SCOL).

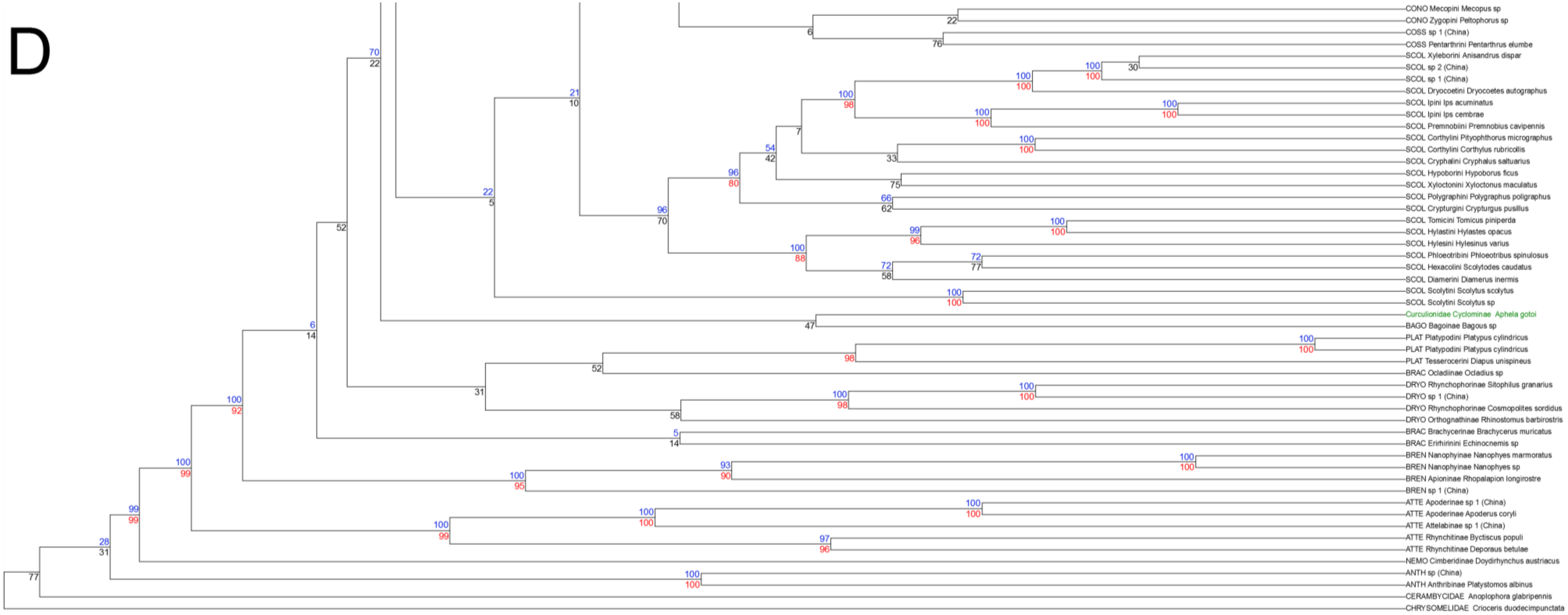


Appendix 5.3 Overview of tree, indicating sections **A-D**, enlarged on the following four pages


A
Entiminae Laparocerini Laparocerus calcatrix
Entiminae Laparocerini Laparocerus waterhousei
ENTI Laparocerini Laparocerus freyi
Entiminae Laparocerini Lichenophagus fritillus
Entiminae Laparocerini Laparocerus morio
Entiminae Laparocerini Lichenophagus tesserula
Entiminae Laparocerini Laparocerus excavatus
Entiminae Laparocerini Laparocerus oromii
Entiminae Sciaphilini Barypeithes araneiformis
Entiminae Sciaphilini Sciaphilus asperatus
Entiminae Trachyphloeini Trachyphloeus scabriusculus
Entiminae Eudiagogini Promecops AMV2011
Entiminae Polydrusini Polydrusus cervinus
Entiminae Entimini Pandeleteius nodifer
ENTI Polydrusini Liophloeus tessulatus
ENTI Polydrusini Polydrusus marginatus
Entiminae Otiorhynchini Otiorhynchus sulcatus
ENTI Otiorhynchini Otiorhynchus sp
ENTI Otiorhynchini Otiorhynchus globulus
ENTI Otiorhynchini Otiorhynchus rugosostriatus
Cyclominae Cyclomini Bronchus DDM2009
Entiminae Tropiphorini Tropiphorus carinatus
ENTI Tropiphorini Tropiphorus bertolini
Entiminae Otiorhynchini Otiorhynchus armadillo
Entiminae Otiorhynchini Otiorhynchus nodosus
Entiminae Brachyderini Strophosoma melanogrammum
Entiminae Brachyderini Strophosoma capitatum
Cyclominae Aterpini Acalonoma pusilla
ENTI sp 5 (China)
Cyclominae Listroderini Germainiellus salebrosus
Cyclominae Listroderini Germainiellus MSL2007
Cyclominae Listroderini Antarctobius falklandicus
Cyclominae Listroderini Antarctobius vulsus
Hyperinae Hypera meles
ENTI Myorhinini sp (RSA)
Entiminae Tanymecini Chlorophanus viridis
Entiminae Tropiphorini Malvinius nordenskioeldi
Entiminae Tropiphorini Malvinius MAM1
Entiminae Tropiphorini Malvinius compressiventris
ENTI Trachyphloeini Trachyphloeus sp
Entiminae Celeuthetini Phraotes EMC2007
Entiminae Rhyncogonini Rhyncogonus pleuralis
Entiminae Rhyncogonini Rhyncogonus brevis
Entiminae Elytrurini Desmelytrurus bicolor
Entiminae Rhyncogonini Rhyncogonus tuberculatus
Entiminae Elytrurini Elytrurus greenwoodi
Entiminae Elytrurini Elytrurus expansus
Entiminae Geonemini Barynotus moerens
Entiminae Celeuthetini Coptorhynchus EMC2007
Entiminae Celeuthetini Heteroglymma EMC2007
Entiminae Celeuthetini Samobius puncticollis
Entiminae Celeuthetini Platysimus insularis
Entiminae Celeuthetini Cnemidothrix EMC2007
Entiminae Celeuthetini Levoecus EMC2007
Entiminae Celeuthetini Sphaerorhinus villulosus
Entiminae Cyphicerini Myllocerus undecimpustulatus
ENTI sp 2 (China)
Entiminae Cyphicerini Myllocerus subfasciatus
ENTI sp 1 (China)
Entiminae Brachyderini Brachyderes incanus
ENTI Brachyderini Brachyderes rugatus
ENTI Brachyderini Strophosoma melanogrammum
ENTI Brachyderini Strophosoma sp


B
Entiminae Brachyderini Brachyderes incanus
ENTI Brachyderini Brachyderes rugatus
ENTI Brachyderini Strophosoma melanogrammum
ENTI Brachyderini Strophosoma sp
Entiminae Polydrusini Apodrosus argentatus
Entiminae Polydrusini AMV2011a
Entiminae Tropiphorini Dyslobus EMC2007
ENTI Oosomini Barianus sp
ENTI sp 3 (China)
Entiminae Tanymecini Chlorophanus sibiricus
Entiminae Sitonini Sitona californicus
Entiminae Peritelini Ctenochirus leucogrammus
Entiminae Peritelini Peritelus sphaeroides
Entiminae Sciaphilini Eusomus ovulum
Entiminae Polydrusini Polydrusus undatus
Cyclominae Listroderini Falklandiellus suffodens
Cyclominae Aterpini Rhadinosomus lacordaire
CYCL Aterpini Rhadinosomus sp
CYCL Aterpini Pelolorhinus sp
CYCL Rhythirrinini Cisolea sp
Cyclominae Listroderini Haversiella albolimbata
Cyclominae Listroderini Puranius championi
CYCL Dichotrachelini Dichotrachelus manueli
Entiminae Tanymecini Tanymecus palliatus
ENTI Geonemini Barynotus obscurus
Entiminae Cylydrorhinini Caneorhinus biangulatus
Entiminae Cylydrorhinini Cylydrorhinus caudiculatus
Entiminae Laparocerini Drouetius azoricus
CYCL Rhythirrinini Rhythirrinus sp
Entiminae Geonemini Apotomoderes menocrater
Entiminae Geonemini Apotomoderes sotomayorae
Entiminae Naupactini Artipus monae
Entiminae Entimini Scelianoma elydimorpha
Entiminae Geonemini Ischionoplus niveoguttatus
Entiminae Geonemini Lachnopus coffeae
ENTI Geonemini Lachnopus curvipes
ENTI Ophryastini Ophryastes sp
ENTI sp 4 (China)
ENTI Tanymecini Geotragus sp
Entiminae Tropiphorini Spartecerus DDM2009
ENTI Cratopini Cratopus sumptuosus
ENTI Psallidiini Psallidium sp
Entiminae Eustylini Exophthalmus marginicollis
Entiminae Entimini Diaprepes balloui
Entiminae Entimini Diaprepes abbreviatus
Entiminae Eustylini Exophthalmus cinerascens
Entiminae Tanymecini Pachnaeus marmoratus
Entiminae Eustylini Exophthalmus quadrivittatus
Entiminae Phyllobiini Compsus maricao
ENTI Tropiphorini Catasarcus sp
Entiminae Eupholini Celebia EMC2007
ENTI Tropiphorini Leptopius sp
Entiminae Naupactini Eurymetopus birabeni
Entiminae Naupactini Eurymetopus fallax
Entiminae Naupactini Pantomorus auripes
Entiminae Naupactini Teratopactus nodicollis
Entiminae Naupactini Naupactus minor
Entiminae Naupactini Aramigus tessellatus
Entiminae Naupactini Naupactus navicularis
ENTI Naupactini Naupactus xanthographus
Entiminae naupactini Naupactus cinereidorsum
Entiminae Naupactini Litostylus AMV2011a
ENTI Naupactini Litostylus pudens
Entiminae Naupactini Asynonychus cervinus
Entiminae Naupactini Pantomorus cervinus
Entiminae Naupactini Naupactus affinis
68
44
50
0
0
2
29
0
26
16
8
88
30
5
9
2
1
100
0
14
4
9
0
2
0
0
69
8
17
85
35
5
5
19
20
1
100
1
97
14
1
4
100
97
48
77
63
49
25
0
4
20
1
77
49
8
48
13
0
32
75
14
10
97
7
46
99
100


C
Entiminae Naupactini Asynonychus cervinus
Entiminae Naupactini Pantomorus cervinus
Entiminae Naupactini Naupactus affinis
Entiminae Sitonini Sitona tibialis
Entiminae Sitonini Sitona lineatus
Entiminae Naupactini Pantomorus viridicans
Entiminae Naupactini Pantomorus albicans
Entiminae Naupactini Galapaganus caroli
Entiminae Naupactini Galapaganus galapagoensis
Entiminae Eustylini Xestogaster AMV2011a
Entiminae Eustylini Eustylus hybridus
Cyclominae Falklandius turbificatus
Cyclominae Falklandius kuscheli
ENTI Sitonini Sitona lineatus
Hyperinae Hypera plantaginis
HYPE Hyperini Hypera postica
Entiminae Omiini Omias rotundatus
Entiminae Phyllobiini Phyllobius pyri
Entiminae Sciaphilini Foucartia squamulata
BARI Baridini Melanobaris laticollis
CRYP Camptorhinini Camptorhinus sp
CURC Tychiini Sibinia fulva
CONO Lobotrachelini sp 2 (China)
CONO Lobotrachelini sp 3 (China)
CONO Lobotrachelini sp 1 (China)
MESO Magdalinini Magdalis sp
CURC Storeini Melanterius sp
CURC Mecinini Miarus sp
MESO Laemosaccini Laemosaccus sp
SCOL Coptonotini Coptonotus cyclopus
MOLY sp 2 (China)
MOLY Pissodini Pissodes sp
MOLY Lepyrini Lepyrus sp
CURC Acalyptini Acalyptus sp
MOLY sp 4 (China)
LIXI Lixini Larinus turbinatus
LIXI Rhinocyllini Bangasternus sp
MOLY Hylobini Hylobius abietis
CURC Anthonomini Anthonomus pomorum
CURC Ceratopini Ceratopus sp
CURC Eugnomini Ancyttalia sp
CEUT Mononychini Mononychus punctumalbum
CEUT Ceutorhynchini Ceutorhynchus assimilis
CEUT Phytobini Rhinoncus sp
CURC Cionini Cionus griseus
CURC Cionini Cionus olens
CRYP Cryptorhynchini Perissops sp
CRYP Cryptorhynchini Ouroporopterus sp
CRYP Cryptorhynchinae Pseudomopsis sp
CRYP Cryptorhynchini Acalles aubei
CRYP Cryptorhynchini sp
COSS Neumatorini Brachytemnus porcatus
CURC Cryptoplini Haplonyx sp
MOLY sp 1 (China)
MOLY sp 3 (China)
CONO Mecopini Mecopus sp
CONO Zygopini Peltophorus sp
COSS sp 1 (China)
COSS Pentarthrini Pentarthrus elumbe
100
18
46
99
11
2
85
86
64
78
4
49
97
16
2
95
32
8
67
36
24
100
100
100
99
6
5
6
3
8
1
76
80
16
16
36
17
25
100
100
100
3
6
60
72
10
1
79
69
100
99
81
75
100
100
0
100
100
98
97
76
93
16
1
37
2
100
100
80
3
22
6
76
100
31


D
CONO Mecopini Mecopus sp
CONO Zygopini Peltophorus sp
COSS sp 1 (China)
COSS Pentarthrini Pentarthrus elumbe
SCOL Xyleborini Anisandrus dispar
SCOL sp 2 (China)
SCOL sp 1 (China)
SCOL Dryocoetini Dryocoetes autographus
SCOL Ipini Ips acuminatus
SCOL Ipini Ips cembrae
SCOL Premnobiini Premnobius cavipennis
SCOL Corthylini Pityophthorus micrographus
SCOL Corthylini Corthylus rubricollis
SCOL Cryphalini Cryphalus saltuarius
SCOL Hypoborini Hypoborus ficus
SCOL Xyloctonini Xyloctonus maculatus
SCOL Polygraphini Polygraphus poligraphus
SCOL Crypturgini Crypturgus pusillus
SCOL Tomicini Tomicus piniperda
SCOL Hylastini Hylastes opacus
SCOL Hylesini Hylesinus varius
SCOL Phloeotribini Phloeotribus spinulosus
SCOL Hexacolini Scolytodes caudatus
SCOL Diamerini Diamerus inermis
SCOL Scolytini Scolytus scolytus
SCOL Scolytini Scolytus sp
Curculionidae Cyclominae Aphela gotoi
BAGO Bagoinae Bagous sp
PLAT Platypodini Platypus cylindricus
PLAT Platypodini Platypus cylindricus
PLAT Tesserocerini Diapus unispineus
BRAC Ocladiinae Ocladius sp
DRYO Rhynchophorinae Sitophilus granarius
DRYO sp 1 (China)
DRYO Rhynchophorinae Cosmopolites sordidus
DRYO Orthognathinae Rhinostomus barbirostris
BRAC Brachycerinae Brachycerus muricatus
BRAC Erirhinini Echinocnemis sp
BREN Nanophyinae Nanophyes marmoratus
BREN Nanophyinae Nanophyes sp
BREN Apioninae Rhopalapion longirostre
BREN sp 1 (China)
ATTE Apoderinae sp 1 (China)
ATTE Apoderinae Apoderus coryli
ATTE Attelabinae sp 1 (China)
ATTE Rhynchitinae Byctiscus populi
ATTE Rhynchitinae Deporaus betulae
NEMO Cimberidinae Doydirhynchus austriacus
ANTH sp (China)
ANTH Anthribinae Platystomos albinus
CERAMBYCIDAE Anoplophora glabripennis
CHRYSOMELIDAE Crioceris duodecimpunctata
70
22
22
6
76
100
100
30
21
10
100
100
100
98
100
100
100
100
7
54
42
100
100
33
96
80
75
22
5
66
62
52
96
70
100
100
99
96
100
88
72
77
72
58
100
100
6
14
47
100
100
98
52
31
100
92
100
100
100
98
58
5
14
100
100
100
99
93
90
100
95
99
99
100
100
100
100
100
99
97
96
28
31
77
100
100

Chapter 6

General discussion, future prospects and concluding remarks

“Two weevils crept from the crumbs.

'You see those weevils, Stephen?' said Jack solemnly.

‘I do.'

‘Which would you choose?'

‘There is not a scrap of difference. They are the same species of *Curculio*, and there is nothing to choose between them.'

‘But suppose you had to choose?'

‘Then I should choose the right-hand weevil; it has a perceptible advantage in both length and breadth.'

‘There I have you,' cried Jack. 'You are bit - you are completely dished. Don't you know that in the Navy you must always choose the lesser of two weevils?’ ”

- Patrick O'Brian , 1979

Nemocephalus monilis (Fabricius, 1787) (Brentidae: Brentinae: Trachelizini). Saba, Dutch Caribbean

Chapter 6: General discussion, future prospects and concluding remarks

6.1 Overview

This thesis has investigated the higher-level relationships amongst and within the families of Curculionoidea, in a phylogenetic framework, using molecular sequence data from a variety of mitochondrial and nuclear markers. Strong evidence has been provided that supports sound inferences into some important relationships and clear phylogenetic divergences in the group. Arguably of equal importance, has been the highlighting of substantial sections of the curculionoid tree that remain recalcitrant to robust resolution, and in many cases, imply incongruence to morphological diagnoses of their constituent taxa.

The mitogenomic assembly from direct sequencing of pooled genomic DNA approach successfully developed here is comparatively simple and should be of wider application beyond weevils. Currently this is likely to be successful for organisms possessing a relatively large mitochondrial: nuclear genome size ratio. However, with enrichment of mitochondrial DNA (*e.g*. through hybrid capture) such limitations could theoretically be at least partly overcome, with the disadvantage of increasing protocol complexity and cost.

6.2 Weevil relationships

The unambiguous division of the Curculionidae *s.str.* into two deeply divergent large clades is well supported by the extensive mitogenomic data (including support from compelling tRNA rearrangements) in Chapter 3. One clade represents the broad-nosed weevils, and the other, all other Curculionidae *s. str.* (except Bagoinae and Platypodinae). This division, which has now been repeatedly recovered with molecular datasets should now be fully acknowledged as representing separate diverse and species-rich lineages, each potentially warranting recognition as family-level taxa. However, relationships *within* each of these clades remain largely obscure despite the recovery of a number of potentially monophyletic subfamilies and tribes as assessed through bootstrap support in Chapters 3 and 4, and the AU, SH and KH tests undertaken in Chapter 5 (*e.g.* Lixinae, Corthylini, Peritelini and Cylydrorhinini).

Nodal support for the arrangement of curculionoid families at the base of the weevil tree is generally very high, indicating the following branching order of monophyletic families (oldest to youngest): Anthribidae → Nemonychidae → Attelabidae → Brentidae → Other Curculionoidea. The Brachyceridae are poorly supported in their placement as sister to all Curculionidae *s.str.* + Dryophthoridae, but Dryophthoridae + Platypodinae is strongly supported as being sister to all Curculionidae *s.str.* The latter two relationships are consistent with the more inclusive definition of Curculionidae proposed by Oberprieler *et al.* (2007). Unfortunately two of the ‘primitive’ weevil families (Belidae and Caridae) were unavailable for analysis, and therefore, remain to be placed on the weevil mitogenomic tree.

One of the interesting findings, from a biological perspective, has been the strong support offered by the mitogenomic analyses for the hypothesis that wood-boring habits have evolved independently at least twice in the weevils. Prior to molecular analyses, it was generally thought (with some exceptions, *e.g.* Marvaldi 1997) that the subfamilies Scolytinae, Platypodinae and Cossoninae were closely related. They were recovered as a monophyletic, derived clade by Kuschel (1995) based on two apomorphies, one of which was "fully developed adult capable of feeding inside plant tissues", which is likely to be a homoplasious trait associated with morphological constraints in adaptation for boring into wood. There have also been long standing doubts about the phylogenetic affinities of the Platypodinae (Kuschel *et al.* 2000), which are justified by the mitogenomic results presented here, indicating that this subfamily is distantly related to the rest of the Curculionidae *s. str.*, and is apparently closely related to the Dryophthoridae. That this last relationship was also recovered by Marvaldi (1997), through analysis of larval characters, is strong independent evidence for its validity and that platypodines, scolytines and cossonines have undergone a remarkable parallelism in the evolution of boring behaviour. Relationships between the other wood-boring subfamilies, Cossininae and Scolytinae, appear to be complicated, though the mitogenomic data suggests that the two belong to two evolutionary distinct lineages, with the bulk of the Scolytinae (except Ipini and the aberrant Coptonotini) forming a well-supported clade within the 'long-nosed weevils' clade, whilst the Cossoninae are paraphyletic, in a sister clade to this, which also contains a large number of other (potentially paraphyletic) subfamilies (*e.g.* Molytinae, Curculioninae, Cryptorhynchinae *etc.*).

Lack of reliable and diagnostic morphological characters available to consistently separate several large subfamilies (*e.g.* Molytinae, Curculioninae, Cryptorhynchinae) indicate that members of these have, to date, been grouped by plesiomorphies rather than synapomorphies, and are therefore, in a phylogenetic framework, not representative of natural lineages. It is therefore somewhat inevitable, though still striking, that many of these very same problematic subfamilies are also repeatedly recovered as para- or polyphyletic according to DNA sequence data, as evidenced in the many phylogenetic analyses undertaken for this thesis and in other studies (e.g. McKenna *et al.* 2009). The problem of limited taxon sampling in such a speciose group as the weevils, clearly prohibits definitive conclusions to be drawn, but nevertheless, even with the sampling available in the present study, sufficient evidence from a large number of gene sequences has been presented to at least cast severe doubt as to the monophyly of a number of well-known weevil subfamilies and tribes. It is hoped that this evidence will be used by taxonomists to re-evaluate the constituent taxa of such groups in the light of the molecular conclusions. Clear candidates for this are the three entimine tribes which were rejected as monophyletic following the constraint analyses in Chapter 5. In the same chapter, the discovery of the cyclomine genus *Aphela* outside of the broad-nosed weevils, is strong evidence that its current taxonomic placing warrants revision.

6.3 Mitogenomics

One of the main advances to arise from the research for this thesis was the successful use of a straightforward method for densely sampling mitogenomes with little prior genome knowledge. Whilst this allowed for the identification of 92 mitogenomic sequences containing eight or more genes, a large number of assemblies remained unused in analyses. These included 244 of lengths between 1000 and ~8000 bp (46 of which > 4000 bp) and it is anticipated that identification of some of these will potentially allow for an even greater number of taxa to be included in a future analysis, albeit at the cost of increased missing data in the matrix. It will be interesting to see if the tRNA rearrangements that apparently characterise the broad-nosed weevils will hold with the addition of further taxa.

A perhaps unexpected result presented in Chapter 4, is that the addition of nuclear markers to the mitogenomic data provides little added value, in terms of nodal support across the reconstructed trees, over the mitogenomic data alone. Accordingly, an argument has been presented suggesting that faced with a choice, expanded taxon sampling may be preferable to increasing sequence data, especially in the case of diverse animal groups such as the weevils.

In Chapter 5, an expanded taxon sampling based upon obtaining shorter sequences from public databases and adding them to the mitogenomic 'backbone' resulted in the identification of some intriguing relationships and allowed for statistical testing of monophly of some higher taxa, although it must be emphasised that the success of this approach is clearly reliant upon a well sampled 'backbone'.

6.4 Future prospects and concluding remarks

Unfortunately it has not been possible here to 'unlock' the valuable genetic resources of the NHM Coleoptera collections through the current SPIA methodology, although the potential problems of DNA degradation brought about by alkylating fumigants, suspected of contributing to this, is unlikely to affect all natural history collections equally. However, NGS technologies, which are innately designed for the sequencing of short DNA fragments, promise to offer a potential solution to this obstacle. Existing methodologies have now been adapted and improved (especially in regard to library preparation) to allow for sequencing of degraded ancient DNA (Knapp & Hofreiter 2010). Indeed, so called 'museum genomics' has already been successfully employed on old museum specimens of mammal skins and bone (Rowe *et al.* 2011), although it has yet to see widespread use on DNA extracted from archival entomological specimens.

Significant progress in understanding the detailed phylogeny of weevil tribes and subfamilies can only be realistically achieved through increased taxon sampling. The methodology devised for this thesis is suited to rapidly achieving this, given availability of specimens. Direct sequencing from pooled DNA can possibly be simplified even further through elimination of the PCR 'bait' sequences, used here for identifying mitogenomes. One possible avenue to pursue, which might achieve this, will be intelligent sample pooling, wherein, for example, only known divergent taxa belonging to clearly different families and subfamilies (potentially also tribes) are pooled together. In so doing, it is possible that identification of resulting assemblies can be achieved through direct BLAST searches against mitochondrial sequences already available on public databases. This could provide sufficiently close identification to allow for confident assignment of each assembly to a specimen, where the family/subfamily/tribe is known previously.

One important aspect of mitogenome assembly remains to be ascertained – namely the lower limit of divergence between pooled genomes which will allow for successful assembly to be possible. This could potentially be either investigated empirically, with real samples of a range of closely related (congeneric) and divergent species, or *in silico* with simulated genomes.

Alternative techniques involving the use of restriction enzymes on genomic DNA for complexity reduction, and capitalising on NGS technologies, can also obtain large numbers of sequences suitable for phylogenetic studies. These include restriction site-associated DNA sequencing (RAD-seq), a procedure related to that used to obtain amplified fragment length polymorphisms (AFLPs) (Vos *et al.* 1995), and which can identify thousands of genetic markers across target genomes (Davey and Blaxter 2011). RAD-seq, which identifies polymorphic variants adjacent to restriction enzyme digestion sites, has been used for both genome assembly (Willing *et al.* 2011) and single nucleotide polymorphism (SNP) marker discovery (Pegadaraju *et al.* 2013) without the need for a reference genome. Whilst RAD-seq has seen most use in population-level studies, its suitability for *de novo* assembly of extended contigs flanking the restriction site can enable large sections of the nuclear genome to be assembled and identified. Such extensive nuclear sequences will be of interest in deeper-level phylogenetic analyses, including those in combination with mitogenomic sequences.

Another burgeoning area of technical development in recent years has been in the field of phylotranscriptomics (Ozsolak & Milos 2011), whereby, hundreds or even thousands of expressed RNA molecules can be sequenced through shot-gun NGS, and corresponding orthologues determined through bioinformatics, for use in phylogenetic reconstructions (Oakley *et al.* 2013). Undoubtedly this approach will eventually be used in Coleoptera phylogeny and it will be interesting to see whether relationships based on mitogenomic and/or 'traditional' nuclear loci will be supported by the new data.

Whilst this thesis has predominantly been concerned with the testing of phylogenetic hypotheses, the techniques employed, and to some extent the data generated, could potentially be used to investigate aspects of weevil biology not yet explored. Because weevils are reliant on plants, some interesting questions that can be addressed pertain to this close relationship. This could include investigation of adaptation in weevil lineages to feed on specific plant tissues/structures and to test for convergence of associated morphological traits across the tree, which may also help explain the current morphological confusion in classifcation.

An example of an interesting morphological character that has proved difficult to use in delimiting higher taxa and would be worthy of study is the sclerolepidia, scale-like structures located along the metepisternal suture on the weevil thorax and found across a number of diverse subfamilies and tribes (Lyal *et al.* 2006). Their function remains incompletely known, although their common occurrence in mostly wood-feeding species, and in taxa with an inability to fly has been noted. It is likely that the function of the sclerolepidia may differ from group to group, which can be ideally investigated after a sound phylogenetic basis for their distribution across the weevil lineages is established.

Of allied interest are biogeographic questions that can also be addressed once a detailed phylogeny is known. It will be interesting to discover the possible geographic origins of important clades and to infer processes leading to the distribution of extant weevil lineages. Similarly, relative species richness measures for different clades may be able to provide evidence as to whether certain lineages were able to radiate more rapidly upon entering new areas, which may ultimately shed light on the role of adaptation to newly encountered plant groups.

6.5 References

Davey JW, Blaxter ML (2011) RADSeq: next-generation population genetics. *Briefings in Functional Genomics* **9**, 416-423.

Knapp M, Hofreiter M (2010) Next generation sequencing of ancient DNA: requirements, strategies and perspectives. *Genes* **1**, 227-243.

Kuschel G (1995) A phylogenetic classification of Curculionoidea to families and subfamilies. *Memoirs of the Entomological Society of Washington* **14**, 5-33.

Kuschel G, Leschen RAB, Zimmerman EC (2000) Platypodidae under scrutiny. *Invertebrate Taxonomy* **14**, 771-805.

Lyal CHC, Douglas DA, Hine SJ (2006) Morphology and systematic significance of sclerolepidia in the weevils (Coleoptera : Curculionoidea). *Systematics and Biodiversity* **4**, 203-241.

Marvaldi AE (1997) Higher level phylogeny of Curculionidae (Coleoptera : Curculionoidea) based mainly on larval characters, with special reference to broad-nosed weevils. *Cladistics-the International Journal of the Willi Hennig Society* **13**, 285-312.

McKenna DD, Sequeira AS, Marvaldi AE, Farrell BD (2009) Temporal lags and overlap in the diversification of weevils and flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7083-7088.

Oakley TH, Wolfe JM, Lindgren AR, Zaharoff AK (2013) Phylotranscriptomics to Bring the Understudied into the Fold: Monophyletic Ostracoda, Fossil Placement, and Pancrustacean Phylogeny. *Molecular Biology and Evolution* **30**, 215-233.

Oberprieler RG, Marvaldi AE, Anderson RS (2007) Weevils, weevils, weevils everywhere. *Zootaxa* **1668**, 491-520.

Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* **12**, 87-98.

Pegadaraju V, Nipper R, Hulke B, Qi L, Schultz Q (2013) *De novo* sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. *BMC Genomics* **14**: 10.1186/1471-2164-14-556.

Rowe KC, Singhal S, Macmanes MD*, et al.* (2011) Museum genomics: low-cost and high-accuracy genetic data from historical specimens. *Molecular Ecology Resources* **11**, 1082-1092.

Vos P, Hogers R, Bleeker M, Reijans M, Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **23**, 4407-4414.

Willing E, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011) Paired-end RAD-seq for *de novo* assembly and marker design without available reference. *Bioinformatics* **27**, 2187-2193.