Developing MMG: A method for the study of biodiversity linking taxonomy, phylogeny and ecology

Alexandra Lauren Crampton-Platt

A thesis submitted for the degree of

Doctor of Philosophy

University College London

Department of Genetics, Evolution and Environment

September 2015

I, Alexandra Lauren Crampton-Platt confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been identified in the thesis.

Abstract

High-throughput sequencing technologies are changing the way in which diversity is studied at all scales and has the greatest potential to facilitate studies of taxa that are intractable to other methods. Insect ecology is one such field, with great abundance and diversity combining with incomplete taxonomic knowledge to hamper studies of diversity at large spatial and temporal scales. A new high-throughput method has recently been proposed to address such issues within a self-contained phylogenetic framework that is linkable with existing biological knowledge via Linnaean taxonomy. This method, 'mitochondrial metagenomics' (MMG), has already been the subject of a number of proof-of-concept studies, frequently focussed on Coleoptera. These studies are unified here with additional similar datasets for the first time to draw together the lessons to be learnt from the results obtained to date and infer the immediate methodological questions that remain to be answered. Particular attention is paid to the prospect of bulk sequencing of mixed specimens and the associated bioinformatics challenges. Consideration is given to mitochondrial phylogeny reconstruction with the prospect of rapidly increasing taxon sampling and the potential for phylogeny-based taxonomy assignment for otherwise uncharacterised communities. Mitochondrial metagenomics is then applied to a landscape-level assessment of the response of the leaf litter beetle communities to habitat differences, taking a combined compositional and phylogenetic perspective. Finally, the results are synthesised for a perspective on the remaining methodological impediments to the further development of MMG, and the future prospects for synthetic analyses of diversity are considered.

Acknowledgements

I thank my supervisor, Alfried, for his continual support and encouragement throughout this project. In particular I thank him for obstinately believing in MMG in spite of the long and bumpy developmental journey. I also thank my secondary supervisors, Paul and David for their support. I am particularly indebted to Paul for allowing me to destroy specimens collected by the New Forest Quantitative Initiative and for his sampling advice both while in the New Forest and in Panama. The foundations for the methodology presented here were conceived with the assistance of Martijn Timmermans, whose advice and experience have been invaluable throughout. I am grateful to all who allowed me to use their datasets for the analyses presented in Chapter 2 and I would like to thank all members of the Volger lab both past and present for their help, humour and companionship throughout these last few years. In particular Benjamin Linard for his enthusiasm for obscure bioinformatics problems. There are many other people who have provided essential assistance throughout: my field assistants Emma Holmes and Cian O'Connor; Dan Carpenter and the NFQI volunteers; Vladimir Blagoderov at the Sackler Biodiversity Imaging Laborotory, NHM, for access to cameras and training; all the molecular lab staff at the NHM for their patience and advice, particularly Alex Aitken and Stephen Russell; Jackie McKenzie-Dodds and Fiona Allen at the Molecular Collections Facility, NHM, for advice and assistance with destructive DNA extractions; the staff at the sequencing facilities at the NHM and Cambridge University, particularly Kevin Hopkins and Shilo Dickens respectively. Computing resources used herein were made available by the Department of Life Sciences, NHM, and maintained by Peter Foster who must be commended for his patience through frequent installation requests and excessive resource consumption. Ladislav Bocak kindly provided the specimen line drawings seen in Chapters 3 and 4. Lastly I thank UCL and NHM for funding this project, and additionally the Entomology Department (NHM) for the allocation of additional DIF funds for sequencing.

Table of Contents

Table of Figures

Table of Tables

Chapter 1 Introducing Mitochondrial Metagenomics

1.1 Molecular Methods for Insect Biodiversity

1.1.1 PCR and the emergence of DNA barcoding

Molecular data has been used for the study of insect diversity since the 1970s, developing from allozyme electrophoresis and restriction mapping to Southern blot and Restriction Fragment Length Polymorphism (RFLP; and allied techniques), and finally the explosion of nucleotide sequencing made possible by the advent of PCR (Polymerase Chain Reaction; Mullis & Faloona 1987) and automated sequencing. More recently the dramatically increased throughput made possible by 'next-generation' sequencing technologies (hereafter HTS, high-throughput sequencing) and associated increases in computing power have further widened the range of questions that can be addressed economically with DNA sequences. Thousands of specimens are now readily analysed in a single study, allowing the diversity of whole communities of insects to be quantified and compared at large scales. Molecular data have been used to address a broad range of questions, from early differentiation between morphologically similar sister species (e.g. Eisses et al. 1979) and associating larval and adult stages (e.g. Berlocher 1980), to genome evolution (e.g. Burke et al. 2010), and analyses of ancient environmental DNA (e.g. Willerslev et al. 2003). Molecular methods have fundamentally changed understanding of the Tree of Life and the distribution of diversity, from the discovery of Archaea based on early rRNA sequences (Woese and Fox 1977) to the routine discovery of hundreds of micro-organisms in small samples of environmental substrates (e.g. Venter et al. 2004; Sogin et al. 2006; Fonseca et al. 2010; Lecroq et al. 2011).

Estimates of global species richness have varied widely and often conflict (Caley et al. 2014), ranging from 3 to 100 million eukaryotes (May 2010) and up to 1 trillion microbes (Locey and Lennon 2016). Recent estimates place total eukaryotic diversity between 3 and 10 million species (Mora et al. 2011; Costello et al. 2013), of which approximately 1.5 million have been described (Costello et al. 2012). A significant portion of these species are insects, with approximately 1 million species described of an estimated 1-5 million (Costello et al. 2012). In many ways researchers in insect diversity face similar problems as microbiologists in that the taxa under study are frequently small, diverse, and hyper-abundant, however, unlike microbial groups, insects generally have good morphological characters useful for delimiting species, classifying them into higher taxa, and testing evolutionary hypotheses. Thus, where microbiologists have embraced each revolution in molecular methodology, entomologists have generally been slow to follow. However, by the late 1990s and early 2000s the number of studies applying DNA sequences to long-standing questions in insect systematics (e.g. the ‘Strepsiptera problem’, Whiting et al. 1997) was growing steadily and expected to continue to do so, although concerns were raised about the lack of synergy due to the wide array of markers used and the corresponding difficulty of synthesising the results (Caterino et al. 2000).

Hebert et al. (2003) brought a new impetus to biodiversity research in general, and insect diversity in particular, by introducing ‘DNA barcoding’ as an efficient and reliable way of assigning specimens accumulated in the course of biodiversity sampling to known species and, potentially, providing a solution to the problem of undescribed diversity by clustering sequences at a level assumed to approximate species (based on pairwise similarity). DNA barcoding can be defined as the use of a single, standardised gene region to identify specimens belonging to a given taxon using distance methods on pairwise similarity measures (Rubinoff et al. 2006). For Metazoa, ~650 bp of the 5’ end of cytochrome oxidase *c* subunit I (*cox1*) was suggested for three main reasons: a) a greater range of phylogenetic signal than many other markers, b) availability of robust universal primers, and c) ease of alignment (Hebert et al. 2003). These ideas were not in themselves new but the greater emphasis on the need for standardisation to accelerate species inventories and the suggestion that DNA barcoding could be a solution to the ‘crisis’ in taxonomy generated a great deal of attention, both positive and negative. Concerns were raised regarding the application of distance methods to an evolutionary problem, the biological interpretation of ‘barcode clusters’, the failure of barcoding to correctly diagnose recognised species for many groups, the narrow focus on a single mitochondrial marker for delimiting species, the diversion of funding from taxonomic research to DNA barcoding, and even whether the results presented in this and subsequent papers supported the authors’ claims regarding the potential of barcoding (e.g. Will & Rubinoff 2004; Rubinoff et al. 2006; Taylor & Harris 2012).

In spite of the debate, and perhaps because of the publicity this generated within the scientific community, the number of studies applying DNA barcoding to a broad range of questions quickly multiplied and the 5’ portion of *cox1* became established as the standard ‘barcode’ marker for molecular biodiversity research in the majority of Metazoa, much in the same way that SSU rRNA was already established as the standard for Bacteria (16S) and Nematoda (18S). This rapid standardisation and growing acceptance of sequence-based solutions to a broad range of questions was significant in driving a growth in the use of molecular methods in ecology, even though large-scale studies on thousands of insect specimens remained rare (e.g. Baselga et al. 2013; Baselga et al. 2015) outside of heavily subsidised projects from the CBOL initiative (Consortium for the Barcode Of Life; e.g. Janzen et al. 2005). Uses of DNA barcodes (in the looser sense of partitioning mitochondrial sequence variation into clusters approximating species) have ranged from specimen identification (e.g. Hebert et al. 2003), screening for cryptic diversity (e.g. Hebert et al. 2004), and linking of life stages (e.g. Ahrens et al. 2007), to population genetics (e.g. Craft et al. 2010), integrative taxonomy (e.g. Montagna et al. 2016), phylogenetics (e.g. Quicke et al. 2012) and phylogeography (e.g. Emerson et al. 2011). Complex host-parasitoid interactions have been elucidated (e.g. Hrcek et al. 2011), insect prey has been identified from predator faeces (e.g Zeale et al. 2011), and insect-plant association established from gut contents (e.g. Navarro et al. 2010). Ecologists have partitioned sampled diversity into molecular MOTU (Molecular Operational Taxonomic Units; Floyd et al. 2002), as an alternative to parataxonomic morphospecies sorting, prior to assessing species presence-absence and abundance at different sites for analyses of richness and turnover between communities, which in turn may be used to rapidly determine conservation priorities (e.g. Smith et al. 2005). In 'haplotype macroecology' barcodes are used to quantify both inter- and intra-specific diversity to test for differences in community structure between ecological groups (e.g. Papadopoulou et al. 2011; Baselga et al. 2013).

Regardless of the arguments surrounding it, DNA barcoding has in just over a decade, already produced a significant legacy for ongoing biodiversity research. It has led to the generation of millions of new sequences for thousands of species sampled from across the globe. These are brought together in the Barcode Of Life Database (BOLD; www.boldsystems.org; Ratnasingham & Hebert 2007), a web-based tool for uploading sequences compliant with barcode standards (including metadata), identifying unknown sequences, and the automated clustering of sequences into approximately species-level groups (labelled with BINs, Barcode Identification Numbers) which are constantly revised with the addition of new data (Ratnasingham and Hebert 2013). Notably, insect sequences make up approximately 75% (>300,000) of these BINs but only around one third of clusters are associated with a recognised species name. Importantly, the value and utility of existing sequences increases as the database grows and in turn encourage further growth in data acquisition. At a time when new sequencing technology is prompting another step-change in approaches and attitudes towards the use of sequence data in biodiversity research it is this database of identified sequences which is the most valuable outcome of the barcoding initiative, as this huge resource for specimen identification will continue to be as relevant for HTS studies as it is for individual DNA barcoding.

1.1.2 Challenges and opportunities of high-throughput sequencing

Although DNA barcoding was initially hailed as an efficient an inexpensive approach to obtain species identifications, current costs of DNA extraction, PCR and Sanger sequencing were recently estimated at $7 (~£5) per individual (without labour; Shokralla et al. 2015). For studies of natural insect communities in which thousands of specimens are routinely collected these costs clearly remain prohibitive in the majority of cases. The advent of HTS and recent advances in multiplexing to separate hundreds of individual samples can now reduce these costs to an estimated $1.5 (~£1) per barcode per specimen (Shokralla et al. 2015), offering a significant opportunity to dramatically increase the rate at which barcodes are generated. At the same time, the advent of HTS has brought about a second revolution in the application of molecular methods to biodiversity, with massively increased throughput and reduced costs per base promising to make genome-level diversity routinely available for non-model organisms.

HTS platforms have been available since the mid-2000s, however it has taken almost a decade for them to become widely used in biodiversity studies and for economical solutions to single-specimen and highly-multiplexed amplicon (i.e. short PCR products) sequencing to start emerging. This lag is in part attributable to the time taken for new technology to be proven reliable and adopted as mainstream, but more challenging for biodiversity researchers was how to economically obtain manageable amounts of useful data for hundreds to thousands of individuals on platforms designed to generate millions of base pairs for a small number of samples. As previously, microbial ecologists led the way in using HTS to directly sequence 16S amplicons from environmental samples (Sogin et al. 2006; Caporaso et al. 2011), and 18S amplicons from mixed nematode samples (Porazinska et al. 2009). This 'metagenetics' (Creer et al. 2010) approach was an ideal solution to the problem of microbial diversity, allowing whole communities to be sequenced simultaneously in a much more efficient manner than was previously possible. In analogy to this, 'metabarcoding' techniques have more recently been adopted by non-microbial ecologists (Taberlet et al. 2012; Yu et al. 2012). Here, a standardised barcode region is amplified from environmental samples or mixtures of specimens that are collected and processed in bulk, without the need for sorting and individual DNA extraction and amplification. As with conventional DNA barcoding, metabarcoding aims to identify the species present in the sample and therefore relies on databases of sequences from identified specimens, although this possibility may be limited for environmental metabarcoding (Taberlet et al. 2012; see below).

For metabarcoding of arthropods the relative ease of amplifying the *cox1* barcode region with degenerate Folmer primers (Yu et al. 2012; Ji et al. 2013) allows the link between DNA barcoding and metabarcoding to be maintained, but application of a combination of primer sets and/or target genes is not uncommon ('metasystematics'; Gibson et al. 2014). Metabarcoding of environmental DNA (eDNA) is more challenging as DNA degradation requires the use of shorter 'mini-barcode' regions, increasing the complexity of primer design further and leading to a lack of standardisation in amplified regions (Taberlet et al. 2012; Cristescu 2014). However, even where the *cox1* barcode (or part thereof) is amplified the incompleteness of sampling at the species-level in existing databases precludes specific identification of most sequences. Thus in the majority of metabarcoding studies sequence diversity is binned into Operational Taxonomic Units (OTUs) or MOTU (molecular operational taxonomic units; Floyd et al. 2002) that are either completely separated from the Linnaean classification or only assigned to higher clades (Deagle et al. 2014), leading to a great loss of taxonomic resolution. In other cases, dedicated reference libraries of identified sequences are developed for a particular project to allow species-level identification (e.g. Bienert et al. 2012). At the same time, the challenge of designing minimally biased primers for equal amplification success of species in mixed samples, the potential for PCR-introduced errors to inflate estimates of diversity, and the loss of the link between biomass and read number, has led to calls for exploration of PCR-free approaches to studying biodiversity (Taberlet et al. 2012).

In parallel, the decreasing cost and increasing capacity of HTS is fuelling genome-scale sequencing and efforts are underway to sequence 5000 insect genomes (i5K; i5KConsortium 2013) and 1000 insect transcriptomes (1KITE). While these projects are critical for increasing knowledge of insect genome structure and are already helping to resolve deep phylogenetic relationships (Misof et al. 2014) they currently have little bearing on the study of natural communities. At a smaller scale, the rise of HTS has facilitated direct shotgun sequencing of environmental DNA, metagenomics, followed by *de novo* assembly of microbial genomes without the need for amplification (Venter et al. 2004; Iverson et al. 2012). With increasing insect genome availability there is some potential for an analogous insect metagenomics, metagenome skimming (MGS), whereby shotgun sequencing is applied to recover the most conserved and repetitive genomic elements in pools of DNA which are then profiled against existing genome scaffolds (Linard et al. 2015), although the resolution that can currently be obtained is limited.

Metagenome skimming builds on the principle of genome skimming (Straub et al. 2012) in which low coverage shotgun sequencing is applied to individual species to simplify and accelerate the process of obtaining multiple markers for phylogenetics without PCR. This is related to both whole-genome sequencing and transcriptomics in that sequencing occurs at the genome scale and does not involve any pre-selection of markers. It is however, significantly less data intensive, requiring only shallow sequencing because the aim is to assemble only the high-copy number portions of the genome such as nuclear rDNAs, repetitive elements, and partial plastid and mitochondrial genomes. Whilst genome skimming itself is not directly useful for studying insect communities, scaling up to metagenome skimming of mixed samples, in analogy to scaling up from DNA barcoding to metabarcoding, is a potentially powerful method to bypass the PCR step.

In insects, genome skimming has already been used to obtain the complete mitochondrial genome of several species (Berman et al. 2014; Kocher et al. 2014; Kocher et al. 2015) and it is this portion of the metagenome that will be of most interest for biodiversity studies, at least in the medium term. This is because a direct link can be made between assembled barcode sequences and already-barcoded species via BOLD, allowing sequencing to be done ‘blind’ without sacrificing the link with taxonomy and the associated wealth of biological information. At the same time, when shotgun sequencing is applied to pools of DNA physically unlinked loci from the same species can no longer be associated, precluding multi-locus phylogenetics (Papadopoulou et al. 2015). However, the presence of multiple linked loci on the mitochondrial genome allows for more accurate phylogeny reconstruction than the barcode alone, whilst simultaneously facilitating post-assembly de-multiplexing and identification using barcodes or other mitochondrial loci as ‘bait’ sequences (Timmermans et al. 2010). The identified mitogenome sequence can then be considered a ‘superbarcode’ that can be used in turn for species identification and phylogenetics. Importantly, the phylogenetic placement of an assembled mitogenome does not require any external information, allowing the integration of unidentified sequences into a unified analytical framework and the inference of higher-level taxonomy when analysed simultaneously with identified sequences (Crampton-Platt et al. 2015). Thus this ‘mitochondrial metagenome skimming’ (MMGS) or ‘mitochondrial metagenomics’ (MMG) could represent an economical opportunity to integrate arthropod ecology and phylogeny at a broad spatial and taxonomic scale.

1.2 Introducing Mitochondrial Metagenomics

Indeed, in the last three years (as anticipated by Taberlet et al. 2012) there has been a small but concerted effort to develop such a method for the economical generation of mitochondrial genomes for 'mito-phylogenomics' and biodiversity studies. Insect-focussed work has been centred in two groups with different approaches and primary motivations yet the degree of consistency in results hints at the flexibility and robustness of such a method. At this early stage in the history of MMG there remain many practical and logistical issues to resolve and many questions remain to be answered. All studies thus far have ostensibly been proofs-of-principle of MMG for various applications ranging from PCR-free barcoding (Zhou et al. 2013) and generation of reference libraries (Tang et al. 2014) for biodiversity research and monitoring of wild populations (Tang et al. 2015), to mito-phylogenomics (Rubinstein et al. 2013; Gillett et al. 2014; Timmermans, Viberg, et al. 2016), the mitochondrial tree-of-life (Crampton-Platt et al. 2015), community ecology (Gómez-Rodríguez et al. 2015) and community phylogenetics (Andújar et al. 2015). These studies position MMG as an important tool in the ecologist's arsenal and resolve the major methodological barriers to its widespread application, although further work is needed. The development of MMG and the integration of these various topics into a synthetic framework for insect biodiversity is discussed below and further elaborated upon throughout this thesis.

1.2.1 What is Mitochondrial Metagenomics?

Throughout the studies to date there has been an inconsistency in terminology and methodology and as yet there is no formal definition of what constitutes mitochondrial metagenomics. Herein the term 'mitochondrial metagenomics' refers to any study whereby sequence data of mitochondrial origin is obtained by shotgun sequencing of genomic DNA from mixtures of specimens for use in analyses of (genetic/species/phylogenetic) diversity either directly (with or without assembly) or as a means to assemble a library of superbarcodes for such analyses. This definition therefore does not include genome skimming of single-specimen libraries for mitochondrial reads (Guschanski et al. 2013; Tilak et al. 2014), but does include studies which attempt to enrich the mitochondrial fraction of mixed samples (Zhou et al. 2013). MMG is therefore a loose term for methods which facilitate PCR-free gathering of mitochondrial data from mixed samples, in much the same way that metabarcoding is a loose term for a collection of methods which aim to obtain a 'barcode' sequence (not necessarily *cox1-5'*) from such samples. MMG does not imply any particular analysis of the mitochondrial data generated, nor is it specific to any particular taxonomic group, although thus far the majority of work has been on insects. Given the allencompassing nature of the term and the wide range of existing and possible use cases, distinctions must be made between the main sources of input DNA (sample types) and the two major types of analysis that the resulting data are used for.

These distinctions and the terminology that will be used throughout this thesis are illustrated in Figure 1.1. The first distinction to make is between the two possible types of sample for MMG. Although all MMG is performed on mixed samples of DNA, those mixtures can be obtained either by designing specific mixtures with known compositions (upper left) or by sampling natural populations and applying MMG to the mix of specimens obtained (upper right). Henceforth the former shall be referred to as 'voucher MMG' while the latter shall be referred to as 'bulk MMG'. It is important to note that this does not necessarily imply that the specimens in the former case are retained as vouchers, nor does it imply that in the latter case the samples are processed directly for DNA extraction without any intervening sorting steps. Rather these should be seen as two alternative approaches to pooling DNA, reflecting

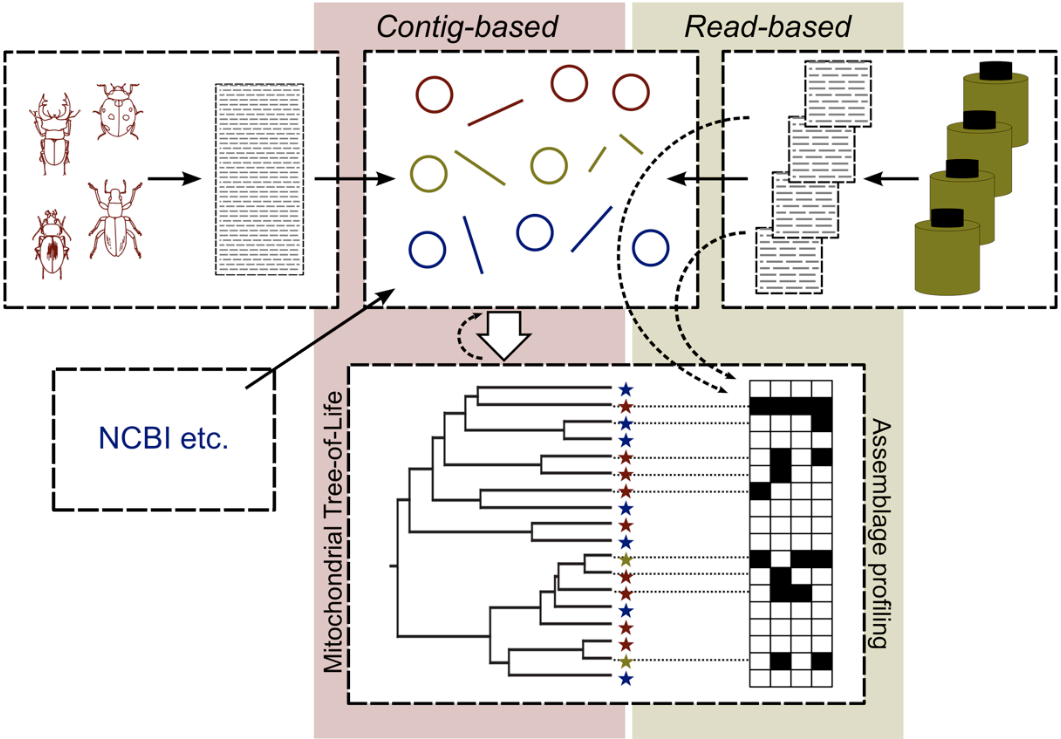


Figure 1.1 A general outline of mitochondrial metagenomics showing the two main sequencing strategies, 'voucher MMG' (top left) and 'bulk MMG' (top right). Both of these can then be applied to 'contig-based' analyses involving *de novo* mitogenome assembly and, potentially, identification of assembled contigs ('superbarcodes') and phylogeny reconstruction. Bulk MMG samples may also be used for 'read-based' analyses whereby the unassembled reads are matched against a contig/superbarcode library for assemblage profiling and possibly estimates of species biomass and intraspecific genetic diversity. In general voucher MMG samples will not be used for read-based analyses, but see text for a contrasting example.

the two main motivations for MMG which are highlighted in the mid-section of Figure 1.1.

In all MMG studies (both biodiversity/ecology and mito-phylogenomics) the initial step will be the construction of a database of mitogenome superbarcodes for species of particular relevance. Some superbarcodes may already be available in public repositories such as GenBank but these will usually be supplemented with additional MMG sequencing for targeted species. The aim of this targeted sequencing step is to maximise the completeness of the species matrix to be used in subsequent steps and as such the most appropriate pooling strategy is to include equal amounts of DNA per species for even sequencing and optimal assembly conditions. It is this pooling strategy to which 'voucher MMG' refers. The data obtained is assembled into contigs that are then linked to morphological identifications via Sanger 'bait' sequences, either from public databases or generated within the same study. These new superbarcodes are then potentially used to generate phylogenetic trees, with or without external data.

In contrast with these latter 'contig-based' analyses, bulk MMG samples will in most cases be used for 'read-based' analyses. This requires a database of contigs against which reads can be matched to obtain an assemblage profile of species presences for each sample. These profiles, when linked to a phylogenetic tree of the contigs enable analyses of phylogenetic community ecology in addition to those of species composition. In the simplest case this approach is used for biodiversity monitoring of mass-trapped arthropods based on presence-absence of a small subset of species of particular interest. In the most complex case, holistic analyses of diversity at multiple hierarchical levels are envisaged, incorporating relative species biomass and their genetic, species and phylogenetic diversity. At both ends of this spectrum, the trap sample (or pooled trap samples from a single site) is the natural unit of analysis. The high-throughput of NGS encourages its application to such samples without the need for intermediate sorting steps and therefore allows samples to be processed rapidly whilst preserving the true (multi-hierarchical) diversity of the sample. Thus, read-based analyses will in most cases be based on pools of DNA from co-collected specimens, with no adjustments made to the pool composition prior to sequencing. For maximal simplicity and cost-effectiveness these pools of DNA will be derived from bulk extraction of tissue in the relevant samples, but 'bulk MMG' also covers artificially pooled DNA or specimens where no adjustment for relative DNA contribution or genetic divergence has been made. In the present work, no true bulk MMG samples are presented but in all Chapters either all or part of the data conform to the principle of bulk MMG in that the biomass ratios and sequence divergences within the target group, Coleoptera, are maintained. Finally, it should be noted that voucher MMG is not necessarily synonymous with contig-based analyses and bulk MMG is not necessarily synonymous with read-based analyses, and indeed the majority of the existing studies do not conform to such expectations. This will be elaborated upon below through the discussion of these existing studies. Also note that all superbarcodes are contigs but not all contigs are superbarcodes. The latter term implies a high-confidence species-level identification based on bait sequences from vouchered specimens but the absence of such identification does not preclude the use of assembled contigs for either phylogeny reconstruction or read-based analyses.

1.2.2 Mitochondrial Metagenomics: The Story so Far

The following discussion will differentiate broadly between the application of MMG to assemble superbarcodes alone and its application to studying *in situ* diversity. The simplest application of MMG is to economically generate large libraries of vouchered superbarcodes as a natural evolution of the DNA barcoding concept, and indeed there is an obvious opportunity to exploit already-barcoded DNA collections for precisely this purpose (Dettai et al. 2012; Taberlet et al. 2012). It is with this kind of application that the simulation studies of Dettai et al. (2012) and the more recent work of Tang et al. (2014) are primarily concerned. In these cases, strict pooling for voucher MMG requires the expected sequence divergence between species to be taken into account in addition to attempting to equalise DNA contribution per species. For truly high-throughput 'superbarcoding' large collections of DNA would be available, allowing pools to be designed in such a way that expected sequence divergence does not drop below a given threshold (e.g. 15% in Dettai et al. 2012). For simplicity, pools would not include more than one species from the same clade (e.g. family, Tang et al. 2014) or could use sequence divergence in the barcode region as a proxy for whole-mitogenome divergences (Dettai et al. 2012). Such use-cases present minimal assembly and analytical complexity and could be easily scaled and standardised for rapid and broad-ranging superbarcode sequencing given sufficient resources. For example Tang et al. (2014) calculated that 1000 mitogenomes could realistically be generated per lane of Illumina HiSeq 2000 even without enrichment, reducing costs to approximately 2 USD per mitogenome. In such cases the availability of DNA and bioinformatics resources become the limiting steps.

Slightly more complex is the use of voucher MMG to generate superbarcodes for mito-phylogenomics. Here, the focus will usually be on a more limited range of taxa and the opportunity for maximising divergences within the pool will be low. Early work on

Ascidians pooled only five species but even at this small scale found that *de novo* genome assemblers were unable to obtain complete mitogenome sequences (Rubinstein et al. 2013). *De novo* transcriptome assemblers were more successful at dealing with the observed variation in coverage which was ascribed to variability in input DNA quality, mt:nuclear ratio and genome size. Subsequent work on insects has had success pooling larger numbers of species and future studies are likely to follow Gillett et al. (2014), aiming to economically generate superbarcodes for 100-200 species simultaneously. Multiple libraries may be prepared at additional cost to ensure that close relatives are not pooled together, although Tang et al. (2014) showed that congeners can be successfully separated, at least at a small scale. Perhaps more challenging for mito-phylogenomics is the quality and quantity of available DNA. Species are chosen for inclusion based on specific hypotheses about the underlying phylogeny and it is not unusual for the corresponding DNA to derive from different sources (Gillett et al. 2014). Some specimens might be freshly collected specifically for the study, whilst others might be pinned museum specimens from which a small amount of tissue is made available for DNA extraction. Equally, DNA extracts may already exist but be of variable age, quality, and quantity. In such cases there are many uncontrolled variables that will affect the likelihood of assembly for each species and ideally multiple libraries would be constructed to minimise bias within any single pool. In the case of Gillett et al. (2014) a single voucher MMG library was constructed, although equal DNA input was not possible for all species. Variability in coverage and assembly success was observed, with just over 50% of input species included in subsequent phylogenetic analyses, however these did not correlate closely with the amount of input DNA suggesting that this may be a crude measure for determining pooling ratios for DNA from different sources.

Perhaps more challenging is the application of voucher MMG to DNA derived only from pinned museum material of various ages and preservation (Timmermans, Viberg, et al. 2016). Such specimens may produce very small amounts of degraded and highly contaminated DNA that present problems for PCR-amplification and Sanger sequencing. Timmermans et al. (2015) avoided the latter issues by shotgun sequencing of DNA extracts (from a single leg per species) but the low quantity of DNA obtained called for pooled sequencing to attain the minimum input requirements for library preparation. The assembled mitogenome contigs were identified against the BOLD database rather than with specimen-derived baits, simultaneously validating the identifications of matching sequences on BOLD with a curated morphological identification. Given the quality of the source material and the short read lengths obtained, the success rate in this study was unsurprisingly lower than that of Gillett et al. (2014) but these sequences would have been difficult to obtain with other methods, making this a particularly important strategy for integrating recently or locally extinct species into molecular phylogenies using existing material. Two congeneric species were found to form a chimeric assembly, however this can be resolved by adjusting the pooling strategy in cases where fragment lengths are particularly short to ensure that sequence divergences within this smaller window are maintained (Dettai et al. 2012).

Another type of complexity in voucher MMG arises when pools include one representative per morphospecies found to co-occur in a particular ecological community. In these cases the likelihood of including closely related species in the pool is high but this may not be known *a priori* unless the species are identified. As mentioned above, small numbers of congeners were successfully pooled previously with otherwise highly divergent species but the relevance of this to finding to real assemblages is unclear (Tang et al. 2014). The extent of this problem was tested tangentially by Gómez-Rodríguez et al. (2015) by checking for chimeras in the more challenging assembly of contigs from bulk MMG samples (see below) against circular mitogenomes assembled from voucher MMG (assumed to be non-chimeric). They found that chimeras did form, but infrequently and unpredictably with respect to breakpoint location, assembly program, and contig length. The sequence divergence between close relatives was not specified but all specimens had been identified and in all observed cases chimeras occurred between congeners. Whether a similar rate of chimera formation would be expected in the voucher MMG sample of the same assemblage remains unknown, but is unlikely to exceed 1%.

The final example of voucher MMG is that of Andújar et al. (2015) where DNA from each morphospecies encountered in each sample was pooled equally, for a total of six libraries. These libraries were then used in both contig- and read-based analyses for community phylogenetics. Lack of existing data for the encountered species required the generation of the reference library from the sampled specimens, while the phylogenetic focus required only the correct assignment of species presence-absence in each sample, allowing the application of a multi-library voucher MMG approach with a single round of sequencing. The assembly of these libraries individually is not expected to have been any more difficult than the voucher MMG sample of Gómez-Rodríguez et al. (2015), however the inclusion of multiple samples which are likely to overlap in species composition (with the same species potentially assembling more completely in some samples than others) introduces complexity to the subsequent steps. Read-based assemblage profiling requires that reads are matched from each sample against the same reference database and it is the merging of multiple assemblies to make this database that can be challenging and has been overcome in a number of ways (see Chapter 3 for a detailed discussion). The read matching step in itself is less challenging, although the threshold at which a species is determined to be 'present' has varied between all three studies that have included this step thus far (Andújar et al. 2015; Gómez-Rodríguez et al. 2015; Tang et al. 2015).

For bulk MMG the simplest use-case is purely for read-based analyses against a reference database generated by other means. Tang et al. (2015) generated a superbarcode library for 48 bee species of interest for monitoring wild populations by genome skimming of individual libraries and subsequently matched reads from bulk MMG samples to obtain the presence-absence and relative biomass of each species in each sample. Similarly, Gómez-Rodríguez et al. (2015) matched reads from bulk MMG samples against their reference database generated by voucher MMG. Alternatively, bulk MMG samples can be used directly for contig-based analyses (Zhou et al. 2013; Crampton-Platt et al. 2015) or a combination of contig- and read-based analyses (Gómez-Rodríguez et al. 2015). Assembly of mitogenomes from bulk MMG samples adds the challenge of uneven sequencing depth and variable intra-specific divergences to the unknown inter-specific divergences in an ecological voucher MMG sample. To date the only example of bulk MMG on a bulk DNA extraction was that of Zhou et al. (2013) wherein 73 insects were homogenised prior to differential centrifugation to enrich for intact mitochondria, DNA extraction, and sequencing. In this case assembly of the barcode region was highly successful but the recovery of long mitochondrial scaffolds was limited, leading to the suggestion that this approach would be applied as a PCR-free alternative to metabarcoding that would potentially allow analyses of relative abundance. Subsequent application of bulk MMG to a sample of nearly 500 tropical beetles was significantly more successful at obtaining long contigs (53% of species with ≥10 protein-coding genes), allowing a shift in emphasis from biodiversity discovery and richness estimation to obtaining a robust phylogenetic tree for the sampled community (Crampton-Platt et al. 2015). In the case of Gómez-Rodríguez et al. (2015) assembly of bulk MMG samples was less successful than the latter, particularly when compared with the voucher MMG equivalent. However, subsequent read mapping against the two resulting alternative reference libraries gave similar results, suggesting that the main biodiversity patterns (including biomass) were recoverable even against a highly incomplete database. This indicates that it is possible to apply both contig- and read-based analyses directly to bulk MMG samples derived from unsorted trap-catch. It was for this kind of synthetic and simultaneous analysis that MMG was originally envisioned; speeding up and simplifying wet-lab protocols and requiring only one round of sequencing to capture the full and unbiased diversity of a sample. The extent to which this is truly feasible with current costs and bioinformatics procedures will be further explored and discussed in the following Chapters.

1.2.3 Mitochondrial Metagenomics and Metabarcoding

Whilst the focus of this thesis is on MMG and exclusively comprises such data, the relationship between MMG and metabarcoding is an important one to consider. Whilst MMG offers several advantages, particularly the opportunity to integrate samples into a common phylogenetic tree and potentially retain biomass and genetic diversity information, it remains an inefficient and expensive method relative to metabarcoding as the mitochondrial fraction is generally no more than ~1% of the sequence data obtained. The assembly of these data into mitochondrial genomes is also not exhaustive, with success rates varying between species within a sample and between samples, precluding complete assembly for all species in any single study thus far. Increasing sequencing depth, read lengths and methods for mitochondrial enrichment, combined with improved assemblers may resolve some of these early problems and costs will decrease with further improvements in sequencing technology. However, the costs of sequencing and data analysis for metabarcoding will always be vastly less due to the reduced data volume requirements, and in particular the cost per species recovered will be greatly lower. Thus future studies could conceivably combine MMG and metabarcoding to maximise sequence length for some species (and hence obtain a robust phylogeny), whilst maximising species recovery from shorter metabarcodes that can then be placed in the tree relative to the species for which MMG was successful.

1.3 An Inordinate Fondness for Beetles

While the methodological focus of this thesis is MMG, the taxonomic focus is beetles (Coleoptera). This is the largest order of insects by number of described species and also encompasses great morphological and ecological diversity. With over 386,000 extant species described in 176 families and four suborders (Slipinski et al. 2011), beetles represent approximately 25% of described eukaryotic species. Estimates of the total number of extant species vary widely (e.g. 870,000 to 4.7 million) but are likely to be in the order of 1 million (Oberprieler et al. 2007). Beetles range in size from some of the smallest (e.g. Ptiliidae are generally <1mm long) to largest (e.g. *Macrodontia cervicornis* (Cerambycidae) may exceed 170mm as adults and 200mm as larvae) insects known today, while also varying in shape, colour, sclerotisation and the presence of extreme morphological structures (frequently sexually selected). Beetles can be found in most terrestrial and freshwater habitats and are highly diverse ecologically. Phytophagy, entomophagy, mycophagy, xylophagy, saprophagy and coprophagy are all present, among others. The order includes important pollinators and many agricultural and silvicultural pests, as well as parasitoids and other natural enemies. The order also includes mimetic and aposematic species and rare examples of parental care and eusociality.

Although highly diverse morphologically and ecologically, the order Coleoptera is well defined and its monophyly is not disputed. The most important feature distinguishing adult beetles from other insects is the sclerotisation of the forewings, known as elytra, which protect the body and hindwings. This can be considered a 'key innovation', a feature likely to be at least in part responsible for the success of this group (McKenna et al. 2015). The four extant suborders and the majority of major lineages within them are also well-defined morphologically yet their phylogenetic relationships are still disputed. Phylogenetic analyses at the order level are hampered by the sheer diversity of the group (with character selection being particularly challenging for morphological analyses at this scale) and are therefore rare relative to studies focussed on particular subgroups of Coleoptera. When such analyses are performed at order level the proportion of total species richness included is inevitably very small and the trade-off for molecular studies attempting to maximise taxon sampling will be the use of a small number of loci and a reliance on data matrices with a large proportion of missing data (e.g. Hunt et al. 2007; Bocak et al. 2014).

Early ancestors of modern Coleoptera, of the extinct suborder Protocoleoptera, first appear in the fossil record in the Early Permian (~280 to 270 Ma) while representatives of all four extant suborders appear in the Triassic (~240 Ma). Based on the similarity between modern Archostemata and the oldest known fossils this group has traditionally been considered the oldest extant lineage and sister to the other three suborders, wherein Adephaga was sister to Myxophaga + Polyphaga (Crowson 1960; also recovered analytically by Beutel & Haas 2000). The most significant alternative hypothesis places Polyphaga as sister to the other three suborders (all possible configurations of these three have been proposed, e.g. Kukalová-Peck & Lawrence 2004; McKenna et al. 2015; Timmermans, Barton et al. 2016), but all configurations of two pairs of sister taxa have also been proposed (Hunt et al. 2007; Pons et al. 2010; Song et al. 2010; Lawrence et al. 2011). Molecular analyses based on nuclear rRNAs, with or without mitochondrial loci, have tended to find a sister relationship between the two largest suborders, Adephaga and Polyphaga (Shull et al. 2001; Caterino et al. 2002; Hunt et al. 2007; Bocak et al. 2014), whereas recent analyses with nuclear protein-coding genes (McKenna et al. 2015) and mitochondrial genomes (Timmermans, Barton, et al.

2016) have favoured the sister relationship of Polyphaga to the other three suborders. In contrast, earlier work with mitogenomes recovered (Myxophaga + Adephaga) (Archostemata + Polyphaga) (Pons et al. 2010; Song et al. 2010).

The most recent molecular analyses finding Polyphaga as sister to the other three suborders are supported by some morphological analyses based on hindwing characters (Kukalová-Peck and Lawrence 1993; Kukalová-Peck and Lawrence 2004), and a recent analysis of the insect phylogeny using transcriptome data that included representatives of all four suborders and also firmly established Strepsiptera (twisted-wing parasites) as the sister taxon of Coleoptera (Misof et al. 2014). This positioning of the Polyphaga (~335,000 species) also minimises the imbalance in species diversity at the base of the tree (Adephaga: ~45,500 species; Archostemata: ~40 species; Myxophaga: ~100 species). Factors contributing both to the huge species richness of the order as a whole, and the imbalance between major lineages, are still largely unknown. Total richness is not explained by species radiations in association with angiosperms, although herbivory is likely to have contributed to the success of some lineages (Hunt et al. 2007). Overall, net diversification rates are high relative to related lineages but within the Coleoptera some groups show significant increases in diversification rates while others show significant decreases (McKenna et al. 2015). The origin of most major modern lineages in the Jurassic, their survival, and their diversification into many ecological niches with repeated invasions in different lineages are together the most likely major contributing factors to the success of Coleoptera overall (Hunt et al. 2007; McKenna et al. 2015), although other mechanisms will have been important within various lineages and it is as yet unknown why beetles were able to diversify so readily into such a range of niches.

1.4 Outline of Thesis

Two commonalities run throughout this thesis, firstly the study group, beetles, and secondly all data in the main analyses derive from MMG, i.e. shotgun sequencing of total DNA from mixtures of beetle specimens and subsequent bioinformatics extraction of the mitochondrial portion. The MMG data in Chapter 2 derive from a variety of experiments, some of which were undertaken by the author and some of which were undertaken by colleagues. In all cases the data sources are clearly stated, including any associated publications. The existence of such a wealth of MMG datasets for beetles (a total of 42 libraries across 12 experiments) now allows a meta-analysis to explore the effects of various experimental parameters on the quality and quantity of the data obtained (both reads and contigs), with a view to generating a set of recommendations for future experimental design and highlighting the most critical areas for further study and optimisation. Chapter 3 moves beyond the data generation step of MMG to ask how best to optimise the assembly and use the resulting data for characterising bulk samples of tropical diversity, with a view towards integrating such data into a growing mitochondrial phylogeny for beetles. Chapter 4 incorporates elements of the preceding two chapters to present a case study for the application of bulk MMG to landscape community ecology, using the New Forest National Park, UK, as a model system. Here, the beetle assemblage in leaf litter is characterised across the landscape for two different woodland types with different management histories. Finally, in Chapter 5, the lessons learnt from the development and application of MMG are drawn together and the future direction of the field is imagined.

Chapter 2 Experimental Design for MMG

Summary

This Chapter introduces the current mitochondrial metagenomics protocol and exploits the large number of existing samples from Coleoptera to examine the effect of experimental design on the results obtained. Previous studies even within Coleoptera have applied different variations of this protocol to a range of samples, limiting the opportunity to draw direct comparisons and assess the downstream effects of sample preparation. Here, the protocol is applied in a standard way to all datasets, allowing underlying differences in data quality and assembly behaviour to be exposed. Such a synthesis is a timely contribution to the growing mitochondrial metagenomics literature and provides some clear recommendations for future studies whilst also highlighting targets for further exploratory sequencing and analysis. The datasets used herein come from various projects, both published and unpublished. The source of all datasets is clearly stated.

2.1 Introduction

As seen in Chapter 1, mitochondrial metagenomics (MMG) can and has been applied to a variety of sample types and the resulting data used to answer a range of different questions. Thus MMG as a concept is somewhat nebulous, however the high degree of success encountered in existing studies demonstrates that the fundamental underlying strategy (shotgun sequencing of mixtures of total DNA) and the technology that it currently utilises (Illumina Solexa sequencing) are, together, hugely flexible. At the most basic level all MMG experiments, regardless of their ultimate aims, are concerned with a relatively small number of technical questions related to the efficiency (relative amount of mitochondrial data obtained) and success (length and number of contigs (contig-based analyses) or detection sensitivity (read-based analyses)) thereof. In the studies to date there have been a small number of consistent observations, particularly that the amount of mitochondrial data obtained varies between just 0.5 and 1.5% (Zhou et al. 2013; Crampton-Platt et al. 2015), and that a coverage of ~10x is sufficient for complete mitogenome assembly (Zhou et al. 2013; Gillett et al. 2014; Crampton-Platt et al. 2015) while increasing coverage above this threshold is not economical and even perhaps harmful (Gillett et al. 2014; Crampton-Platt et al. 2015). Beyond this, and the finding that contig-based analyses are maximally effective from voucher MMG samples (rather than bulk samples; Gómez-Rodríguez et al. 2015), few parameters affecting experimental design have been highlighted and no fully replicated and controlled experiments have been undertaken.

An early study based on simulations indicated that mitochondrial genomes might be successfully and cost-effectively assembled from mixtures of DNA and posited the potential for such an approach to complement PCR-based metagenomics (Dettai et al. 2012). Detailed consideration was given to various parameters of pooled mitogenome sequencing to aid subsequent experimental design, particularly focussing on optimal pooling strategies and is therefore of some relevance to the design of voucher MMG experiments. However, the simplified assemblies based on simulated HTS data and assumptions of high levels of mitogenome enrichment have not translated into real-world scenarios, particularly for natural samples where species composition is not known *a priori*. There remains a wide gap between the expectations derived from this study and the performance observed in other studies to date, in spite of increases in available read lengths and sequencing capacity, and the variability in bioinformatics procedures compounds the difficulty of making realistic predictions about the success of any planned experiment. With the increasing evidence pointing towards generally low efficiency of MMG for arthropods and the lack of any serious attempt at mitochondrial enrichment for natural samples, either a revision of expectations or a significant improvement in MMG methods is required.

Updating the estimates made by Dettai et al. (2012) to reflect the current maximum Illumina MiSeq output (15 Gb) and realistic mitochondrial data proportions (1% as opposed to 50% after enrichment) suggests a maximum capacity per run of 469 species at 20x for equally pooled DNA and an average mitogenome length of 16 kb. These calculations of course assume no data loss and do not incorporate estimates of assembly efficiency; instead this is simply the number of mitogenomes that could be covered to 20x with this amount of data. Even under the simplified scenarios considered by Dettai et al. (2012) complete assembly of all species was not achieved, indicating that such calculations are of limited practical value. MMG would benefit from detailed and rigorous experimentation on a range of DNA pools sequenced under different strategies and assembled with a full range of programs and parameter settings, however the high cost and potential stochasticity of sequencing (requiring replication), the complexity of such an analysis and the uncertain real-world relevance of conclusions drawn from artificial samples make this a remote prospect. Instead, the present work applies a standardised procedure to a variety of available datasets in an attempt to assess the effect of experimental variation, with the caveat that the effect of sample composition itself (quality and quantity of mtDNA per species; intra- and interspecific sequence divergence) is impossible to account for. The procedure described herein is certainly not optimal, either in general or for each specific dataset, however it has been developed to obtain good results on average across a broad range of MMG samples sequenced with Illumina MiSeq technology, and as such represents a current 'best-practice' of sorts, at least as a starting point for more detailed optimisation for individual experiments.

2.1.1 The State-of-the-Art

2.1.1.1 Data Volume

Whilst the number of MMG studies to date is relatively small and only two main bioinformatics pipelines have been used each study presents a slight variant, pursues different objectives, and uses different benchmarks of success. The main methodological outcomes are discussed here to assess the extent to which the conclusions drawn are common across multiple studies and between pipelines. Unfortunately no direct experimental comparisons between the two existing pipelines and broad sequencing strategies (HiSeq ultra-deep sequencing versus MiSeq low coverage sequencing) have been made at this time so the following synthesis is somewhat speculative. The most striking difference between the eight current studies, other than the sequencing platform used, is the ratio of input species to total sequencing volume. For the HiSeq studies using MMG for contig assembly (Zhou et al. 2013; Tang et al. 2014), 15.5 and 35 Gb of raw data were generated for 37 and 49 species respectively, while the MiSeq studies assembled contigs from between 1.8 and 16.9 Gb of raw data for between 27 and 232 species. This equates to approximately one order of magnitude difference in the amount of raw data generated per species on average (HiSeq: 0.42 and 0.71 Gb; MiSeq: 0.03 to 0.09 Gb). At the other end of the analysis, the overall reported success rate of species recovery is also highly variable. Zhou et al. (2013) recovered 34 of 37 MOTUs (91.2%) based on overlap with a portion of the *cox1* barcode region (13 did not extend further) but only five of these comprise 8 or more genes. Tang et al. (2014) had more success, with *cox1*-inclusive scaffolds containing a minimum of 7 genes obtained for all 49 taxa (20 circularised) after merging of multiple assemblies. Gene completion was considerably higher when non-overlapping scaffolds were linked based on BLAST matches to NCBI or targeted bait sequences.

For the studies using the MiSeq, reported success rates for ecological studies (based on recovery of GMYC groups) were 58.0%, 63.8%, 43.9%, and 88.6% (Andújar et al. 2015; Crampton-Platt et al. 2015; Gómez-Rodríguez et al. 2015 (*DeNovoRL* and *MitoRL*) respectively), with the reported rate of circular assembly similarly variable (Gillett et al. 2014: 19.1%; Crampton-Platt et al. 2015: 33.2%; Gómez-Rodríguez et al. 2015: 26.9% (*DeNovoRL*) and 48.3% (*MitoRL*)). Viewing these figures in the light of input data volume suggests that the pipeline reliant on MiSeq data is more efficient but with an overall lower success rate than the more data-intensive HiSeq pipeline. Where the optimum between these two lies is open to debate given the vast differences in compositional complexity and DNA fragment sizes between the various samples, although it appears that the ‘low coverage’ MiSeq strategy tends to err on the side of ‘too low’. Note that the high rate of completion at the gene level in the Tang et al. study (2014) relied upon making links between non-overlapping scaffolds, mainly based on inferring higher-level taxonomic identifications from existing GenBank data. This was made possible by the pooling strategy employed (one species per family in most cases) and thus would be a greatly uncertain step for more complex samples representing real assemblages, particularly given the wide variation in taxonomic representation between mitochondrial loci on GenBank. Non-overlapping contigs have plausibly been combined based on relative positions in the tree topology (Andújar et al. 2015; Gómez-Rodríguez et al. 2015), although no external data was used to confirm that this strategy was reliable.

Table 2.1 The main experimental design features of the six arthropod MMG studies to date wherein shotgun sequencing was applied to mixtures of genomic DNA and subsequently used for mitogenome assembly.

	Zhou and colleagues[1]	**Vogler and colleagues**[2]
Illumina platform	HiSeq 2000	MiSeq
Library type	TruSeq	TruSeq, TS PCR-free
Read length	100 and 150 bp PE	250-300 bp PE
Insert size	200 and 250 bp (reported)	307-560 bp (estimated herein)
Raw data volume	15.5 and 35 Gb	1.8-13.3 Gb (per library)
Species per library	37 and 49	27-232
Assemblers	SOAP*denovo*2, SOAP*denovo*Trans, IDBA-UD	Celera Assembler, IDBA-UD, Newbler

2.1.1.2 Sequencing Platform and Library Preparation

The differences between the sequencing platforms used are considerable, with HiSeq machines achieving far higher data volume at a cost of shorter read length and increased run times. The reported insert lengths are also shorter than those calculated herein for the MiSeq libraries (Table 2.2) but it remains unclear what effect these lengths have on the quality of the resulting assemblies, if any. The reported proportions of mitochondrial data also vary from 0.53% (following enrichment; Zhou et al. 2013) to 1.43% (Crampton-Platt et al. 2015), with the latter study hypothesising that longer insert sizes may lead to a slight enrichment of the mitochondrial fraction, for reasons unknown. Equally, the difference observed within that study may have simply reflected stochastic variation between libraries, and the apparent disparity between these two studies may arise simply from different methods for calculating the proportion of mitochondrial data. All but two of the existing studies have used TruSeq libraries and therefore in this respect are broadly comparable, although the reported rates of data loss due to quality filtering are highly variable. Of the remaining two, one used a mix of TruSeq and TruSeq PCR-free libraries, with the data retention and assembly success greatly improved in the latter (these two factors are likely linked, alongside the effect of sample type (bulk and voucher MMG respectively); Gómez-Rodríguez et al. 2015). Lastly, Timmermans et al. (2015) used a TruSeq Nano library with a high rate of data retention but relatively low assembly success, presumably related to the use of highly degraded DNA from museum specimens (average mitochondrial read length 167 bp after stitching pairs).

[1] Zhou et al. 2013; Tang et al. 2014

[2] Gillett et al. 2014; Andújar et al. 2015; Crampton-Platt et al. 2015; Gómez-Rodríguez et al. 2015

2.1.1.3 Assembly and Re-assembly

A range of assembly programs have been used in the course of MMG experiments (Table 2.1; also MIRA and IDBA_tran in Timmermans et al. 2015), with strong selection bias between the two pipelines preventing meaningful assessment of overall performance. Even studies using multiple assemblers do not necessarily draw direct comparisons between them so opportunities to examine success rates with a variety of datasets have been missed. Generally, assembled sequences are presented from a single set of program parameters so there is little information available on the effect of changing these parameters for MMG studies. Presumably, within each study the parameters have been largely optimised for the dataset in hand but whether the same settings are optimal for all samples is unclear. IDBA-UD is the only program to have been applied to both HiSeq and MiSeq data, although in the relevant HiSeq study there was no explicit discussion of the relative merits of the three assemblers used (Tang et al. 2014). In addition, HiSeq studies in all cases have used the assembled scaffolds for analysis whilst the MiSeq studies have used the contigs, due to uncertain confidence in the scaffolding step with low coverage sequencing. Programs such as IDBA-UD also introduce Ns to pad the gaps between scaffolded contigs, leading to problems with re-assembly and alignment in later steps. In the Tang et al. (2014) study final scaffold quality was assessed by comparison against the original versions and read mapping was used to highlight potentially erroneous low coverage regions. However, for bulk MMG samples, mapping quality with a range of current tools was found to be too variable to be certain that such mappings accurately reflected the assembly, precluding the application of read mapping as an assembly curation step for these samples at the current time (Crampton-Platt et al. 2015). In the latter study, combining two assemblies in Geneious (with manual curation) was found to have a positive effect on both the length distribution of the final contig set (skewed more towards long contigs) and the total number of unique sequences included in the alignments for each gene. Gillett et al. (2014) also observed the positive effect on the contig length distribution of automated merging with Minimus2. Three other studies also merged multiple assemblies but did not elaborate on the efficacy or utility of this step (Tang et al. 2014; Gómez-Rodríguez et al. 2015; Timmermans, Viberg, et al. 2016).

2.1.1.4 Pooling Strategies: Sequence Identity

From the work of Dettai et al. (2012) a pairwise divergence of at least 15% in *cox1* was recommended between all multiplexed species to aid unequivocal assembly, although correct assembly below this threshold was observed. More recently, Tang et al. (2014) successfully assembled full length mitogenomes for three *Drosophila* species to demonstrate that congeneric species could be pooled. However, the author observes that the *cox1* pairwise identity between these three sequences (KM244689, KM244693, KM644700) was still relatively high, ranging between 88% and 92%. Gómez-Rodríguez et al. (2015) observed that the number of congeneric species in the pool did not have a significant effect on the likelihood of recovery (based on GMYC analyses centred on *cox1*) but where chimeric contigs were identified they were formed between congenerics. However, the level of sequence divergence between congenerics in that study was not considered directly. Timmermans et al. (2015) observed a chimeric sequence between two close relatives and the breakpoint was traced to a region of low sequence divergence. However, in the light of Dettai et al.'s (2012) simulations, it was suggested that the short read lengths available from the museum specimens in that study exacerbated the risk of chimeric assembly as these are more likely to be fully contained within conserved regions. Thus the risk of chimeric assembly between close relatives is likely to decrease with increasing read lengths and insert sizes, and the identity threshold at which similar sequences can be reliably assembled into independent contigs is likely to increase further.

2.1.1.5 Pooling Strategies: Input DNA per Species

Finally, one fundamental issue facing MMG is the uneven recovery of mitochondrial data between species within a pool and the related question of how to predict the sequencing volume required for optimal assembly with any given combination of sequencing strategy and bioinformatics pipeline. This issue is the most significant for bulk MMG samples where the input biomass per species would usually be unknown, but it is also a challenge for voucher MMG samples due to the variation in mtDNA content within (life stage, age, tissue type) and between species. Even where an attempt has been made to equalise the amount of DNA per species assembly success has varied considerably. In general, the species that fail to assemble tend to have the lowest input DNA, but the reverse is not necessarily true.

While it is evident that long mitogenome sequences can be assembled with as little as 10x coverage (Gillett et al. 2014; Crampton-Platt et al. 2015) in practice this is not a useful guide when calculating the amount of sequencing required for any given experiment due to the wide variation in sequencing depth observed even within studies where DNA was equilibrated at the pooling step (Gillett et al. 2014; Tang et al. 2014). For bulk MMG no such steps are taken and calculations of mean expected data proportions per species or specimen are likely to significantly underestimate the amount of data required (even where the number of species/specimens is known) due to the variation in species biomass in such samples, on top of any underlying intrinsic differences in mitochondrial proportion. This, in addition to uncertainty in the precise amount of data obtained, drastically different rates of data loss due to low quality base calls, variation in overall mitochondrial proportion and differences between conspecifics hamper effective experimental design. Although voucher MMG has been shown to be more efficient for the assembly of long mitogenome sequences than bulk MMG (Gómez-Rodríguez et al. 2015) it is worth noting that no study to date has obtained anywhere near complete recovery of input species. In addition, the observed variation in coverage even in the most controlled experiment thus far (100 ng gDNA per species; Tang et al. 2014) shows that equilibration of genomic DNA translates poorly into equal mitogenome sequencing. More positively, the latter study found no evidence that the quality of an individual DNA sample affected the likelihood of assembly, and the successful assembly of long contigs from degraded DNA from dried museum specimens (Timmermans, Viberg, et al. 2016) goes some way to allaying fears that variation in DNA degradation in mass-trapped arthropod samples might bias the outcome of bulk MMG.

2.1.1.6 Chapter Aims and Expectations

It is clear that the further development of MMG would benefit from a concerted experimental effort to explore the boundaries and effect of some of the points highlighted above to determine which factors are the most critical for success and the steps that are most in need of re-evaluation. Such an experiment is not imminently foreseeable, however the rapid growth in MMG datasets available to the author present a significant but limited opportunity for an in-depth exploration of some of the issues discussed. These analyses are limited both taxonomically and methodologically, covering only beetles and the Illumina MiSeq platform respectively. However, this narrow focus allows the most important common factors affecting these datasets to be identified and will hopefully facilitate the design of simple confirmatory experiments in other systems. Areas under investigation include the extent of data loss due to read pre-processing, observed mitochondrial data proportions and the effect of library preparation, variation in assembly performance, and the differential assembly behaviour of voucher and bulk MMG samples.

Read processing steps are expected to remove a low proportion of reads overall, with TruSeq (TS) libraries probably suffering more from proportional data loss at this step that TruSeq Nano (TSN) and TruSeq PCR-free (TSP) libraries. This is because the effect of quality control is likely to be greater in the older TS libraries as the quality of Illumina MiSeq data has theoretically improved with each new kit release and the majority of TS libraries herein predate the current MiSeq v3 chemistry. Additionally, the corresponding shorter read lengths of the TS libraries (250 bp, c.f. 300 bp) allows less margin for quality trimming when using a fixed minimum read length requirement for retention (150 bp). In contrast, the adapter removal step should affect all library types similarly as the majority of adapter sequences will be removed by the MiSeq software before it reaches the end-user. The insert sizes of TSN and TSP libraries are expected to be the same as all were made using the 550 bp kit, whereas the TS libraries will show a greater range due to variation in user-requested sizes, but overall the mean is expected to be lower than that for TSN/TSP as the recommended (i.e. default) length was 300 bp. Choice of library should have no effect on the proportion of mitochondrial data obtained. Insert size has previously been hypothesised to have an effect on the proportion of mitochondrial data (Crampton-Platt et al. 2015) but there is no clear reason why this would be the case. If indeed such a pattern is identified, the larger insert libraries (TSN and TSP) are expected to have a greater mitochondrial proportion than TS libraries on average, but TS libraries with insert sizes in the range of TSN/TSP should have a similar proportion. Any response of mitochondrial proportion to insert size is expected to be the same for all three library types.

No systematic differences in assembly behaviour are expected *a priori* between the three assemblers trialled herein. Crampton-Platt et al. (2015) previously observed that IDBA-UD produced more short (<5 kb) and more long (≥15 kb) than Celera Assembler for the *BorneoCanopy* dataset, but this may not be observed repeatedly or more widely. However, in all cases longer insert sizes are expected to aid assembly of long contigs as the likelihood of spanning regions of low interspecific divergence increases with fragment size, allowing the sequences to be resolved correctly. Mean sequencing depth per species is expected to have a significant effect on the likelihood of long contig assembly and a mean coverage of ~10x is expected to be required for the assembly of complete mitogenomes (Gillett et al. 2014; Crampton-Platt et al. 2015).

Voucher and bulk MMG samples are expected to show divergent assembly behaviour, with bulk samples overall less efficient with respect to sequencing effort. Variable biomass in bulk samples leads to a highly uneven distribution of reads between species and this, combined with a low coverage sequencing strategy, reduces the likelihood of contig assembly for low biomass species while high biomass species will still assemble successfully at reduced sequencing volume. In contrast, input DNA is equalised between species as far as possible in voucher MMG samples and so the likelihood of assembly is expected to be similar for all species and to be more closely dependent on overall sequencing volume. In addition, voucher MMG samples eliminate intraspecific genetic variation whereas bulk MMG samples are likely to contain variable levels of intra- and interspecific variation. This has previously been hypothesised to complicate the assembly of bulk MMG data, leading to reduced assembly success (observed as multiple short contigs) even at high coverage for some species (Crampton-Platt et al. 2015), whereas contig length is expected to correlate closely with coverage for voucher MMG.

2.2 Materials and Methods

2.2.1 Data Description

The analyses presented herein were applied to a range of Illumina MiSeq datasets sequenced over approximately two years (December 2012-December 2014). These datasets derive from experiments with different aims and amounts of available DNA, and thus vary in experimental design but all are comprised exclusively of Coleoptera, except for a few instances of misidentified non-beetle larvae. The length of the reads obtained is generally (but not always) a reflection of when the library was sequenced, as the MiSeq v3 chemistry and 600-cycle kit was introduced in August 2013. Similarly, by March 2014, the original TruSeq library kits had been phased out and largely replaced with TruSeq Nano and TruSeq PCR-Free kits, thus library type is loosely related to the date of library preparation. A description of each experiment is given below and the associated publication listed where appropriate. Experimental design details are summarised in Table 2.2.

2.2.1.1 BorneoCanopy

Experiment conducted by the author (Crampton-Platt et al. 2015). A sample of 477 beetle individuals representing approximately 209 morphospecies, derived from rainforest canopy fogging in Danum Valley, Sabah, Malaysia. DNA was extracted destructively from each individual separately and then pooled in equal volumes. This is the only experiment where two libraries of the same type were made from the same DNA pool, attempting to test the effect of increasing insert size on assembly success. DNA barcodes (*5'-cox1*) are available for 327 of 477 specimens, or 161 of 209 morphospecies. In the published analysis of this dataset, combining the DNA barcodes and contigs gave an estimate of 232 species.

2.2.1.2 IberSoils

Experiment conducted by Carmelo Andújar and Paula Arribas (Andújar et al. 2015). Six libraries from three locations in southern Spain, each of which was comprised of specimens representing each morphospecies extracted from up to 28 soil pit samples per location. Samples were split into 'superficial' soil (leaf litter and up to 5cm depth of topsoil) and 'deep' soil (2500 cm^3 up to a depth of 40 cm) in each case and the soil fauna extracted with Berlese apparatus. DNA was extracted individually from up to three specimens per morphospecies (total number of individuals: 535 adults and 959 larvae) and pooled such that each morphospecies was represented by approximately the same amount as estimated with a NanoDrop Spectrophotometer. DNA barcodes are available for 288 GMYC species out of an estimated total of 324 (contig and Sanger barcodes combined).

2.2.1.3 ChrysIber

Experiment conducted by Carola Gómez-Rodríguez (Gómez-Rodríguez et al. 2015). This experiment had two components; firstly, a set of ten natural samples where DNA from all Chrysomelidae specimens collected at each of ten protected areas throughout Spain was pooled in equal volumes (herein '*ChrysoAL*'). The DNA had previously been extracted from the prothorax of all individuals collected from a total of 20 sites across the Iberian Peninsula (Baselga et al. 2015). All specimens were identified morphologically to species. The second component was a reference library of DNA from one representative specimen per morphological species known to be present in the ten natural samples, plus an additional 5 species from an adjacent locality (herein '*ChrysoRL*'). In this case, the volume of DNA pooled was based on specimen size (in four classes) to approximately equilibrate the amount per species, with the greatest volume of eluate taken from the smallest specimens. All 11 libraries from this experiment were sequenced twice. DNA barcodes are available for 165 of 171 *ChrysoAL* species (170 of 176 *ChrysoRL* species).

2.2.1.4 UK-BI

Experiment conducted by author on behalf of the NHM Biodiversity Initiative. An equilibrated sample (based on Qubit fluorometer measurements of individual DNA extractions) comprising a single specimen for each of 165 (morpho)species sampled in three sites in the United Kingdom by the NHM Biodiversity Initiative (Wytham Wood, Oxfordshire; New Forest, Hampshire; Epping Forest, Essex). The sample was prepared and sequenced twice, once with each of the TruSeq Nano and PCR-free kits. No DNA barcodes were generated for this dataset; instead contigs were identified via publicly available sequences (GenBank and BOLD).

2.2.1.5 FrenchGuianaFIT

Experiment conducted by Julia Lipecki (MSc Applied Biosciences and Biotechnology, Imperial College London, 2014) on behalf of the NHM Biodiversity Initiative. Two libraries from each of two sites in Nouragues National Nature Reserve, French Guiana, each comprising a single representative specimen of each morphospecies identified from a flight-intercept trap sample (FIT). Morphospecies sorting was undertaken independently for each

Table 2.2 Experimental design details and data volume for each library in the present study.

Experiment	Library	No. individuals	(Est.) No. species	Pooling	Library	Read length (bp)	Mean insert size (bp)	Raw reads (pairs)	Pairs for assembly	Est. 'true' mito. pairs (% of QC)
BorneoCanopy	BC-short	477	(232)	Volume	TruSeq	250	425	16,996,158	833,709	157,909 (1.86)
	BC-long	477	(232)	Volume	TruSeq	250	440	16,898,216	1,257,165	224,507 (1.98)
IberSoils	Cadiz-Deep	327	(138)	Equi.	TruSeq	250	441	11,910,681	1,100,149	202,845 (2.45)
	Cadiz-Supr.	471	(104)	Equi.	TruSeq	250	342	9,362,853	605,055	101,658 (1.33)
	Ciudad-Deep	166	(72)	Equi.	TruSeq	250	377	5,851,175	444,153	102,456 (2.29)
	Ciudad-Supr.	170	(43)	Equi.	TruSeq	250	336	7,275,223	448,747	60,737 (1.05)
	Cordoba-Deep	203	(91)	Equi.	TruSeq	250	377	7,102,170	376,901	58,200 (1.09)
	Cordoba-Supr.	157	(35)	Equi.	TruSeq	250	348	4,788,824	269,420	30,305 (0.79)
ChrysIber	ChrysoRL [1]	176	176	Equi.	TS PCR-free	300	561	3,547,641	430,663	51,915 (1.73)
	ChrysoRL [2]	176	176	Equi.	TS PCR-free	300	560	22,142,793	2,595,762	310,375 (1.67)
	AllLoc-ADS	273	41	Volume	TruSeq	250	347	3,826,264	164,202	32,510 (1.18)
	AllLoc-ANC	327	67	Volume	TruSeq	250	340	3,818,855	182,606	30,488 (1.11)
	AllLoc-EUM	223	41	Volume	TruSeq	250	338	3,650,663	163,717	26,355 (1.02)
	AllLoc-HOR	156	27	Volume	TruSeq	250	334	3,897,855	176,161	37,070 (1.28)
	AllLoc-JCB	206	36	Volume	TruSeq	250	308	3,718,361	174,807	30,855 (1.19)
	AllLoc-LAS	336	56	Volume	TruSeq	250	348	3,953,642	203,824	35,185 (1.20)
	AllLoc-MAC	232	49	Volume	TruSeq	250	321	4,639,716	217,340	40,461 (1.20)
	AllLoc-OMA	299	45	Volume	TruSeq	250	331	4,348,412	142,421	27,330 (0.95)
	AllLoc-SAN	252	47	Volume	TruSeq	250	317	4,300,812	124,019	26,837 (0.84)
	AllLoc-TUE	303	48	Volume	TruSeq	250	317	4,045,998	149,492	31,537 (1.00)
UK-BI	UK-BI-Lib1	165	165	Equi.	TS PCR-free	300	521	3,348,719	260,300	42,347 (2.11)
	UK-BI-Lib2	165	165	Equi.	TS Nano	300	461	12,433,632	938,670	114,849 (1.06)
FrenchGuianaFIT	FG-site1	163	(163)	Equi.	TS Nano	300	522	10,400,450	744,819	109,673 (1.31)
	FG-site2	216	(216)	Equi.	TS Nano	300	504	11,804,601	881,492	150,184 (1.57)
PanamaVane	P-Wk1	96	(96)	Equi.	TS Nano	300	518	9,734,394	789,021	142,904 (1.78)
	P-Wk2	96	(96)	Equi.	TS PCR-free	300	533	9,537,654	436,393	44,237 (0.58)
	P-Wk4	226	(226)	Equi.	TruSeq	250	487	13,546,286	863,136	180,977 (1.91)
RichmondPark	RP-Water	21	21	Equi.	TruSeq	250	499	15,000,088	2,485,220	178,329 (1.36)
	RP-Ground	24	24	Equi.	TruSeq	250	520	6,931,625	523,320	82,114 (1.52)
Curculionoidea	Curculionoidea	173	173	Equi.	TruSeq	250	355	18,341,901	1,086,684	238,716 (1.57)

Scolytinae	Scolytinae	72	72	Equi.	TS Nano	300	529	7,654,569	565,507	140,741 (2.38)
Staphyliniformia	Staphyliniformia	148	148	Equi.	TS Nano	300	509	14,366,793	1,639,046	187,214 (1.52)
Scarab. & Chryso.	Scarabaeinae [1]	49	49	Equi.	TS Nano	300	423	4,858,513	235,318	36,256 (1.59)
	Scarabaeinae [2]	49	49	Equi.	TS Nano	300	475	6,015,065	478,818	69,801 (1.58)
	Chrysomelidae [1]	79	79	Equi.	TS Nano	300	446	20,342,150	1,165,676	139,473 (1.35)
	Chrysomelidae [2]	79	79	Equi.	TS Nano	300	484	17,427,851	1,615,669	180,839 (1.27)
	ChrysoScarab [1]	127	127	Equi.	TS Nano	300	448	10,647,749	604,772	80,905 (1.58)
	ChrysoScarab [2]	127	127	Equi.	TS Nano	300	478	12,217,914	1,134,589	143,989 (1.58)
ReferenceSet	Run1[3]	479	479	Volume	TruSeq	300	340	24,114,781	1,508,581	429,525 (2.78)
	Run2-Lib1	153	153	Equi.	TS Nano	300	455	13,575,074	994,189	88,938 (0.84)
	Run2-Lib2	81	81	Equi.	TS Nano	300	479	9,122,099	648,743	54,011 (0.67)
	Run2-Lib3	78	78	Equi.	TS Nano	300	500	8,951,062	606,371	56,271 (0.72)

Table 2.3 Input data volumes and main mitogenome assembly results per dataset.

Dataset	**(Est.) No. species**	**Raw reads (pairs)**	**Pairs for assembly**	**IDBA-UD *cox1-5'***	**IDBA-UD >10 kb**	**Newbler *cox1-5'***	**Newbler >10 kb**	**Celera *cox1-5'***	**Celera >10 kb**
BorneoCanopy	(232)	33,894,374	2,090,874	161	110	142	98	174	77
IberSoils	(324)	46,290,926	3,244,425	252	113	212	86	263	91
ChrysIber (RL)	176	25,690,434	3,026,425	173	144	164	131	161	130
ChrysIber (AL)	171	40,200,578	1,698,589	141	31	104	39	104	42
UK-BI	165	15,782,351	1,198,970	94	56	82	53	103	62
FrenchGuianaFIT	Unknown	22,205,051	1,626,311	150	108	132	95	155	108
PanamaVane	Unknown	32,818,334	2,088,550	244	141	229	138	243	139
RP-Water	21	15,000,088	2,485,220	23	20	21	15	21	17
RP-Ground	24	6,931,625	523,320	22	15	20	17	20	17
Curculionoidea	173	18,341,901	1,086,684	122	86	124	79	120	57
Scolytinae	72	7,654,569	565,507	63	56	61	61	63	58
Staphyliniformia	148	14,366,793	1,639,046	94	68	84	63	98	68
Scarabaeinae	49	10,873,578	714,136	33	22	32	22	32	16
Chrysomelidae	79	37,770,001	2,781,345	77	56	68	50	78	50
ChrysoScarab	127	22,865,663	1,739,361	101	65	94	55	112	65
ReferenceSet	538	55,763,016	3,757,884	229	117	220	98	244	57

[3] Included in *ReferenceSet* assembly but not in read-based analyses.

sample. The volume of DNA pooled per specimen was equilibrated based on Qubit fluorometer measurements of individual extractions. DNA barcodes and/or cytochrome b (*cob*) sequences are available for 224 and 302 specimens respectively.

2.2.1.6 PanamaVane

Experiment conducted by Kirsten Miller. Three samples of non-scolytine/platypodine beetles sampled by vane trapping over three weeks at two heights (1m and 10m) in tropical forest on Barro Colorado Island, Panama. Each library comprises DNA from a single representative per morphospecies, with the volume of DNA equilibrated based on Qubit fluorometer measurements of individual extractions. Available DNA barcodes are not presently useful as 'bait' sequences, as the specimens have not been identified.

2.2.1.7 RichmondPark

Experiment conducted by Paula Arribas and Carmelo Andújar. Two samples of specimens hand-collected in Richmond Park SSSI, Greater London. One sample comprises all terrestrial beetles (adults and larvae) found in the environs of Adam's Pond (grassland and woodland) and the other comprises all aquatic beetles (adults and larvae) found in Adam's Pond. DNA was extracted from each specimen individually and an equilibrated pool generated based on Qubit fluorometer readings. One or two bait sequences (*cox1-5'* and *cob*) are available for each specimen.

2.2.1.8 Curculionoidea

Experiment conducted by Conrad Gillett (Gillett et al. 2014). One equilibrated library (based on Qubit fluorometer measurements of individual DNA extracts) of 173 species of Curculionoidea assembled to increase sampling of the mitochondrial phylogeny of this superfamily. 31 species had too little DNA available for equilibration so these were added to the pool by volume. Between one and three bait sequences (*cox1-3'*, *cob*, *16S*) were available for each species.

2.2.1.9 Scolytinae

Experiment conducted by Kirsten Miller. One equilibrated library (based on Qubit flurometer measurements of individual DNA extracts) comprised of one individual for each of 45 and 25 (morpho)species of Scolytinae/Platypodinae from Barro Colorado Island, Panama, and coniferous forest in the United Kingdom, respectively. DNA barcodes are available for all species.

2.2.1.10 Staphyliniformia

Experiment conducted by Emeline Favreau (MRes Biosystematics, Imperial College London, 2014). One equilibrated library (based on Qubit flurometer measurements of individual DNA extracts) of 148 Staphyliniformia species generated to increase taxon sampling of the mitochondrial phylogeny of this infraorder. No bait sequences were generated specifically for this experiment, instead contig identification relied upon existing publicly available sequences (GenBank and BOLD).

2.2.1.11 Scarabaeinae and Chrysomelidae

Experiment conducted by Thijmen Breeschoten (MSc Biology, Universiteit Leiden, 2015). Three equilibrated libraries (based on Qubit flurometer measurements of individual DNA extracts) including species of Scarabaeinae (mostly Onthophagini) and Chrysomelidae. One library consists of exclusively of Scarabaeinae, one of Chrysomelidae, and one comprises a mixture of the two. Each library was sequenced twice.

2.2.1.12 ReferenceSet

Experiment conducted by Amie Hunter (MRes Biodiversity Informatics and Genomics, Imperial College London, 2014). Four libraries comprising DNA donated by Zoologische Staatssammlung München. An initial round of sequencing was based on one library comprised DNA from 479 species pooled by volume. A second round of sequencing was based on three equilibrated libraries with 153, 81 and 78 species respectively. There was significant species overlap between the two rounds of sequencing. The initial library is the only example of TruSeq library sequenced with the 600-cycle kit and thus was not included in the read processing analyses outlined below, however it was combined with the other three libraries to maximise the amount of data available for assembly. DNA barcodes were kindly provided by Jérôme Morinière (German Barcoding-of-Life, ZSM) ahead of publication on BOLD.

2.2.2 Mitogenome Assembly

The same data processing and assembly procedures were applied to all libraries, regardless of sample type. Initially the forward and reverse reads for each library were processed with Trimmomatic (v0.30; Lohse et al. 2012) to remove any sequencing adapters not previously detected and removed by the MiSeq software. Default clipping settings were used

(ILLUMINACLIP:2:30:10) but in all cases the applicable index was included in the indexed adapter sequence template file provided to the program. Paired reads passing this step were further filtered with Prinseq-lite (v0.20.4; Schmieder and Edwards 2011) to remove low quality sequences (-min_len 150 –min_qual_mean 25 –trim_qual_right 20 –ns_max_n 0). Only pairs where both reads passed quality control were retained. These reads were then filtered independently (in FASTA format) against a database of 245 coleopteran mitochondrial genomes (*MitoDB*; Timmermans et al. in review) using BLAST (-task blastn –evalue 1e-5 –max_target_seqs 1 –dust no; Altschup et al. 1990) to retain only 'mitochondrial-like' reads. All pairs where at least one read returns a hit of any length with an E-value of 1e-5 against the *MitoDB* were retained, using cdbfasta/cdbyank (The Institute for Genomic Research, Available from: http://sourceforge.net/projects/cdbfasta/) to extract these pairs from the quality-controlled FASTQ files. This step applies only a loose filter to the data to minimise the loss of truly mitochondrial reads which are divergent from the *MitoDB* sequences, thereby functioning primarily as a data reduction step to minimise the computational demands of *de novo* assembly. A more conservative method for estimating the number of mitochondrial reads is detailed below.

The paired, quality-controlled, 'mitochondrial-like' reads for each experiment were then assembled using three programs: IDBA-UD (--mink 80 --maxk [read length] --similar 0.98; (Peng et al. 2012), Newbler (-mi 98 –ml 150 -rip; Margulies et al. 2005), and Celera Assembler (doOverlapBasedTrimming=0 doToggle=1 toggleUnitigLength=1000 unitigger=bogart; Myers et al. 2000). IDBA-UD requires paired reads to be interleaved in FASTA format whilst Newbler requires paired reads to be interleaved in FASTQ format with pre-Casava 1.8 style read headers. Celera Assembler reads in FASTQ data through a FRG wrapper containing information about the library. In all cases the technology was specified as 'illumina-long', quality type 'sanger', read orientation 'innie', and insert size of 500 bp (±200 bp), for paired reads in separate files.

The contigs assembled by each program were filtered by length using samtools and bedtools to retain only those ≥1 kb. These were then further filtered with BLAST (-task blastn –evalue 1e-5 –max_target_seqs 1) against *MitoDB* and the results processed to extract probable mitochondrial contigs with cdbfasta/cdbyank. A BLAST hit-length of 1 kb was previously found to balance the correct removal of non-mitochondrial contigs with the retention of mitochondrial contigs of a useable length (Crampton-Platt et al. 2015).

'True' Mitochondrial Reads

The number of reads expected to truly originate from the mitochondrial genome, as opposed to 'mitochondrial-like' reads, was estimated from the blastn results from the read filtering step, by manipulating the hit tables with *awk*. Pairs where both reads returned a hit of a least 100 bp with E≤1e-5 were assumed to be truly mitochondrial. Both the raw number of 'true' mitochondrial reads and their proportion of the quality-controlled reads were used in later analyses.

2.2.3 Read Processing and Mitochondrial Proportions

All of the following analyses were conducted in R using the core packages unless otherwise stated (R Core Team 2015). All plots were produced with the *lattice* package (Sarkar 2008). Variability was observed between libraries in the number and proportion of reads discarded at each read-processing step (adapter removal and quality control) and after filtering for 'mitochondrial-like' reads. The possibility of a systematic effect of library preparation (TruSeq, TS PCR-free, TS Nano) was tested for by analysis of deviance, using generalised linear models (glm; function *glm*) with quasibinomial errors (logit link) to account for overdispersion in the response variable (number of reads retained, treated as a proportion). F tests (function *anova*, test="F") were used to test the significance of the models.

An estimate of the insert size for each library was obtained by read mapping to the IDBA-UD contigs (see below) with SMALT, requiring 98% identity (-y 0.98; v 0.7.6; Wellcome Trust Sanger Institute, Available from: https://www.sanger.ac.uk/resources/software/smalt/). The SAM alignment files produced by SMALT were converted to BAM with samtools (Li et al. 2009) and parsed with a Python script (https://gist.github.com/davidliwei/2323462#file-getinsertsize-py) to obtain the insert size estimate. Where multiple libraries from the same experiment were combined for assembly, both the combined set of reads and the individual libraries were mapped against the contigs derived from the combined IDBA-UD assembly to estimate the average insert size for assembly (see below) and per library respectively. The effect of library preparation on insert size and mitochondrial proportion was assessed by one-way ANOVA (analysis of variance) and logistic regression respectively (function *aov*; function *glm*, family="quasibinomial"). The combined effect of library preparation and insert size on mitochondrial proportion was assessed by logistic analysis of covariance (ANCOVA; function *glm*, family="quasibinomial").

2.2.4 Defining Assembly Success for MMG

In assessing the assembler performance the primary aims of contig-based mitochondrial metagenomics should be considered, namely to obtain as complete a representation of the species in the pool as possible. Completeness can be judged in one of two ways, firstly in terms of the proportion of species recovered and secondly in terms of the completeness of the contigs representing those species. For a highly fragmented assembly the number of contigs will be a very poor indicator of the number of species recovered as most species will be represented by multiple non-overlapping contigs, requiring a gene-centred approach to species richness estimates to ensure orthology (Chapter 2; Andújar et al. 2015; Crampton-Platt et al. 2015; Gómez-Rodríguez et al. 2015). Analogous to this, the success of species recovery for each assembler can be assessed by comparing the number of assembled contigs which contain a particular locus, for example the *cox1*-5' 'barcode' region. In some studies the barcode region may be the main target for assembly (e.g. Zhou et al. 2013), however in most cases (e.g. phylogenetics and ecological reference libraries) maximising sequence lengths and contiguity will be the priority, with low levels of species recovery addressed either by higher coverage sequencing *a priori* or additional *post hoc* sequencing to obtain missing species. Thus the most relevant measure of assembly success for MMG is the length distribution of the resulting contigs, with the most successful assembly of any given dataset considered to be the one producing the most complete and nearly-complete mitogenome sequences. This, and the ratio of long contigs (≥10 kb) to the number of input species (where known) is therefore the most relevant benchmark for success for MMG. Herein, mitochondrial contigs ≥15 kb are considered complete (likely to contain all 13 protein-coding genes and 2 rRNAs) whilst those 10-15 kb are considered nearly-complete (likely to contain at least 8 of these 15 genes).

The number of *cox1-5'* 'barcode' sequences generated by the assemblers can also be tracked as a secondary measure of success. The number of reads required to assemble this region (approximately 660 bp) will be lower than the number required to assemble a contig ≥10 kb, thus these are expected to accumulate more rapidly and should be a more complete representation of the number of input species where sequencing depth is insufficient to assemble a single long contig for each. In most cases the barcode region will be the main 'bait' sequence for linking an assembled mitogenome to a particular species and thus the assembly rate of this marker is of particular interest. The barcode region is extracted bioinformatically with cdbfasta/cdbyank based on the co-ordinates of hits from BLAST searches of the contigs against a small database of sequences for the target taxonomic group (-task blastn –evalue 1e-5 –max_target_seqs 1; filter hit table for hits ≥250 bp).

2.2.5 Assembler Performance and Insert Size

The contig length distributions produced by each assembler for all datasets combined were compared by two-sample Kolmogorov-Smirnov tests (function *ks.test*) and each was tested for unimodality with Hartigan's dip test (function *dip.test*, package *diptest* (Martin Maechler, Available from: https://cran.r-project.org/web/packages/diptest/index.html)). Comparisons were also made between assemblers for each dataset individually in the same way. The variation between the assemblers in the proportion of contigs in various length classes was assessed by analysis of deviance (function *glm*, family="quasibinomial"). Lastly, logistic regression was used to assess the effect of insert size on the contig length distribution by modelling the proportion of contigs in each of four size classes (1-5 kb, 5-10 kb, 10-15 kb, ≥15 kb) as a function of average insert size and assembler (function *glm*, family="quasibinomial").

2.2.6 Sequencing Effort and Species Recovery

The effect of data volume on species recovery was analysed by logistic regression after normalising by the number of input species (for libraries where this is known) to control for variation in sequencing effort between experiments. The number of assembled sequences as a proportion of input species (*cox1* and contigs ≥10 kb) was modelled as a response to the number of 'true' mitochondrial pairs normalised by number of input species (function *glm*, family="quasibinomial").

2.2.7 Voucher MMG versus Bulk MMG

Thus far, the sample type has been ignored in the analyses, however any differences between the behaviour of bulk MMG (variable input DNA per species) and voucher MMG (input DNA equalised per species as far as possible) will have important implications for future ecological experiments. The success of assembly from natural samples will determine whether such experiments can be completely '*de novo*', with the assembly of contigs and the assessment of species presence-absence by read-mapping against those contigs achievable with the same samples, or whether a reference library for the anticipated species must be constructed separately first. Gómez-Rodríguez et al. (2015) compared these two approaches and found that natural samples produced assemblies that were more fragmented and incomplete than a reference library for the same species. The data from this experiment (*ChrysIber*) is re-examined here with a focus on the effect of data volume on assembly success in these two sample types. Library type, mean insert size and read length all differed between the two datasets and thus the effect of these on assembly success cannot be controlled for. Differences in overall assembly success due purely to the smaller number of reads available for assembly in *ChrysoAL* were accounted for with additional assemblies for the *ChrysoRL* data subsampled to the number of reads in *ChrysoAL*. Whilst this is a fairer comparison than the assembly of the full dataset, *ChrysoRL* reads are also longer (300 bp vs. 250 bp) and therefore the subsampled assembly still included a greater data volume.

To assess the assembly behaviour of the two sample types in response to increasing sequencing effort subsamples of the quality-controlled, 'mitochondrial-like' reads were taken every 100,000 pairs from each dataset and assembled with IDBA-UD using the parameters outlined previously. The numbers of *cox1-5'* barcodes and long contigs across successive subsamples were plotted as proxy measures for species accumulation. Note that this is not a true species accumulation curve because the identity of the sequences was not compared between subsamples, although as each successively larger subsample included the reads from the smaller subsamples the same sequences are likely to be assembled repeatedly.

The effect of sequencing depth on contig length for the two sample types was assessed visually by plotting contig length against mean coverage, estimated by read mapping with SMALT, as above. Such plots were made for all three assemblers for each of the two sample types, and additionally for the subsampled *ChrysoRL* IDBA-UD assembly and an IDBA-UD assembly of the *ChrysoAL* data with the minimum contig length parameter set to 1 kb (IDBA-1k; default: --min_contig 200). Lastly, equivalent plots were made for the contigs published by Gómez-Rodríguez et al. (2015) to assess whether the conclusions drawn in that study were biased by differential assembly success in the two datasets.

2.3 Results

2.3.1 Mitogenome Assembly

The results of the read processing steps and mitogenome assembly for each library or experiment are summarised in Table 2.2 and Table 2.3 respectively. Note that there were only four samples prepared as TruSeq PCR-free libraries (c.f. 23 TruSeq and 15 TruSeq Nano), and all libraries sequenced with the 500 cycle MiSeq v2 kit were TruSeq, and all but one of the TruSeq libraries were sequenced with the 500 cycle kit. Thus the effect of kit cannot be separated from the effect of library type and has been ignored in the analyses. For several experiments, multiple library types were combined for assembly so the effect of library on assembly success cannot directly be assessed. Instead, variation measured from the data themselves such as insert size and the number and proportion of mitochondrial reads were correlated with variation in assembly success. In all cases the three assemblers each produced a large number of mitochondrial contigs of varying lengths, with a tendency towards assembling the smallest and largest contigs in most cases, creating apparently bimodal length distributions. In all cases the number of short contigs far exceeded the number of long contigs, as illustrated by the differential recovery of *cox1* sequences and long contigs highlighted in Table 2.3.

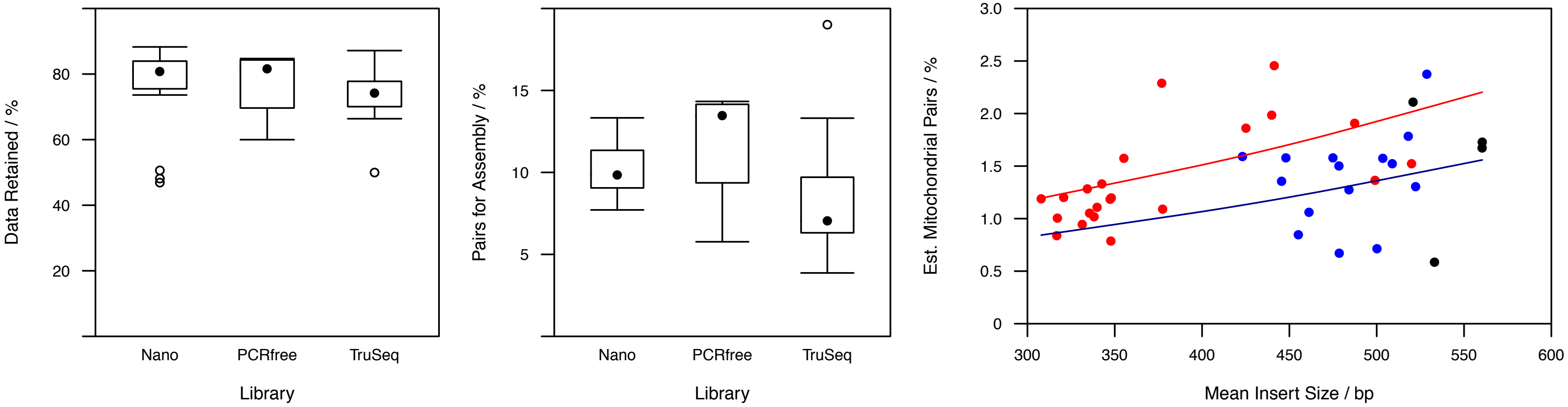


Figure 2-1 Library type, read processing and mitochondrial proportions. Left: The percentage of input read pairs retained following adapter removal and quality control for each library type. Middle: The percentage of read pairs retained for assembly after filtering against *MitoDB* for each library type. Right: The effect of insert size and library type on the percentage of quality controlled reads that are estimated to be truly mitochondrial. TruSeq libraries are shown in red, TruSeq Nano libraries in blue, and TruSeq PCR-free libraries in black. The corresponding fitted lines are for TruSeq libraries (red) and TSN+TSP libraries (dark blue) respectively. Note that in the two boxplots all datasets have been included for illustration purposes but the outliers were removed prior to analysis.

2.3.2 Read Processing and Mitochondrial Proportions

Taking the removal of adapter contamination and quality control as a single 'read processing' step, there was no significant difference between the library types or kits when considering all datasets. The oldest TS library (500-cycle) and a single MiSeq run (600-cycle) with three TSN libraries resulted in unusually high data loss at the adapter removal and quality control steps respectively and thus appeared as outliers overall. A second run with the same three TSN libraries resulted in quality control losses more similar to other TSN libraries, indicating that there was a technical problem with the first run rather than with the libraries or samples themselves. After removing these four outliers there was a significant effect of library on the proportion of data retained, with TS samples performing significantly worse than others. TSP and TSN libraries were not significantly different and were therefore combined in the minimum adequate model (Figure 2-1; $F_{1,36}$= 10.47, p=0.003; TS: μ=75.6%; TSN/TSP: μ=81.7%). Treating these as a single step is justified in that these two processes are always likely to be applied together to MMG samples, and the main concern is to minimise data loss overall. Taking each step individually indicated that the first (removing adapter contamination) was the main driver behind the observed differences, with a significant increase in data loss found in TS libraries as compared with the other two types (TSN and TSP combined; $F_{1,36}$= 7.52, p=0.01; TS: μ=86.6% retained; TSN/TSP: μ=92.2%). No significant difference between library types was observed when applying quality control to the post-Trimmomatic reads.

When filtering the quality-controlled reads from all datasets against *MitoDB*, no significant differences were observed between libraries. However, this finding was strongly influenced by a single TS sample (*RP-Wate*r: 19.01% reads retained) that is known to include several non-beetle larvae at high biomass that may have an unforeseen effect at this step. After removing this library from the analysis there was a significant difference between TS libraries and TSN/TSP combined in the proportion of quality-controlled reads retained for assembly (Figure 2-1; $F_{1,39}$= 9.32, p=0.004; TS: μ=8.3%; TSN/TSP: μ=10.6%). When considering only the portion of quality-controlled reads that were 'truly' mitochondrial (all datasets) there was no significant effect of library type and no clear outliers that might have affected this result. However, when including insert size as a predictor a significant positive correlation was observed, with a significantly greater response of TS libraries to insert size than TSN/TSP libraries (Figure 2-1; $F_{2,40}$= 5.17, p=0.010). There is a clear difference in insert sizes between the three library types, with TSP libraries found to have significantly longer inserts than TSN libraries when this was analysed independently ($F_{2,40}$= 29.97,

$p<0.001$; TS: μ= 375 bp; TSN: μ=482 bp; TSP: μ=544 bp). The difference between TSP/TSN libraries and TS libraries is to be expected due to the development of the former for the 600-cycle kit (in all cases the 550bp version of TSP/TSN was used). The greater observed variation in TruSeq insert sizes is due to the greater fragment length flexibility afforded by the gel-based size selection, allowing users to request non-standard sizes. The significant difference between TSP and TSN libraries is a potentially interesting finding, assuming that this persists with increased TSP sampling.

2.3.3 Assembler Performance and Insert Size

Overall the cumulative contig length distributions produced by the three assemblers were significantly different from one another (Figure 7-1; Figure 7-2), although the difference between IDBA and Newbler was lower than between CA and either of these (CA vs. IDBA: D=0.068, $p<0.001$; CA vs. Newbler: D=0.065, $p<0.001$; IDBA vs. Newbler: D=0.033, p=0.003). All three were found to differ significantly from a unimodal distribution and therefore are at least bimodal (CA: D=0.024, $p<0.001$; IDBA: D=0.043, $p<0.001$; D=0.035, $p<0.001$). On a dataset-by-dataset basis there was a tendency for CA assemblies to differ significantly from one or both of the other two but this was not always the case, and only for one dataset did all three differ significantly from one another (*Scolytinae*; Table 7.1). The majority of assemblies were found to be non-unimodal and there was only one dataset for which all three assemblers produced a unimodal length distribution (*ChrysoAL*; Table 7.2).

Differences between the assemblers were found in the proportion of contigs 5-10 kb (CA significantly greater: $F_{1,47}$= 36.09, $p<0.001$; CA: μ=13.4%; IDBA/Newbler: μ=9.2%) and 10-15 kb (IDBA significantly lower: $F_{1,47}$= 6.77, p=0.012; IDBA: μ=3.9%; CA/Newbler: μ=5.6%) but not in the classes 1-5 kb and ≥15 kb. When considering only the longer contigs IDBA tended to outperform the other two assemblers, with a significantly greater proportion of ≥5 kb contigs that were ≥10 kb ($F_{1,47}$= 8.76, p=0.005; IDBA/Newbler: μ=67.9%; CA: μ=56.0%) and a significantly greater proportion of ≥10 kb contigs that were ≥15 kb ($F_{1,47}$= 10.89, p=0.002; IDBA: μ=80.3%; CA/Newbler: μ=69.2%). Thus, for any contig at least 5 kb in length, the likelihood of that contig being nearly-complete (≥10 kb) was greater for IDBA assemblies on average, as was the likelihood of any contig at least 10 kb in length being approximately full-length (≥15 kb).

The apparent variation in behaviour between the three assemblers in the four size classes was examined further with respect to insert size, with striking results (Figure 2-2). The proportion of contigs in the smallest size class (1-5 kb) showed a strong negative correlation with increasing insert size ($F_{1,47}$= 75.9, p<0.001) whilst the largest size class (≥15 kb) showed a strong positive correlation ($F_{1,47}$= 57.80, p<0.001), with no significant difference in the behaviour of the three assemblers in each case, as seen above. In the 5-10 kb size class there was no response to insert length so the minimum adequate model included just CA and IDBA+Newbler as previously. In the 10-15 kb size class there was a significant positive response to increasing insert length and a slight difference in the behaviour of the three assemblers as seen above, such that IDBA responded less strongly than CA+Newbler ($F_{2,47}$= 17.08, p<0.001).

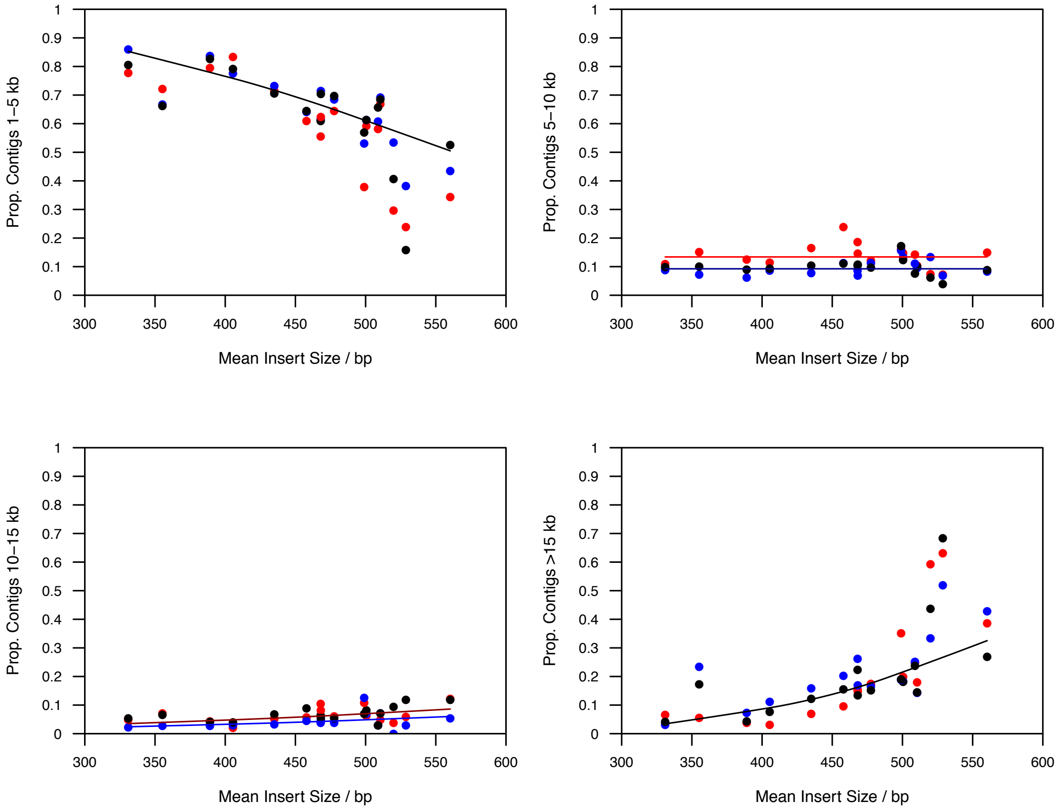


Figure 2-2 Effect of insert length on the proportion of assembled mitochondrial contigs in each of four size classes, coloured by assembler (CA: red; IDBA: blue; Newbler: black). TL: Proportion of contigs 1-5 kb. The three assemblers did not behave significantly differently, hence a single fitted line for the response to insert size alone. TR: Proportion of contigs 5-10 kb. There is no response to insert size but IDBA+Newbler (dark blue line) have a significantly lower proportion of contigs in this size class than CA (red line). BL: Proportion of contigs 10-15 kb. Fitted lines are for CA+Newbler (dark red) and IDBA (blue). BR: Proportion of contigs ≥15 kb. Fitted line for the response to insert size only, no effect of assembler.

2.3.4 Sequencing Effort and Species Recovery

A strong positive relationship between sequencing effort ('true' mitochondrial pairs per input species) and species recovery (assembled sequences per input species) was observed for both *cox1* barcodes and contigs ≥10 kb (Figure 2.3). Two datasets were not included in this analysis as the number of input species was unknown. One further dataset was excluded after initial data inspection as sequencing effort was more than double that of any other library (*RP-Water*; μ=8492 mitochondrial pairs per species). The minimum adequate model for both *cox1* and long contigs included only sequencing effort as an explanatory variable as no significant effect of any assembler was observed (*cox1*: $F_{1,38}$= 45.25, p<0.001; long contigs: $F_{1,38}$= 17.03, p<0.001). Unsurprisingly, the rate of recovery of long contigs is lower than that for the shorter *cox1* barcodes.

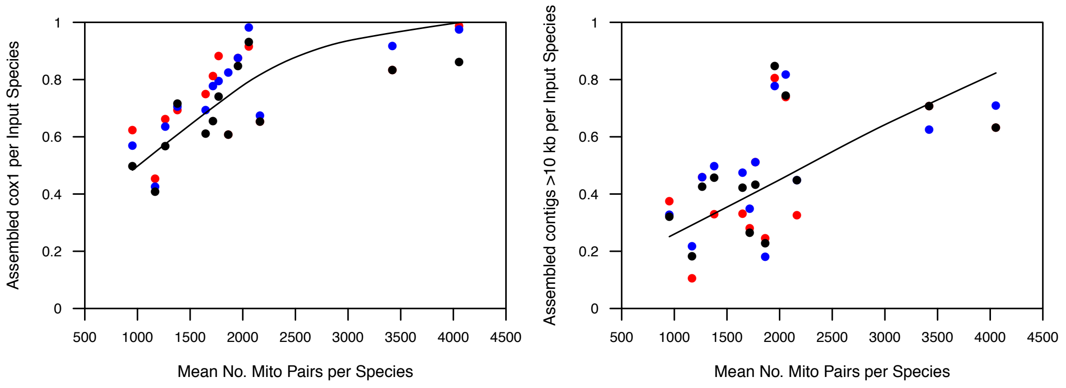


Figure 2.3 The proportion of species recovered with respect to sequencing effort. Sequencing effort measured as mean number of mitochondrial read pairs per input species. L: *cox1* barcodes assembled as a proportion of input species. R: contigs ≥10 kb assembled as a proportion of input species. Points represent data from CA (red), IDBA (blue) and Newbler (black). Fitted lines based on the models described in the text.

2.3.5 Voucher MMG versus Bulk MMG

The *ChrysoRL* and *ChrysoAL* samples behaved differently from the read-processing step through to assembly and the inference of species diversity. Whilst the *ChrysoAL* libraries overall had more raw reads than the *ChrysoRL* libraries, the proportional loss of data due to read processing was greater (22.3% c.f. 16.0%) and the proportion of quality-controlled reads retained for analysis was much lower (5.8% c.f. 14.0%), such that the *ChrysoAL* assembly was finally based on less than 60% of the number of reads than the *ChrysoRL* assembly (and used shorter 250 bp reads) (Table 2.1; Table 2.2). These differences are at least partly explained by the use of TS library preparation for the *ChrysoAL* samples and TSP for *ChrysoRL*, and the corresponding differences in insert size (*ChrysoRL* μ=560 bp ;

ChrysoAL µ=331 bp). Additionally, MiSeq v3 chemistry is expected to produce higher quality data than the v2 chemistry and thus a higher rate of data loss due to read processing for *ChrysoAL* is unsurprising. The observed differences are unlikely to be a product of the quality of the original DNA extractions as the *ChrysoAL* libraries included all specimens in *ChrysoRL* except five, and all DNA was extracted in the same way on specimens that were directly killed in absolute ethanol.

Tracking the increase in the number of 'species' (*cox1* and contigs ≥10 kb) recovered with increasing sequencing effort showed clear differences between the two datasets (Figure 2.4). The accumulation of long contigs in *ChrysoRL* approximately followed that of *cox1*, albeit consistently lower. In contrast the slopes were strongly divergent for *ChrysoAL*, with the number of long contigs almost constant between 0.6 and 1.6 million pairs whereas *cox1* sequences continued to accumulate. These different behaviours are assumed to derive from differences in the distribution of reads between species in the two sample types, with the more even representation of species in voucher MMG allowing continual accumulation with increasing sequencing effort whereas for bulk MMG the dominant species are assembled rapidly with low effort (n.b. slightly higher rate of long contig recovery by *ChrysoAL* up to 0.4 million pairs) but each subsequently less abundantly represented species requires deeper sequencing. Interestingly, the equivalent plot for *ChrysoAL* assemblies under the alternative IDBA parameters (--min_contig 1000) shows a different pattern from either of the other two, with different slopes again observed for the two sequence types but with a reduced accumulation of the shorter *cox1* sequences and an increased accumulation of long contigs relative to the original IDBA assemblies (Figure 7.4).

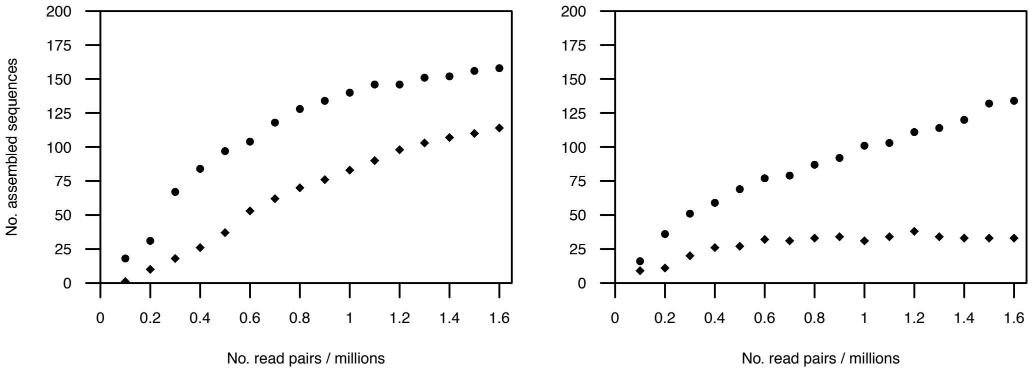


Figure 2.4 The accumulation of *cox1* barcodes (dots) and long contigs (diamonds) in assemblies of subsampled reads for the *ChrysIber* experiment. L: *ChrysoRL*; R: *ChrysoAL*.

Similarly, plots of contig length against mean coverage showed great differences in behaviour between the two datasets for all three assemblers (IDBA: Figure 2.5; CA and Newbler: Figure 7.5) and this persisted when *ChrysoRL* was subsampled (Figure 7.6). All *ChrysoRL* assemblies show a distinct pattern whereby contig length increases rapidly as mean coverage increases, with full-length contigs (≥15 kb) frequently assembled above approximately 10x. No further increase in contig length is observed with increasing coverage as the full mitochondrial genome is generally 15-18 kb long and therefore there is no clear benefit derived from sequencing to a depth greater than ~20x. In all cases there are several persistently short contigs with coverage greater than 25x, although this is more prevalent in CA and Newbler than IDBA. These are unlikely to be incompletely assembled as a result of insufficient sequencing and presumably present some particular idiosyncratic challenge, the severity of which varies between the three programs. However, overall the *ChrysoRL* assemblies behave broadly as would be expected, with IDBA appearing to be particularly efficient. When comparing the *ChrysoRL* plots for the IDBA and Newbler assemblies herein with the corresponding plot for the published set of contigs (*MitoRL*; contigs ≥3 kb only; Figure 7.7) it appears likely that at least some of the cases of short high coverage contigs were resolved by the reassembly step (IDBA and Newbler contigs reassembled in Geneious).

In contrast, in all three *ChrysoAL* assemblies the previously observed increase in contig length in response to increasing coverage is only apparent for a small subset of the total number of contigs. In the majority of cases coverage is a poor predictor of contig length, particularly in the IDBA assembly, although all three appear to cope poorly at high coverage. Comparing the IDBA and Newbler plots with the equivalent for the published *DeNovoRL* indicates that reassembly went part way to resolving this issue, with a large increase in the number of long contigs and a general shift towards longer contigs, although the large number of remaining incomplete contigs with coverage >20x indicates that this process was not as efficient as hoped. Notably, the additional *ChrysoAL* IDBA-1k assembly behaved more closely to the *ChrysoRL* dataset than the original *ChrysoAL* assembly, suggesting that assembly efficiency for bulk MMG samples could be improved slightly with further parameter optimisation.

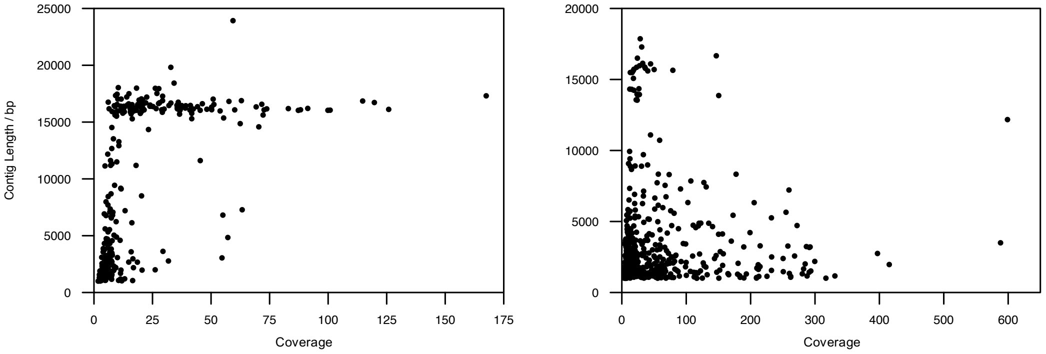


Figure 2.5 Assembled contig length as a response to mean coverage with IDBA-UD for voucher versus bulk MMG experiments. L: *ChrysoRL*; R: *ChrysoAL.*

2.4 Discussion

2.4.1 Read Processing and Mitochondrial Proportions

The first step in any NGS analysis is pre-processing the raw reads to remove adapter contamination and low quality bases. No exhaustive test of the effect of different programs has been undertaken herein; instead the response of different library preps to uniform settings has been assessed. Overall TS libraries performed significantly worse than TSN/TSP libraries (Figure 2-1), mainly due to higher losses due to adapter contamination.

Whilst the results of this analysis suggest that TS libraries should no longer be used, the uneven distribution of libraries over time is problematic. The two most recently sequenced TS libraries (500-cycle; v2 chemistry) lost only 12.9% and 22.2% of reads overall (*RP-Water* and *RP-Ground* respectively) and thus were more similar to TSN/TSP libraries than the majority of other TS libraries. This, combined with a tendency for the greatest losses with each library type and chemistry to be seen in the oldest samples (e.g. v2 TS: *BC-short*, 50.0%; v3 TSP: *UK-BI*, 40.0%) suggests that the age of the chemistry or library preparation kit is an important factor affecting data loss due to read processing. Whether this is related to the timing of machine upgrades, variation in kit quality or operator experience is currently unknown however, anecdotally, being an early adopter may not be a good strategy. Thus, where the greater data volume and longer reads of the 600-cycle kit are not required and TS library preps are still available there is probably no strong argument against using this method based on rates of data retention. There is currently little evidence to differentiate the two newer library preparation kits at this step. TSN libraries appear to vary less than TSP libraries in the amount of data lost (after excluding outliers) but the sample size for the latter was very small (n=4 c.f. n=12). The significantly lower proportion of quality-controlled 'mitochondrial-like' reads in the TS libraries reinforces the choice of TSN/TSP, indicating that these are likely to maximise the retention of high quality reads for assembly (Figure 2-1).

Surprisingly, the TSP libraries were found to have significantly longer insert sizes on average than TSN libraries in spite of the expectation that both kits produce 550 bp insert sizes (Illumina 2013). Confirmation of this observation will require sequencing of additional TSP libraries but this is a clear target for future experiments given the observed effect of insert size on assembly success (see below). Additionally, a significant effect of insert size on the proportion of mitochondrial reads was observed with respect to library type, particularly for TS libraries (Figure 2-1). This supports the observations of Crampton-Platt et al. (2015) but while the observed increases are proportionately large, all libraries included fewer than 2.5% mitochondrial reads after quality control and thus overall efficiency is low. Where a TS library must be used, there are clear gains to be made from maximising insert length, at least to ~450 bp. Again, the small number of TSP libraries and the statistically insignificant difference between these and the TSN libraries precludes their separation. However, the apparently greater insert size and relatively high mitochondrial proportion in three of four cases warrants further investigation. Whether or not the fourth sample should be considered an outlier will likely have a significant effect on the modelled relationship for TSP libraries.

2.4.2 Assembler Performance and Insert Size

All three assemblers behaved differently to one another when their cumulative length distributions were compared across all datasets simultaneously, although when plotted they appeared similar (Figure 7-1), possibly with a slightly increased bimodality in IDBA and Newbler (more rapid accumulation of the shortest and longest contigs than CA). Following this, when considering cumulative length distributions within datasets there was a tendency for CA to be significantly different from the other two. Of the three assemblers IDBA consistently produced significantly non-unimodal distributions, reflecting the second peak in the corresponding histograms at around 15 kb (Figure 7.3). When not considering the shortest contigs (1-5 kb) there was a significant increase in the proportion of long contigs (≥10 kb) and in the proportion of long contigs that were complete, indicating that IDBA is better at maximising contig length but only once a contig reaches approximately one third of the length of the full mitogenome. On a dataset-by-dataset basis IDBA was not always the optimal choice, not assembling the greatest number of long contigs and *cox1*-5' sequences in five and nine of the sixteen datasets respectively. Thus, if only a single assembly is undertaken IDBA would generally be preferred but the addition of at least one other assembler is recommended to maximise the likelihood of obtaining a long contig for each species when the results are combined (Chapter 3; Crampton-Platt et al. 2015).

In addition to having a significant effect on the proportion of reads retained for assembly, insert size was shown to affect the distribution of contigs between four length categories, with increasing insert size associated with a decrease in the proportion of the smallest contigs and an increase in the proportion of the most complete ones, while the proportion of contigs 5-15 kb were largely unaffected (Figure 2-2). Thus increasing insert size has a significant effect on assembly efficiency by biasing the length distribution towards contigs ≥15 kb and away from contigs <5 kb. Given that maximising sequence contiguity is the primary aim at the assembly step for MMG this finding has significant implications for the design of future experiments, in particular because insert length can be controlled relatively easily by adjusting library preparation protocols. The observed relationships appear exponential, however there are relatively few samples with insert sizes 500-600 bp and in particular there are no samples with insert sizes 530-560 bp. The sample with the largest insert size (*ChrysoRL*, μ=560 bp) deviates from the general trend and this, when combined with the apparent increase in variation between assemblers above ~500 bp, calls for additional sampling within this interval to ascertain whether the observed trend holds, and to check that there is no negative effect of insert sizes >530 bp.

2.4.3 Sequencing Effort and Species Recovery

Increasing data volume is assumed to maximise the likelihood of species recovery and the length of the corresponding contigs and although this was supported by the bulk of the samples herein, the three datasets with the greatest sequencing effort did not assemble *cox1* as completely as would have been predicted based on the less deeply sequenced samples (only two of these libraries were included in the analysis; Figure 2.3). This deviation was not observed in the equivalent analysis for the recovery of long contigs, wherein the maximal recovery rate was 84.7% (*Scolytinae*, Newbler). This difference in behaviour may be indicative of a saturation effect whereby above a certain sequencing depth assembly efficiency decreases. The threshold at which this happens will vary with the length of the marker, with a clear decrease in return on sequencing effort apparent for the barcode locus (~660 bp) above ~2500 mitochondrial pairs per species but no such decrease observed for the long contigs (≥10 kb) within the sampled range. However, sampling density above this threshold was limited to two datasets in the present analysis. These of course may not be indicative of the general assembly behaviour of datasets with this level of sequencing effort, requiring a significant increase in sampling in the range 2000-4000 mitochondrial pairs per species for confirmation. Additional experiments aiming to optimise long contig recovery in particular should target this range of sequencing depth to assess whether the current results can be improved upon. It is clear that the current level of sequencing effort is far from ideal and in most cases should at least be doubled to maximise the rate of recovery with any single assembler.

2.4.4 Voucher MMG versus Bulk MMG

Clear differences in assembly behaviour and efficiency were observed between the voucher and bulk MMG samples from the *ChrysIber* experiment, with the former clearly the preferred approach for efficient assembly of long contigs even in the face of variable sequencing depth. However, the improvements in contig length observed in the reassembled *ChrysoAL* data presented by Gómez-Rodríguez et al. (2015; *DeNovoRL* Figure 7.7) and the increased sequence contiguity observed with the additional IDBA assembly indicate that the assembly of these datasets can be greatly improved with careful curation and, potentially, alternative assembly parameters; although in this case the divergent species accumulation behaviour would need to be addressed by combining both IDBA assemblies. Notably, even when attempting to equilibrate the amount of DNA per species for voucher MMG there can still be an approximately 10-fold difference in sequencing depth between species (Figure 2.5) and it is therefore not surprising that no instance of complete species recovery was observed for any of the datasets analysed herein. The even greater disparity in sequencing depth in the bulk samples clearly contributes to incomplete assembly, as data insufficiency is likely to be a genuine constraint on the assembly of contigs for low biomass species. Although excessive data also appear to cause problems for the assemblers in some instances this can probably be resolved to a large extent be by reassembly or perhaps subsampling. The latter may be appropriate in cases where the species is represented by a mix of haplotypes that create ambiguity during manual curation of reassembled contigs. In these cases subsampling may help to restrict the assembly to only the most abundant haplotypes and aid contig extension by the assembler. The lack of data for low biomass species will therefore be the primary limitation for bulk MMG for the foreseeable future, particularly as bioinformatics steps are further refined and assembly programs better suited to the particular challenges of MMG are developed. This insufficiency is difficult to address at the current time. Maximising insert size and data quality appear to increase efficiency for any single sample, however the approximately twofold increase in mitochondrial proportion encountered herein only has a limited effect on cost-effectiveness.

For studies only concerned with maximising sequence contiguity, perhaps for phylogenetics or to generate superbarcode reference libraries, voucher MMG is the clear choice. However, for ecological studies the choice of sequencing strategy remains somewhat uncertain as two independent steps are required, firstly to generate a reference database and secondly to obtain assemblage profiles by read mapping against that database. The published results from the *ChrysIber* datasets and the analyses presented here suggest that the solution is to generate a reference library of all species likely to be encountered within a given study and then apply low coverage sequencing to bulk samples for assemblage profiling against the complete reference set. However, determining an appropriate level of low coverage sequencing to maximise species detection at this step has not thus far been explored and will presumably vary significantly between assemblages, making it almost impossible *a priori* to differentiate the boundary between the sequencing effort required for profiling and that required for effective *de novo* assembly of the reference set from the bulk samples themselves. If the amount of sequencing effort required for complete assemblage profiling is not much less than that required for assembly and the assembly of bulk MMG data can be further optimised the requirement for additional sequencing for the reference set is negated, particularly when considering the additional effort required to make a sufficiently complete species inventory and generate the reference library within any single study. These issues are further discussed in Chapter 4 in the light of the assembly results obtained therein.

2.4.5 Conclusions

In spite of the unbalanced selection of samples and the confounded distribution of libraries between MiSeq chemistries, a small number of conclusions can be drawn from the present study and additional areas in need of further work can be highlighted. Firstly, the choice of library type should be between TruSeq Nano and TruSeq PCR-free to maximise data quality, insert size and mitochondrial proportion. Unfortunately, the effect of library type herein cannot be extricated from the effect of MiSeq chemistry and the observed differences in data quality are possibly not relevant for future studies. However, the longer default insert sizes of the newer library preparation kits and the corresponding tendency for increased mitochondrial proportion and improved assembly of long contigs make these an obvious choice. The longer insert size TSN protocol requires significantly less input DNA (200ng, 550 bp insert) than either TSP (2µg, 550 bp insert) or TS (1µg, 300 bp insert) and therefore will be the most relevant for many studies. However, further exploration of the possible differences between TSN and TSP highlighted herein may prove fruitful for studies where DNA availability is not limiting, e.g. bulk MMG on homogenised samples. Where possible insert size should be maximised, although any biasing effect that this may have on the composition of the resulting data is currently unknown. Within the TSN and TSP libraries the effect of longer insert size on increased mitochondrial proportion was less than for TS libraries, although the sampled size range was smaller. Even without this, the clear effect of insert size on the assembly makes it worthwhile. More TSP libraries are required both to discriminate from TSN and to assess whether there is an upper limit on the positive effect of insert length as it is TSP libraries that are most likely to sample in the 500-600 bp range. Of the assemblers used in the present study there is relatively little to discriminate between them and the better-performing program varied between datasets, thus no clear recommendation can be made at the present time. As such it is advisable to use more than one program within any particular study and assess their relative behaviours before making a final choice. Further to this, combining the contigs from multiple programs by reassembly is likely to improve the results of MMG (see Chapter 3 for more details) although the added complexity of this step may be undesirable in some cases. Clearly there is a much wider range of potential assembly programs that could be used than the selection presented here, although in the author's experience the majority of genome assemblers perform poorly on MMG data and the lack of metagenomic assemblers for Illumina paired-end reads is currently limiting. However, this is clearly a dynamic field and new programs are published frequently. The assembly of multiple (circular) orthologous sequences from mixtures presents a specific problem that is currently not addressed in the literature but with the increasing profile of MMG this will hopefully be solved in the medium term. Finally, although the voucher MMG strategy presents a simplified assembly challenge and is more data-efficient than bulk MMG, at least some of the issues associated with the latter are likely to be resolvable. The main limitation in all cases is the amount of mitochondrial data obtained and this is clearly exacerbated by the uneven distribution of species biomass in bulk MMG. Thus for generating superbarcode libraries, a voucher MMG approach where DNA is equalised as far as possible between species is preferable to blind pooling. However, for ecological studies the relative merits of the two approaches may be less clear-cut than initially thought and thus the optimal strategy may vary between systems.

Chapter 3 Characterising Communities in a Phylogenetic Framework with MMG

Summary

This Chapter applies MMG to a sample of tropical beetles obtained via canopy fogging and seeks to characterise that sample in terms of species richness, taxonomic composition, and phylogenetic relationships. External superbarcodes are incorporated for phylogeny reconstruction to act as a taxonomic scaffold from which the sample can be characterised at the family level. Practical issues related to building community phylogenies and the wider beetle phylogeny with respect to rapidly increasing taxon sampling are addressed using maximum likelihood analyses in RAxML. The choice of data coding (all nucleotides, protein-coding genes translated, protein-coding genes with 3rd position removed and 1st position RY-coded) is assessed with respect to two levels of taxon sampling and the effect of intermediate taxon sampling is assessed for the preferred matrix coding with a reduced superbarcode set. When constructing the community phylogeny an important question is whether a standalone tree with only the contigs derived from the sample is sufficient or whether external references are required to counterbalance the uneven taxon sampling encountered in a local sample. The effect of various strategies, with and without the use of backbone trees and with and without superbarcodes is assessed. A related question is the choice of locus for gene-centred analyses to ensure orthology where the presence of non-overlapping contigs complicate richness estimates, thus the taxonomic profiles obtained by a matrix centred on the *cox1* barcode region (allowing comparison with available morphological identifications) is compared with that obtained by the most abundant locus, *nad4l*. Lastly, the effect of combining the results of different assembly programs on sequence contiguity and diversity representation is discussed with respect to the challenges presented by bulk MMG samples that were previously highlighted in Chapter 2. The dataset used here has previously been published in a different form, but all analyses presented here are new.

3.1 Introduction

The primary motivation behind the development of MMG has been to facilitate large-scale and integrative analyses of arthropod diversity that are not hindered by the taxonomic impediment and are comparable between studies. In highly diverse and poorly characterised systems the gold standard approach, namely a complete inventory requiring morphological identifications of all sampled individuals, is time consuming and requires significant input from expert taxonomists (Basset et al. 2012). In most instances this is impractical and leads to reductive approaches either with respect to taxonomic resolution (e.g. parataxonomy, metabarcoding) or ecological breadth (e.g. surrogate taxa). In the latter case the lack of congruence in diversity patterns between taxonomic and ecological groups makes the choice of indicator taxa an uncertain step with potentially serious effects on study conclusions (Lawton et al. 1998). Meanwhile for parataxonomy and particularly metabarcoding, 'species' diversity may only be measured at order level due to limits on the obtainable resolution (e.g. Gibson et al. 2014). However, while metabarcoding protocols vary with respect to wet-lab protocols and delimitation of species-level sequence clusters it is possible for the raw data to be re-analysed repeatedly, allowing third-party verification of results and the simultaneous analysis of samples from multiple studies with a single protocol to test new hypotheses (Ji et al. 2013). This ability to redefine species groups as new data become available is one significant advantage of DNA based methodologies for the large-scale study of arthropod diversity and is particularly relevant for the on-going development of MMG.

As outlined in Chapter 1, the rationale for focussing on the mitochondrial fraction of the metagenome is the twofold offer of phylogeny and species identification (in Metazoa), to analyse patterns of diversity at a range of hierarchical levels whilst maintaining a link with existing taxonomic and biological knowledge. In all MMG studies to date, the mitochondrial contigs generated by any single assembly have been observed to partition the diversity of the sample approximately at the level of species, thus the number of orthologous sequences assembled can be considered an estimate of species diversity (although the choice of sequence has a large effect on these conclusions, see Chapter 2). A phylogeny generated from the orthologs from a single sample therefore approximates the phylogenetic relationships between the species present in the community, facilitating community phylogenetic analyses (Andújar et al. 2015). Beyond this, the rapid accumulation of mitogenome sequences from MMG data (both voucher MMG and bulk MMG) will facilitate both increasingly densely sampled mito-phylogenomic analyses for systematics and biogeography, and more precise characterisation of existing bulk MMG data and newly sampled communities. The present Chapter is primarily concerned with the application of bulk MMG to the accurate characterisation of a single community, however the robustness of the results under variable taxon sampling is a primary concern with this latter expectation of increasingly large-scale analyses of diversity with MMG data.

Mitogenome sequences have a relatively long and controversial history in phylogenetics and their ability to recover deep evolutionary relationships has been debated. In insects mito-phylogenomics has been applied at a variety of levels, studying anything from interfamilial (Gillett et al. 2014) to interordinal relationships (Simon and Hadrys 2013) and using a wide range of phylogenetic methods (reviewed in Cameron 2014). Whilst there is increasing acknowledgement that the challenges presented by among-site rate heterogeneity and biased nucleotide composition can be overcome by appropriate model choice (Talavera and Vila 2011) and careful investigation of problematic placements (Cameron 2014), the majority of studies to date have suffered from limited taxon sampling due to the expense and difficulty of generating mitogenome sequences. Following the demonstration of a pooled-sequencing approach for long-range PCR products it was clear that such issues could now be overcome, using the power of next-generation sequencing platforms to cheaply generate mitogenomes for potentially hundreds of species simultaneously (Timmermans et al. 2010). However, whilst this was followed by several studies in Coleoptera using the same methodology (Timmermans and Vogler 2012; Haran et al. 2013; Timmermans, Barton, et al. 2016) and more recently the PCR-free equivalent, voucher MMG (Gillett et al. 2014; Timmermans, Viberg, et al. 2016), an equivalent increase in mitogenome sequencing has not been seen in other insect groups. Thus the currently available literature regarding mito-phylogenomics is out of step with the new opportunities and challenges associated with rapidly increasing mitogenome availability.

Obtaining an accurate community phylogeny and the evolutionary relationships between co-occurring species in the context of the wider (and growing) mitochondrial phylogeny are two complementary foci of MMG that are addressed in this Chapter. In addition to characterising a sample phylogenetically, the taxonomic composition of MMG samples may also be characterised using external information. The assembly of the barcode locus and the comparison of these sequences against existing databases provides an immediate inventory of already-barcoded species. In the absence of existing records the phylogeny itself can be used to broadly characterise the taxonomic composition of the community, whilst also quantifying evolutionary relationships (Crampton-Platt et al. 2015). Placing the assembled mitochondrial contigs in a phylogeny with superbarcodes allows the assignation of higher-level (usually family) taxonomic ranks to those contigs found to be in monophyly with two or more superbarcodes. Even at low rates of superbarcode taxon sampling, Crampton-Platt et al. (2015) assigned over 60% of tested contigs to superfamily and family level, increasing to >98% when superbarcode sampling increased approximately seven times. Thus, even in the absence of barcode records, an MMG approach is likely to allow an increasingly fine-grain description of uncharacterised communities as superbarcode sampling improves. In contrast, the resolution of taxonomy assignment based on BLAST searches against thousands of sequences on GenBank (e.g. with MEGAN), as commonly applied in metabarcoding, is highly dependent on the composition of the database used and the variability of the chosen marker (Gibson et al. 2014), while the accuracy of any single assignment has been observed to be both unpredictable and unquantifiable (Crampton-Platt et al. 2015).

An implicit assumption of MMG is that all contigs are placed correctly in the phylogeny and for the longest sequences this does not appear to be problematic (Crampton-Platt et al. 2015). However, to date no assessment has been made to determine if there is a lower length limit below which placement becomes unreliable, or whether any of the mitochondrial loci (or combinations thereof) are more or less likely to be reliable than others in this respect. Recent mitogenome phylogenies have tended to require a minimum number of loci for inclusion but for bulk MMG samples many species are represented by short contigs and such cut-offs greatly reduce the number of species that are represented in the community phylogeny. At the same time, species for which the mitogenome sequence are only partially assembled are likely to be represented in the dataset by multiple short but unlinked sequences. These issues have given rise to various strategies to maximise the number of species included in community analyses whilst ensuring orthology to prevent richness inflation. Each of the three current examples applying MMG to ecological communities have used a different approach, with variability in the required length and gene composition of included contigs, but in all cases the addition of shorter sequences required the presence of one or more chosen loci. The extent to which these different approaches produce consistent and reliable results remains untested, however in two of the examples a minimum sequence length was required for inclusion to maximise the likelihood of correct phylogenetic placement, at the cost of several species only represented by shorter sequences in the target region (Crampton-Platt et al. 2015; Gómez-Rodríguez et al. 2015). In the third example a complex iterative approach was used whereby a PhyloBayes analysis on the amino acid alignments (CAT model) for the longest sequences were used as a backbone constraint topology for the addition of short barcode-centred contigs, this topology was then used as a constraint for the addition of PCR barcodes (Andújar et al. 2015). The latter strategy maximised the number of species represented in the phylogeny by MMG data by requiring as little as 100 bp overlap between sequences in the barcode region yet the addition of the PCR barcodes further increased observed species richness. This, alongside the recovery rates observed in Chapter 2 (Figure 2.3), illustrates that even under permissive conditions, MMG is unlikely to recover all species in the sample, an issue that will only worsen with the uneven representation of DNA between species in bulk MMG samples. Minimising such limitations in bulk MMG analyses is the primary aim of the current Chapter and is a key step towards the wider application of bulk MMG to larger-scale biodiversity questions.

Building on the results of Chapter 2, the effect of reassembling contigs from three different assemblers will be assessed with respect to the contig length distribution and ‘completeness’ of assembly. This step is expected to shift the length distribution towards longer contigs and reduce the number of contigs in the final non-redundant dataset, while also resolving the bulk MMG problem of short, high coverage contigs identified in Chapter 2 (Crampton-Platt et al. 2015). Following Chapter 2, the three individual assemblies are expected to be similar with respect to their length distribution and to overlap significantly in their contig composition, although no single assembly is expected to fully recover all sequences, requiring multiple assemblies to maximise community representation from MMG (assessed with respect to unique gene sequences; Crampton-Platt et al. 2015). *A priori* no single assembler is expected to outperform the other two in the recovery of unique sequences and it is not known to what extent the use of three assemblers will improve community representation over the use of two assemblers.

For the phylogenetic analyses, increasing taxon sampling is expected to improve and stabilise tree topology. Increasing the sampling of identified superbarcodes will also improve the resolution of taxonomic assignments inferred for the *BorneoCanopy* contigs by monophyly with superbarcodes (Crampton-Platt et al. 2015). Matrix coding is expected to have an effect on the basal relationships and, perhaps, the monophyly of major clades, however no effect is expected at the tips of the tree (i.e. the most closely related species will always appear as sister taxa). By reducing the effects of compositional heterogeneity, reductive coding (e.g. removal of 3rd positions, RY-coding of 1st positions, translation to amino acids) is expected to outperform the nucleotide matrix with respect to the recovered relationships between major clades and the monophyly thereof. For the community phylogeny, the choice of gene required for contigs to be included in the analysis is not expected affect overall tree topology or the accuracy of placement, even for short contigs. If this is the case, using the most frequently assembled gene will be a justifiable strategy, allowing the maximum number of species to be represented in the phylogeny. Overall, the placement of all contigs is expected to be correct and therefore the taxonomic profile inferred from the phylogeny will closely match the composition identified from specimen morphology.

3.2 Materials and Methods

3.2.1 Sample Description

The data used in this Chapter are derived from the study by Crampton-Platt et al. (2015) but have been re-analysed herein. A sample of 477 beetle individuals representing approximately 209 morphospecies was obtained by rainforest canopy fogging in Danum Valley, Sabah, Malaysia (henceforth *BorneoCanopy*). DNA was extracted destructively from each individual separately and then pooled in equal volumes. Two TruSeq libraries were prepared aiming for insert sizes of 480 and 850 bp respectively and each was sequenced on a full Illumina MiSeq run (500-cycles; 250 bp paired-end).

Morphological identifications based on specimen images are available to complement and verify the tree-based approach to describing the assemblage. These identifications serve as a baseline assemblage profile against which the tree-based profile can be compared. In addition, *cox1-5'* barcodes are available for 329 individuals following two rounds of PCR and Sanger sequencing. These barcodes provide both a minimum DNA-based species-richness estimate when combined with the equivalent data from the assembled contigs, and a way to link the assembled contigs with the morphological identifications allowing a subset of the phylogeny-based taxonomy assignments to be verified (see below). In the previous analysis, GMYC for the combined PCR- and contig-derived barcodes gave an estimated total richness of 232 species, of which 75% were represented in the phylogeny for the assemblage (Crampton-Platt et al. 2015).

3.2.2 Mitogenome Assembly

Illumina data pre-processing, filtering for ‘mitochondrial-like’ reads and mitogenome assembly with three programs (Celera Assembler (CA), IDBA-UD (IDBA), Newbler) was undertaken as part of the analyses presented in Chapter 2. Substantial overlap is expected between the contigs assembled independently by each program; however, no single assembler is expected to find the optimal solution. Thus, to maximise sequence contiguity and the species representation in the final dataset, the three sets of contigs were combined by re-assembly to obtain the most complete set of contigs possible.

All mitochondrial contigs ≥15kb were manually checked for identical or near-identical terminal regions in Geneious R6.1 (Biomatters 2013). Such regions indicate that the complete mitochondrial genome has been assembled and the duplicated region was trimmed from one side to allow circularisation of the contig later. This step is particularly important where the assembly terminates within a gene to prevent the duplication of homologous sequences in the alignments. Subsequently, all mitochondrial contigs ≥1 kb were re-assembled in Geneious in four steps to generate a non-redundant set. Firstly, the circularisable contigs were assembled together to remove redundancy in this set. Secondly, all linear contigs were mapped against the non-redundant circular set to remove incompletely assembled contigs that were fully recovered by one or two of the other assemblers. This step resulted in a non-redundant set of linear contigs that were then assembled together to maximise sequence contiguity. In the first instance contigs ≥5 kb were assembled together, with contigs 1-5 kb added in the subsequent step. Within each of these four steps two assembly iterations were performed. Firstly high-stringency 'custom' assembly settings (overlap ≥500 bp; overlap identity ≥98%; mismatches per read ≤2%; gaps ≤5%; gap size 3 bp) were used to identify the homologous overlapping contigs assembled with the greatest consistency by the different programs. Secondly, contigs which were 'unused' by Geneious in the high-stringency iteration were re-mapped ('medium sensitivity / fast') or re-assembled with the curated set ('highest sensitivity / slow') as appropriate to maximise the likelihood of detecting homologous contigs which have high identity overlapping regions but exhibit disagreement between assemblers at the termini, reducing overall similarity. At each step in this re-assembly procedure the contigs were checked manually and edited where necessary to resolve discrepancies between the different assemblers. This manual curation was necessary to minimise the incorporation and perpetuation of assembly errors in the final set of contigs.

3.2.3 Annotation, Gene Extraction and Dataset Refinement

The non-redundant set of contigs was checked for tRNA sequences using COVE v2.4.4 (Eddy and Durbin 1994; Coleoptera covariance models Timmermans and Vogler 2012) and hits above a score threshold of 40 were parsed with a Perl script (Crampton-Platt et al. 2015) to generate GenBank format files from the input FASTA formatted sequences and convert the hit co-ordinates to tRNA annotations where applicable. These files were opened in Geneious and contigs flagged for circularisation were circularised with the first residue of tRNA-Ile used as the starting co-ordinate. Where tRNA-Ile was not annotated the last residue of tRNA-Gln was used instead (reverse orientation). All contigs were then re-exported in both FASTA and GenBank format for a BLAST-based (Altschup et al. 1990) gene annotation step.

Protein-coding (PCGs; tblastx) and rRNA genes (blastn) were identified by querying a database of contigs with representative sequences for each gene obtained from 51 annotated mitochondrial genomes downloaded from GenBank (not generated by Timmermans and colleagues). The database size parameter (-dbsize) was tuned for each gene to minimise spurious hits and maximise correct recovery using a training dataset where the expected recovery rate per gene was known. For the majority of genes a -dbsize of 100,000 was used, with the exception of *atp8* (reduced to 1000 to minimise spurious hits) and *nad6* (increased to 1,000,000 to increase the likelihood of a hit being accepted). For both tblastx and blastn an e-value of 1e-5 was used and query sequence filtering was disabled (-seg no and -dust no respectively). Additionally, for tblastx the genetic code for both the database and the query sequences was set to invertebrate mitochondrial (translation table 5). Results were output in tabular format and subsequently sorted and filtered to retain only the longest hit per contig for each gene. These hits were further filtered by length to minimise the inclusion of spurious short hits, at a cost of the loss of partial gene sequences at the ends of contigs. To achieve this, the annotated *Tribolium castaneum* (NC_003081) mitochondrial genome was used as a template to determine the approximate expected length of each gene and the length cut-off was set to be approximately 50% of this value in each case. Cdbfasta (The Institute for Genomic Research, Available from: http://sourceforge.net/projects/cdbfasta/) was then used to extract the corresponding contig sequences based on the filtered hit table for each gene. Finally, the sequences for each gene were aligned with MAFFT v7 allowing the direction of sequences to be adjusted (--auto --adjustdirection; Katoh and Standley 2013) to ensure all sequences for each gene were in the same orientation. The hit table used to generate the co-ordinates for cdbfasta was further used to annotate the same regions in the GenBank formatted sequence for each contig with a custom Java script (Benjamin Linard, 2015).

The initial alignments from MAFFT were checked by eye for poorly aligning sequences in Geneious. The corresponding annotations were checked and deleted where clearly erroneous. Such problems were confined to shorter contigs where additional genes were annotated despite being absent (particularly problematic for *atp8*). The cleaned alignments were exported, unaligned, and realigned using transAlign (PCGs) (translation table 5; invertebrate mitochondrial code; Bininda-Emonds 2005) and MAFFT (rRNAs) (E-INS-i). PCG alignments were then checked for frame shifts and trimmed to start and stop codons where possible but in all cases to start and end with complete triplets and translate in the forward direction. Two divergent sequences were identified during this curation step and discarded after further investigation suggested that these were not of arthropod origin (non-arthropod hits to *nt* database by blastn and incomplete gene annotation). Poorly aligning terminal regions were trimmed from the rRNA alignments but otherwise not edited. Several identically duplicated sequences in each of the fifteen alignments were identified in Geneious, indicating that the contig re-assembly step was not exhaustive. In each case the affected contigs were assembled together and the longer of the two was retained in the alignments. In the majority of cases these high-identity contigs had not previously been identified due to terminal disagreements. One pair of contigs were found to overlap almost identically at both ends, creating a new circular contig. This process also identified one apparent chimera in a re-assembled contig that included an identical *nad2* and partial *cox1* sequence to a Newbler contig but was otherwise highly divergent. Further investigation revealed that the problem originated in a single CA contig that was included in the re-assembled contig, rather than deriving from the re-assembly process itself. This chimeric region was trimmed from the re-assembled contig and the affected sequences removed from the alignments.

3.2.4 Assessing the effect of combining assemblies

The effect of the re-assembly step was assessed in three ways. Firstly by comparing the length distributions of the contigs in each of the three assemblies with the non-redundant set, both with pairwise Kolmogorov-Smirnov tests (two-sample, two-sided) and visually by plotting their respective (cumulative) length distributions in R v3.2.1 (R Core Team 2015). Secondly, contig length was plotted against mean coverage in R, following the method outlined in Chapter 2, to visualise the effect of reassembly on contiguity. In brief, quality-controlled and BLAST-filtered reads were mapped to each of the four sets of contigs with SMALT (-y 0.98; v 0.7.6; Wellcome Trust Sanger Institute, Available from: https://www.sanger.ac.uk/resources/software/smalt/) and mean coverage per contig obtained from the resultant SAM file with Qualimap v2.0 (García-Alcalde et al. 2012) after conversion to BAM (samtools; Li et al. 2009). Lastly, the curated alignments were used to assess the extent of redundancy between the different assemblers, at the level of unique gene sequences. This was measured by searching the unique sequences for each gene against a database of the raw contigs for each of the three assemblies using megablast (-perc_identity

98 -max_target_seqs 1; -word_size 5 for *atp8*, *nad4l*, *nad6*). A unique sequence was considered as 'recovered' by an assembler when the BLAST alignment length for the contig was ≥50% of the length of the unique sequence that returned the hit. Each unique sequence was then scored as present in one assembly or a combination of all three. Mean coverage per gene was also estimated to assess the extent to which overall variation in gene frequency was correlated with the number of reads aggregating these regions. Mean coverage per gene was measured as the total number of BLAST aligned bases (over a minimum hit length threshold) divided by the total number of bases in each gene alignment. Megablast searches of the unique gene sequences were made against a database of the quality-controlled, BLAST-filtered reads (-perc_identity 98 -max_target_seqs 1000000). A hit length of 200 bp was required for all genes except the three shortest (all <400 bp) where the threshold was set as 50% of the length of the respective gene in the *Triboilum castaneum* reference genome (NC_003081) (*atp8*: 78 bp; *nad3*: 179 bp; *nad4l*: 144 bp). The correlation between gene frequency and average coverage was measured with Pearson's product-moment correlation coefficient in R.

3.2.5 Supermatrices

All curated alignments were exported, unaligned and combined with the equivalent data for all coleopteran mitochondrial contigs available on GenBank (expanded-MitoDB; *exMitoDB*) and eight Neuroptera outgroup sequences. This combined dataset was then aligned as previously with transAlign and MAFFT for a final set of alignments. Genes were concatenated with a Perl script (Bocak et al. 2014) to generate supermatrices under various criteria. To estimate the phylogeny of beetles a minimum of 8 genes (PCGs and/or rRNAs) were required per contig for both the *exMitoDB* and the *BorneoCanopy* contigs for a total of 278 (270 Coleoptera, 8 Neuroptera; *Mito270*) and 146 respectively (*8+ contigs*). Three versions of this supermatrix were made with different treatment of the PCGs in each case: all nucleotides (allNuc), 1st position RY-coded and 3rd position removed (1RY2), amino acid (AA). Matrix manipulation and translation was undertaken in Mesquite (Maddison and Maddison 2011).

For a parallel assessment of the effect of matrix coding under reduced taxon sampling the same treatments were applied to generate three supermatrices for the *8+ contigs* alone (with the 8 outgroup sequences). Additionally, three 1RY2 supermatrices (with outgroups) were used to assess the effect of reference taxon sampling on tree topology (stability and monophyly of major clades) and taxonomy assignment: 1) *Mito270* with *8+ contigs*; 2)

Mito46 with *8+ contigs* - this matrix simulates the effect of limited reference taxon availability (reduced *MitoDB* size) by including only the 46 circular coleopteran mitogenomes from *exMitoDB*; 3) *8+ contigs* alone.

A gene-centred approach was taken for estimating the tree for the assemblage to ensure orthology of incomplete contigs. The third most abundant gene in the dataset, *nad4l*, was chosen to maximise the number of overlapping contigs in the analysis. The two longest genes, *nad5* and *cox1*, were observed more frequently but both included partial sequences such that no alignment position was covered by all sequences. This allows the possibility that non-overlapping contigs from the same species would be included in the "orthologous" set, potentially inflating estimates of species richness. The *nad4l*-centred matrix (no minimum contig length cut-off or minimum number of loci) contained 203 contigs (including all but six *8+ contigs*) plus the 8 outgroups. For comparison and to allow validation against the morphological identifications associated with the DNA barcodes a requirement of a 100bp overlap in the *cox1* barcode (following Andújar et al. 2015) retained 168 contigs. To mitigate against possible erroneous placement of short contigs due to low phylogenetic power, a second *nad4l*-centred supermatrix was generated including 275 *MitoDB* sequences (no minimum number of loci) with this gene and the 8 outgroups. The equivalent *cox1*-centred supermatrix included 119 *MitoDB* sequences and the 8 outgroups.

3.2.6 Phylogenetic Inference

Nucleotide supermatrices were partitioned by gene and codon position for PCGs and by gene for rRNAs. The amino acid analysis was partitioned by gene for both PCGs (MTART substitution matrix) and rRNAs. All analyses were run using RAxML v8 (Stamatakis 2014) on the Cipres Science Gateway (Miller et al. 2010) with maximum likelihood tree estimation and 100 rapid bootstraps conducted in a single analysis under the GTRCAT model. For the *nad4l*-centred and *cox1*-centred datasets initial trees (without *MitoDB* sequences) were inspected for short branch lengths indicating closely related contigs. Five such cases were detected in each tree and investigated in Geneious. Three and two contigs with >98% identity to longer sequences were discarded, reducing the number of *BorneoCanopy* contigs to 200 and 166 for *nad4l* and *cox1* respectively. For the assemblage trees, four analyses were run on each dataset: 1) the gene-centred *BorneoCanopy* contigs alone (with outgroups); 2) the gene-centred *BorneoCanopy* contigs plus the (gene-centred) *Mito275* and *Mito119* sequences as appropriate; 3) the topology for the *8+ contigs* alone (taken from the matrix coding and taxon sampling analyses) was used as a binary backbone (the *8+ contigs* without the relevant locus were pruned) for the addition of the shorter *BorneoCanopy* contigs (i.e. no external reference sequences were used apart from the 8 outgroup taxa); 4) an initial tree was generated with only the 8+ subset of the contigs included in (2), this tree was then used as a binary backbone (option -r in RAxML) for the addition of the shorter gene-centred *BorneoCanopy* and *MitoDB* contigs. All trees were visualised in Dendroscope (Huson and Scornavacca 2012) and rooted *post hoc* with the outgroups**.** Tree topologies were compared by calculating the Robinson-Foulds (RF; Robinson and Foulds 1981; Steel and Penny 1993) distance (for trees pruned to include the same number of tips) to assess the stability of the branching pattern between analyses. For comparison of these distances between analyses with different datasets the normalised RF score for each tree is calculated as RF/number of tips. These analyses were run in R using the *phangorn* package (*RF.dist*; Schliep 2011) on 'multiPhylo' objects.

3.2.7 Contig Identification and Species Richness Estimate

Contig-derived *cox1* sequences were searched against a database of 329 *cox1-5'* 'baits' derived from PCR and Sanger sequencing of individual specimens, using megablast (-perc_identity 98 -max_target_seqs 1), to link each contig with a morphological identification where applicable. These identifications were compared with the assignations made based on phylogenetic placement of the *8+ contigs* in the trees with and *Mito270* to examine the effect of taxon sampling on the number and resolution of identifications achieved. Tree-based identifications required monophyly of the contig with the reference sequences and were made to the lowest rank available (generally family or above). The same approach was used to characterise the contigs in the *nad4l* and *cox1* phylogenies and Spearman's rank correlation coefficient was computed (*cor.test*, method="spearman") to test whether the distribution of contigs between superfamilies was comparable with that inferred from the morphological identifications.

Contig and Sanger barcode sequences were aligned in Geneious with MUSCLE (Edgar 2004) with two outgroup sequences (NC_011277 and NC_011278) and trimmed to a matrix length of 648 bp. Sequences less than 320 bp were discarded. The remaining sequences were collapsed to unique haplotypes (allowing up to one mismatch) with a Perl script (Douglas Chesters, Available from: http://sourceforge.net/projects/collapsetypes/) and used for phylogeny reconstruction with RAxML (GTR+I+G, 100 rapid bootstraps). The tree was made ultrametric using r8s v1.8 (Sanderson 2003) after rooting with and then pruning the outgroup taxa. Putative species were delimited with GMYC under the single threshold model

Table 3.1 Data volume at each read processing step and the estimated percentage of quality-controlled pairs that were truly mitochondrial, for each of the two BorneoCanopy libraries.

Read pairs	*BC-short*	*BC-long*	Total
Raw	16,996,158	16,898,216	33,894,374
Adapters removed	10,701,469	11,961,260	22,662,729
QC	8,492,740	11,310,264	19,803,004
Blast filtered	833,709	1,257,165	2,090,874
Est. mitochondrial (%)	157,909 (1.86)	224,507 (1.98)	382,416 (1.93)

Table 3.2 Mitochondrial contigs obtained from the three assemblers and in the non-redundant set (after re-assembly), in each of four size classes. The number of contigs ≥15 kb that were circularised is also indicated.

Assembly	1-5 kb	5-10 kb	10-15 kb	≥15 kb (circular)
CA	456	105	33	44 (25)
IDBA	422	45	19	91 (54)
NWBL	365	54	35	63 (39)
NR	346	38	21	111 (80)

using the package *splits* (Tomochika Fujisawa and Thomas Ezard, Available from: http://r-forge.r-project.org/projects/splits/).

3.3 Results

3.3.1 Mitogenome (Re)-Assembly

The effect of the data processing steps undertaken in Chapter 2 are summarised in Table 2.1. The short- and long-insert length TruSeq libraries were approximately the same size, with 17.0 and 16.9 million read pairs respectively, however *BC-short* was worse affected by adapter contamination and low quality base calls and consequently was reduced more in size by the adapter removal and quality control steps. A similar proportion of the quality controlled reads were retained by the BLAST-filtering step in each case (9.82% and 11.12% respectively) and the estimated proportion of 'true' mitochondrial reads was also similar (1.86% and 1.98%). Although the two libraries were prepared to have different insert sizes (estimated by the sequencing provider to be on average 480 bp and 850 bp respectively), read mapping to the mitochondrial contigs ≥1 kb from the IDBA assembly indicated that the insert sizes of this portion of the reads were similar, averaging 425 bp and 440 bp respectively.

Assemby of the BLAST-filtered reads (both libraries combined) with three different programs (CA, IDBA and Newbler) and the subsequent contig-filtering step against *MitoDB* gave broadly similar results, with all three programs assembling >500 mitochondrial contigs ≥1 kb, of which the majority were <5 kb (>70% in all cases). IDBA has both the largest proportion of contigs <5kb and the largest number and proportion of 'complete' (≥15 kb) and 'nearly-complete' (10-15 kb) contigs of all three assemblers. The proportion of contigs ≥15kb that were circularised was similar between assemblies. The cumulative contig length distribution for each of the three assemblies and the non-redundant set is shown in Figure 3-1 and the equivalent frequency distribution is shown in Figure 8.1. All four datasets show positive skew towards contigs 1-2 kb and appear to be bimodal due to the presence of a second peak in the range 15-18 kb. However, the results of Hartigan's dip test for unimodality (Table 7.2) show that the CA distribution is not significantly different from a unimodal distribution (D=0.0162, p=0.3214) whereas the other datasets are at minimum bimodal (IDBA: D=0.0491, p<0.001; Newbler: D=0.0364, p<0.001; NR set: D=0.0719, p<0.01). Following this, the CA distribution was found to be significantly different from each of the other three with pairwise Kolmogorov-Smirnov tests (CA vs. IDBA: D=0.143, p<0.001; CA vs. Newbler: D=0.111, p=0.002; CA vs. NR set: D=0.148, p<0.001). IDBA was not significantly different from either Newbler or the NR set (IDBA vs. Newbler: D=0.052, p=0.456; IDBA vs. NR set: D=0.080, p=0.060), although Newbler was significantly different from the latter (D=0.095, p=0.019), as suggested by Figure 3-1. When considering all three sets of raw mitochondrial contigs together, there was an overall significant difference between their combined cumulative length distribution and that for the non-redundant set (D=0.103, p<0.001). Strikingly, the third quartile in the non-redundant set was longer than in all three individual assemblies, indicative of a shift towards longer contigs (10900 bp c.f. 5573-6088 bp). Reassembly led to overall reduction in the number of contigs, with an associated increase in the number of long (>10 kb) contigs and the proportion of circularised ≥15kb contigs to 72% (c.f. 57% CA; 59% IDBA; 62%, Newbler).

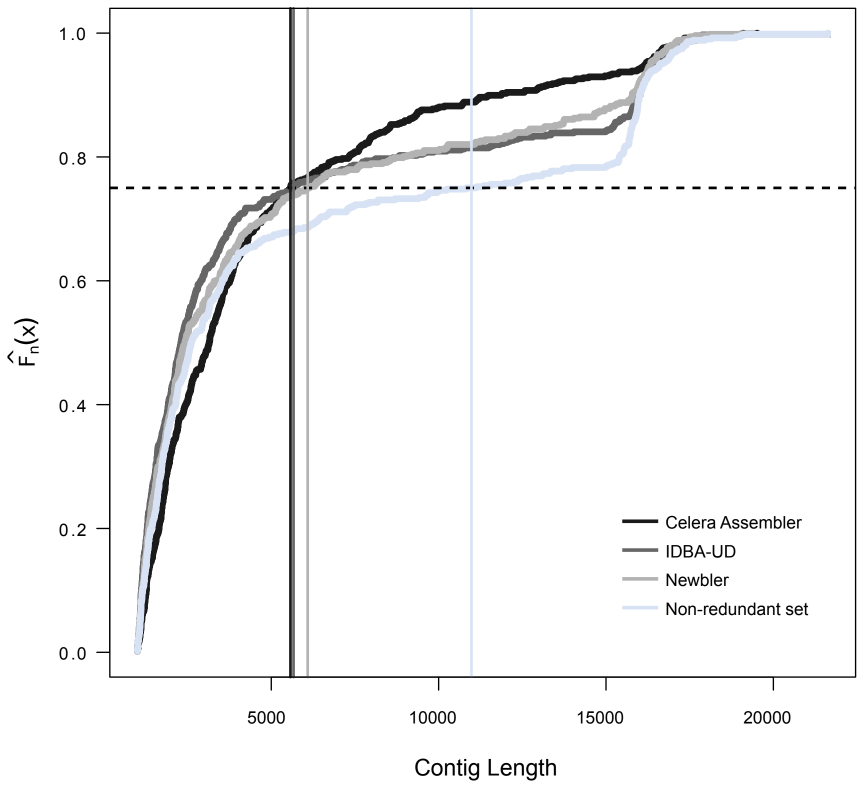


Figure 3-1 Cumulative length distributions for mitochondrial contigs in each of the three raw assemblies and the non-redundant set.

Figure 3.2 shows the length of the contigs as a function of their calculated mean coverage, for each of the three assemblies and the non-redundant set. In all three of the original assemblies contig length generally increases rapidly between 0 and 10x, with the majority of the long contigs clustered between 10 and 50x. However, in all cases there are several short contigs with very high coverage (up to ~200x). After reassembly the frequency of these short high-coverage contigs decreases while the number of long high-coverage contigs increases. This suggests that sequence contiguity for several species has been improved; yet the several remaining short high-coverage contigs indicate that this process is not completely effective.

In spite of the increase in the number of long contigs in the non-redundant set, the number of genes per contig retained a strong bimodal distribution, tending to be either complete or highly incomplete. For example, 112 contigs comprised a single gene whilst 111 contigs comprised all 15 genes (Figure 8.2). Overall 59.4% of contigs in the curated alignments contained 1-3 genes, yet these incomplete sequences represent just 17.4% of the aligned nucleotides, compared with 60.8% in the contigs with all 15 genes (22.1%). The number of sequences in the curated alignments varied between 178 (*nad2*) and 217 (*nad5*) (Figure 3.3) and the estimated mean coverage of those sequences ranged between 17x (*cox3*) and 23x (*nad2*, *atp8*, *nad5*), with an overall mean of 20x. The number of sequences per alignment was not found to be correlated with the mean coverage (r=-0.133, t=-0.482, d.f.=13, p=0.638). While Newbler performed well when considering the contig length distribution (more contigs >10 kb than CA and fewer contigs <5 kb than either CA or IDBA), it performed poorly when considering the number of unique gene sequences that were contributed to the final alignments (Figure 3.3). In this respect CA outperformed both other assemblers combined in the majority of instances. Overall Newbler performed relatively poorly, with 10-23% of sequences in each alignment not recovered (c.f. 2-9% and 4-11% for CA and IDBA respectively). The inclusion of both IDBA and CA, over CA alone, resulted in a net gain of 1.6-8.6%, adding Newbler gave a net gain of 0-1.5%. Strikingly, for all genes except *cox1*, over 90% of unique sequences were recovered by at least two of the assemblers (86% for *cox1*).

3.3.2 Matrix Coding for Maximum Likelihood

The recovered topologies from the three analyses of the *8+ contigs* with *Mito270* differed from one another in various respects. At the suborder level, the all-nucleotide (allNuc) and amino acid (AA) analyses recovered the same relationships, notably with basal (Myxophaga+Adephaga) and a paraphyletic Polyphaga with the inclusion of Archostemata between the two scirtoid branches. In contrast, the RY-coded matrix with the 3rd codon position removed (1RY2) recovered Myxophaga as the basal branch and Adephaga as sister to (Archostemata+Polyphaga), with the Polyphaga found to be monophyletic (Figure 3.4). Within Adephaga the Geadephaga was not monophyletic in any analysis due to the placement of the single Cicindelidae superbarcode within the Hydradephaga. Within the Polyphaga the inferred topologies varied greatly at the superfamily level, with the allNuc analysis in particular recovering alternative superfamily and infraorder placements such as a sister relationship between Bostrichoidea and Elateriformia, and a polyphyletic Staphyliniformia resulting from the placement of Histeroidea as basal to Polyphaga[-Scirtoidea]. Neither allNuc nor AA recovered Chrysomeloidea or Curculionoidea as monophyletic whereas 1RY2 did. As a result of these high-level differences between the topologies, allNuc and AA were found to be more similar based on the Robinson-Foulds metric (normalised RF = 0.27) than either was to 1RY2, and the latter was more similar to AA (0.30) than to allNuc (0.32).

The topologies from the equivalent analyses with the *8+ contigs* alone were more similar to one another but in no case were all of the component major clades found to be monophyletic.

In all three trees the Archostemata was placed within the Polyphaga with Scirtoidea as the basal branch of (Archostemata+Polyphaga). The three analyses also recovered the same relationships between the non-Cucujiform polyphagan superfamilies, in contrast with the results from the full tree. Within the Cucujiformia the Cucujoidea were in all cases recovered as three lineages rather than two (with variable placements) and in no case were Chrysomeloidea or Phytophaga (Chrysomeloidea+Curculionoidea) monophyletic. Overall, the AA and 1RY2 topologies were more similar to each other (normalised RF = 0.13) than either was to the allNuc topology (both 0.16). Comparing the topology of the 8+ contigs between the two sets of trees showed that the two 1RY2 topologies were the most similar (normalised RF = 0.13) and the *8+* 1RY2 and *Mito270* allNuc topologies were the most divergent (0.22). As a result of these analyses, the 1RY2 strategy was chosen for all subsequent phylogenetic analysis as the one most likely to recover all suborders, infraorders and superfamilies as monophyletic.

3.3.3 Effect of Taxon Sampling

To further examine the effects of taxon sampling on tree topology and the placement of the *BorneoCanopy 8+ contigs* in particular, three analyses were compared: *8+ contigs* alone, *8+ contigs* with reduced reference set (*Mito46*), *8+ contigs* with expanded reference set (*Mito270*). All three trees recovered different relationships between the suborders, with Adephaga as the basal group in the *8+* topology (no Myxophaga) and a paraphyletic Polyphaga with the insertion of the single Archostemata contig between the (single) scirtoid lineage and the rest of the Polyphaga. The suborders were all monophyletic in the trees with *Mito46* and *Mito270*, with the former showing a sister relationship between (Myxophaga+Adephaga) and (Archostemata+Polyphaga) (c.f. with *Mito270* above, (Myxophaga+(Adephaga+(Archostemata+Polyphaga)))). Within the Polyphaga, the relationships between superfamilies in the *8+* topology tended to be more similar to that with *Mito270* than with *Mito46*, for example both recovering a sister relationship between Bostrichoidea and Cucujiformia, and between Cleroidea and Tenebrionoidea. Some nodes were aided by the addition of the reduced reference set, e.g. Chrysomeloidea was recovered as monophyletic in the *Mito46* analysis with the inclusion of the single Hispinae contig that was placed as sister to Anthribidae in the 8+ analysis, whilst others were hampered, e.g. the paraphyly of Staphyliniformia by Bostrichoidea. With the further increase in taxon sampling this latter issue were resolved once more. Additionally, Cucujoidea was recovered as two major lineages rather than three and as a result the sister relationship of Chrysomeloidea and Curculionoidea was recovered (‘Phytophaga’). Overall, the symmetric difference between

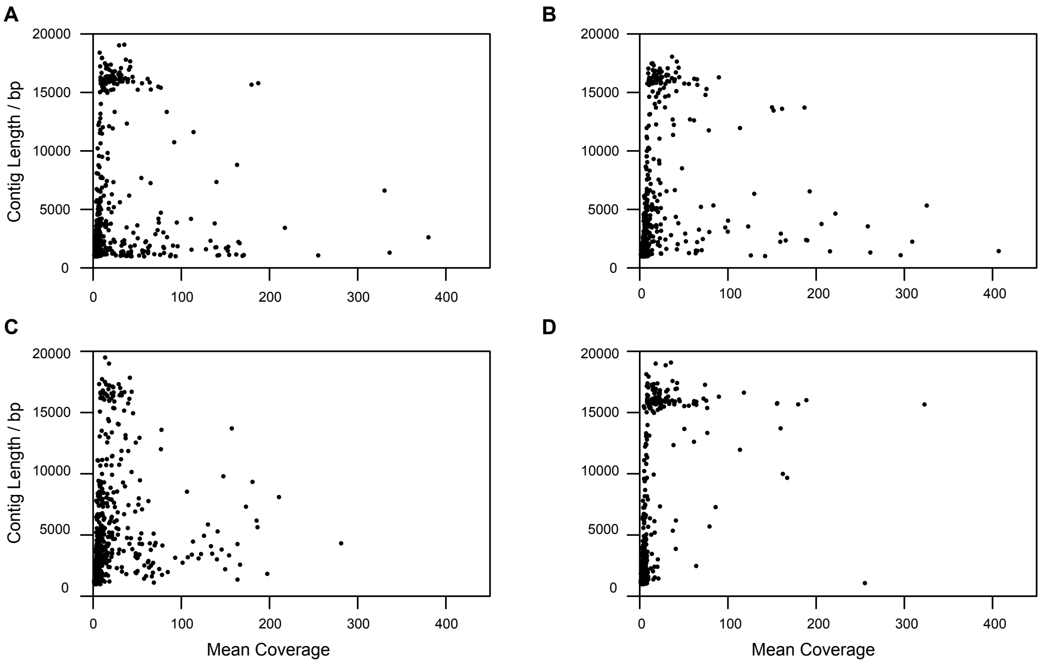


Figure 3.2 Assembled contig length as a response to mean coverage for each of the three assemblies and the non-redundant set. a) IDBA-UD. b) Newbler. c) Celera Assembler. d) Non-redundant set.

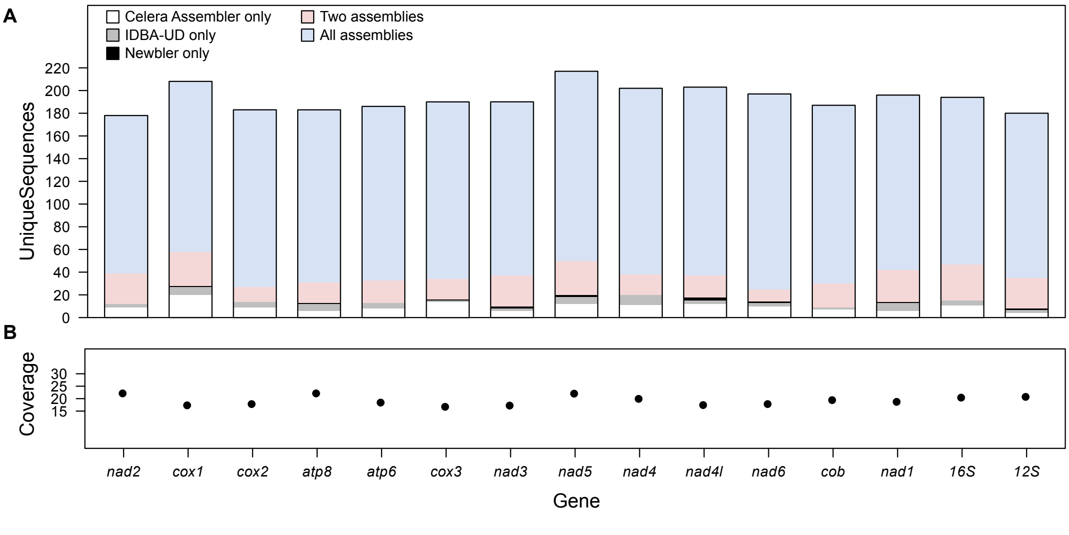


Figure 3.3 Redundancy between assemblers and variation in coverage between genes. a) The frequency of each gene in the final alignments and the extent of redundancy between assemblers in each case. b) Mean coverage per nucleotide in each alignment.

the three trees (for the *8+* contigs) was very similar, with slightly increased similarity observed between the reference-inclusive topologies (normalised RF = 0.11) than between either of these and the *8+* topology alone (0.13 in both cases). Following Crampton-Platt et al. (2015), increasing the number of reference sequences included in the phylogeny also improved the achieved resolution of taxonomic assignation, with 84.2% of contigs identified to family or better with *Mito270* compared with 34.2% with *Mito46* (Table 3.3). All verifiable (based on morphological identifications linked via DNA barcode baits; 100 of 146 contigs) assignments were accurate at the level at which they were made in the tree with *Mito46* as were all but six of the assignments made in the tree with *Mito270*. In the latter cases, four contigs identified morphologically as Cleridae were mis-assigned to Melyridae (Cleroidea), however the internal relationships within the Cleroidea were poorly resolved, with the reference sequences for this family forming a paraphyletic clade with Prionoceridae and Phycosecidae, while no Cleridae references were available. However, for the most part it appears likely that these issues arise from the instability of the taxonomy within this superfamily rather than the incorrect placement of the contigs (see Discussion). The two remaining conflicting identifications were two contigs assigned to Brentidae based on tree topology that were identified as Curculionidae based on morphology (both Curculionoidea).

3.3.4 Building a Community Phylogeny

Four community phylogenies were generated for each of the two gene-centric datasets, *nad4l* and *cox1*, of which two included superbarcode reference sequences and two were generated in a two-step process using a topology from sequences with ≥8 genes as a backbone constraint for the addition of the shorter sequences. The *cox1* barcode is missing from many of the superbarcode sequences and thus their number is greatly reduced in the *cox1* analyses compared with the *nad4l* analyses (119 c.f. 275). Following this, the achieved resolution of identifications is lower in the *cox1* with superbarcode analyses than the equivalent for *nad4l* (consistent with the results outlined above) whereas the analyses without the superbarcodes are unaffected due to the reliance on the identifications for the 8+ contigs derived from the *Mito270* topology.

In all four trees without superbarcodes the single Archostemata contig was placed between Scirtoidea and the rest of the Polyphaga, making the latter paraphyletic. Of the trees with superbarcodes, the two *nad4l* analyses recovered (Myxophaga+(Adephaga+(Archostemata+Polyphaga))) whereas the two *cox1* analyses recovered ((Myxophaga+Adephaga)+(Archostemata+Polyphaga)). This difference in the inferred relationships between the four suborders follows that observed between the *Mito270* and *Mito46* topologies respectively and thus is likely to reflect the differences in taxon sampling rather than gene choice. Within the Polyphaga the results are more variable, with only the *nad4l* analysis with superbarcodes based on an *8+* backbone recovering all superfamilies as monophyletic (with the expected split of Cucujoidea into two lineages). In general, when comparing the equivalent topologies between the two datasets the *nad4l* analysis recovers more monophyletic superfamilies than the *cox1* analysis, again likely reflecting the effect of (superbarcode) sampling. The exception to this are the analyses without superbarcodes constrained with the *8+* topology alone, wherein the superfamily relationships are identical apart from the placement of Bostrichoidea as sister to Staphylinoidea (*cox1*) or Cucujiformia (*nad4l*). The similarity between the two is due to the use of the same topology as a backbone constraint (pruned to include only the relevant contigs) and the one major discrepancy is due to the fact that only one *8+* contig was assigned to Bostrichoidea (by the *Mito270* topology) and this contig contained *nad4l* but not *cox1* and thus was not included in the backbone topology for the latter analysis. The position of Bostrichoidea was therefore unconstrained in this case. For each dataset, the topology of the *BorneoCanopy* contigs was most similar in the superbarcode-inclusive analyses (normalised RF = 0.05 (*cox1*) and 0.10 (*nad4l*)), and in all cases the pairwise *cox1* comparisons were more similar than the equivalent *nad4l* comparisons (normalised RF = 0.05-0.19 vs. 0.10-0.26). The greatest dissimilarity was observed between the '+superbarcode -backbone' and '-superbarcode +backbone' topologies in the *cox1* analyses, whereas for *nad4l* it was between the two unconstrained analyses (+/- superbarcode). When comparing the topology of the *BorneoCanopy 8+ contigs* present in both datasets between the two sets of analyses, the most similar (excluding the '-superbarcode +backbone' comparison discussed above) were the *cox1* '-superbarcode -backbone' and *nad4l* '-superbarcode +backbone' topologies (normalised RF = 0.09), followed by the two '+superbarcode -backbone' topologies (0.10).

ID resolution	Mito46 (all)	Mito270 (all)	Mito46 (verifiable)	Mito270 (verifiable)
Subfamily	0	26	0	16
Family	50	97	30	69
Superfamily	28	19	20	14
Suborder	67	4	50	1
Order	1	0	0	0

Table 3.3 Resolution of identifications for 8+ contigs in the Mito270 1RY2 tree, both overall (n=146) and for the subset of contigs for which placement could be verified against their respective morphological identifications with barcode baits (n=100).

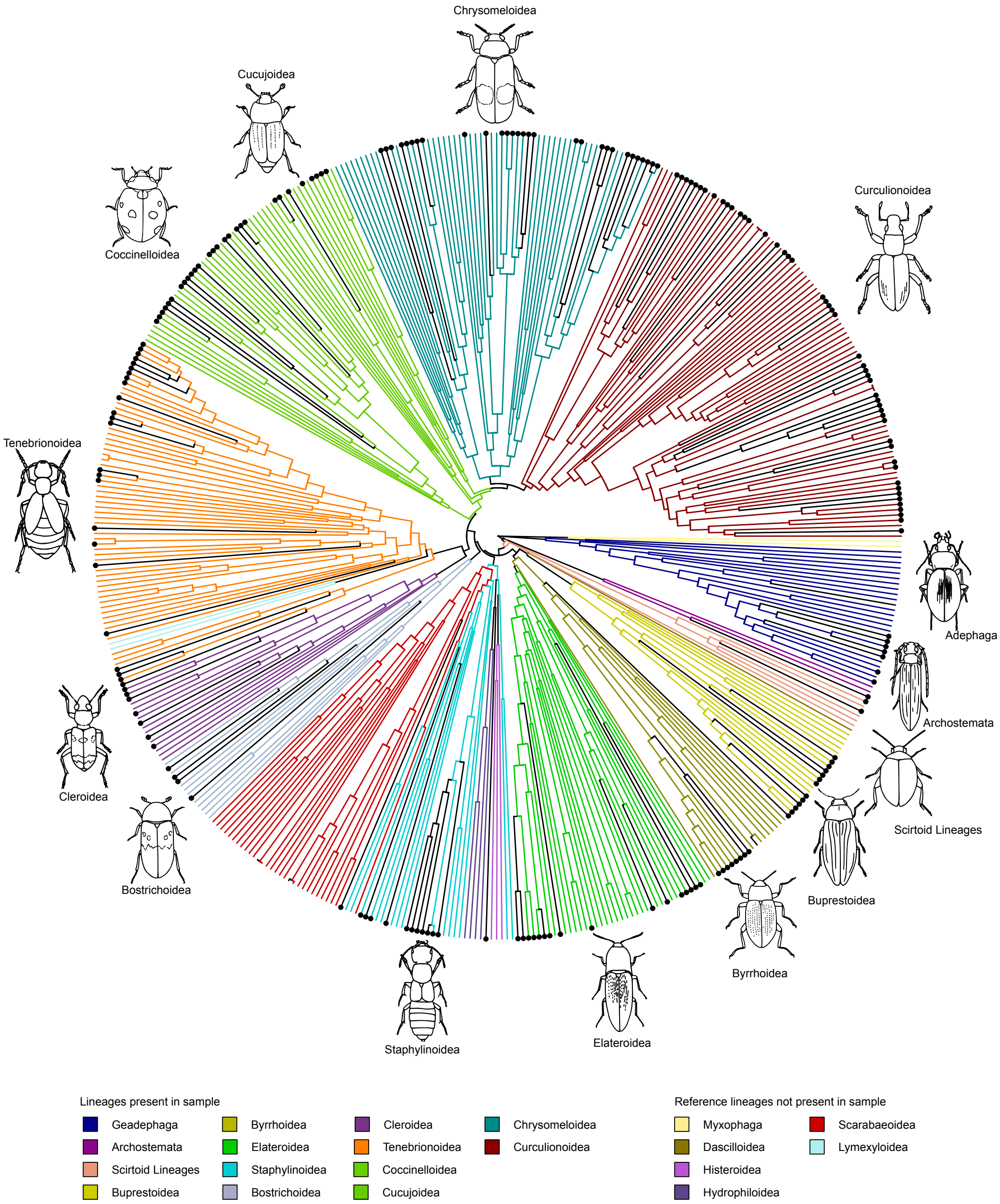


Figure 3.4 Mitochondrial phylogeny of beetles centred on *nad4l*. 275 reference sequences (*Mito275*) and 200 *BorneoCanopy* contigs including *nad4l*. *BorneoCanopy* contigs are marked with filled black circles. *Mito275* sequences and *BorneoCanopy* contigs with bait identifications are coloured with respect to superfamily. *BorneoCanopy* contigs without identifications are highlighted in black. Note that a single short *BorneoCanopy* contig has been incorrectly placed in Scarabaeoidea.

Table 3.4 Resolution of identifications in the community phylogenies, with and without superbarcodes and backbone topologies, for each gene-centric dataset. Numbers in parentheses are for the full dataset in each case (i.e. including *8+ contigs*), numbers outside parentheses are for the shorter contigs. Note that the *8+ contigs* in the two -superbarcode analyses derive their identifications from the *8+* with *Mito270* topology and these are in turn used to infer identifications for the shorter contigs.

ID resolution	-superbarcode -backbone		+superbarcode -backbone		-superbarcode +backbone		+superbarcode +backbone	
	nad4l	***cox1***	***nad4l***	***cox1***	***nad4l***	***cox1***	***nad4l***	***cox1***
Subfamily	7 (31)	4 (28)	6 (27)	2 (13)	7 (31)	4 (28)	6 (24)	2 (13)
Family	35 (128)	14 (101)	37 (127)	6 (74)	34 (127)	14 (101)	34 (122)	6 (74)
Superfamily	14 (33)	11 (28)	16 (44)	13 (49)	15 (34)	11 (28)	19 (51)	13 (49)
Suborder	4 (8)	7 (9)	1 (2)	14 (29)	4 (8)	7 (9)	1 (3)	14 (29)
Order	0 (0)	0 (0)	0 (0)	1 (1)	0 (0)	0 (0)	0 (0)	1 (1)

When comparing the resolution of taxonomic assignments between the various analyses (Table 3.4), the *nad4l* analyses always made a higher proportion of (sub)family assignations than the equivalent *cox1* analyses, both when considering all *BorneoCanopy* contigs and only the shorter ones (<8 genes). Overall these rates were similar between the four *nad4l* analyses and the differences seen when comparing the shorter subset and the full set were lower than in the equivalent *cox1* comparisons. These differences appear to be an artefact of taxon sampling rather than the choice of gene as the assignment rates for the full *cox1* dataset without superbarcodes is similar to those for the full *nad4l* dataset and in both cases these figures are driven by the *8+ contig* assignments made in the *Mito270* analysis, i.e. in the comparison where the effect of reference taxon sampling is minimised (130 vs. 140 identified *8+ contigs*) the overall rate of assignment to (sub)family is almost identical and the rate for short contigs is similar whereas when the effect of reference taxon sampling is maximised (119 vs. 275 superbarcodes) both the overall and short contig assignment rates are very different.

In the *cox1* analyses, the two +superbarcode and the two -superbarcode topologies resulted in the same taxonomic assignments for each contig, such that the use of a backbone topology had no effect on *BorneoCanopy* contig placement for this gene whereas all four analyses with *nad4l* were slightly different. Five disagreements between analyses were observed in the *nad4l* set (n=200) and none in the *cox1* set (n=166), although in the latter case the overall lower assignment resolution may also have reduced the chance of observing disagreements between the analyses. In all cases where the placement of a *cox1* contig could be verified against the morphological identification the two were in agreement at the level at which the tree-based assignation was made (except for one possible case of barcode sequence mix-up and the apparent paraphyly within Cleroidea), suggesting that the placement of even the shortest contigs was correct in all analyses. The same verification was not possible for the *nad4l* analyses, although the high degree of consistency between the taxonomic profiles (see below) and consistent placement for the majority of contigs suggests that these results are likely to be reliable on average. However, in the '+superbarcode -backbone' *nad4l* analysis one short contig (*nad4l* and *nad6*) was placed in the Scarabaeoidea (consistent with Passalidae) although the support values for both the placement of the contig itself and of the Passalidae as the basal branch of this superfamily were extremely low (<10). This placement is known to be incorrect due to the absence of any Scarabaeoidea in *BorneoCanopy* and in the three other analyses this contig was placed within the Staphylinidae.

3.3.5 Taxonomic Composition and Species Richness

The morphological assessment of the specimen images estimated that there were 209 species in 34 families within the *BorneoCanopy* sample, plus 3 species that were not identifiable to family level (Crampton-Platt et al. 2015). These 212 species covered three of the four suborders of Coleoptera and 13 of the 16 recognised polyphagan superfamilies. The distribution of these morphospecies between the superfamilies is shown in Figure 3-5 alongside the equivalent results for the 200 and 166 contigs respectively in the *nad4l* and *cox1* community phylogenies. The '+superbarcode -backbone' results are presented in each case as this is the simplest analysis and the one in which *nad4l* achieved the highest rate of (sub)family-level taxonomy assignment for the short contigs. The low rate of success with *cox1* is likely to be in large part an artefact of the reduced superbarcode availability with this locus and thus is presented mainly for completeness, and also to assess whether the distribution of contigs between superfamilies is similar to that observed with the morphological identifications and *nad4l* analysis, in spite of their reduced number, i.e. whether the same pattern of diversity was observed even at a reduced density. The most striking result, apart from the large number of *cox1*-based assignations only to Polyphaga, is the greatly reduced rate of recovery for Staphylinoidea in both contig-based analyses compared with the number of morphospecies. There were also five superfamilies where the *nad4l* analysis indicated more species than expected from morphology. The distribution of contigs between superfamilies in both analyses was highly correlated with the true distribution inferred from the morphological identifications, although the *nad4l* correlation was higher (*cox1*: $S=144.78$, $\rho=0.79$, $p<0.001$; *nad4l*: $S=59.17$, $\rho=0.91$, $p<0.001$). The *cox1* result was more highly correlated with *nad4l* than the morphology, indicating that the contig-

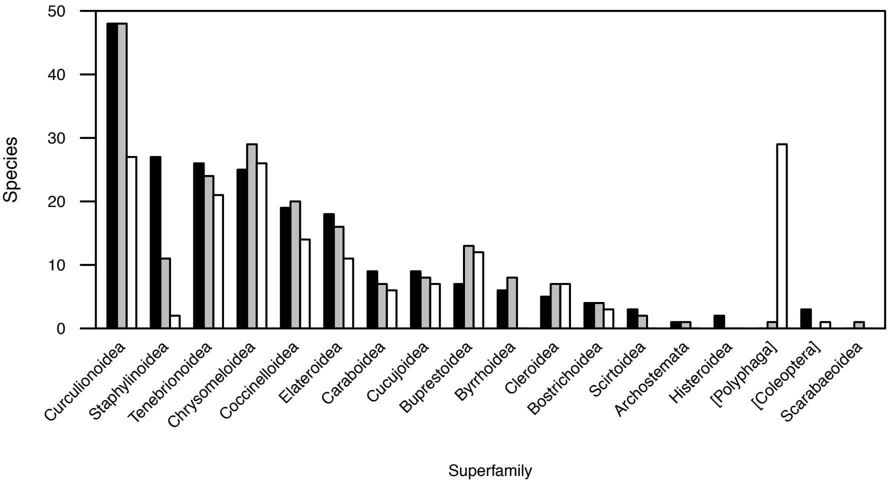


Figure 3-5 Superfamily-level taxonomic profiles inferred from the morphology (black) and the two community phylogenies (*nad4l*: grey; *cox1*: white; +superbarcode -backbone).

based analyses gave equivalent results even though the assignment resolution was much lower in the *cox1* analysis (S= 71.20, ρ= 0.90, p<0.001).

Total species richness within the *BorneoCanopy* sample was estimated by combining contig-derived barcode sequences with those obtained via PCR for a maximally inclusive diversity assessment. In total 185 contigs contained the *cox1* barcode region and were aligned with the Sanger barcodes for an initial matrix of 514 sequences, including two Neuroptera outgroups. Partial contig-derived sequences were discarded and the remaining 493 sequences were further collapsed to retain one representative for each of 336 unique haplotypes (334 ingroup haplotypes). GMYC analysis on an ultrametric phylogeny delimited 232 putative species in the canopy sample. Of these, 129 (55.6%) were shared by both the contigs and the Sanger sequences, 69 (29.7%) were recovered by the Sanger sequences alone, and 34 (14.7%) were recovered only by the contigs. Thus, the 164 contig-derived sequences included in the analysis represented 163 GMYC species. The two contigs which were delimited to the same GMYC group were 98% similar in the barcode alignment and 97.8% similar overall (both <2.5 kb).

3.4 Discussion

3.4.1 MMG for Bulk Samples

The assembly results for the bulk *BorneoCanopy* sample are consistent with those obtained in Chapter 2 when contrasting the 'voucher MMG' and 'bulk MMG' samples from the *ChrysIber* study (Gómez-Rodríguez et al. 2015) in that the contigs produced by each of the three assemblers tended to be short (Figure 8.1) and exhibited a similar pattern of increasing and then decreasing contig length with increasing coverage (Figure 3.2). The same behaviour was observed for all three assemblers and contrasted strongly with the equivalent pattern for the non-redundant set wherein the majority of short high-coverage contigs had been replaced by long high-coverage contigs, giving the appearance of an asymptotic relationship between coverage and contig length (due to the maximum mitogenome size of ~18 kb) and indicating that there is no benefit to the assembly of increasing average coverage per species above ~20x (Figure 3.2). The same effect of combining assemblies was observed when comparing the cumulative length distributions of the four datasets, with a shift in the third quartile to >10 kb and a sharp increase in the distribution function at around 15 kb such that 21.5% of the contigs were over this threshold in the NR set, compared with 6.9-15.8% in the individual assemblies (Figure 3-1). Thus, for this sample the re-assembly process resulted in a significant shift towards longer, more complete contigs as a result of the assembly of multiple shorter contigs, and in some cases allowed the merging of high coverage contigs that were problematic for the various assemblers. The remaining cases of short high-coverage contigs (Figure 3.2) indicates that this process has not been fully effective, although overall sequence contiguity increased greatly for a final dataset where 60% of the nucleotides available for analysis were contributed by the most complete 22% of contigs, while 59% of contigs were highly incomplete (≤3 genes) but contributed only 17% of the aligned data. This suggests that the difficulty for assembly presented by bulk MMG samples can at least in part be overcome by combining multiple assemblies in the absence of any further improvements in the current assembly algorithms or the development of new programs to deal specifically with the issue of assembling multiple (circular) orthologous contigs from complex mixtures containing variable read depth, interspecific divergences and intraspecific genetic variation.

In addition to the effect on sequence contiguity, combining multiple assemblies increased the number of unique gene sequences included in all fifteen gene alignments, with each assembly providing a small number of novel sequences in almost all cases (Newbler did not for 6 of 15 genes). Thus incorporating multiple assemblies improves the rate of recovery, with the greatest benefit derived from combining Celera Assembler and IDBA-UD. Taken at face value the results shown in Figure 3.3 indicate that CA would be the program of choice in cases where only a single assembly is desirable, due to the high rate of gene recovery. In particular if the assembly of *cox1* is the main aim of the study (e.g. Zhou et al. 2013) then CA would be the clear choice. However, this assembly was also dominated by short contigs, more so than either of the other two, indicating that many of the recovered genes will be non-contiguous and thus generate a highly incomplete nucleotide matrix. The similarity between the contigs length distributions for the IDBA assembly and the non-redundant set indicates that the former provides the main scaffold into which the other two assemblies are incorporated for an overall increase in sequence lengths. IDBA was the most successful single assembler in terms of the numbers of long and circularised contigs obtained thus for voucher MMG and bulk MMG concerned with phylogenetic analyses this assembler may be the most efficient choice for a single assembly. This tension between the increased contig lengths achieved by IDBA and the increased gene recovery achieved by Celera is resolved by the recommendation that these two programs are always applied to MMG samples and the resulting contigs subsequently re-assembled. The Newbler assembly provided relatively little novel data, however the availability of a third contig in many cases aided the decision making process when manually checking the Geneious re-assemblies making its inclusion worthwhile where possible, despite the limited novelty and overall low recovery rates (smallest number of raw contigs, greatest number of unique gene sequences missed). Although differences in gene recovery were observed between the three assemblers it should be noted that in all cases at least 72% of the sequences in any single alignment were recovered by all three programs indicating that the assembly of mitogenome data from MMG samples is highly repeatable, although not yet fully optimised with any of the programs used herein.

3.4.2 Building a Beetle Tree-of-Life

There is much debate in the literature about how particular features of insect mitogenome sequences (compositional and among-site rate heterogeneity) should be dealt with in phylogenetic analyses (reviewed in Cameron 2014), although the hierarchical levels at which the various studies have focussed and the extent of taxon sampling varies widely. Often, the results of analyses with the CAT site-heterogeneous mixture model on amino acid alignments in the program PhyloBayes (Lartillot et al. 2009) are the preferred choice, especially at inter-ordinal levels and above, as this model tends to minimise the effects of long-branch attraction (LBA) (Talavera and Vila 2011). Mitochondrial genomes have been successfully applied to intra-ordinal relationships using other methods and the need for reductive coding of matrices has been refuted (Cameron et al. 2007), however the majority of studies to date have involved extremely limited taxon sampling (generally fewer than 30 in-group sequences) and the effect that this might have on the analysis has not been considered. In the present study, relatively modest changes in taxon sampling (n=146, 192, 416) within a single analysis type (1RY2) and different matrix coding (all nucleotide, 1RY2, amino acid) for the same taxa (n=146 and 416) produce highly variable topologies with RAxML. It is likely that other available methods are better able to deal with the heterogeneity of the current dataset and thus these effects might be reduced by using a different analytical strategy (Sheffield et al. 2009; Talavera and Vila 2011), although at a likely cost of significantly increased analysis times. Here the main concern is to efficiently obtain an acceptable tree topology with increasingly large and complex datasets, hence the focus on the effects of taxon sampling and matrix coding in RAxML analyses. Bayesian analyses rapidly become impractical with increasing dataset size using current implementations and thus realistically will not be used to generate very large trees directly. While the routine use of PhyloBayes for large MMG datasets is unlikely in the foreseeable future, using smaller PhyloBayes analyses to generate incomplete backbone constraint trees for RAxML may be a good compromise solution. Constraining major groups to be monophyletic is a common strategy to simplify analyses and obtain the expected topology in cases where certain nodes are known to be difficult to recover, however it is important to note that this is not a useful strategy for bulk MMG samples where many contigs are unlinked to a morphological identification. For example, the single Cassidinae contig in the *BorneoCanopy* set was placed within the Curculionoidea in the majority of analyses and thus any constraint on Chrysomeloidea would not have included this sequence. In the analyses where this contig was correctly placed this made the superfamily monophyletic and thus there was no need to constrain it.

For the full dataset requiring a minimum of eight genes (n=416), matrix coding had a large effect on tree topology, with the allNuc and AA topologies more similar to each other than either was to 1RY2 but the latter recovered all major clades as monophyletic, exhibited the highest bootstrap support values at major nodes and gave the topology most similar to the PhyloBayes (amino acid alignment, CAT model) topology obtained by (Timmermans, Barton, et al. 2016). This was the only analysis of the three to obtain monophyly of the four suborders and all superfamilies (with the traditional Cucujoidea split into two of the currently recognised lineages - ‘Coccinelloidea’ and ‘Cucujoidea *s.s.*’ (McKenna et al. 2015)). In the same comparisons under reduced taxon sampling (n=146, *8+ contigs*) there were fewer differences between the three analyses, with the same inferred relationships between all major clades except within the Cucujiformia. Here, the 1RY2 topology was marginally preferred as Anthribidae (with the single Cassidinae contig) was recovered as the basal branch of Curculionoidea. Thus, it appears that choice of matrix coding has an increasing impact on the likelihood of recovering high-level lineages as monophyletic as taxon sampling increases, presumably because sequence heterogeneity increases with dataset size and leads to increased noise in the full alignment. Comparing the topologies of the *8+ contigs* between the six analyses indicated that the 1RY2 trees were the most similar and therefore analyses with this coding appear to be the most robust to variation in taxon sampling. Thus, for the subsequent maximum likelihood analyses with variable taxon sampling and contig lengths, 1RY2 coding was considered to be the most appropriate choice.

Comparing the 1RY2 analyses for the *8+ contigs* under three levels of taxon sampling indicated that the topology generally improves as sampling increases (no superbarcodes; with 46 superbarcodes; with 270 superbarcodes), although the effect as measured with the Robinson-Foulds metric was relatively low. The number of observed differences was lower in the comparison of the reference-inclusive trees than between either of these and the tree with the *BorneoCanopy* contigs alone, indicative of a slight stabilising effect as sampling increases. The two reference-inclusive topologies were however quite different, both when considering only the *8+ contigs* and when assessing the overall topologies, but in general increasing taxon sampling facilitated the recovery of additional monophyletic clades. Thus as a general rule it appears that including additional taxa improves tree topology both for the subset of contigs of interest and the wider mitochondrial tree-of-life (MTOL).

In all four trees where it was assessed (allNuc, 1RY2 and AA with *Mito270*; 1RY2 with *Mito46*), the placement of the *BorneoCanopy 8+ contigs* (n=100) was consistent both with the morphological identifications where available (except for within Cleroidea) and between the analyses. This indicates that the placement of individual contigs in relation to their closest relatives is robust to variation in taxon sampling and matrix coding even though overall the topologies vary with respect to the monophyly of major clades and the relationships between them. That the high-level topology of the *8+ contigs* is not greatly worse in the absence of external reference sequences is promising for the generation of robust and reliable community phylogenies even where their veracity cannot be assessed due to a dearth of superbarcodes.

3.4.3 Building a Community Phylogeny

There are two main approaches available for generating a phylogeny for an ecological sample or community with a gene-centred analysis. The simplest option is to undertake a single analysis with all sequences that contain the locus of interest, alternatively an initial tree can be generated using only the longest (assumed to be the most informative) sequences that contain the locus and use the resulting topology as a backbone constraint for the addition of all shorter sequences. Within these two approaches there is also the possibility of including external superbarcodes (if available) or not. Following the previous section, the use of reference sequences is expected to improve the overall topology due to the increase in taxon sampling, particularly where the sampled community is taxonomically unbalanced, and also aid characterisation of the sample(s) (see next section). For analyses of community phylogenetic diversity the reference sequences would subsequently be pruned for a final community phylogeny. There are a number of questions that could be asked in relation to these strategies, including whether there should be a minimum contig length and/or loci number cut-off for inclusion in the analysis (e.g. Crampton-Platt et al. 2015; Gómez-Rodríguez et al. 2015), or whether shorter contigs should be added in one or more steps (e.g. Andújar et al. 2015). Here the aim was to assess whether the choice of gene affects tree topology and downstream analyses when allowing maximal sequence inclusion (no minimum length or number of loci). Two loci were used for these analyses, one of which is highly conserved at the amino acid level (*cox1*) and one of which is highly variable (*nad4l*). The latter was the most frequently recovered locus where all contigs overlapped, while the former was used to maximise the number of contig placements that could be verified against the morphological identifications. The dense superbarcode sampling available for Coleoptera is primarily due to the data generated by Timmermans and colleagues using long-range PCR (Timmermans et al. 2010; Timmermans and Vogler 2012; Haran et al. 2013; Timmermans, Barton, et al. 2016). Unfortunately this method involves amplification of two main fragments which overlap in the middle of *cox1* and the fragment which spans the control region and includes the barcode is more difficult to amplify, thus many of the available sequences include only the *cox1-3'* to *cob* fragment. As a result, the differences observed in the gene-centred analyses with superbarcodes are likely to result primarily from the reduced taxon sampling in the *cox1* dataset rather than gene choice *per se*. Therefore the effect of gene choice is difficult to assess from the present analyses, although the similarity in the results where no superbarcodes were used (minimising the difference in taxon sampling) suggests that it has little or no effect. Overall the placement of short contigs was not problematic, with no *cox1* cases found to be incorrect where verifiable against morphology.

Additionally, in both datasets the placement of short contigs was generally consistent in all four trees, indicating that this is neither greatly affected by the choice of gene nor the type of analysis, although resolution of achieved identifications did vary (see next section) and one incorrect identification was made in the '+superbarcode -backbone' *nad4l* analysis.

In both sets of gene-centric analyses the topology of the *BorneoCanopy* contigs was most similar in the presence of superbarcodes, a finding that is consistent with the earlier observation that increasing taxon sampling had a stabilising effect on the topology of the *8+ contigs*. The unconstrained analysis without superbarcodes in both cases produced the worst topology and therefore this appears to be the least useful strategy for obtaining an accurate community phylogeny, although the rate of taxonomy assignment based on the position of the *8+ contigs* in the *Mito270* analysis was unaffected. The three remaining topologies in each case were broadly similar and indeed for *cox1* the '-superbarcode +backbone' topology was preferred. Thus in the absence of suitable superbarcodes the community phylogeny should be generated in two steps, firstly by analysing the longest sequences in isolation (≥8 genes herein) and subsequently using this topology as a backbone constraint for the addition of the shorter gene-centric contig set. Even when suitable superbarcodes are available their utility for generating the community phylogeny is not clear cut - the preferred topology under reduced superbarcode sampling (*cox1*) was '-superbarcode +backbone' whereas under dense taxon sampling (*nad4l*) it was '+superbarcode +backbone', indicating that the number of available superbarcodes will influence the strategy for generating a community phylogeny. However, the most consistent analysis type between the two datasets was '+superbarcode - backbone', i.e. the strategy most robust to variation in taxon sampling and/or gene choice. The boundary between 'sparse' and 'dense' superbarcode sampling lies is somewhere between 119 and 275 for the current dataset and thus is far beyond what is currently available on NCBI for other insect groups. The '-superbarcode +backbone' *nad4l* topology was not much worse than the preferred one and thus in most instances this is likely to be the most practical, realistic and robust strategy for generating community phylogenies from MMG data for the foreseeable future. However, where possible it is desirable to compare at least two topologies to mitigate against the effects of taxon sampling and gene choice and maximise the likelihood of uncovering incorrect contig placements. From these results the optimum suggested combination for future studies is '-superbarcode +backbone' and '+superbarcode -backbone'.

3.4.4 Characterising Communities with MMG

For downstream community ecology analyses maximising the number of species recovered from MMG data is critical to identifying true patterns of diversity and detecting real differences between communities (Gómez-Rodríguez et al. 2015). This question is dealt with in greater detail in the following chapter, however the completeness of the data obtained from any bulk MMG sample will always be a critical benchmark for success, albeit one that is difficult to assess for true bulk samples which have not be characterised *a priori*. The current dataset is a good test case for the application of MMG to bulk samples of tropical diversity both because the morphological and PCR barcode-based characterisations offer a baseline against which observed species richness can be judged, and because the results presented herein can be compared with those presented in Crampton-Platt et al. (2015) which were derived from an alternative treatment of the same raw data. In both MMG studies the estimated number of species was slightly lower than the conservative morphological richness estimate (212 morphospecies) and similar to the richness recovered by the barcode data. However, the barcode data is incomplete (<70% of specimens) and thus both sequence-based methods underestimate diversity (~200 species in all cases). When the MMG and barcode data were combined the total number of species estimated for the sample was the same in both analyses (232 GMYCs) in spite of the slightly increased assembly success herein. The consistency in the results suggests both that the current estimate of diversity for this sample is largely complete and that MMG analyses are both highly repeatable and robust to variation in the precise protocol used. In both cases approximately 85% of the predicted number of species were recovered by MMG analyses focussed on the most frequently assembled gene, although in the previously published analysis the requirement for contigs to be ≥2 kb reduced the proportion included in the community phylogeny to 75%.

Although the present assembly is slightly more complete than the previous version, two major deficiencies remain when comparing the taxonomic profiles obtained from the community phylogenies with the morphological profile, namely the failure to include any Histeroidea contigs in either community phylogeny and the great discrepancy between the expected and observed Staphylinidae richness (morphology: 27; *nad4l*: 11; *cox1*: 2). In the latter case gross morphology is usually expected to underestimate species richness and thus the true discrepancy is likely to be greater. However, in the current *nad4l* analysis the observed Bostrichoidea species richness matches that expected from the morphology, whereas previously no contigs were identified as Bostrichoidea, although the position of one short contig was consistent with this identification in the community phylogeny (Crampton-Platt et al. 2015, Fig.3b).

Neither community phylogeny completely represented the diversity of the sample expected from the morphological analysis, however in both cases the observed superfamily-level profile was strongly correlated with the expected profile and so the results can be considered broadly equivalent. The *nad4l* tree incorporated a greater proportion of the expected species richness (200 contigs vs. 166 in *cox1*) and also benefitted from greater superbarcode sampling with respect to the achieved resolution of taxonomy assignments. These two factors combined resulted in a much closer correlation between the taxonomic profile obtained from this dataset and the morphological profile than was achieved with the *cox1* dataset, although the latter was highly correlated with the *nad4l* profile. This and the greater similarity in the profiles obtained from the superbarcode-exclusive topologies show that the relatively low representation of the barcode region in the superbarcode set has artificially reduced the success of taxonomic profiling for the *cox1*-centred analysis. This issue is expected to be most severe in Coleoptera and will become less problematic as the number of available superbarcodes increases, especially with the increasing uptake of voucher MMG over LR-PCR.

3.4.5 Conclusions

This Chapter builds on the foundation laid in Chapter 2 to consider a re-assembly step to optimise the initial set of contigs, considering both the effect of this step on the contiguity of the final sequences and the relative contribution of each of the three original datasets. This step is expected to continue to be important for both voucher and bulk MMG in the medium term but the extent of its efficacy is rarely considered. Here this step has a significant effect on the length distribution obtained and is seen to resolve many cases of incomplete assembly of high coverage species, in agreement with the observations of Crampton-Platt et al. (2015). The inclusion of multiple assemblies also assists recovery of unique gene sequences, although the effect of adding the third assembler, Newbler, was limited. Therefore a minimum of two assemblies should always be undertaken and combined wherever possible and a third assembly included as an additional check if desired. The re-assembly process for this tropical sample is found to be effective but not complete, leading to further consideration of this problem in the next Chapter for a more complex case, but there does not appear to be any significant impediment to the application of bulk MMG to tropical samples. The findings in the present Chapter with respect to the strategy for phylogeny reconstruction support the use of reductive coding and maximum likelihood analyses as an effective method for obtaining a satisfactory phylogeny under variable and increasing superbarcode sampling. For the community phylogeny the results are less clear-cut, with variation in the quality of the topologies obtained when the shortest sequences are included and a small number of apparently erroneous placements. The effect of gene choice to ensure orthology in these analyses was difficult to assess due to the effect of differential superbarcode sampling in the *nad4l* and *cox1* sets. Notably, the analyses with the most similar level of taxon sampling also gave the most similar result, indicating that gene choice may be unimportant. Overall the accuracy of contig placement appeared high even for the shortest sequences, indicating that these could routinely be included in community phylogenies, although the misplacement of one short sequence in only one of the analyses led to the suggestion that short contig placements should be compared between at least two different topologies where possible. Unsurprisingly the reduced number of available superbarcodes in the *cox1* phylogeny had a large effect on the resolution of taxonomic profiling, although the pattern was similar to that obtained with *nad4l* and that based on morphology. Thus the observed patterns were congruent in spite of the incomplete description of the *cox1* set and the potential for a phylogenetic approach to describing the broad taxonomic composition of uncharacterised communities is largely confirmed.

Chapter 4 Landscape Ecology and MMG: a case study in New Forest NP

Summary

This Chapter draws on the results of Chapter 3 to find an optimum solution to the re-assembly of bulk MMG samples from a UK terrestrial beetle community and use a phylogenetic approach to taxonomic description whilst also taking advantage of the availability of bait sequences for species-level identifications where possible. A landscape-level application of bulk MMG is seen as an important test for assessing the sensitivity of the approach and its utility for conservation planning at this scale. A combined compositional and phylogenetic perspective is taken, unifying the key motivations for MMG-based community ecology for the first time. Bulk MMG is used to assess the extent to which two woodland habitats with different management histories in the New Forest National Park differ in their leaf litter beetle communities, with mixed results. Expected differences in alpha diversity are not recovered but beta diversity both within and between habitats conforms to expectation and the phylogenetic analyses point towards some unique differences between the two communities that are not detected by other means. The prospects of bulk MMG for temperate community ecology are discussed, particularly with respect to the observed pattern of accumulation and apparent insufficiency of the current sequencing depth to recover the true alpha diversity.

4.1 Introduction

Knowledge of the distribution and structure of biodiversity is a key requisite for effective conservation planning and subsequent monitoring of the effect of implemented management actions, however, in practice, detailed data are usually lacking at all taxonomic, spatial and temporal scales (Favreau et al. 2006). In particular, the abundance and diversity of invertebrates, and the corresponding knowledge gaps with respect to taxonomy, distribution, spatial and temporal dynamics, and ecological function have a knock-on effect on their conservation (Cardoso et al. 2011). Thus, in spite of the long-recognised advantages of including terrestrial arthropod assemblages in conservation planning and monitoring (Kremen et al. 1993; Hughes et al. 2000), such steps are rarely taken. The potential for next-generation sequencing (NGS) to increase the inclusivity of ecosystem assessments is slowly being realised, with increasing calls for the use of metabarcoding to assess otherwise intractable diversity via environmental DNA (Hajibabaei et al. 2011; Thomsen and Willerslev 2014) and bulk arthropod samples (Yu et al. 2012; Ji et al. 2013) across a range of terrestrial and aquatic systems. In addition to the higher throughout of such methods as compared with morphological surveys, one major advantage of such data is the possibility for results to be verified by external observers and compared between areas, potentially making decisions regarding funding allocations more transparent and effective (Ji et al. 2013). Mitochondrial metagenomics (MMG) may also be an effective tool for biodiversity monitoring in such situations although it is currently significantly more expensive due to the lack of PCR-enrichment. However, concerns over the potentially biasing effect of PCR on inferred taxonomic composition and genetic diversity, and the loss of the link between biomass and read numbers may be sufficient in some cases to justify the use of MMG over, or in combination with, metabarcoding. Indeed, a direct comparison of MMG and metabarcoding for monitoring wild bee populations found that MMG had a higher profiling success whilst also producing species richness, community structure and biomass patterns closely correlated with the morphological results (Tang et al. 2015). In addition, MMG offers the potential for phylogenetic diversity to be measured in baseline surveys for the entire invertebrate community which can later be taken into account when prioritising areas for conservation (Faith 1992).

MMG has previously been found to reliably replicate species diversity patterns established from species-level morphological identifications, although the relative merit of bulk MMG when compared with voucher MMG for generating the reference library against which read-based assemblage profiling is undertaken was debatable (Gómez-Rodríguez et al. 2015; but see Chapter 2). The results obtained in the latter study against a reference library generated via voucher MMG, alongside those of Tang et al. (2015), indicate that variation in species richness and biomass between samples can be effectively recovered, although the required depth of sequencing is uncertain. In the study of Gómez-Rodríguez et al. (2015) the results obtained against the reference library generated by bulk MMG were generally correlated with expectations from morphology but the failure to assemble contigs for many low biomass species hampered the recovery of some patterns. In this context, and following the clear differences in assembly behaviour of the voucher and bulk MMG samples illustrated in Chapter 2, there is a need for the limits of bulk MMG to be further assessed.

In the present Chapter, bulk MMG is used to assess landscape-level patterns of beetle diversity in leaf litter. This is the spatial scale at which many conservation decisions are made and thus the sensitivity of MMG to variation at this level is a key test. The landscape in question is the New Forest National Park, bounded mostly within Hampshire on the south coast of England, United Kingdom. The National Park was designated in 2005 and covers approximately 57,100 hectares of enclosed and pasture woodland, wet and dry heathland, and grassland. The National Park incorporates an array of existing protected areas, the most significant of which is the New Forest Special Area for Conservation. The New Forest landscape has been described as the largest area of semi-natural vegetation in lowland Britain (Tubbs 1968) and includes habitats that are otherwise rare, most notable for this study being the extensive pasture woodland, thought to be the largest in north-west Europe (JNCC 2015). The unique mosaic of habitats derives from its position on relatively poor soils and a long and complex history of human exploitation. This mosaic is thought to contribute to a relatively high diversity of invertebrates throughout the New Forest area although increasing grazing pressure appears to be causing declines in some habitats (Pinchen and Ward 2010). However, baseline data appear to be extremely limited and sampling effort across the landscape tends to be sporadic and localised, making it difficult to draw firm conclusions about the health of the invertebrate communities. Of particular significance are the ancient unenclosed pasture woodlands (a.k.a. Ancient and Ornamental Woodlands (A&O)), which have the highest density of invertebrate species of conservation concern in the National Park (Pinchen and Ward 2010). These are recognised as an exceptionally important European stronghold of saproxylic beetle diversity alongside Windsor Great Park, yet many species known to have been present historically have not been recorded for several decades making the current status of this guild uncertain (Alexander 2010).

4.1 Introduction

The pasture woodland has survived largely intact since at least the 1600s and presumably since the instigation of forest law under William I, although the extent of woodland cover has varied historically and is much debated (Newton et al. 2010). Interspersed between fragments of ancient woodland and the more open habitats are the woodland 'inclosures' which have been felled and replanted at various points in their history and were enclosed to deter browsing. Since the Second World War and the formation of the Forestry Commission to replenish forestry reserves the number of inclosures expanded, with a focus on exotic conifers for rapid timber production that is only recently being reversed (Smith and Burke 2010). Active inclosures tend to be fenced against livestock whereas animals are free to move unimpeded throughout the pasture woodland and other habitats. As a result, deciduous inclosures generally have a greater availability of understory vegetation when compared with pasture woodland sites, although livestock exclusion is not always effective and fences do not provide any barrier against deer.

The Park is thought to be at a critical moment in its history, with rising concern about resilience in the face of climate change, the upward trend in stocking rates to unprecedented levels, and the shift in woodland management practices (Newton 2010). In spite of the long tradition of natural history and ecology in the New Forest area there is surprisingly little quantitative data collected at the landscape scale, particularly for invertebrates, limiting understanding of their diversity and dynamics across the landscape and the interaction between adjacent habitat patches. One recent study by Carpenter et al. (2012) sought to rectify this with a benchmark survey of soil macrofauna covering all habitat types with a spatially replicated 'parcelled' sampling design. This study revealed a distinct leaf litter community associated with wooded habitats, of which the ancient woodlands in the 'core' of the National Park were particularly diverse but otherwise there was no clear separation between ancient woodlands and inclosures (Carpenter et al. 2012). Whether these two habitats should be considered as distinct and therefore managed independently is a key question at a time in which the opening up of significant numbers of inclosures to grazing is being planned (Smith and Burke 2010). There is particular concern that this will lead to the further deterioration in food plant availability for pollinators (Pinchen and Ward 2010) but there are no clear predictions as to the effect of such management changes on the wider invertebrate community. The effect on the community in leaf litter may be potentially significant due to an increase in trampling by large mammals and the removal of surface vegetation, altering the microclimate of the newly exposed leaf litter. Thus in the present study bulk MMG is used to assess the extent to which these two habitats are currently distinct, with respect to the beetle community in leaf litter. Using beetles as a proxy for the diversity of the entire community is not ideal, however this group is consistently the most abundant in New Forest leaf litter samples and had the highest species richness of the groups studied by Carpenter et al (2012), making this a good test case in the first instance.

Following the potential bulk MMG challenges highlighted in Chapter 2 and the contrasting results of Chapter 3, there still remain some questions regarding the success of assembly from such samples. There are many potential reasons for the disparity in assembly success for the *ChrysIber* and *BorneoCanopy* data, not least the expected differences in species abundance distributions between temperate and tropical communities. The high species to specimen ratio (232 to 477) in the *BorneoCanopy* sample may have facilitated assembly by reducing the disparity in DNA contribution per species to the pool. In contrast, the 2607 specimens in the *ChrysIber* study were drawn from just 171 species (Gómez-Rodríguez et al. 2015). The latter situation is typical of temperate samples of beetle diversity and thus the observed discrepancy in assembly behaviour (Chapter 2) and subsequent ecological analyses (Gómez-Rodríguez et al. 2015) between bulk and voucher MMG is a cause for concern. Consequently, the application of the lessons learnt from Chapter 3 to a temperate system is one of the major themes of the current Chapter. Nevertheless, while the observed discrepancies highlight an important focus for further optimising the performance of bulk MMG on real world samples, it should be highlighted that even with the greatly reduced database size (96 species missing from *DeNovoRL* c.f. 24 from *MitoRL*) the most important ecological patterns were still recovered by bulk MMG in the study of Gómez-Rodríguez et al. (2015) and profiling success was reasonable. Thus even with an incomplete inventory, bulk MMG is expected to be sufficiently sensitive to detect variation in diversity between samples, although the significance of such differences may be reduced.

4.1.1 Chapter aims

Here, bulk MMG is applied to twenty samples of leaf litter beetle diversity from paired woodland sites across the New Forest National Park. Ten ancient woodland and five inclosure plots sampled by Carpenter et al. (2012) are revisited along with an additional five inclosures to complete the spatially replicated paired sampling design. All adult specimens from each site are homogenised for DNA extraction, making these the only truly bulk samples presented in this thesis. Samples are sequenced on the Illumina MiSeq platform after TruSeq PCR-free library preparation and the data obtained is pooled to assemble a single set of reference contigs for the global community which are subsequently used for phylogeny reconstruction and read-based assemblage profiling to infer the incidence of each species across the landscape. The profiling results are used in combination with the phylogeny to establish patterns of compositional and phylogenetic diversity across the landscape and between habitats. Methodological points under consideration include the exhaustiveness of assembly and the extent to which database completion and marker choice affects the observed patterns.

Following Carpenter et al. (2012), the leaf litter community in the ancient woodlands, particularly the core sites, is expected to be more species-rich than that in the inclosure woodlands but turnover in community composition is expected to be low between habitats overall and similarly high between sites in both habitats. If differences are found, the expectation would be for lower rates of turnover between inclosure sites relative to ancient woodlands due to more recent and frequent habitat disturbance (caused by clear felling for timber), resulting in communities of mostly vagile (dispersive) and widely distributed species which vary little in composition between sites. No specific prediction can be made about the phylogenetic diversity of these communities based on previous work, but will plausibly follow the pattern observed at the species level. Similarly, phylogenetic structure has not previously been examined in the New Forest, therefore evidence pointing to different community assembly processes in the two habitats would potentially be significant for future conservation planning. Higher disturbance levels in the inclosure woodlands may favour species with particular traits, such as high vagility, ecological generalism, and greater tolerance of variable abiotic conditions. Such traits are likely to be clustered at the tips of the phylogeny in groups of closely related species, hence the community associated with inclosure woodlands may show signs of ‘habitat filtering’ in their phylogenetic structure. Lastly, Gómez-Rodríguez et al. (2015) found that the most important ecological patterns were recovered even against a highly incomplete reference database, thus the results of the various analyses are expected to vary little with respect to database size.

4.2 Materials and Methods

4.2.1 Sampling

Ten pairs of sites within the New Forest National Park were sampled between May and July 2011 (Table 4.1; Figure 4-1). Each pair of sites was composed of one ancient pasture woodland (A&O) and one inclosure. Classification of habitats was based upon Forestry Commission records (Carpenter et al. 2012). Sites tend to be dominated by oak (mostly *Quercus robur*) and/or beech (*Fagus sylvatica*). Of the sites sampled, five pairs (core sites) and an additional five A&O woodlands (peripheral sites) had previously been selected and sampled as part of the New Forest Quantitative Initiative (NFQI). An additional five inclosures were selected to pair with the existing peripheral A&O woodlands and these ten were sampled by the author. The ten core sites were sampled as part of the NFQI annual survey cycle. All sites were sampled following the standardised protocol detailed below.

Fifteen 1m^2 quadrats were sampled at even distances along a 100m transect centred on the middle of each site. Leaf litter and superficial soil was collected and sifted by shaking through a 1cm^2 litter sieve to remove the largest organic and inorganic components. The residue from each quadrat was suspended in mesh bags inside Winkler extractors in order to passively extract live invertebrates as the residue dried naturally over the course of 3 days. Winkler extractors work both through the random movement of invertebrates through the leaf litter residue and directed movement downwards as this substrate dries out over time, in each case causing invertebrates to fall into collecting pots filled with absolute ethanol suspended below the mesh bags. Winkler extractors have been shown to be ~70% efficient for recovery of Coleoptera from leaf litter over this time period (Krell et al. 2005). NFQI volunteers sorted samples from all 20 sites, with adult and larval Coleoptera removed and stored separately in absolute ethanol at -20C.

Table 4.1 Site details, including habitat and positional classifications. Each pair of sites includes one A&O woodland and one inclosure. Additionally each pair is classified as occurring in the core woodland block or as a peripheral woodland patch.

Label	Pair	Habitat	Position	Co-ordinates	Site Name
BWW	1	Ancient	Core	N50.84550 W1.69617	Berry Wood
MAW	2	Ancient	Core	N50.86767 W1.65377	Mark Ash Wood
TTW[4]	3	Ancient	Core	N50.83527 W1.48021	Tantany Wood
WWW	4	Ancient	Core	N50.84989 W1.57546	Whitley Wood
ANW[5]	5	Ancient	Peripheral	N50.91061 W1.67423	Anses Wood
BSW	6	Ancient	Peripheral	N50.94817 W1.62907	Bramshaw Wood
HLW[5]	7	Ancient	Peripheral	N50.80612 W1.61535	Hincheslea Wood
PHW	8	Ancient	Peripheral	N50.79558 W1.70199	Pigsty Hill Wood
RSW	9	Ancient	Peripheral	N50.87748 W1.72989	Red Shoot Wood
SWW	10	Ancient	Peripheral	N50.90586 W1.59178	Shaves Wood
SOI	1	Inclosure	Core	N50.84100 W1.68586	South Oakley Inclosure
HWI	2	Inclosure	Core	N50.87523 W1.65234	Highland Water Inclosure
DLI[4]	3	Inclosure	Core	N50.83923 W1.51533	Denny Lodge Inclosure
NPI	4	Inclosure	Core	N50.84746 W1.58524	New Park Plantation
SBI[5]	5	Inclosure	Peripheral	N50.91502 W1.66836	South Bentley Inclosure
BSI	6	Inclosure	Peripheral	N50.95241 W1.63637	Bramshaw Inclosure
STI[5]	7	Inclosure	Peripheral	N50.79538 W1.62963	Set Thomas Inclosure
HLI	8	Inclosure	Peripheral	N50.80901 W1.68545	Holmsley Inclosure
GLI	9	Inclosure	Peripheral	N50.87014 W1.74146	Great Linford Inclosure
BHI	10	Inclosure	Peripheral	N50.90343 W1.57463	Brockishill Inclosure

[4] Classified as a peripheral site in the study of Carpenter et al. (2012)

[5] Classified as a core site in the study of Carpenter et al. (2012)

4.2.2 DNA Extraction and Sequencing

Adult specimens were air-dried and imaged using an SLR camera on a quadrat-by-quadrat basis. The specimens from each site were combined, dried at 36°C to remove any remaining ethanol, and stored at -80°C. Specimens were then ground to a fine powder using a pestle and mortar. To maximise grinding efficiency, the equipment was cooled at -80°C, placed on dry ice and filled with liquid nitrogen immediately prior to the addition of the specimens. Specimens were ground rapidly and the powder transferred to 20ml of CTAB buffer and mixed thoroughly by inversion. The samples were then incubated at 56°C overnight and DNA extracted following an isopropanol clean-up protocol. The DNA concentration of each sample was estimated using a Qubit Fluorometer High Specificity kit. These measurements were used to determine the pooling ratio used for sequencing, such that sequencing volume was approximately proportional to biomass. TruSeq PCR-free libraries (550 bp insert kit) were constructed for each of the 20 samples. The libraries were sequenced across 2 Illumina MiSeq runs (600-cyles; Illumina MiSeq v3 chemistry) to obtain 300 bp paired-end reads. Library preparation and sequencing was undertaken at the University of Cambridge DNA Sequencing Facility, Department of Biochemistry. Library details are shown in Table 4.2.

Table 4.2 Number of adult Coleoptera collected at each site and corresponding library preparation details. Superscript number indicates whether the respective library was sequenced on the first or second MiSeq run.

Site	No. Individuals	DNA Conc$^{n.}$ (µg/ml)	Insert Size	Raw Pairs (millions)	HQ Pairs (millions)	'Mito.-like' Pairs (k)	Mitochondrial Pairs (k) (%)
ANW^{2}	139	72.9	583	1.57	1.32	172.04	27.83 (1.06)
BWW^{2}	253	54.5	568	0.56	0.45	51.42	6.90 (2.11)
HLW^{2}	180	70.6	580	2.13	1.77	164.05	17.78 (1.53)
MAW^{2}	125	64.7	596	1.43	1.10	152.45	29.95 (1.01)
WWW^{1}	230	92.6	432	2.44	2.16	186.93	25.42 (2.73)
BSW^{2}	127	47.1	588	1.16	0.94	142.04	20.69 (1.18)
PHW^{2}	166	51.0	574	1.53	1.32	133.33	18.62 (2.19)
RSW^{2}	175	87.6	579	1.30	1.01	147.27	23.96 (1.41)
SWW^{2}	85	42.3	598	1.18	0.99	122.37	18.07 (2.37)
TTW^{1}	537	120	425	3.02	2.69	266.05	40.98 (1.82)
SBI^{1}	146	499	418	16.32	15.14	886.48	74.13 (1.53)
SOI^{2}	105	33.8	559	1.12	0.90	86.04	11.38 (0.49)
STI^{2}	218	55.5	571	1.11	0.93	101.16	13.16 (1.27)
HWI^{2}	121	14.4	563	0.45	0.37	34.50	4.50 (1.42)
NPI^{2}	125	45.3	576	1.36	1.16	129.81	25.07 (1.23)
BSI^{2}	64	34.9	563	1.07	0.99	69.37	5.13 (2.17)
HLI^{1}	372	115	427	3.38	3.00	296.24	42.24 (0.52)
GLI^{2}	108	106	577	2.61	2.24	266.92	38.44 (1.41)
BHI^{2}	81	29.9	542	1.00	0.82	82.71	15.99 (1.72)
DLI^{2}	103	36.4	569	0.82	0.67	76.46	11.90 (1.95)

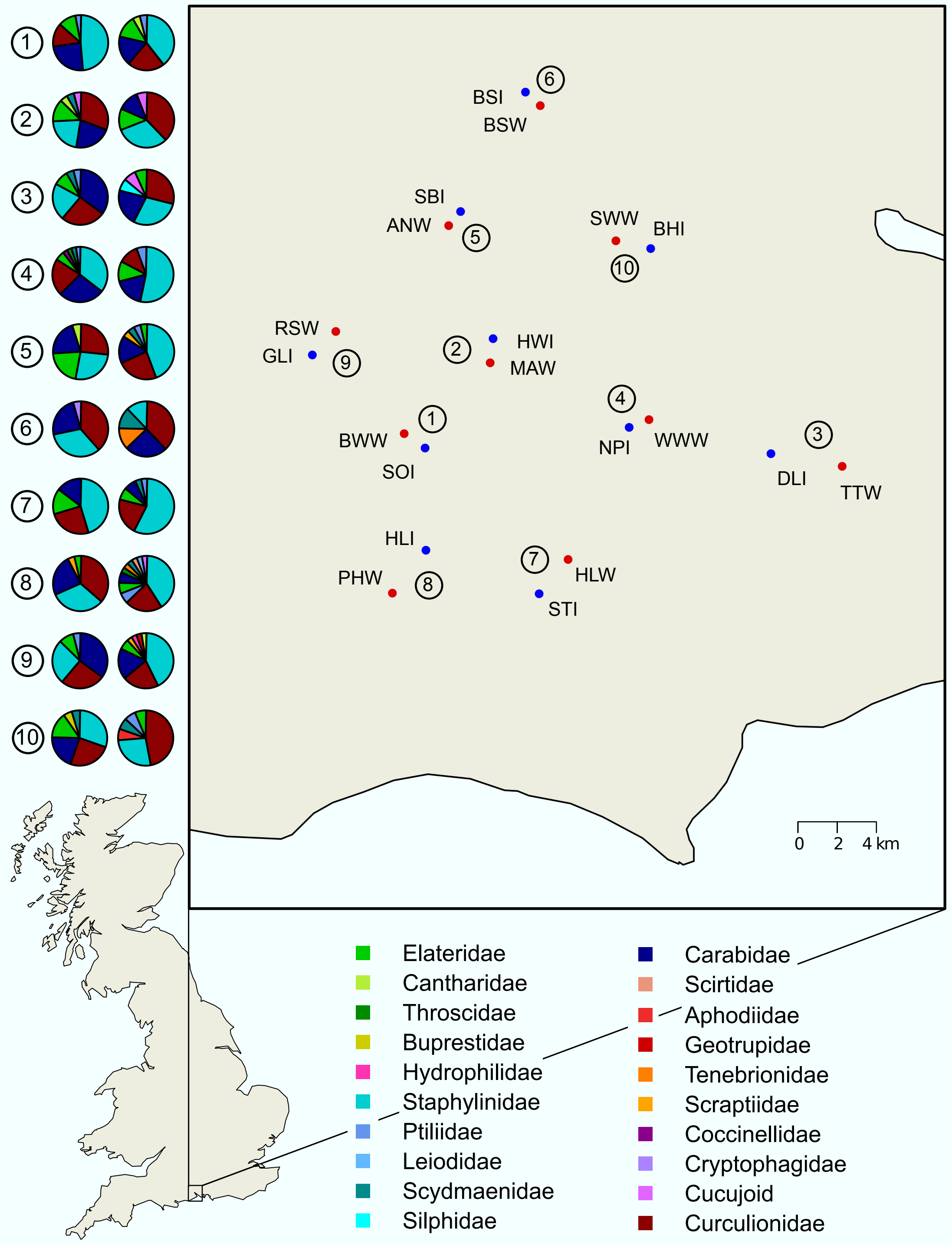


Figure 4-1 Location of sites in the present study and family-level taxonomic profiles based on the *cox1*+BOLD analysis (see Materials & Methods and Results). Each pair of sites is numbered following Table 4.1, with A&O sites as red dots and inclosure sites as blue dots.

4.2.3 Mitogenome Assembly

Mitogenome assembly was undertaken following the same steps as Chapters 2 and 3. In brief, the raw data was filtered to remove adapter sequences and low quality bases with Trimmomatic (Lohse et al. 2012) and Prinseq-lite (Schmieder and Edwards 2011) respectively. Quality-controlled reads were then filtered against a database of 245 mitogenome sequences (MitoDB; Timmermans, Barton et al. 2016) with blastn (Altschup et al. 1990) to retain 'mitochondrial-like' reads for assembly. All twenty libraries were combined for assembly to maximise the likelihood of assembling a mitogenome sequence for low biomass species. The reads were assembled with Celera Assembler (Myers et al. 2000), IDBA-UD (Peng et al. 2012) and Newbler (Margulies et al. 2005), with 98% similarity required with the latter two programs and minimum and maximum kmer lengths of 80 and 300 bp respectively for IDBA-UD. An additional IDBA-UD assembly was undertaken requiring a minimum contig length of 1 kb at each iteration (henceforth IDBA-1k). The resulting contigs from all assemblies were filtered by length to remove contigs <1 kb and then further filtered against MitoDB with blastn, requiring a minimum 1 kb hit length, to limit the inclusion of non-coleopteran sequences in the subsequent steps.

All contigs ≥15 kb from all four assemblies were checked for circularity in Geneious and trimmed as appropriate. Contigs from the three initial assemblies were then merged following the same procedure as Chapter 3, with IDBA-1k contigs ≥5 kb added to this non-redundant set as a final step. As previously, the circularisable contigs were assembled together to remove redundancy in this set. Linear contigs ≥1 kb were mapped to the non-redundant circular set and the remaining linear contigs were assembled in two steps, initially taking contigs ≥5 kb and subsequently adding contigs 1-5 kb. The IDBA-1k contigs ≥5kb were then added to the non-redundant set in two steps. Firstly, the circular IDBA-1kb contigs were assembled with the non-redundant set to highlight any contigs that were circularisable in the IDBA-1k assembly which were linear in the non-redundant set. Linear IDBA-1k contigs ≥5 kb were then added by re-assembly for a final non-redundant set.

The quality-controlled 'mitochondrial-like' reads for each library were mapped against the IDBA contigs with SMALT (-y 0.98; Wellcome Trust Sanger Institute, Available from: https://www.sanger.ac.uk/resources/software/smalt/) and the mean insert size was estimated in each case, as in Chapter 2. This was combined with the equivalent data from Chapter 2 for an updated analysis of the effect of insert size and library type on the proportion of mitochondrial reads (logistic ANCOVA; function *glm*, family=”quasibinomial”) in R (R

Core Team 2015). The same reads were also combined and mapped against the mitochondrial contigs from each of the four assemblies and the two iterations of the non-redundant set (with and without IDBA-1k) and the mean coverage for each contig estimated with Qualimap (v2.0; García-Alcalde et al. 2012). This was plotted against contig length for each dataset in R for a visual assessment of the efficiency of the various assemblies and the efficacy of re-assembly. Lastly, the rate of species accumulation (*cox1* barcodes and contigs ≥10 kb) in successively larger subsamples was assessed following Chapter 2 (IDBA-UD assembly, increments of 100k 'mitochondrial-like' pairs).

4.2.4 Annotation, Gene Extraction and Dataset Refinement

Annotation, gene extraction and dataset refinement followed the approach used in Chapter 3. In brief, tRNA genes were annotated with COVE (Eddy and Durbin 1994), followed by BLAST-based annotation for PCGs (tblastx) and rRNAs (blastn). Putative PCG and rRNA sequences were aligned with MAFFT (Katoh and Standley 2013) and checked for erroneously included sequences. Cleaned sequences were subsequently re-aligned with transAlign (translation table 5; Bininda-Emonds 2005) and MAFFT for PCGs and rRNAs respectively and manually curated. All *cox1-5'*, *cox1-3'*, *cob*, *16S* and *12S* sequences were queried against GenBank (megablast, 98% identity) and the *cox1-5'* sequences were additionally queried against BOLD (http://www.boldsystems.org) to identify the corresponding contigs to species where possible. Checking for exactly duplicated sequences in each alignment revealed five pairs of sequences requiring further investigation. This led to the removal of two short contigs and two pairs of contigs were merged to form longer sequences. In the fifth case the overlapping region in the rRNAs was almost identical between the two sequences but the PCGs were divergent (92% pairwise identity). Both contigs were identified as *Barypeithes pellucidus* via the barcode region on BOLD and were retained for the next step.

The alignments for the three most frequently recovered genes (*cox1*, *nad5*, *nad4*) were checked to find the 100 bp region with maximal overlap between contigs. A region within *nad4* was found to retain 99 contigs while the optimal *nad5*, *cox1-5'* and *cox1-3'* regions retained 92, 88 and 89 respectively. Two parallel approaches were therefore taken to generating a community phylogeny and assemblage profiles, one using the maximal *de novo* contig set centred on *nad5*, and one centred on the *cox1* barcode incorporating both the assembled contigs and sequences downloaded from BOLD (see Assemblage Profiling). The two gene-centred alignments were curated via visual assessment of phylogenetic trees including these sequences and the eight neuropteran outgroups to ensure only one contig was retained per species. These trees were generated with RAxML (Stamatakis 2014) following the procedure in Chapter 3 (no minimum number of loci, 1RY2 coding, partitioned by gene and position, RAxML: -f a -N 100) and contigs separated by short branch lengths were checked by reassembly in Geneious (Biomatters 2013). For any pair of contigs with ≥98% identity in the relevant region (*cox1* or *nad4*) the most complete contig was retained in all cases except one where the two overlapped at both ends and were thus collapsed to form a new circular contig.

4.2.5 Phylogeny Reconstruction

The curated alignments were then re-aligned with the equivalent data from an appropriate set of superbarcodes, using transAlign and MAFFT. The superbarcode set included the expanded MitoDB sequences from Chapter 3 (exMitoDB) and the two UK datasets available from Chapter 2 (*UK-BI* and *RichmondPark*, identified to species where possible via *cox1-5'*, *cox1-3'*, *cob*, and *16S* against GenBank and BOLD). Where more than one superbarcode was available for the same species the most complete sequence was used. Final superbarcode-inclusive tree topologies were then generated for each gene-centred dataset with RAxML (no minimum number of loci, 1RY2 coding, partitioned by gene and position, RAxML: -f a -x 100) and inspected for short branch lengths between *de novo* contigs and superbarcodes which might allow the transfer of species level identifications from the latter to the former. Where species level identifications were not possible, higher-level classifications were made for the *de novo* contigs based on monophyly with identified superbarcodes (following Chapter 3). The *cox1*-centred topology was further used as a binary backbone for the addition of BOLD barcode sequences identified as being present in the samples (see Assemblage Profiling; all nucleotides, partitioned by gene and position, RAxML: -r -f a -N 100 -m GTRCAT). The three trees were pruned in R to retain one tip per species and remove all superbarcodes and outgroup sequences. The branch lengths of the resulting community phylogenies were re-estimated for the included sequences (all nucleotides, partitioned by gene and position, RAxML -f e -t -m GTRCAT).

4.2.6 Assemblage Profiling

For assemblage profiling the quality-controlled 'mitochondrial-like' reads were queried independently with megablast against a database of all curated protein-coding gene sequences (-perc_identity 98) and another of all coleopteran *cox1* barcode sequences downloaded from BOLD (-perc_identity 99; 171,501 sequences, downloaded 20th August

2015). In both BLAST searches the longest hit ≥100 bp was retained for each read. The total number of accepted hits for each sequence was then collated for each site. At least 1% of the total number of hits for each contig were required to accept it as present at any given site, following Gómez-Rodríguez et al. (2015). For the BOLD sequences, hits accruing to different sequences with the same morphological identification were collapsed and a minimum of two reads per site for each species was required to accept it as present. The contig-based assemblage profile was further filtered to retain only those contigs in the *nad4* and *cox1* centred datasets and these were used to generate three alternative species presence-absence tables (*nad4*, *cox1*, *cox1*+BOLD). For *cox1*+BOLD, *cox1* contig results were combined with those from the barcode sequences (for maximally inclusive assemblage profiles) after adding the latter to the *cox1*-centred phylogeny and collapsing contig and barcode profiles with the same morphological identification and zero branch lengths. The addition of the barcode sequences to the phylogeny highlighted a small number of cases where published barcodes with different morphological identifications were nearly identical at the sequence level, producing ambiguous identification results when queried against BOLD. In these cases one sequence was selected for retention in the tree (preference was given to the *de novo* contigs) and identification was made to genus only.

4.2.7 Ecological Analyses

All analyses were undertaken for the *nad4*, *cox1*, and *cox1*+BOLD assemblage profiles. In each case the phylogeny was pruned to retain one tip per species in the respective profiles. The phylogenies were made ultrametric by penalised likelihood (Sanderson 2002) using the *chronopl* function in R (package *ape*; Paradis et al. 2004), with the optimal value of lambda selected by cross-validation. Note that for these analyses the assignation of several site pairs as 'core' or 'peripheral' has been adjusted with respect to the initial sampling design such that there are four core pairs and six peripheral pairs (see Discussion). Analyses were undertaken in R using packages *vegan* (Oksanen et al. 2015), *ape*, *picante* (Kembel et al. 2010), and *betapart* (Baselga and Orme 2012) unless otherwise stated.

4.2.7.1 Alpha Diversity

Alpha diversity was measured simply as the species diversity at each site (analogous to species density), based on the presence-absence matrices (function *specnumber*). Phylogenetic diversity (PD) was estimated from the community matrices and ultrametric trees (function *pd*). PD is expected to correlate positively with species diversity and therefore may be misleading. This correlation was tested with Pearson's product-moment correlation co-efficient (function *cor.test*) and rarefied PD was estimated for all sites by limiting species diversity to that of the least rich community (function *phylorare*; subsample by species; Nipperess and Matsen 2013). Differences in the mean diversity observed in the two habitats (Ancient vs Inclosure) and in core and peripheral sites were assessed with t-tests (function *t.test*). Equivalence between the results obtained by the three datasets was assessed with Pearson's product-moment correlation coefficient (function *cor.test*).

4.2.7.2 Beta Diversity

Compositional dissimilarity between sites was estimated with the Sørensen index and decomposed to differentiate between turnover and nestedness (Baselga 2010; function *beta.pair*). Analogous phylobetadiversity estimates were made using the *1-Phylosor* index which can also be reduced to its turnover and nestedness components (function *phylo.beta.sor*). Values of both indexes range from 0 (complete identity) to 1 (complete dissimilarity). Mantel tests were used to check for an effect of habitat or position on the change in composition between sites (function *mantel*). The significance of the correlation between distance matrices generated from each of the three datasets was also assessed with Mantel tests. Multi-site compositional and phylogenetic beta diversity were similarly computed within and between compartments (function *beta.multi*; function *phylo.beta.multi*).

For an assessment of phylogenetic community structure, species were classified as occurring exclusively in ancient or inclosure sites or in both and the phylogenetic diversity and clustering of these groupings were tested by comparing observed PD, mean pairwise distance (MPD, analogous to −1(NRI)) and mean nearest taxon distance (MNTD, analogous to −1(NTI)) to a null model based on community randomisations (functions *ses.pd*, *ses.mpd*, *ses.mntd* respectively; independent swap, 999 randomisations).

4.3 Results

4.3.1 Sequencing and Mitogenome Assembly

The total number of adult Coleoptera recovered from each of the twenty sites is listed in Table 4.2, alongside information for the corresponding library prepared from the total DNA thereof. Although all samples were prepared as TruSeq PCR-free libraries a clear difference was observed between the mean insert sizes of the libraries in the first and second runs (Figure 4-2). Combining these twenty libraries with the equivalent data from Chapter 2 does not affect the previous result, such that both insert size and library have a significant effect

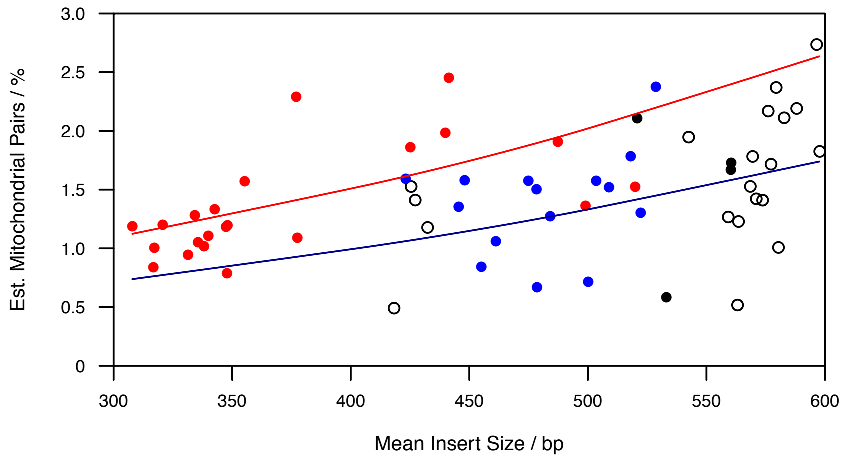


Figure 4-2 Updated assessment of the effect of insert size and library type on the estimated percentage of mitochondrial reads obtained. Libraries added in this study are shown as open circles. Note that the four libraries sequenced on the first MiSeq run have much a shorter insert size than those sequenced on the second run. TruSeq (red); TruSeq Nano (blue); TruSeq PCR-free (black).

on the proportion of mitochondrial reads, with the response curve for TSP and TSN libraries combined significantly different from that for TS libraries ($F_{2,60}$=11.15, p<0.001).

The number of mitochondrial contigs in each of four length categories for the four assemblies and the two iterations of the non-redundant set are summarised in Table 4.3. Of the three initial assembles (CA, IDBA, NWBL), CA assembled the greater number and proportion of long (≥10 kb) and circularised contigs. In contrast with the Chapter 3, CA also assembled the fewest short contigs (<5 kb) and the initial non-redundant set included more contigs than two of the three component assemblies (CA and NWBL). The IDBA-1k assembly alone recovered the same number of ≥15 kb contigs as were included in the initial non-redundant set and more were circularised, however the number of short contigs was greatly reduced. Combining this with the initial non-redundant set increased the number of long and circularised contigs and reduced the number of short contigs. The differences in assembler behaviour are also visible from the plots of contig length against mean coverage shown in Figure 4.3. Of the three initial assemblies, the number of short high coverage contigs is lowest in CA, suggesting that this program has dealt with the variability in the datasets more successfully than either of the other two. There is a striking difference between the two IDBA assemblies, with IDBA-1k assembling ≥15 kb contigs in all but 3 cases where mean coverage was >30x and exhibiting a pattern similar to that seen in the non-redundant sets. The plot for the initial non-redundant set shows a clear improvement over the three initial assemblies with an overall reduction in short high-coverage contigs. This was improved further with the addition of the IDBA-1k set, with only four contigs <15 kb with mean coverage >20x. These four were subsequently found to overlap identically in the gene alignments and were collapsed to form two nearly complete contigs. Thus the final non-redundant set appears to fully optimise contig length with respect to mean coverage for the current level of sequencing.

To further assess assembly completion, the number of *cox1* barcode and long (≥10 kb) contigs assembled by IDBA-UD (default --min_contig) was tracked with increasingly large subsamples of input reads (Figure 4.4). The results for the two markers are similar and in both cases the rate of accumulation is slow with several step-wise increases in the numbers recovered followed by stable recovery, indicating that each additional gain in species recovery requires a significant increase in sequencing effort.

Table 4.3 *NewForest* assembly results in four size classes. Includes the initial non-redundant set generated from the three standard assemblies and the additional IDBA assembly and final non-redundant set.

Assembly	1-5 kb	5-10 kb	10-15 kb	≥15 kb (circular)
CA	185	24	11	42 (23)
IDBA	232	29	11	37 (17)
NWBL	203	13	19	27 (13)
NR v1	214	16	8	50 (29)
IDBA_1k	92	11	4	50 (33)
NR v2	199	17	7	55 (38)

4.3.2 Phylogenies

The *cox1* and *nad4*-centred datasets each contained 88 contigs, of which 61 were in both datasets. The *cox1* phylogeny included just 203 superbarcodes, compared with 350 in the *nad4* analysis due to the absence of the barcode region in many of the exMitoDB sequences. For the sequences common between the two trees (252 in total), the tree topologies were largely congruent (RF = 56 of 498) and recovered the same relationships between the four suborders, with Myxophaga as the basal coleopteran branch and Adephaga sister to (Archostemata+Polyphaga). Contig placement in both trees was consistent with identifications made based on external databases (GenBank, BOLD) in all cases and when the BOLD barcodes (110 in the first instance, 17 in the final tree) were added to the *cox1* topology their placement was consistent with their identifications except for two Cryptophaginae (Cucujoidea *sensu strico*) sequences which were placed as sister to Ciidae (Tenebrionoidea) and basal to all Curculionoidea (Figure 4.3). Both trees were similar to that recovered by the 1RY2 analysis with the greatest taxon sampling in Chapter 3, although

Elateroidea was placed as sister to (Byrrhoidea+Buprestoidea). In both trees Scarabaeoidea and Staphylinoidea were polyphyletic, with Passalidae placed as sister to Histeroidea or Hydrophiloidea and Ptiliidae placed as sister to Passalidae or Histeroidea (*cox1* and *nad4* respectively in each case). The remaining differences between the two topologies at the superfamily level were the recovery of three rather than two Cucujoidea lineages and the paraphyly of Byrrhoidea with Dascilloidea by *nad4*, in both cases resulting from the placement of a single superbarcode not present in the *cox1* dataset; and the placement of Cleroidea at the base of Cucujiformia by *nad4* in comparison with a sister relationship between (Cleroidea+Tenebrionoidea) and the rest of the cucujiform lineages by *cox1*. Finally, *nad4* recovered both Curculionoidea and Chrysomeloidea as monophyletic and as sister lineages, forming the clade 'Phytophaga', whereas in the *cox1* topology Cerambycidae formed a clade with the cucujid lineages, making Chrysomeloidea polyphyletic.

4.3.3 Compositional Diversity

4.3.3.1 Alpha Diversity

In the following sections the statistics presented in the text are for the *cox1*+BOLD dataset unless otherwise stated. Results for *cox1* and *nad4* can be found in Table 9.1:Table 9.4 but in all cases, except where otherwise indicated in the text, the results were consistent between all three datasets. Species richness per site is shown in Figure 4.4 split by habitat (left panel) and by habitat and position (core vs. peripheral, right panel). These indicate that the range of species richness is lower in the ancient woodlands than the inclosure woodlands and that this is consistent between the core and peripheral plots. Overall mean species richness is higher in the ancient woodlands (μ=24.0 vs. μ=21.1) but not significantly so (t=0.92, d.f.=18, p=0.369). No significant differences were observed between core and peripheral woodlands either overall or within habitat types. However, species richness was strongly correlated with the number of individuals per site (t=4.94, d.f.=18, 0<0.001, r=0.759). Species richness per site is shown in Table 4.4 and in all cases these results were strongly correlated between the three datasets (vs. *cox1*: ρ=0.987, t=26.42, d.f.=18, p<<0.001; vs. *nad4*: ρ=0.927, t=10.47, d.f.=18, p<<0.001).

4.3.3.2 Beta Diversity

Values of total beta diversity, measured as multi-site Sørensen dissimilarity and its turnover (Simpson dissimilarity) and nestedness components, were computed from all sites within the various compartments (all sites, within habitats, within positions) and between them. Overall multi-site dissimilarity was high, with a dominant turnover component (>95%; β_{SOR}=0.883,

β_{SIM}=0.843, β_{SNE}=0.040; Table 4.5). Within the various compartments total dissimilarity was higher for inclosures and peripheral sites as compared with ancient woodlands and core sites respectively, with turnover again the dominant component. Overall when considered at the compartment level, beta diversity was much lower between compartments (Table 4.6) than multi-site beta within compartments (Table 4.5), indicating that the same species were encountered in both habitats and both core and peripheral plots, with slightly higher differences between the latter than the former. Turnover was the dominant component explaining differences between the habitats (>90%; β_{SOR}=0.291; β_{SIM}=0.273; β_{SNE}=0.018) but both turnover and nestedness were important between core and peripheral plots (β_{SOR}=0.316; β_{SIM}=0.197; β_{SNE}=0.119) (Table 4.6). Mantel tests showed that pairwise dissimilarity was significantly greater between sites in different habitats than between sites in the same habitat for both total beta diversity and turnover in all cases, although the amount of variance explained was low (β_{SNE}: r=0.134, p=0.007; β_{SIM}: r=0.116, p=0.023). In contrast, no effect of site position was found. The dissimilarity matrices obtained for each dataset were highly correlated in pairwise Mantel tests (r>0.89, p=0.001 in all cases).

4.3.4 Phylogenetic Diversity

4.3.4.1 Phylo-alpha Diversity

Phylogenetic diversity was significantly correlated with species density (r=0.975, t=18.81, d.f.=18, p<<0.001) but after rarefaction based on the lowest observed species density in each case no correlation was observed (r=-0.070, t=-0.30, d.f.=18, p=0.769). No significant differences in mean phylogenetic diversity were observed between habitats or core and peripheral plots, or between core and peripheral plots within habitats, either before or after rarefaction (Figure 4.4). Raw and rarefied PD per site are shown in Table 4.4. Both phylogenetic diversity (vs. *cox1*: r=0.973, t=18.01, d.f.=18, p<<0.001; vs. *nad4*: r=0.934, t=11.11, d.f=18, p<<0.001) and rarefied PD (vs. *cox1*: r=0.506, t=0.25, d.f.=18, p=0.023; vs. *nad4*: r=0.945, t=12.27, d.f.=18, p<<0.001) estimates were strongly correlated between the three datasets.

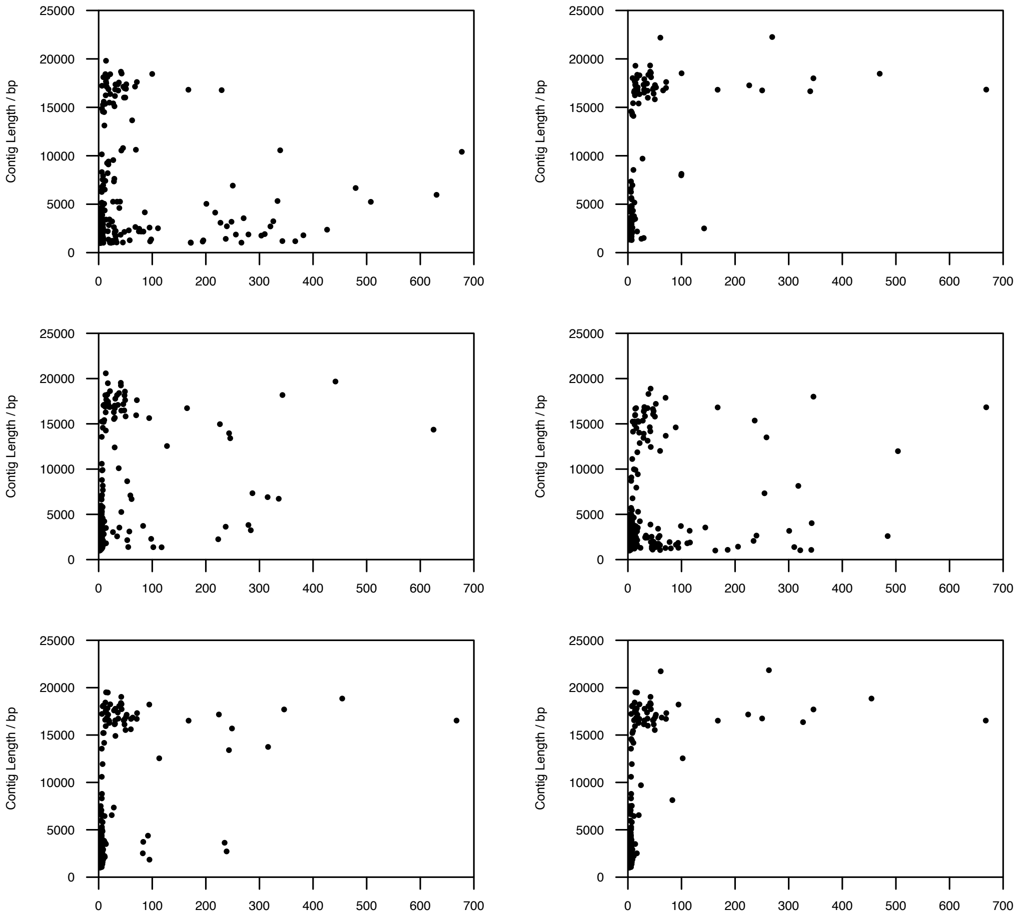


Figure 4.3 Coverage plots for each assembly and two iterations of the non-redundant set: a) IDBA; b) IDBA-1k; c) CA; d) NWBL; e) initial NR set; f) final NR set.

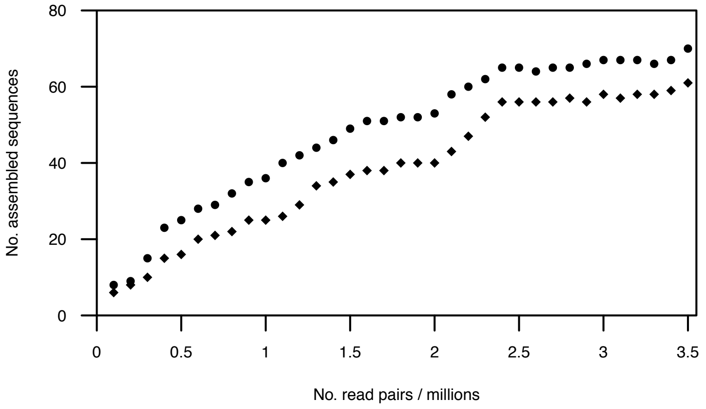


Figure 4.4 Sequence accumulation for cox1 (dots) and long contigs (diamonds) with IDBA assemblies of subsets of the NewForest data.

4.3.4.2 Phylo-beta Diversity

Following the results for compositional beta diversity, multi-site phylogenetic dissimilarity was high overall ($p\beta_{SOR}$=0.836, $p\beta_{SIM}$=0.776, $p\beta_{SNE}$=0.059) in all three datasets (Table 4.5; Table 9.2), with turnover the dominant component (>92%). Multi-site dissimilarity and the proportional contribution of nestedness were slightly higher in inclosure and peripheral sites than in ancient woodlands or core sites. In all cases, phylogenetic dissimilarity and the proportional contribution of turnover were slightly lower than for the corresponding values of compositional dissimilarity, possibly reflecting a tendency for changes to occur at the tips level rather than between deeper lineages. Differences between compartments were lower, following the compositional results, with a greater role of nestedness explaining dissimilarity between core and peripheral plots than between habitats (Table 4.6). Following the compositional results, Mantel tests for the effect of habitat or positional turnover on phylogenetic dissimilarity indicated a slightly significant effect of habitat difference on total beta diversity and turnover in all cases ($p\beta_{SOR}$: r=0.127, p=0.004; $p\beta_{SIM}$: r=0.139, p=0.012), but no effect of position. The phylogenetic dissimilarity matrices obtained for each dataset were found to be highly correlated in pairwise Mantel tests (r>0.84, p=0.001 in all cases).

The phylogenies were also used to assess the extent to which habitat associations were non-random. For this each species was classified as exclusive to one habitat or neither (Figure 4-5) and the significance of the observed distribution across the tree was tested using standard effect sizes (SES) of the measured parameters. The results were somewhat inconsistent between the three datasets, presumably reflecting differences both in the assemblage profiles obtained and the branch lengths of the phylogenies. When viewing the pattern of these associations between the various tree topologies (Figure 9.1) there is a consistent cluster of Carabidae that are exclusively found in ancient woodlands in all three trees. In contrast there is just one carabid exclusive to inclosures, again in all trees. While ancient-exclusive species are likely to be carabids, inclosure-exclusive species are likely to be drawn from Staphyliniformia (*cox1*+BOLD). Importantly, the identity of the species that are habitat-specific is maintained between the datasets where this could be verified from species-level identifications. From the various SES results (Table 4.7) there is little consistent significant evidence of non-random phylogenetic structure between the two sets of species. For ancient-exclusive species the values of PD_{SES}, MPD_{SES} and $MNTD_{SES}$ are negative in the *cox1* tree (clustering) and positive in the *cox1*+BOLD tree (overdispersion), possibly resulting from a stochastic effect of reduced taxon sampling in the *cox1* tree.

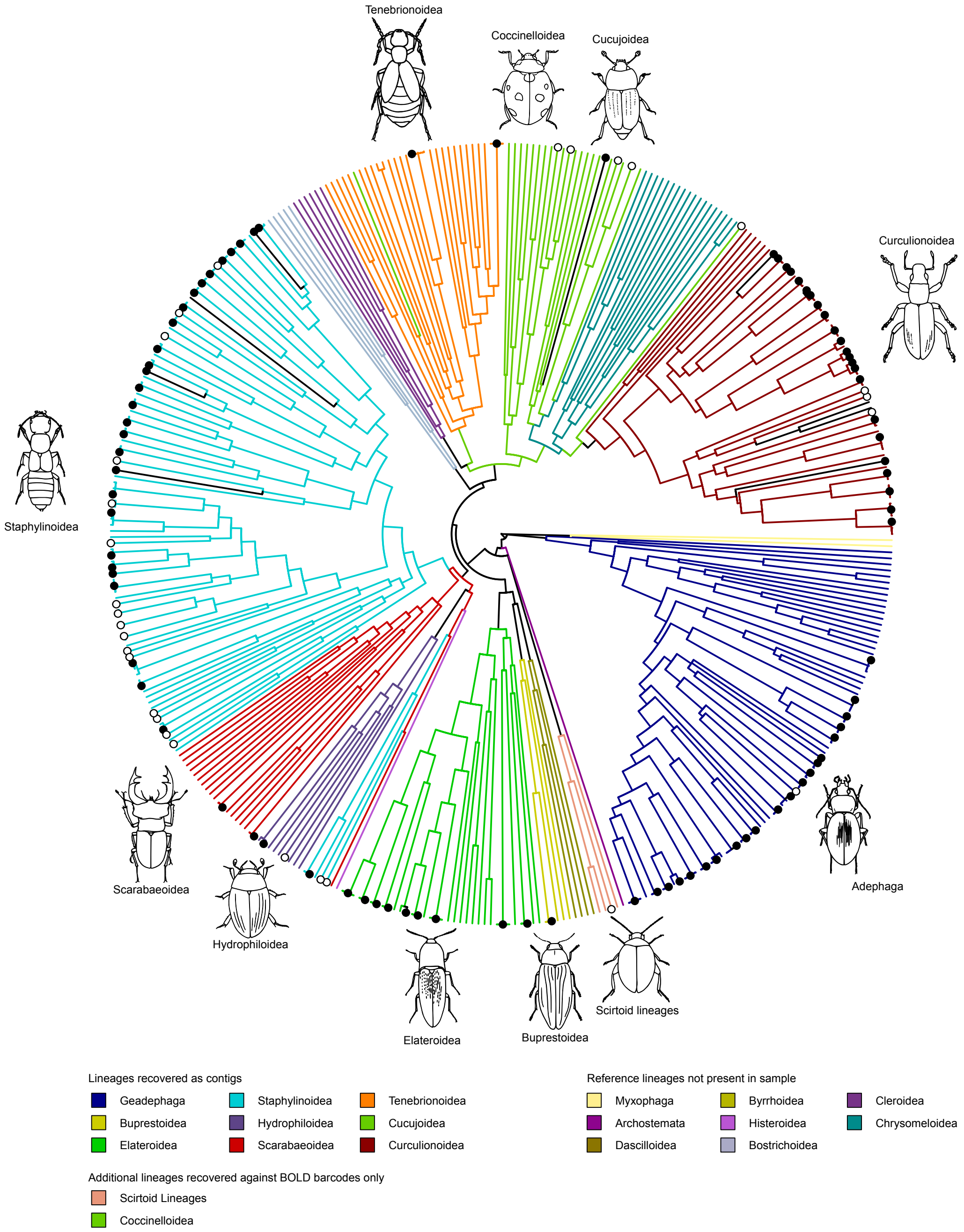


Figure 4.3 Beetle mitochondrial phylogeny including reference sequences, *NewForest cox1*-centred contigs, and BOLD barcodes for species found to be present. Two superfamilies were recovered against BOLD only. Filled circles indicate species represented by *NewForest* contigs, open circles indicate species recovered against BOLD only. Coloured tips indicate identified contigs, black tips indicate unidentified contigs.

In each case only the MPD_{SES} value is significant, indicating that the differences observed mainly derive from changes deep in the tree rather than at the tips. Close inspection of the two trees indicates that this is possibly due to the addition of two novel lineages with relatively deep divergences in the *cox1*+BOLD tree that are not present in the *cox1* tree (Cryptophagidae and Coccinellidae). In contrast, all values but one (MPD_{SES}, *cox1*+BOLD) were positive for inclosure-exclusive species across all three datasets, possibly indicating that the observed pattern was less sensitive to taxon sampling. However, in the two smaller datasets (*cox1* and *nad4*) values of PD_{SES} and $MNTD_{SES}$ are positive and significant whereas they are positive but non-significant in the *cox1*+BOLD dataset, in which the previously positive but non-significant values of MPD_{SES} became significant and negative. This supports the observation that clustering appears much greater in the *cox1*+BOLD tree for the inclosure-only species and the concentration of these species in the Staphylinidae. In all cases observed values are negative for the species occurring in both habitats, with PD_{SES} and $MTND_{SES}$ consistently significant whereas MPD_{SES} is only significant for the *nad4* tree. These results suggest that PD is relatively low for these species and they tend to be clustered

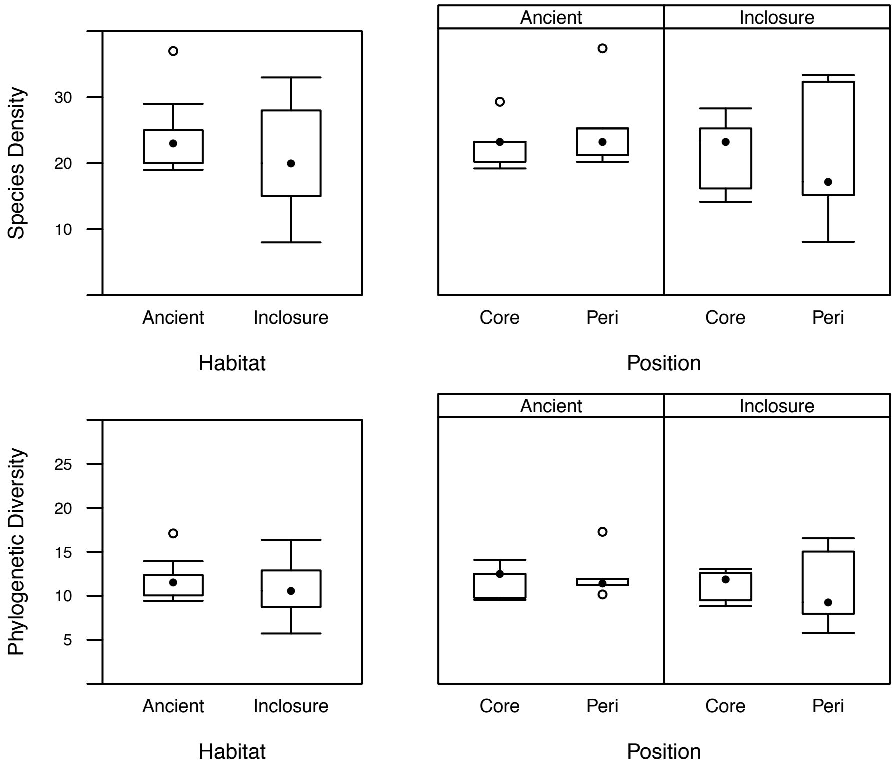


Figure 4.4 Species richness and phylogenetic diversity in the *cox1*+BOLD analysis. Top panel: Species density per site by habitat (left) and by habitat and position (right). Bottom panel: Phylogenetic diversity per site by habitat (left) and by habitat and position (right). No significant differences observed.

towards the tips of the tree. Clustering deeper in the tree is reduced in the *cox1* trees presumably as a result of the differential recovery of several lineages between the two genes.

4.4 Discussion

4.4.1 Compositional Diversity Patterns

The leaf litter collected from these twenty woodland sites across the New Forest produced a large number of invertebrates, totalling approximately 30,700 individuals in 25 invertebrate orders (Paul Eggleton, personal communication). Herein the focus was on one of the most abundant groups, Coleoptera (3379 adults, ~1100 larvae not included), although the same analyses could equally have been applied to the raw bulk samples without sorting for an analysis of total invertebrate diversity, albeit requiring much greater sequencing effort. A minimum of 88 species were represented by the assembled contigs (between the *cox1* and *nad4* datasets) and this increased to 102 when the analysis was expanded to include sequences available on BOLD. In all analyses the three datasets performed similarly, recovering the same patterns in almost all cases and estimated diversity measures were significantly correlated between them. The equivalence of the different datasets will be discussed further below (see Landscape Ecology and MMG), however in the present section the discussion will be confined to the results obtained from *cox1*+BOLD, the most inclusive community matrix.

The patchy distribution of woodland sites throughout the New Forest within a diverse matrix of open habitats may confound pure habitat effects with those of patch size and isolation. There is a distinct spatial structure in woodland habitats even at this small scale, with a central belt of continuous canopy cover surrounded by open habitats within which there are 'satellite' patches of woodland. Woodland habitats of either type within the central belt may be expected to be more homogenous in species composition than otherwise expected and exhibit increased species richness due to the greater connectivity of these sites. In contrast, isolation of peripheral sites may limit immigration and over time stochastic changes within each patch may lead to divergent community composition and lower species richness. The positional categories used herein differ somewhat from those of Carpenter et al. (2012) in which the classification of woodlands sites is based on their respective 'parcels' rather than being a proxy for habitat continuity. Thus for the present study TTW and DLI have been reclassified as "core" sites whereas ANW, SBI, HLW and STI have been reclassified as "peripheral" sites, leading to a slightly unbalanced design.

Table 4.4 Species richness, phylogenetic diversity and rarefied phylogenetic diversity for each site.

	Ancient	SR	PD	PD$_{rare}$	Inclosure	SR	PD	PD$_{rare}$
Core	BWW	29	13.9	5.4	SOI	23	11.8	5.5
	MAW	23	12.4	5.6	HWI	16	9.4	5.4
	TTW	37	17.1	5.5	DLI	17	9.2	5.3
	WWW	23	12.4	5.7	NPI	14	8.7	5.6
Peripheral	ANW	19	9.7	5.4	SBI	25	12.5	5.2
	BSW	21	10.0	5.1	BSI	8	5.7	5.7
	HLW	20	9.4	5.0	STI	28	12.9	5.0
	PHW	25	11.8	5.2	HLI	32	16.4	5.4
	RSW	23	12.3	5.4	GLI	33	14.9	5.4
	SWW	20	11.1	5.6	BHI	15	7.9	4.9

Table 4.5 Multi-site compositional and phylogenetic beta diversity from the *cox1*+BOLD analysis. Values shown for total beta diversity (β_{SOR}) and its components, turnover (β_{SIM}), and nestedness (β_{SNE}).

	Multi-site beta			Multi-site phylo-beta		
	β_{SOR}	β_{SIM}	β_{SNE}	$p\beta_{SOR}$	$p\beta_{SOR}$	$p\beta_{SNE}$
Total	0.88	0.84	0.04	0.84	0.78	0.06
Ancient	0.79	0.74	0.04	0.70	0.63	0.07
Inclosure	0.83	0.76	0.07	0.75	0.65	0.10
Core	0.80	0.75	0.04	0.7	0.64	0.06
Peripheral	0.83	0.75	0.07	0.76	0.66	0.10

Table 4.6 Compositional and phylogenetic beta diversity between habitat and positional compartments. Values shown for total beta diversity (β_{SOR}) and its components, turnover (β_{SIM}), and nestedness (β_{SNE}).

	Compositional			Phylogenetic		
	β_{SOR}	β_{SIM}	β_{SNE}	$p\beta_{SOR}$	$p\beta_{SIM}$	$p\beta_{SNE}$
Habitat	0.29	0.27	0.02	0.29	0.27	0.02
Position	0.32	0.20	0.12	0.29	0.18	0.12

Table 4.7 Standardised effect sizes for measures of phylogenetic community structure in the *cox1*+BOLD analysis. Phylogenetic diversity (PD); mean pairwise distance (MPD); mean nearest taxon distance (MNTD). Significant positive values indicate overdispersion; significant negative values indicate clustering. Significant results are highlighted in bold.

	PD$_{SES}$		MPD$_{SES}$		MNTD$_{SES}$	
	z	p	z	p	z	p
Ancient	1.26	0.90	1.70	**0.99**	1.06	0.85
Inclosure	0.48	0.69	-2.33	**0.02**	0.64	0.74
Both	-4.23	**0.001**	-1.04	0.16	-3.86	**0.001**

4.4 Discussion

In the present study no differences were observed in alpha diversity between the two habitat types or between core and peripheral plots. Variation in richness appeared greater for inclosure samples than A&O woodlands although any such difference was not found to be significant. This contrasts starkly with the findings of Carpenter et al. (2012), wherein core A&O sites were significantly richer than either core inclosure sites or peripheral A&O woodlands. In spite of this, the latter study found relatively low turnover between the three woodland habitats analysed (~0.3) and marginally significant differences in pairwise turnover within habitats. Here, dissimilarity in community composition between the habitats was similar, with turnover by far the dominant component rather than nestedness. Thus, when taken at the landscape level, the community composition of these two habitats is very similar with the majority of species shared between them. The observed differences are mainly due to species which are confined to one habitat or the other, rather than resulting from species loss in one relative to the other. When comparing core and peripheral plots the overall difference is similar to that between habitats, however nestedness plays a more important role, possibly indicating a loss of species in the more isolated peripheral plots. When viewed at the local level (i.e. considering the compositions of individual sites) total dissimilarity and turnover values were very high both overall and within each habitat, indicating that species in each habitat pool are patchily distributed throughout the landscape leading to a high level of species replacement between any pair of sites. As seen by Carpenter et al. (2012), these pairwise dissimilarities were significantly different between the two habitats for both total dissimilarity and the turnover component, with inclosure sites tending to be more dissimilar from one another than A&O sites. This shows that inclosure sites tend to be more distinct from one another than A&O sites, and may imply that the contribution of novel species from subsequent inclosure sites might be greater than from subsequent A&O sites. This finding is likely to be related to the greater variability in understory vegetation in inclosures, in terms of availability, structure, and floristic composition.

As outlined above, the results obtained in the present study differ slightly from those obtained by Carpenter et al. (2012). While the beta diversity results are similar between the two studies, the non-significant differences in alpha diversity obtained in the present study as a result of the lower observed species diversity in core A&O sites (*sensu* Carpenter et al. (2012); median 22 c.f. 39) are problematic. Whether these differences are attributable to incompleteness of the assemblage profiles obtained by bulk MMG or temporal stochasticity in the diversity and composition of the sampled communities cannot be determined directly from this dataset. The consistency in the results between the different datasets (variable

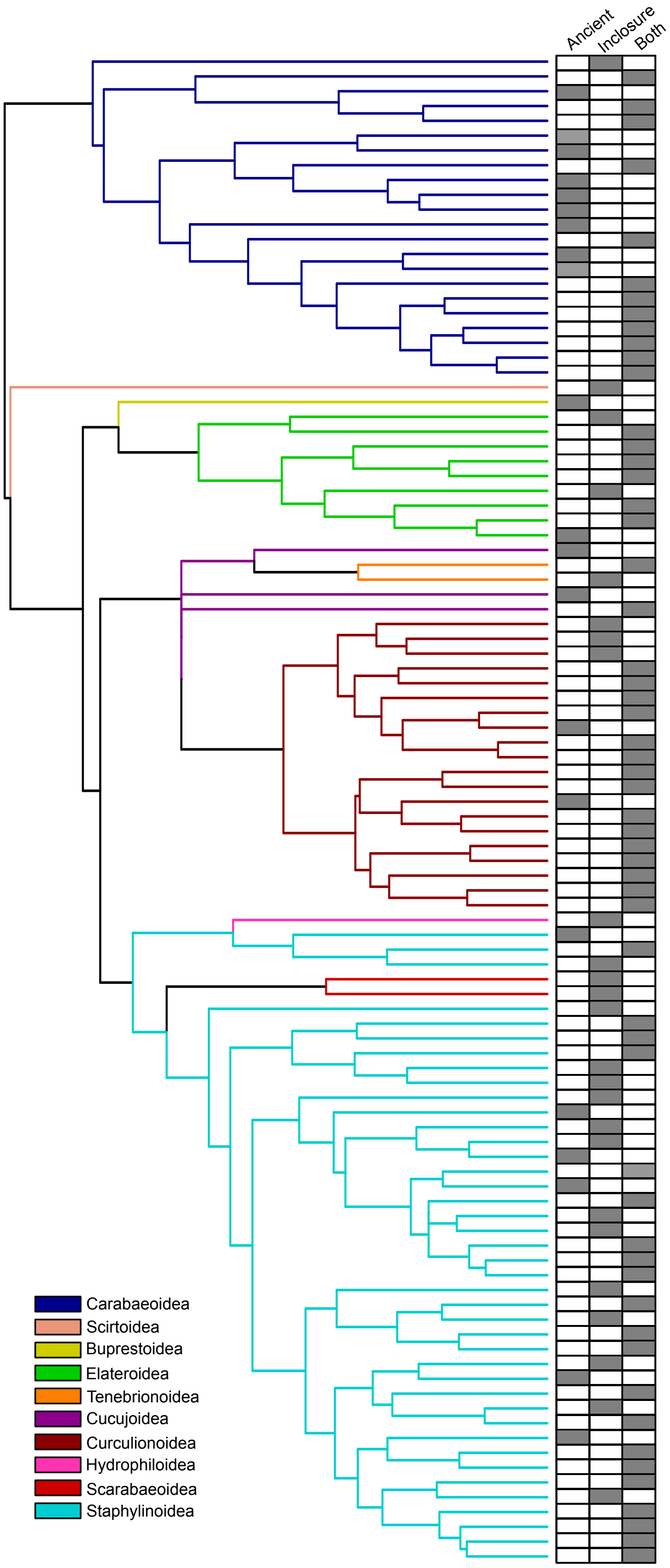


Figure 4-5 Community phylogeny based on the *cox1*+BOLD analysis, indicating species found only in one habitat or both

levels of completion) and the detection of similar beta diversity patterns as Carpenter et al. (2012) may suggest that undersampling is not major problem. Alternatively, the incompleteness in these datasets is so extreme that major differences in alpha diversity have not been detected whilst not hampering the recovery of the major beta diversity patterns. The fifteen sites included in both studies were sampled by Carpenter et al. in May 2010, while for the present study the ten core plots (*sensu* Carpenter et al.) were sampled by the NFQI in May 2011 and the ten peripheral plots were subsequently sampled by the author. Thus temporal turnover in the communities is likely to account for some differences between the studies and may also account for some of the differences observed within the present study.

A morphology-based point of reference is available for one of the sites, Whitley Wood (WWW), which is sampled monthly as part of an on-going long-term monitoring project (Eggleton et al. 2009). The results for this sample can be used as a benchmark to infer where the likely differences between the two studies have occurred, although this is only one sample of twenty. In May 2010 54 morphological species were identified from 1048 specimens at this site (773 *Acrotrichis* spp.), while 31 species were identified from 230 specimens in the present sample (95 *Acrotrichis* spp.), illustrating the potential for large inter-annual differences in snapshot samples taken from the same locality. Comparing the morphological results for May 2011 with those obtained from the *cox1*+BOLD dataset shows a clear bias in the MMG data against small species, particularly when occurring at low frequency. In the MMG dataset as a whole and for WWW in particular there is a lack of very small species in groups such as Scydmaenidae or Latriididae even though these are evident from the specimen images and were known from the WWW benchmark, although at low frequency. Notably, the smallest species encountered in these samples are *Acrotrichis* spp. (~1mm) which has a tendency to form aggregations and thus appear infrequently but in large numbers, increasing the likelihood of detection with MMG. In the WWW sample there were 95 *Acrotrichis* specimens but only 148 reads were recovered. In this light it is unsurprising that MMG failed to detect several small species in this sample and presumably this pattern was repeated across the landscape.

4.4.2 Phylogenetic Diversity Patterns

The unique contribution of the present study to the assessment of diversity patterns in the New Forest National Park is the analysis of phylogenetic diversity. When viewed in the context of the full coleopteran phylogeny (Figure 4.3) the majority of species encountered in this study are split between three clusters associated with the families Carabidae, Staphylinidae and Curculionidae. These families are also dominant in their contribution to total species richness at each individual site (Figure 4-1).

In general the results obtained with respect to phylogenetic diversity were consistent with the results based on species composition alone, with no significant differences in alpha diversity, low dissimilarity between the species pool in the two habitats but high dissimilarity and turnover between individual sites. More interestingly, the phylogenetic analysis gave an alternative perspective on the uniqueness of the two habitats, with over 20 species unique to each in the *cox1*+BOLD tree (Figure 4-5). Whilst this observation is in itself important, this would also have been uncovered by direct inspection of the community matrices. The visualisation of this pattern in the context of the phylogeny does however provide unique information regarding the relatedness of these species and therefore greater insight into differences between the two habitats that are not picked up by other measures. However, the variability in these results between datasets is much greater than for the other analyses and thus should be interpreted with caution pending further sampling.

The *cox1*+BOLD tree offers the most complete representation of the total diversity encountered between the twenty samples but for low biomass species (approximately frequency x size) the short length of the barcode sequences reduces the likelihood of detection relative to species for which a longer contig is available. Without the inclusion of the barcodes these species are not detected because pooled read numbers across all sites are still insufficient for assembly, thus species diversity is maximised in the *cox1*+BOLD dataset but possibly at the cost of stochastically incomplete detection of low biomass species across the landscape. This is important to bear in mind when examining the phylogenetic distribution of species that appear to be limited to one or other habitat. However, the striking phylogenetic segregation between ancient-exclusive and inclosure-exclusive species is unlikely to be completely random. More detailed investigation of the ecology of the relevant species would be required to hypothesise the causes of this pattern but this is an interesting observation to bear in mind for future studies in the New Forest. The significant overdispersion and clustering in the MPD parameter for the A&O and inclosure-specific species respectively could indicate that the factors controlling community assembly in these two habitats are different. Overdispersion is often interpreted as indicative of a strong effect of competitive interactions between co-occurring species whilst clustering may indicate habitat filtering. Such observations fit broadly with the patterns that might be predicted for these habitats based on the differences in the disturbance regime, but the current analysis will need to be confirmed with additional sampling and a focussed effort to minimise detection bias against low biomass species.

4.4.3 Landscape Ecology and MMG

Further to the few results available for TruSeq PCR-free libraries from Chapter 2, the present study contributes additional data points to further establish the relationship between insert size and the proportion of mitochondrial data obtained. Of particular note are the four libraries which were prepared for the first run, having both a shorter insert size (by approximately 100 bp) than those prepared for the second run and a lower estimated proportion of mitochondrial reads, reinforcing the findings from Chapter 2. When these samples were included with those other datasets the result seen Chapter 2 is repeated, with a significant effect of both insert size and library on mitochondrial proportion. The effect of insert size is greater for TS than either TSP or TSN libraries, although the range of observed insert sizes is reduced in the latter two whilst being longer on average. The reduced mitochondrial proportion in the four libraries with a shorter insert size in the current study may indicate that the same relationship would be observed for TSP as TS libraries if the range of insert sizes sampled increased but there are no such signs in the TSN libraries. Of the 24 TSP libraries now available, three appear to have a much lower mitochondrial proportion than would otherwise have been predicted from their insert sizes. This was not observed in the TruSeq Nano libraries in Chapter 2, for which there was a similar level of sampling, and thus may be a stochasticity exclusively associated with the TSP method, although a sample-specific effect cannot be excluded.

The New Forest samples also present the opportunity to increase the number of TSP libraries for further investigation of the relationship between insert size and assembly efficiency. However, this was not undertaken herein as it would have required the separate assembly of each library and this did not appear to be necessary after optimisation of the non-redundant set following the addition of the IDBA-1k assembly. As was previously seen in Chapter 3, the re-assembly of the three sets of raw contigs (CA, IDBA, NWBL) had a dramatic positive effect on the observed relationship between mean coverage and assembled contig length. Again, this step was not fully effective but the number of contigs that were clearly not optimal was low. The IDBA-1k assembly was added following the noted impact that this had on the *ChrysIber ChrysoAL* assembly in Chapter 2, with a small but valuable further increase in the number of long contigs and a reduction in the shortest ones. As previously, the IDBA-1k coverage plot indicated that contig length was better optimised with respect to coverage than in the three other assemblies, although possibly at the cost of a loss of additional diversity represented only by short contigs (Table 4.3), making the combination of all four assemblies worthwhile. The final plot in the series indicates that the final non-redundant set is fully optimised with respect to sequencing depth in the current samples (Figure 4.3). Given the large number of remaining short low coverage contigs it is clear that these samples have not been sequenced to a sufficient depth for maximal assembly length for all species present. However, the current study demonstrates that the apparent assembly challenge presented by deeply sequenced species can be overcome by careful re-assembly of several datasets, indicating that additional sequencing to increase contig length for superficially sequenced species should be easily accommodated.

Viewing the question of assembly completion from another angle, the rate of accumulation in the two markers in the subsampled IDBA assemblies suggests that the long contigs obtained represent a significant proportion of the true diversity of the sample (Figure 4.4). In Chapter 2 the equivalent plots for the *ChrysoAL* data showed a large divergence between the recovery rate in these two markers. This was thought to be indicative of low assembly quality when compared against the equivalent data for *ChrysoRL*. In the current case, the observed accumulation rates appear to be similar between the two markers, with the long contigs lagging only slightly behind the much shorter barcode sequences. This may indicate that assembly efficiency for the long contigs is high and thus the final dataset is likely to be largely complete. However, the shallow slope and apparent step-wise increases observed in the accumulation of both markers may mean that significant additional sequencing would have recovered a further increase in species recovery. However, from these plots and the relationship between contig length and mean coverage in the final non-redundant set it is likely that the assemblage profiles obtained are as complete as possible for the current level of sequencing.

The possibility that incomplete sampling has compromised the ecological results discussed above cannot be precluded, however the inclusion of the barcode sequences from BOLD did not have a great effect and only increased the number of species included in the analysis by 16%. Of the ~170,000 beetle sequences available on BOLD only a further 13 would have been added if the required number of matched reads had been reduced to one, and thus by definition these species were rare, occurring at a single site each. Whilst it is also likely that there are some species present in the samples that are not currently represented on BOLD and thus could only have been recovered by *de novo* assembly, the fact that they did not assemble even with the high level of assembly effort, indicates that these also are rare across the landscape and do not drive diversity patterns at the community level.

While generating the internal reference contigs from bulk MMG is inefficient compared with a voucher MMG approach, the difficulties encountered by Gómez-Rodríguez et al. (2015) do not appear to be insurmountable with a slight increase in re-assembly effort and thus where a direct bulk MMG approach is desirable for practical reasons the author does not see any intrinsic barrier to its application apart from the requirement for greater sequencing depth. In cases where a combined voucher MMG (for contig-based analyses) and low coverage bulk MMG (for read-based analyses) is contemplated it is worthwhile considering whether or not splitting the planned voucher MMG sequencing effort between the bulk samples would give approximately similar assembly results while also increasing sensitivity for biomass and (potentially) genetic diversity analyses. The answer to such a question will be dependent on a combination of the expected species richness and evenness and the intended sequencing volume. Such issues are further discussed in the final Chapter in the context of the rest of the thesis.

4.4.4 Conclusions

The present study represents the first application of MMG to study landscape-level patterns of beetle diversity. The similarity in the results between the three different datasets and the similarity between the beta diversity patterns recovered herein and those seen by Carpenter et al. (2012) are encouraging, in spite of the differences in recovered alpha diversity from the latter. While these differences are likely to be partly attributable to inter-year variation, the main current limitation for the application of bulk MMG to temperate communities appears to be the loss of low biomass species due to insufficient sequencing depth rather than problems related to incomplete assembly. The inclusion of a phylogenetic perspective generally supports the compositional results and provides a unique opportunity to reveal differences in the lineages that appear to associate with each of the two habitats. Whilst these results are very preliminary and may in part result from incompleteness in the assemblage profiles, there does appear to be some differentiation and may point to different drivers of community assembly even at this small scale and these otherwise similar communities. This demonstrates some of the potential of phylogenetic approaches to uncover differences in communities which appear similar with other metrics, and in this case highlights the fact that the pasture and inclosure woodlands are likely to support subtly different leaf litter communities and therefore both contribute to the gamma diversity of the landscape. If confirmed these results could have implications for the future management of these habitats as their distinctness would argue in favour of maintaining current differences in management strategy.

Chapter 5 Discussion

5.1 A Methodological Perspective: Current Status and Future Prospects

At the beginning of this thesis a new methodology for the study of insect biodiversity was introduced and named ‘mitochondrial metagenomics’. The subsequent Chapters have focused on exploring the limits of the current implementation to access and describe beetle diversity from mixtures of DNA, with a view to expanding this approach to simultaneously analyse all insects obtained by mass-trapping. It was hoped that a shotgun sequencing approach would eliminate the biases associated with the equivalent PCR-based metabarcoding, while the focus on the mitochondrial genome facilitated species identification with respect to existing barcode databases wherever possible, and accurate phylogenetic placement in all cases. The latter is a crucial step towards seamless integration between biodiversity discovery, species description, taxonomy, molecular systematics, ecology, and phylogeography for a truly holistic approach to the ‘problem’ of insect diversity. Such a system is clearly far from being realised, although many of the building blocks exist or are feasible with appropriate application of current technology. Here the focus has been on generating and maximally exploiting HTS data for estimates of compositional and phylogenetic diversity that are not hampered by lack of species-level descriptions of the fauna under study. The main conclusions of this work and the future prospects for MMG are further discussed below.

In Chapter 2 a wealth of existing MMG datasets for beetles were exploited for a timely assessment of the main experimental design steps which should be considered in the future, both for further work on beetles and when expanding to other insect orders. Over the time during which the various experiments have taken place, the Illumina MiSeq chemistry and the associated library preparation kits have changed. This confounded the analysis of the effect of library preparation choice on the data losses expected from downstream read-processing steps and, unsurprisingly, the newer library kits and sequencing chemistry were found to retain significantly more data. As such, the newer technology should generally be preferred for maximal cost-effectiveness. However, there was some indication that the choice of library preparation for current projects is unimportant following high quality recent TruSeq datasets. Thus library preparation choices should primarily be driven by DNA availability and insert size. The TruSeq Nano and PCR-free kits offer two standard insert size options, 350 bp and 550 bp, with the former essentially replacing the original TruSeq kit. In the experiments included in Chapter 2 the 550 bp TSN/TSP kits were used in all cases, leading to a relatively uniform insert size range within each of these, although the TSP libraries were found to have larger inserts on average. In contrast, a much greater range of insert sizes was observed in the TS libraries because in several experiments the sequencing provider was asked to aim for a longer fragment size than was standard. This provided the opportunity to test for a change in the estimated proportion of mitochondrial reads with respect to insert size in each of the library types and, surprisingly, a positive response was found. The reason for this remains unclear but implies that longer insert sizes cause a very slightly increased bias towards sequencing the mitochondrial fraction. The effect of insert size is greater in the TS libraries than for TSN/TSP while the latter two appear to exhibit greater variation in mitochondrial proportion for a given insert size than the former, making the use of insert size to maximise mitochondrial proportion possibly more reliable with TS libraries. Given the observed positive effect of increased insert size on biasing assembly towards longer contigs, maximising this parameter must be seen as beneficial even if the effect on the proportion of mitochondrial reads is negligible or not a primary concern. Based on the modelled relationship in the current analysis (both in Chapters 2 and 4) a TS library with a long insert size (e.g. 600 bp) would be expected to maximise mitochondrial proportion relative to TSN/TSP for the same insert size, whilst also maximising assembly of long contigs, however the lack of TS sampling above 550 bp would need to be addressed to confirm this.

Two important related questions are unanswerable with the current analysis but should be investigated as a matter of priority. The simplest is to determine where the optimum insert size range for mitochondrial proportion and assembly success lies. This limit certainly does not appear to have been reached within the current set of experiments. The second, and somewhat more complex question is whether the choice of insert size has a biasing effect on the species composition of the resulting reads and thus all downstream analyses. Longer insert sizes will bias against more degraded DNA and this will be a significant cause for concern where a mix of quality is expected. For ecological samples any taxonomic bias in the rate of degradation would be particularly difficult to account for, while the time of capture may also lead to variable degradation between specimens in traps that run for several days. Tang et al. (2014) did not observe any effect of DNA quality on assembly/sequencing success, however their analysis was for a short insert size (250 bp) so only significant degradation would have been likely to have a noticeable effect in this case. Further to this, the possibility that there is any intrinsic taxonomic bias not directly related to DNA degradation that is caused by this or any other aspect of experimental design needs to be assessed, although the complexity of such an experiment is likely to be prohibitive. Any bias that is detected could perhaps be mitigated against by sequencing at least two libraries with two different insert sizes in all cases, although this would clearly increase costs. The *BorneoCanopy* experiment presented in Chapter 3 is the only one for which two different insert size libraries were explicitly prepared but it is unclear whether the mix of insert sizes had an additive effect on the assembly over and above that of the equivalent amount of data for a single insert size library.

In all Chapters the same three assembly programs were applied, with Chapters 3 and 4 additionally including a reassembly step. In Chapter 2 no consistent differences between assemblers were observed on a dataset by dataset basis but overall CA tended to behave divergently from IDBA and Newbler. IDBA had a greater tendency to assemble either short or long contigs and thus the frequent assembly of a larger number of long contigs was masked when considering these as a proportion of all contigs assembled. All assemblers showed a response to insert length in three of the four size classes examined, in particular showing the opposite effect in the shortest and longest categories and indicating a significant biasing effect of insert size on assembly success, as discussed above. Whilst there was no conclusive evidence to promote the use of any one of these assemblers, the greater tendency for IDBA and CA to behave differently but in unpredictable directions for each dataset, and Newbler to variously be non-significantly different from each led to the suggestion that as a minimum both CA and IDBA assemblies should be performed where possible and the length distributions compared to assess the extent of differential success before proceeding with a single assembly. Newbler was also found to contribute the smallest proportion of unique gene sequence to the *BorneoCanopy* dataset, reinforcing the prioritisation of IDBA and CA. However, during the re-assembly step the addition of a third assembly provided additional confidence and assisted decision-making in some cases. The inclusion of a third assembly (not necessarily Newbler) is therefore generally useful, although the complexity of this step increases. Where computing resources are limiting, it should be noted that IDBA is significantly faster and more efficient than CA, particularly with respect to disk space and memory consumption, although for especially large datasets the number of CPUs required for a reasonable memory footprint may become limiting.

Performance comparisons between three assemblers used herein and the SOAP*denovo* programs favoured by Zhou and colleagues have not been made and no claim is made that any of these assemblers are the optimal current solution to MMG assembly. However, other assemblers trialled by the author have not been as successful, with attempts to use the SOAP*denovo* suite either producing comparatively very few long contigs or not running to completion due to insufficient computing resources. The only other useful assembler trialled is Ray Meta (Boisvert et al. 2012) which performed similarly to the other three assemblers for the *RichmondPark* experiment, but has not yet been applied to one of the larger datasets. Obviously this remains a highly dynamic area of research and new assemblers are published frequently. Of particular interest in the immediate future is the potential for assemblers that are able to assemble circular genomes natively, and one such program, named Org.Asm, was recently made available online prior to publication (Available from: http://pythonhosted.org/ORG.asm/). This assembler is designed primarily for genome skimming and may not produce good results from mixtures, but this remains to be tested.

In the absence of a clearly optimal assembler for MMG several studies including those presented in Chapters 3 and 4 have used a re-assembly approach whereby multiple assemblies are merged to maximise sequence contiguity. In all cases this has been highly beneficial, increasing the number of complete sequences obtained. The effect of this can be observed most clearly in Chapter 3 from the shift in the cumulative length distribution towards longer sequences and additionally in the 'before and after' plots of contig length as a function of coverage in both Chapters 3 and 4. In these it is clear that re-assembly is able to resolve many cases where high coverage has apparently hampered the assembly. The reasons for the failure to extend these contigs remain unclear but the pattern is observed in all datasets and all three assemblers. From the coverage plots for the *DeNovoRL* assembly (Chapter2; Gómez-Rodríguez et al. 2015) and *BorneoCanopy* (Chapter 3) it is clear that the re-assembly process has not been exhaustive and could perhaps be improved with the addition of further assemblies. In contrast, in Chapter 4 the second iteration of the non-redundant set appeared to be fully optimised with respect to the current level of sequencing after the initial tree-building step. The addition of the extra IDBA assembly had a small but positive effect on this dataset, indicating that this may be a useful general strategy.

At this time the need for the re-assembly is clear as it has a large impact on the length of the contigs available for phylogenetic reconstruction. However, it is hoped that the need for this will diminish with improvements in the available assembly programs. The re-assembly process as currently implemented is less replicable than the assemblies themselves, requiring manual intervention at each step to prevent the perpetuation of errors in the reassembled contigs. It also very time consuming where there are a large number of contigs to be assembled and does not identify all cases where contigs should be combined. Tang et al. (2014) presented an alternative re-assembly approach using TGICL (Pertea et al. 2003) in the first instance that may be more replicable than the method using Geneious herein, however this was still followed by manual inspection and identification of additional overlaps. Either way, this step is unsatisfactory and is a significant bottleneck in the protocol, yet at the current time no better solution has been found. For the time being, this step remains crucial and the use of a single assembly cannot be recommended on the basis of current results. The contrast between the IDBA and IDBA-1k assemblies in Chapter 2 and the effect of adding an IDBA-1k assembly in Chapter 4 suggest that this is a promising strategy to explore further, however the increased rate of long contig assembly comes at a cost of reduced short contig diversity. This is demonstrated by both the reduced rate of barcode accumulation illustrated in Figure 7.4 and the much smaller number of contigs 1-5 kb in Table 4.3. Therefore, unless sequencing depth is sufficient for all species to be assembled into longer contigs it would be detrimental to the estimation of diversity to rely on the IDBA-1k assembly alone.

Insufficient sequencing depth has been a constant theme throughout this thesis. While the failure to recover barcode sequences for all species in Chapter 2 (Figure 2.3) and the inclusion of only 87.5% of the estimated *BorneoCanopy* richness in the *nad4l*-centred analysis may at least partly reflect inefficiencies in the assembly process, the large number of short low coverage sequences in all assemblies indicates that sequencing depth has been too low in all MiSeq experiments to date. The New Forest example in Chapter 4 is particularly striking because the assembly appears to be optimal for the data in hand but there is a clear problem with insufficient sequencing depth preventing complete assembly. The number of species that are unrepresented in the resulting datasets is unknown, as is the extent of incompleteness for each of the assemblage profiles. Additional sequencing will be required to assess this further, but how much more? The slow step-wise accumulation of diversity in Figure 4.4 may indicate that representation is nearly complete, or alternatively that significant increases in sequencing volume are needed. This dataset may be a good test case for combining MMG with metabarcoding. The assembly presented here is based on two full runs of Illumina MiSeq and this level of sequencing may already be difficult to justify for many projects. The need for significantly increased sequencing for a relatively modest increase in species recovery would therefore be problematic. As an alternative, the current level of sequencing could be combined with a small amount of metabarcoding data to fill in the gaps, and these short sequences then be placed in the existing barcode-centred phylogeny.

The relatively slow rate of sequence accumulation was observed in both of the temperate systems assessed (*ChrysIber* and *NewForest*, Chapters 2 and 4), contrasting strongly with equivalent anlayses for the tropical *BorneoCanopy* dataset, which behaved more closely to the *ChrysIber ChrysoRL*. This is accounted for by the much higher species:specimen ratio in this sample, and the same pattern is likely to be repeated in other tropical samples, although this was not tested on the available *FrenchGuianaFIT* or *PanamaVane* samples. This may point to a possible temperate-tropical divide in the utility of MMG in its current form and should be further assessed to determine whether alternative strategies should be devised. A greater bulk MMG efficiency in tropical systems could be exploited to rapidly expand mitogenome sampling of previously unsequenced species, and arguably it is this context that MMG has the greatest potential for integrating the process of biodiversity discovery directly into the construction of a (mitochondrial) tree-of-life. For example, in Chapter 3 146 *8+ contigs* were assembled from ~17 Gb of raw data, all of which are completely novel sequences. In contrast, ~27 Gb of raw data in Chapter 4 produced only 64 *8+ contigs*, of which thirteen are either already published or were also recovered in one (or both) of the *UK-BI* or *RichmondPark* libraries. Whilst this duplication is a positive outcome at this early stage in the development of MMG and demonstrates the repeatability and reliability of the assembly process, continuing to reassemble the same species in many independent studies would be a waste of sequencing effort, especially if a small number of species that are frequently found at high biomass are sequenced repeatedly while continuing to fail to assemble low biomass species. It is perhaps this argument that provides the strongest call for a reference library approach based on voucher MMG for temperate systems, whereas there appears to be relatively little to gain from voucher MMG in tropical systems. However, as pointed out in Chapter 4 the lower limit for detection against a full reference library remains unknown and will be unpredictable *a priori*. This would perhaps require bulk MMG libraries to be sequenced at low coverage repeatedly until species accumulation curves from read-mapping against a reference database approach an asymptote. This would of course not bypass the problem of excessive sequencing of high biomass species, but the uncertainty regarding the number of species that are missed because they fail to assemble would be removed. Some additional experiments in this direction would be beneficial at this point as there is a real possibility that for any 'reference library plus low coverage bulk sequencing' strategy to have a satisfactorily high profiling success might require almost as much sequencing as direct *de novo* assembly.

A first test would be to assess whether the *DeNovoRL* assembly of Gómez-Rodríguez et al. (2015) could be improved upon to the point at which assembly is as optimal as that for *MitoRL*, following the approach in Chapter 4. If this were possible, a simple *in silico* experiment would be able to address the question of how much sequencing would have been required to achieve the full assemblage profile against either reference library. Once that threshold is determined, the profiling success attained against the full reference library at that threshold should be compared with the profiling success of that data volume against an optimised assembly of that same data. If there is a significant difference in the results obtained by each and the data volume required for profiling against the full reference set is greatly reduced, then this would be a strong argument for a reference library approach with low coverage sequencing. If either the profiling results or the required sequencing volumes are not significantly different then there is little to gain from the additional reference library construction step with voucher MMG.

To some extent, many of the issues highlighted in this thesis would be resolved by an effective procedure for unbiased mitochondrial enrichment. If MMG were more efficient in this respect the depth of sequencing required for a similar assembly result would be greatly reduced. Even a relatively modest enrichment from 1% to 10% would be hugely beneficial. Although this may still be insufficient for complete assembly of the lowest biomass species in the current temperate bulk MMG examples, the length of their assembled contigs should increase, maximising the likelihood of being included in the community phylogeny. The prospects for enrichment are however unclear at this time. Zhou et al. (2013) reported an enriching effect of differential centrifugation from an expected (but not measured) 0.05% to 0.5% but whether this difference is truly a result of enrichment is unclear as the reported mitochondrial proportion in all MMG studies to date has been at least 0.5% without enrichment. Differential centrifugation requires intact mitochondria and is usually performed on live or freshly killed tissue and the likelihood of recovering intact mitochondria from alcohol-preserved specimens is low. As an alternative, the author trialled the use of ultracentrifugation on genomic DNA extracts, taking advantage of the high AT-content of the insect mitochondrial genome. The results of these experiments were somewhat mixed and although some enrichment appeared to be possible with this method the increase in protocol complexity, use of non-standard laboratory equipment, and potential for unpredictably induced biases meant that this was not further pursued.

Perhaps more promising is the prospect of using hybrid capture once a sufficiently dense sampling of superbarcodes is available for a particular taxonomic group. For beetles such tests are currently underway but it is unclear how permissive this approach could be with respect to sequence divergence from the probes at such a large scale. A recent study on degraded DNA from museum specimens of Sunda colugo (*Galeopterus variegatus*, Mammalia) was able to capture sequences 10-13% divergent from the probes while maintaining high selection efficiency and genome coverage (Mason et al. 2011), suggesting that there is some scope for such an approach. However, significantly increased divergences would need to be achievable to enable minimally biased capture from bulk MMG samples with probes generated from 100-200 superbarcodes. If this were possible, voucher MMG could then be applied to other taxonomic groups to increase superbarcode sampling to the point at which effective probe sets could be generated for these as well. One could then envisage a future MMG where DNA is extracted in bulk from mixed samples with aliquots then enriched for the target group(s) of interest for data efficient multi-taxon comparisons.

Regardless of whether a hybrid-capture based enrichment is successful, the development of an unbiased solution to the problem of MMG sequencing efficiency is perhaps the most urgent problem that will need to be addressed if this method is to prove useful at a large scale. In the absence of efficient enrichment, sequencing effort will need to increase if the current success rate of the MiSeq-based bulk MMG protocol is to be improved upon. Strategies to size sort specimens for DNA extraction in multiple size classes followed by equimolar pooling of the extracts may be an attractive solution to minimise the problem of variation in biomass for *de novo* assembly for bulk MMG, as hybrid between bulk MMG and the size-sorted voucher MMG approach of Gómez-Rodríguez et al. (2015). While this is likely to improve assembly success, how this would integrate with analyses of relative biomass based on read-mapping is unclear. The extracts from the different size classes could perhaps be kept separate and the biomass results adjusted *post hoc* to reflect size class and read number, although this dramatically increases library costs.

In summary, the analyses presented in this thesis have explored a number of technical aspects of the MMG approach and some new insights with respect to both sequencing and bioinformatics strategies have been revealed. The main conclusions and recommendations that can be drawn from this work are illustrated in Table 5.1. Particularly important is the demonstration that highly complete bulk MMG assemblies can be obtained without an initial step to generate a reference library by voucher MMG. These analyses have generated new questions with important implications for the further development and utility of the approach. Some of these questions could be addressed at least in part with the datasets already available, or with some additional re-sequencing thereof. The assembly challenge in particular is likely to diminish rapidly as metagenome assemblers for Illumina data become more common. One of the great advantages of sequence-based approaches is the possibility to reanalyse the data at a later time as new bioinformatics tools become available and the analytical challenges associated with MMG should be transitory. However the major economic and technical barriers to the wider uptake of MMG are likely be difficult to address, particularly with respect to efficiency and enrichment.

Table 5.1 Recommended strategies for each step in the MMG procedure, based on results from this thesis and the author's personal experience.

	Step	Recommendation
Experimental design	**Pooling strategy**	For targeted sequencing e.g. phylogenetics or building a reference library: voucher MMG
		For biodiversity/ecological studies with mixed field-collected samples: bulk MMG with bulk DNA extraction
		For bulk MMG where have a large variation in body size, consider sorting to two or more size classes and sequencing separately to minimise under-representation of small species
Illumina MiSeq	**Library preparation**	Voucher MMG: TruSeq Nano
		Bulk MMG: TruSeq PCR-free
		Maximise insert size where possible; choose 550 bp TSN/TSP kits
	Sequencing depth	Voucher MMG: approximately 150 species per MiSeq run
		Bulk MMG: as much as possible but not less than 1 MiSeq run per 10 samples
Bioinformatics	**Trim adapters?**	Yes
	Quality control?	Yes
	Filter for mitochondrial reads?	Depends on availability of suitable database to filter against and data volume; ~50 mitogenomes per expected insect order representing all major sub-lineages is probably sufficient (not tested here); data volume has a large effect on computation time for assembly (exponential increase) so the larger the data volume, the greater the benefit of filtering on downstream steps; time for filtering increases linearly with data volume (for a given database size)
	Assembly	For maximal species recovery and contig length, combine the output of 2+ assemblers by reassembly; currently recommend IDBA-UD plus Celera Assembler as a minimum
		For pilot studies etc. where a rapid assessment of success is required, use a single IDBA-UD assembly over alternatives
		For multi-sample bulk MMG always do a combined assembly of reads from all sites to maximise species recovery, especially where sequencing depth is low; additional site-by-site assemblies may be useful (not tested here)
Phylogenetics		Maximise taxon sampling as far as possible by adding published superbarcodes
		In the absence or limited availability of appropriate superbarcodes, a two-step procedure (generate backbone tree with the longest contigs first) may improve topology when many short contigs are analysed
		Use 1RY2 coding with RAxML; better topologies may be obtained with PhyloBayes (not tested here, expected to be unwieldy for larger (~200+ taxa) datasets)
Characterising communities		Assess species presence-absence per site (bulk MMG) by mapping reads from each to the non-redundant contigs from the combined assembly of all sites (i.e. site-by-site assemblies are not necessary to determine species composition)
		For community phylogeny ideally compare placement of each contig in two different topologies to check for inconsistencies; recommended analyses '+superbarcode –backbone' and '-superbarcode +backbone'
		Assess higher-level taxonomic composition by assigning contigs to (e.g.) family based on monophyly with superbarcodes

5.2 A Wider Perspective

In spite of the challenges outlined above, it is important to place this work in context. Even at current levels the efficiency and simplicity of MMG for the large-scale generation of mitogenome sequences is significantly greater than any other available method. The most basic formulation of MMG, namely voucher MMG for generating large libraries of superbarcodes, is arguably the one that will have the greatest impact and uptake in the wider community, both as part of an expanded DNA barcoding concept and for mito-phylogenomics. In the latter case, significant increases in taxon sampling are now possible even with relatively modest effort, as seen in the expansion of mitogenome sampling of Curculionoidea between (Haran et al. 2013; LR-PCR) and (Gillett et al. 2014; voucher MMG). The mitogenome sequences generated during the course of these analyses represent a huge resource for future phylogenetic reconstruction and although the overall number of unique 8+ contigs generated (~1400) is far less than the number of species that was recently obtained by data-mining GenBank (8441 species; Bocak et al. 2014), the majority of species in the latter analysis were represented by one to two of five possible loci (3 mitochondrial; 2 nuclear rDNAs). Alongside additional voucher MMG sequencing for superbarcodes, efforts to generate nuclear markers should increase markedly to facilitate combined analyses. A significant recent contribution by McKenna et al. (2015) sampled eight nuclear loci for 367 beetle species with good coverage of extant families and matching this highly complete matrix with a similar mitogenome matrix could prove extremely powerful. However the overlap between the latter dataset and the mitogenome set herein is currently very limited. In the mean time, the continued application of increasingly densely sampled datasets to resolve the mitochondrial phylogeny of beetles is both an exciting and apparently feasible prospect, and there is no reason to suspect at this time that this would prove different for other insect groups. Initial analyses of the dataset generated herein obtain a topology largely congruent with those seen under dense taxon sampling in Chapters 3 and 4, indicating that the beetle mitogenome phylogeny may become stable at densities of ~500 taxa (Figure 5-1).

Looking beyond beetles, there is little data currently available to conclude how effective MMG is likely to prove for other taxa, although Zhou and colleagues have had success with mixed insect MMG samples and genome skimming of Apocrita (Hymenoptera). Initial assessments of bulk MMG sequencing for the Diptera in the *BorneoCanopy* sample are also promising. The majority of insect mitochondrial genomes sequenced to date conform to a highly conserved gene arrangement that differs little from the ancestral arthropod arrangement, and show relatively little length variation (Cameron 2014). While there are exceptions to this pattern, there is no *a priori* expectation that beetle mitogenomes are easier to assemble from metagenomic data than the majority of other insect orders. Thus the results obtained in the present work should be extensible to other orders. Currently the largest difference in these analyses would be the greatly reduced level of superbarcode sampling in other order relative to Coleoptera, hindering the mitochondrial data-filtering step. However, in the short term assemblies of unfiltered data on voucher MMG samples can be used to increase superbarcode sampling to a useful level, simultaneously facilitating superbarcode-based taxonomic descriptions of uncharacterised samples.

Finally, the great potential for bulk MMG in tropical systems to rapidly increase representation of species that are otherwise unsequenced needs to occur in synergy with traditional taxonomy to maximise the value of the sequences obtained and facilitate the description of new species. For MMG the most efficient approach would be to apply non-destructive DNA extraction methods to unsorted trap-catch, leaving the specimens intact for morphological assessment. Residual DNA could then be extracted individually from specimens that are found to be of particular interest to allow the generation of a bait

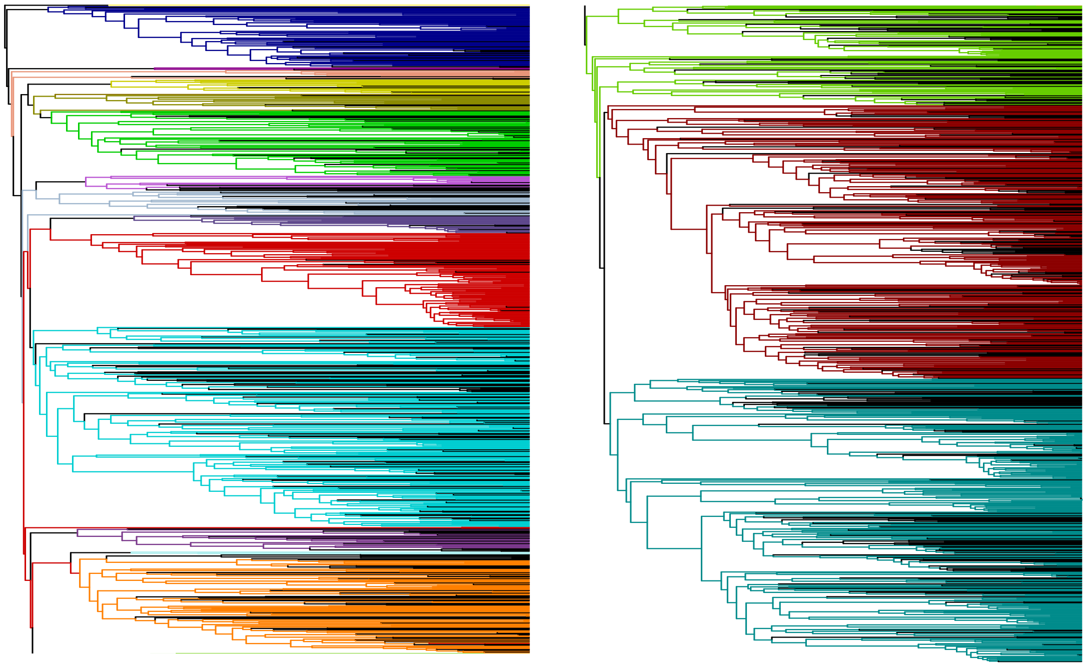


Figure 5-1 Mitochondrial phylogeny for 1529 beetle species with 8+ genes. RAxML analysis for protein-coding genes (1RY2-coding) and rRNAs. All but 278 were assembled in the present thesis (IDBA-UD assemblies; Chapter 2). Identified sequences are highlighted by superfamily following the convention in Chapters 3 and 4.

sequence by PCR and post-assembly identification of the corresponding mitogenome. How realistic such an approach would be remains to be tested, particularly for groups with hard exoskeletons such as Coleoptera, but amplification was possible for a subset of *BorneoCanopy* Diptera after non-destructive extraction.

To conclude, a great deal of progress has been made towards PCR-free analyses of insect biodiversity in a relatively short period of time. Many questions have been resolved and successful protocols established, minimising the methodological barrier to the wider uptake of MMG in the short term. However, a large number of technical questions remain to be answered and the progress made towards answering these in the next few years will likely determine the longevity of this approach. There is evidently great potential for an integrated phylogeny-centred framework for the study of insect diversity at large spatial scales. The present work demonstrates that this is now technically feasible and the methodology to obtain the underlying data for the implementation of such a framework is in place.

Chapter 6 References

Ahrens D, Monaghan MT, Vogler AP. 2007. DNA-based taxonomy for associating adults and larvae in multi-species assemblages of chafers (Coleoptera: Scarabaeidae). Mol. Phylogenet. Evol. 44:436–449.

Alexander K. 2010. Saproxylic beetles. In: Newton AC, editor. Biodiversity in the New Forest. Newbury: Pisces Publications. p. 46–53.

Altschup S, Gish W, Miller W, Myers E, Lipman D. 1990. Basic Local Alignment Search Tool. J. Mol. Biol. 215:403–410.

Andújar C, Arribas P, Ruzicka F, Crampton-Platt A, Timmermans MJTN, Vogler AP. 2015. Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. Mol. Ecol. [Internet] 24:3603–3617. Available from: http://doi.wiley.com/10.1111/mec.13195

Baselga A, Fujisawa T, Crampton-Platt A, Bergsten J, Foster PG, Monaghan MT, Vogler AP. 2013. Whole-community DNA barcoding reveals a spatio-temporal continuum of biodiversity at species and genetic levels. Nat. Commun. [Internet] 4:1892. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23695686

Baselga A, Gómez-Rodríguez C, Vogler AP. 2015. Multi-hierarchical macroecology at species and genetic levels to discern neutral and non-neutral processes. Glob. Ecol. Biogeogr. [Internet] 24:873–882. Available from: http://doi.wiley.com/10.1111/geb.12322

Baselga A, Orme CDL. 2012. betapart : an R package for the study of beta diversity. Methods Ecol. Evol. [Internet] 3:808–812. Available from: http://doi.wiley.com/10.1111/j.2041-210X.2012.00224.x

Baselga A. 2010. Partitioning the turnover and nestedness components of beta diversity. Glob. Ecol. Biogeogr. 19:134–143.

Basset Y, Cizek L, Cuenoud P, Didham RK, Guilhaumon F, Missa O, Novotny V, Odegaard F, Roslin T, Schmidl J, et al. 2012. Arthropod Diversity in a Tropical Forest. Science (80-.). [Internet] 338:1481–1484. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1226727

Berlocher SH. 1980. An Electrophoretic Key for Distinguishing Species of the Genus Rhagoletis (Diptera: Tephritidae) as Larvae, Pupae, or Adults. Ann. Entomol. Soc. Am. [Internet] 73:131–137. Available from: http://aesa.oxfordjournals.org/content/73/2/131.abstract

Berman M, Austin CM, Miller AD. 2014. Characterisation of the complete mitochondrial genome and 13 microsatellite loci through next-generation sequencing for the New Caledonian spider-ant Leptomyrmex pallens. Mol. Biol. Rep. 41:1179–1187.

Beutel R, Haas F. 2000. Phylogenetic Relationships of the Suborders of Coleoptera (Insecta). Cladistics [Internet] 16:103–141. Available from: http://doi.wiley.com/10.1006/clad.1999.0124

Bienert F, DE Danieli S, Miquel C, Coissac E, Poillot C, Brun J-J, Taberlet P. 2012. Tracking earthworm communities from soil DNA. Mol. Ecol. [Internet]:2017–2030. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22250728

Bininda-Emonds ORP. 2005. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. BMC Bioinformatics [Internet] 6:156. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1175081&tool=pmcentrez&rendertype=abstract

Biomatters. 2013. Geneious (version 6.1) created by Biomatters. Available from. Available from: http://www.geneious.com/

Bocak L, Barton C, Crampton-Platt A, Chesters D, Ahrens D, Vogler AP. 2014. Building the Coleoptera tree-of-life for >8000 species: composition of public DNA data and fit with Linnaean classification. Syst. Entomol. [Internet] 39:97–110. Available from: http://doi.wiley.com/10.1111/syen.12037

Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. [Internet] 13:R122. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4056372&tool=pmcentrez&rendertype=abstract

Burke M, Dunham J, Shahrestani P, Thornton K, Rose M, Long A. 2010. Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature [Internet] 467:587–590. Available from: citeulike-article-id:7843688\nhttp://dx.doi.org/10.1038/nature09352

Caley MJ, Fisher R, Mengersen K. 2014. Global species richness estimates have not converged. Trends Ecol. Evol. [Internet] 29:187–188. Available from: http://dx.doi.org/10.1016/j.tree.2014.02.002

Cameron SL, Lambkin CL, Barker SC, Whiting MF. 2007. A mitochondrial genome phylogeny of Diptera: Whole genome sequence data accurately resolve relationships over broad timescales with high precision. Syst. Entomol. 32:40–59.

Cameron SL. 2014. Insect mitochondrial genomics: implications for evolution and phylogeny. Annu. Rev. Entomol. [Internet] 59:95–117. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24160435

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U. S. A. [Internet] 108 Suppl:4516–4522. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3063599&tool=pmcentrez&rendertype=abstract

Cardoso P, Erwin T, Borges P, New T. 2011. The seven impediments in invertebrate conservation and how to overcome them. Biol. Conserv. [Internet] 144:2647–2655. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0006320711002874

Carpenter D, Hammond PM, Sherlock E, Lidgett A, Leigh K, Eggleton P. 2012. Biodiversity of soil macrofauna in the New Forest: a benchmark study across a national park landscape. Biodivers. Conserv. [Internet] 21:3385–3410. Available from: http://www.springerlink.com/index/10.1007/s10531-012-0369-0

Caterino MS, Cho S, Sperling FAH. 2000. The Current State of Insect Molecular Systematics: A Thriving Tower of Babel. Annu. Rev. Ecol. Evol. Syst. 45:1–54.

Caterino MS, Shull VL, Hammond PM, Vogler a P. 2002. Basal Relationships of Coleoptera interred from 18S rDNA sequences. Zool. Scr. 31:41–49.

Costello MJ, May RM, Stork NE. 2013. Can We Name Earth's Species Before They Go Extinct? Science (80-.). [Internet] 339:413–416. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1230318

Costello MJ, Wilson S, Houlding B. 2012. Predicting total global species richness using rates of species description and estimates of taxonomic effort. Syst. Biol. 61:871–883.

Craft KJ, Pauls SU, Darrow K, Miller SE, Hebert PDN, Helgen LE, Novotny V, Weiblen GD. 2010. Population genetics of ecological communities with DNA barcodes: an example from New Guinea Lepidoptera. Proc. Natl. Acad. Sci. U. S. A. 107:5041–5046.

Crampton-Platt A, Timmermans MJTN, Gimmel ML, Kutty SN, Cockerill TD, Chey VK, Vogler AP. 2015. Soup to Tree: The Phylogeny of Beetles Inferred by Mitochondrial Metagenomics of a Bornean Rainforest Sample. Mol. Biol. Evol. [Internet] 32:2302–2316. Available from: http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/msv111

Creer S, Fonseca VG, Porazinska DL, Giblin-Davis RM, Sung W, Power DM, Packer M, Carvalho GR, Blaxter ML, Lambshead PJD, et al. 2010. Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. Mol. Ecol. [Internet] 19 Suppl 1:4–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20331766

Cristescu ME. 2014. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends Ecol. Evol. [Internet] 29:566–571. Available from: http://linkinghub.elsevier.com/retrieve/pii/S016953471400175X

Crowson R. 1960. The phylogeny of Coleoptera. Annu. Rev. Entomol. [Internet] 5:111–134. Available from: http://www.annualreviews.org/doi/pdf/10.1146/annurev.en.05.010160.000551

Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P. 2014. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biol. Lett. [Internet] 10:20140562. Available from: http://dx.doi.org/10.1098/rsbl.2014.0562

Dettai A, Gallut C, Brouillet S, Pothier J, Lecointre G, Debruyne R. 2012. Conveniently pre-tagged and pre-packaged: extended molecular identification and metagenomics using complete metazoan mitochondrial genomes. PLoS One [Internet] 7:e51263. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23251474

Eddy S, Durbin R. 1994. RNA sequence analysis using covariance models. Nucleic Acids Res. [Internet] 22:2079–2088. Available from: http://nar.oxfordjournals.org/content/22/11/2079.short

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Eggleton P, Inward K, Smith J, Jones DT, Sherlock E. 2009. A six year study of earthworm (Lumbricidae) populations in pasture woodland in southern England shows their responses to soil temperature and soil moisture. Soil Biol. Biochem. [Internet] 41:1857–1865. Available from: http://dx.doi.org/10.1016/j.soilbio.2009.06.007

Eisses KT, van Dijk H, van Delden W. 1979. Genetic Differentiation Within the Melanogaster Species Group of the Genus Drosophila (Sophophora). Evolution (N. Y). 33:1063–1068.

Emerson BC, Cicconardi F, Fanciulli PP, Shaw PJA. 2011. Phylogeny, phylogeography, phylobetadiversity and the molecular analysis of biological communities. Philos. Trans. R. Soc. Lond. B. Biol. Sci. [Internet] 366:2391–2402. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3130430&tool=pmcentrez&rendertype=abstract

Faith DP. 1992. Conservation evaluation and phylogenetic diversity. Biol. Conserv. [Internet] 61:1–10. Available from: http://linkinghub.elsevier.com/retrieve/pii/0006320792912013

Favreau JM, Drew CA, Hess GR, Rubino MJ, Koch FH, Eschelbach K a. 2006. Recommendations for assessing the effectiveness of surrogate species approaches. Biodivers. Conserv. 15:3949–3969.

Floyd R, Abebe E, Papert A, Blaxter M. 2002. Molecular barcodes for soil nematode identification. Mol. Ecol. [Internet] 11:839–850. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11972769

Fonseca VG, Carvalho GR, Sung W, Johnson HF, Power DM, Neill SP, Packer M, Blaxter ML, Lambshead PJD, Thomas WK, et al. 2010. Second-generation environmental sequencing unmasks marine metazoan biodiversity. Nat. Commun. [Internet] 1:98. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2963828&tool=pmcentrez&rendertype=abstract

García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A. 2012. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics [Internet] 28:2678–2679. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22914218

Gibson J, Shokralla S, Porter TM, King I, van Konynenburg S, Janzen DH, Hallwachs W, Hajibabaei M. 2014. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proc. Natl. Acad. Sci. U. S. A. [Internet] 111:8007–8012. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4050544&tool=pmcentrez&rendertype=abstract

Gillett CPDT, Crampton-Platt A, Timmermans MJTN, Jordal B, Emerson BC, Vogler AP. 2014. Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). Mol. Biol. Evol. [Internet] 31:2223–2237. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24803639

Gómez-Rodríguez C, Crampton-Platt A, Timmermans MJTN, Baselga A, Vogler AP. 2015. Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. Methods Ecol. Evol. [Internet] 6:883–894. Available from: http://doi.wiley.com/10.1111/2041-210X.12376

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZT, et al. 2013. Next-generation museomics disentangles one of the largest primate radiations. Syst. Biol. [Internet] 62:539–554. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3676678&tool=pmcentrez&rendertype=abstract

Hajibabaei M, Shokralla S, Zhou X, Singer G, Baird D. 2011. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS One [Internet] 6:e17497. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3076369&tool=pmcentrez&rendertype=abstract

Haran J, Timmermans MJTN, Vogler AP. 2013. Mitogenome sequences stabilize the phylogenetics of weevils (Curculionoidea) and establish the monophyly of larval ectophagy. Mol. Phylogenet. Evol. [Internet] 67:156–166. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23319085

Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. Proc. Biol. Sci. [Internet] 270:313–321. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1691236&tool=pmcentrez&rendertype=abstract

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM. 2004. Identification of Birds through DNA Barcodes. PLoS Biol. [Internet] 2:e312. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=518999&tool=pmcentrez&rendertype=abstract

Hrcek J, Miller SE, Quicke DLJ, Smith MA. 2011. Molecular detection of trophic links in a complex insect host-parasitoid food web. Mol. Ecol. Resour. 11:786–794.

Hughes JB, Daily GC, Ehrlich PR. 2000. Conservation of insect diversity: A habitat approach. Conserv. Biol. 14:1788–1797.

Hunt T, Bergsten J, Levkanicova Z, Papadopoulou A, St. John O, Wild R, Hammond PM, Ahrens D, Balke M, Caterino MS, et al. 2007. A Comprehensive Phylogeny of Beetles Reveals the Evolutionary Origins of a Superradiation. Science (80-.). [Internet] 318:1913–1916. Available from: http://www.sciencemag.org/content/318/5858/1913.short

Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. [Internet] 61:1061–1067. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22780991

i5KConsortium. 2013. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J. Hered. [Internet] 104:595–600. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23940263

Illumina. 2013. Comparison of TruSeq ® Sample Preparation Kits. Available from: http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_truseq_comparison.pdf

Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. 2012. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science [Internet] 335:587–590. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22301318

Janzen DH, Hajibabaei M, Burns JM, Hallwachs W, Remigio E, Hebert PDN. 2005. Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360:1835–1845.

Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching R, Dolman PM, Woodcock P, Edwards FA, et al. 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecol. Lett. [Internet] 16:1245–1257. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23910579

JNCC. 2015. The New Forest. Available from: http://jncc.defra.gov.uk/protectedsites/sacselection/sac.asp?EUCode=UK0012557

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. [Internet] 30:772–780. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3603318&tool=pmcentrez&rendertype=abstract

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463–1464.

Kocher A, Guilbert É, Lhuillier É, Murienne J. 2015. Sequencing of the mitochondrial genome of the avocado lace bug Pseudacysta perseae (Heteroptera, Tingidae) using a genome skimming approach. C. R. Biol. [Internet] 338:149–160. Available from: http://www.sciencedirect.com/science/article/pii/S1631069114003072

Kocher A, Kamilari M, Lhuillier E, Coissac E, Péneau J, Chave J, Murienne J. 2014. Shotgun assembly of the assassin bug Brontostoma colossus mitochondrial genome (Heteroptera, Reduviidae). Gene [Internet] 552:184–194. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0378111914010592

Krell F, Chung AYC, Deboise E, Eggleton P, Giusti A, Inward K, Krell-westerwalbesloh S. 2005. Quantitative extraction of macro-invertebrates from temperate and tropical leaf litter and soil : efficiency and time-dependent taxonomic biases of the Winkler extraction. Pedobiologia (Jena). 49:175–186.

Kremen C, Colwell RK, Erwin TL, Murphy DD, Noss RF, Sanjayan M a. 1993. Terrestrial Arthropod Assemblages: Their Use in Conservation Planning. Conserv. Biol. 7:796–808.

Kukalová-Peck J, Lawrence JF. 1993. Evolution of the Hind Wing in Coleoptera. Can. Entomol. [Internet] 125:181–258. Available from: href="http://dx.doi.org/10.4039/Ent125181-2

Kukalová-Peck J, Lawrence JF. 2004. Relationships among coleopteran suborders and major endoneopteran lineages: Evidence from hind wing characters. Eur. J. Entomol. 101:95–144.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics [Internet] 25:2286–2288. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19535536

Lawrence JF, Ślipiński A, Seago AE, Thayer MK, Newton AF, Marvaldi AE. 2011. Phylogeny of the Coleoptera Based on Morphological Characters of Adults and Larvae. Ann. Zool. 61:1–217.

Lawton J, Bignell D, Bolton B, Bloemers G, Eggleton P, Hammond P, Hodda M, Holt R, Larsen T, Mawdsley N, et al. 1998. Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest. Nature [Internet] 391:72–76. Available from: http://people.biology.ufl.edu/rdholt/holtpublications/078.PDF

Lecroq B, Lejzerowicz F, Bachar D, Christen R, Esling P, Baerlocher L, Osterås M, Farinelli L, Pawlowski J. 2011. Ultra-deep sequencing of foraminiferal microbarcodes unveils hidden richness of early monothalamous lineages in deep-sea sediments. Proc. Natl. Acad. Sci. U. S. A. 108:13177–13182.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Linard B, Crampton-Platt A, Gillett CPDT, Timmermans MJTN, Vogler AP. 2015. Metagenome skimming of insect specimen pools: potential for comparative genomics. Genome Biol. Evol. [Internet] 7:1474–1489. Available from: http://gbe.oxfordjournals.org/cgi/doi/10.1093/gbe/evv086

Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. Proc. Natl. Acad. Sci. [Internet]:10.1073/pnas.1521291113. Available from: https://peerj.com/preprints/1451

Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B. 2012. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. [Internet] 40:W622–W627. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3394330&tool=pmcentrez&rendertype=abstract

Maddison WP, Maddison DR. 2011. Mesquite: a modular system for evolutionary analysis. Available from: http://mesquiteproject.org

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben L a, Berka J, Braverman MS, Chen Y-J, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

Mason VC, Li G, Helgen KM, Murphy WJ. 2011. Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. Genome Res. [Internet] 21:1695–1704.

Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3202286&tool=pmcentrez&rendertype=abstract

May RM. 2010. Tropical Arthropod Species, More or Less? Science (80-.). [Internet] 329:41–42. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1191058

McKenna DD, Wild AL, Kanda K, Bellamy CL, Beutel RG, Caterino MS, Farnum CW, Hawks DC, Ivie M a., Jameson ML, et al. 2015. The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution. Syst. Entomol. [Internet]. Available from: http://doi.wiley.com/10.1111/syen.12132

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). New Orleans, LA: Institute of Electrical and Electronics Engineers (IEEE). p. 45–52. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5676129

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science (80-.). [Internet] 346:763–767. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1257570

Montagna M, Mereghetti V, Lencioni V, Rossaro B. 2016. Integrated Taxonomy and DNA Barcoding of Alpine Midges (Diptera: Chironomidae). PLoS One [Internet] 11:e0149673. Available from: http://dx.plos.org/10.1371/journal.pone.0149673

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. 2011. How Many Species Are There on Earth and in the Ocean?Mace GM, editor. PLoS Biol. [Internet] 9:e1001127. Available from: http://dx.plos.org/10.1371/journal.pbio.1001127

Mullis KB, Faloona FA. 1987. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol. 155:335–350.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, et al. 2000. A whole-genome assembly of Drosophila. Science [Internet] 287:2196–2204. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=151187&tool=pmcentrez&rendertype=abstract

Navarro SP, Jurado-Rivera JA, GóMez-Zurita J, Lyal CHC, Vogler AP. 2010. DNA profiling of host-herbivore interactions in tropical forests. Ecol. Entomol. [Internet] 35:18–32. Available from: http://doi.wiley.com/10.1111/j.1365-2311.2009.01145.x

Newton AC, Cantarello E, Myers G, Douglas S, Tejedor N. 2010. The condition and dynamics of New Forest woodlands. In: Newton AC, editor. Biodiversity in the New Forest. Newbury: Pisces Publications. p. 132–146. Available from: http://eprints.bournemouth.ac.uk/13924/1/licence.txt

Newton AC. 2010. Synthesis: status and trends of biodiversity in the New Forest. In: Newton AC, editor. Biodiversity in the New Forest. Newbury: Pisces Publications. p. 218–227. Available from: http://eprints.bournemouth.ac.uk/13924/1/licence.txt

Nipperess D a., Matsen F a. 2013. The mean and variance of phylogenetic diversity under rarefaction. Methods Ecol. Evol. 4:566–572.

Oberprieler RG, Marvaldi AE, Anderson RS. 2007. Weevils, weevils, weevils everywhere. Zootaxa 1668:491–520.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2015. vegan: Community Ecology Package. Available from: http://cran.r-project.org/package=vegan

Papadopoulou A, Anastasiou I, Spagopoulou F, Stalimerou M, Terzopoulou S, Legakis A, Vogler AP. 2011. Testing the species--genetic diversity correlation in the Aegean archipelago: toward a haplotype-based macroecology? Am. Nat. [Internet] 178:241–255. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21750387

Papadopoulou A, Taberlet P, Zinger L. 2015. Metagenome skimming for phylogenetic community ecology : a new era in biodiversity research. Mol. Ecol. 24:3515–3517.

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. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics [Internet] 28:1420–1428. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22495754

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, et al. 2003. TIGR gene indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. Bioinformatics 19:651–652.

Pinchen BJ, Ward L k. 2010. The New Forest cicada and other invertebrates. In: Newton AC, editor. Biodiversity in the New Forest. Newbury: Pisces Publications. p. 58–64. Available from: http://nora.nerc.ac.uk/20867/

Pons J, Ribera I, Bertranpetit J, Balke M. 2010. Nucleotide substitution rates for the full set of mitochondrial protein-coding genes in Coleoptera. Mol. Phylogenet. Evol. [Internet] 56:796–807. Available from: http://dx.doi.org/10.1016/j.ympev.2010.02.007

Porazinska DL, Giblin-Davis RM, Faller L, Farmerie W, Kanzaki N, Morris K, Powers TO, Tucker AE, Sung W, Thomas WK. 2009. Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. Mol. Ecol. Resour. [Internet] 9:1439–1450. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21564930

Quicke DLJ, Smith MA, Janzen DH, Hallwachs W, Fernandez-Triana J, Laurenne NM, Zaldívar-Riverón A, Shaw MR, Broad GR, Klopfstein S, et al. 2012. Utility of the DNA barcoding gene fragment for parasitic wasp phylogeny (Hymenoptera: Ichneumonoidea): data release and new measure of taxonomic congruence. Mol. Ecol. Resour. [Internet] 12:676–685. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22487608

R Core Team. 2015. R: A Language and Environment for Statistical Computing. Available from: http://www.r-project.org

Ratnasingham S, Hebert PDN. 2007. BOLD: The Barcode of Life Data System (www.barcodinglife.org). Mol. Ecol. Notes 7:355–364.

Ratnasingham S, Hebert PDN. 2013. A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. PLoS One 8.

Robinson D, Foulds L. 1981. Comparison of Phylogenetic Trees. Math. Biosci. [Internet] 141:131–141. Available from: http://www.sciencedirect.com/science/article/pii/0025556481900432

Rubinoff D, Cameron S, Will K. 2006. A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. J. Hered. [Internet] 97:581–594. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17135463

Rubinstein ND, Feldstein T, Shenkar N, Botero-Castro F, Griggio F, Mastrototaro F, Delsuc F, Douzery EJP, Gissi C, Huchon D. 2013. Deep sequencing of mixed total DNA without barcodes allows efficient assembly of highly plastic Ascidian mitochondrial genomes. Genome Biol. Evol. 5:1185–1199.

Sanderson M. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics [Internet] 19:301–302. Available from: http://bioinformatics.oxfordjournals.org/content/19/2/301.short

Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19:101–109.

Sarkar D. 2008. Lattice: Multivariate Data Visualization with R. New York: Springer Available from: http://lmdvr.r-forge.r-project.org

Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics [Internet] 27:592–593. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3035803&tool=pmcentrez&rendertype=abstract

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics [Internet] 27:863–864. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3051327&tool=pmcentrez&rendertype=abstract

Sheffield NC, Song H, Cameron SL, Whiting MF. 2009. Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics. Syst. Biol. 58:381–394.

Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen DH, Hallwachs W, Golding GB, Hajibabaei M. 2015. Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform.

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. Syst. Biol. 50:945–969.

Simon S, Hadrys H. 2013. A comparative analysis of complete mitochondrial genomes among Hexapoda. Mol. Phylogenet. Evol. [Internet] 69:393–403. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23598069

Slipinski S, Leschen R, Lawrence J. 2011. Order Coleoptera Linnaeus, 1758. In: Zhang, Z.-Q. (Ed.) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. Zootaxa 3148:203–208.

Smith J, Burke L. 2010. Managing the New Forest's Crown Lands. In: Newton AC, editor. Biodiversity in the New Forest. Newbury: Pisces. p. 212–217. Available from: http://eprints.bournemouth.ac.uk/13924/1/licence.txt

Smith MA, Fisher BL, Hebert PDN. 2005. DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. Philos. Trans. R. Soc. Lond. B. Biol. Sci. [Internet] 360:1825–1834. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1609228&tool=pmcentrez&rendertype=abstract

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc. Natl. Acad. Sci. U. S. A. [Internet] 103:12115–12120. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1524930&tool=pmcentrez&rendertype=abstract

Song H, 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. Syst. Entomol. 35:429–448.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Steel M, Penny D. 1993. Distributions of Tree Comparison Metrics - Some New Results. Syst. Biol. [Internet] 42:126–141. Available from: http://sysbio.oxfordjournals.org/content/42/2/126.short

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. Am. J. Bot. [Internet] 99:349–364. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22174336

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. [Internet] 21:2045–2050. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22486824

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 Evol. Biol. [Internet] 11:315. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3213125&tool=pmcentrez&rendertype=abstract

Tang M, Hardman CJ, Ji Y, Meng G, Liu S, Tan M, Yang S, Moss ED, Wang J, Yang C, et al. 2015. High-throughput monitoring of wild bee diversity and abundance via mitogenomics. Methods Ecol. Evol. [Internet] 6:1034–1043. Available from: http://doi.wiley.com/10.1111/2041-210X.12416

Tang M, Tan M, Meng G, Yang S, Su X, Liu S, Song W, Li Y, Wu Q, Zhang A, et al. 2014. Multiplex sequencing of pooled mitochondrial genomes-a crucial step toward biodiversity analysis using mito-metagenomics. Nucleic Acids Res. [Internet] 42:e166. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25294837

Taylor HR, Harris WE. 2012. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Mol. Ecol. Resour. [Internet] 12:377–388. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22356472

Thomsen PF, Willerslev E. 2014. Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. Biol. Conserv. [Internet] 183:4–18. Available from: http://dx.doi.org/10.1016/j.biocon.2014.11.019

Tilak M-K, Justy F, Debiais-Thibaud M, Botero-Castro F, Delsuc F, Douzery EJP. 2014. A cost-effective straightforward protocol for shotgun Illumina libraries designed to assemble complete mitogenomes from non-model species. Conserv. Genet. Resour. [Internet] 7:37–40. Available from: http://link.springer.com/10.1007/s12686-014-0338-x

Timmermans M, Barton C, Haran J, Ahrens D, Culverwell C, Ollikainen A, Dodsworth S, Foster P, Bocak L, Vogler A. 2016. Family-Level Sampling of Mitochondrial Genomes in Coleoptera: Compositional Heterogeneity and Phylogenetics. Genome Biol. Evol. [Internet] 8:161–175. Available from: http://gbe.oxfordjournals.org/lookup/doi/10.1093/gbe/evv241

Timmermans M, Dodsworth S, Culverwell C, Bocak L, Ahrens D, Littlewood D, Pons J, Vogler A. 2010. Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. Nucleic Acids Res. [Internet] 38:e197. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995086&tool=pmcentrez&rendertype=abstract

Timmermans M, Viberg C, Martin G, Hopkins K, Vogler AP. 2016. Rapid assembly of taxonomically validated mitochondrial genomes from historical insect collections. Biol.

J. Linn. Soc. 117:83–95.

Timmermans M, Vogler AP. 2012. Phylogenetically informative rearrangements in mitochondrial genomes of Coleoptera, and monophyly of aquatic elateriform beetles (Dryopoidea). Mol. Phylogenet. Evol. [Internet] 63:299–304. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22245358

Tubbs CR. 1968. The New Forest: An Ecological History. Newton Abbott: David and Charles

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen J a, Wu D, Paulsen I, Nelson KE, Nelson W, et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science [Internet] 304:66–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15001713

Whiting MF, Carpenter JC, Wheeler QD, Wheeler WC. 1997. The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. Syst. Biol. 46:1–68.

Will KW, Rubinoff D. 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics [Internet] 20:47–55. Available from: http://doi.wiley.com/10.1111/j.1096-0031.2003.00008.x

Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MTP, Shapiro B, Bunce M, Wiuf C, Gilichinsky DA, Cooper A. 2003. Diverse Plant and Animal Genetic Records from Holocene and Pleistocene Sediments. Nature 300:791–795.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U. S. A. 74:5088–5090.

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z. 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol. Evol. [Internet] 3:613–623. Available from: http://doi.wiley.com/10.1111/j.2041-210X.2012.00198.x

Zeale MRK, Butlin RK, Barker GLA, Lees DC, Jones G. 2011. Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. Mol. Ecol. Resour. 11:236–244.

Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q. 2013. Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. Gigascience [Internet] 2:4. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3637469&tool=pmcentrez&rendertype=abstract

Chapter 7 Appendix 1

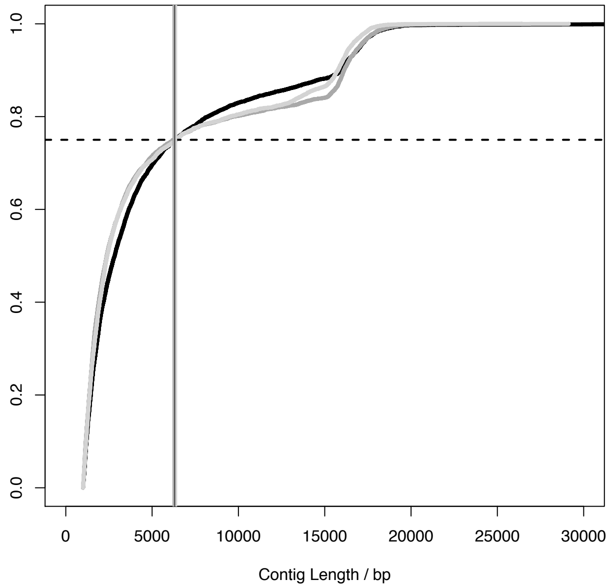


Figure 7-1 Cumulative length distribution for contigs assembled by each program across all datasets. CA: black; IDBA: dark grey; NWBL: light grey. Vertical lines indicate 3rd quartile length in each case.

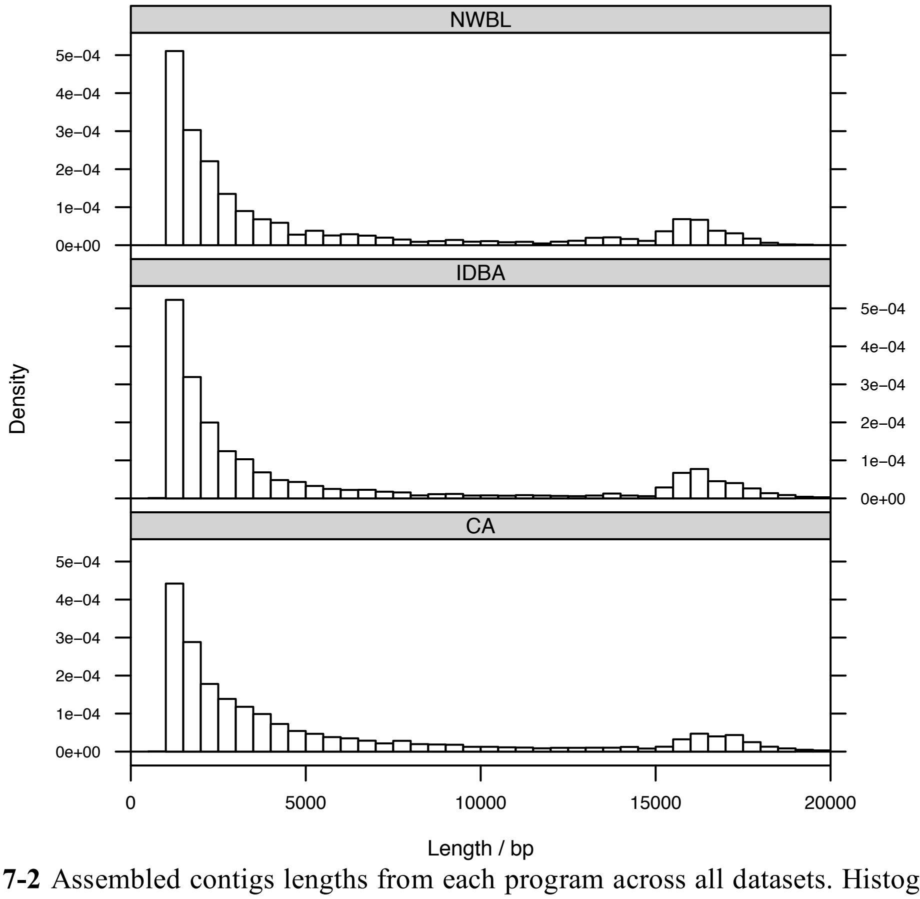


Figure 7-2 Assembled contigs lengths from each program across all datasets. Histogram bins 500bp.

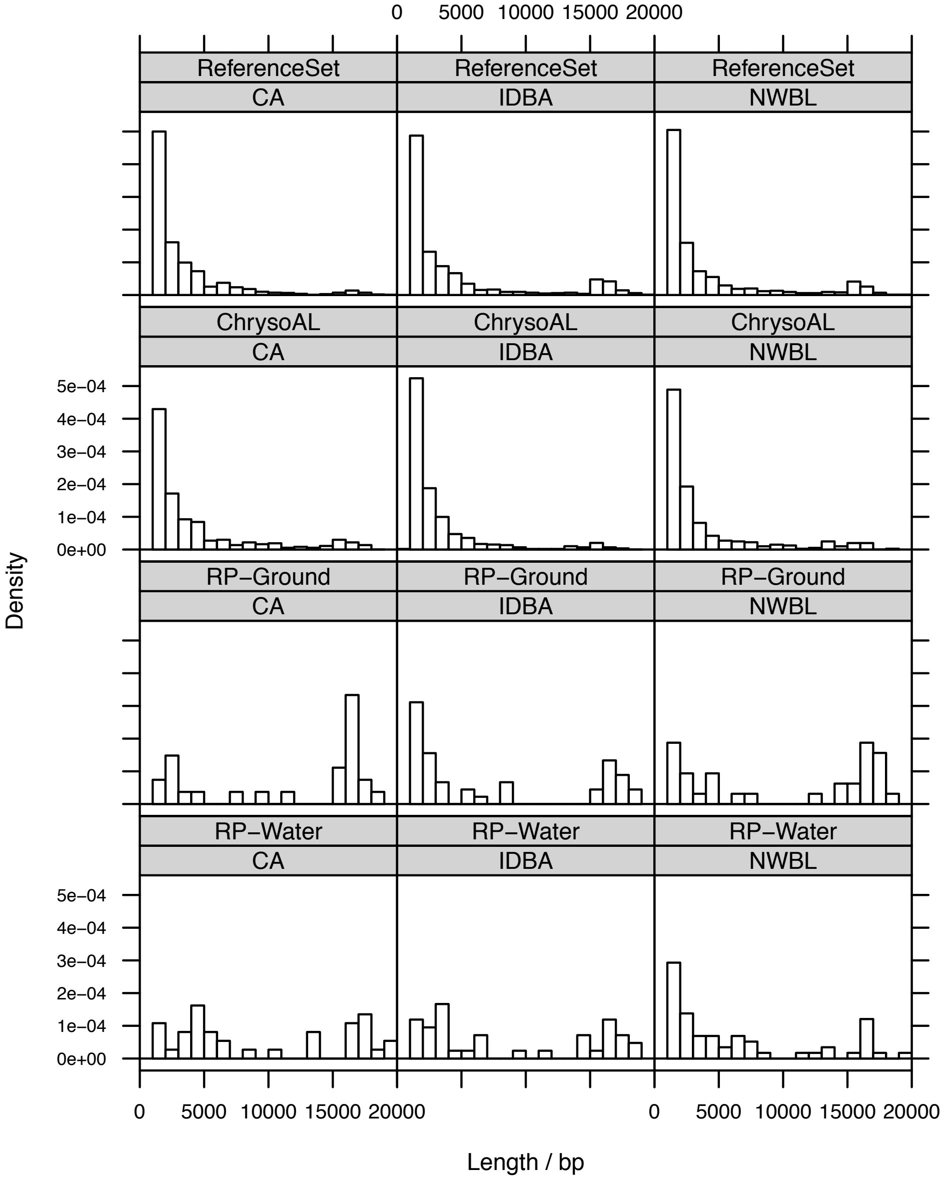
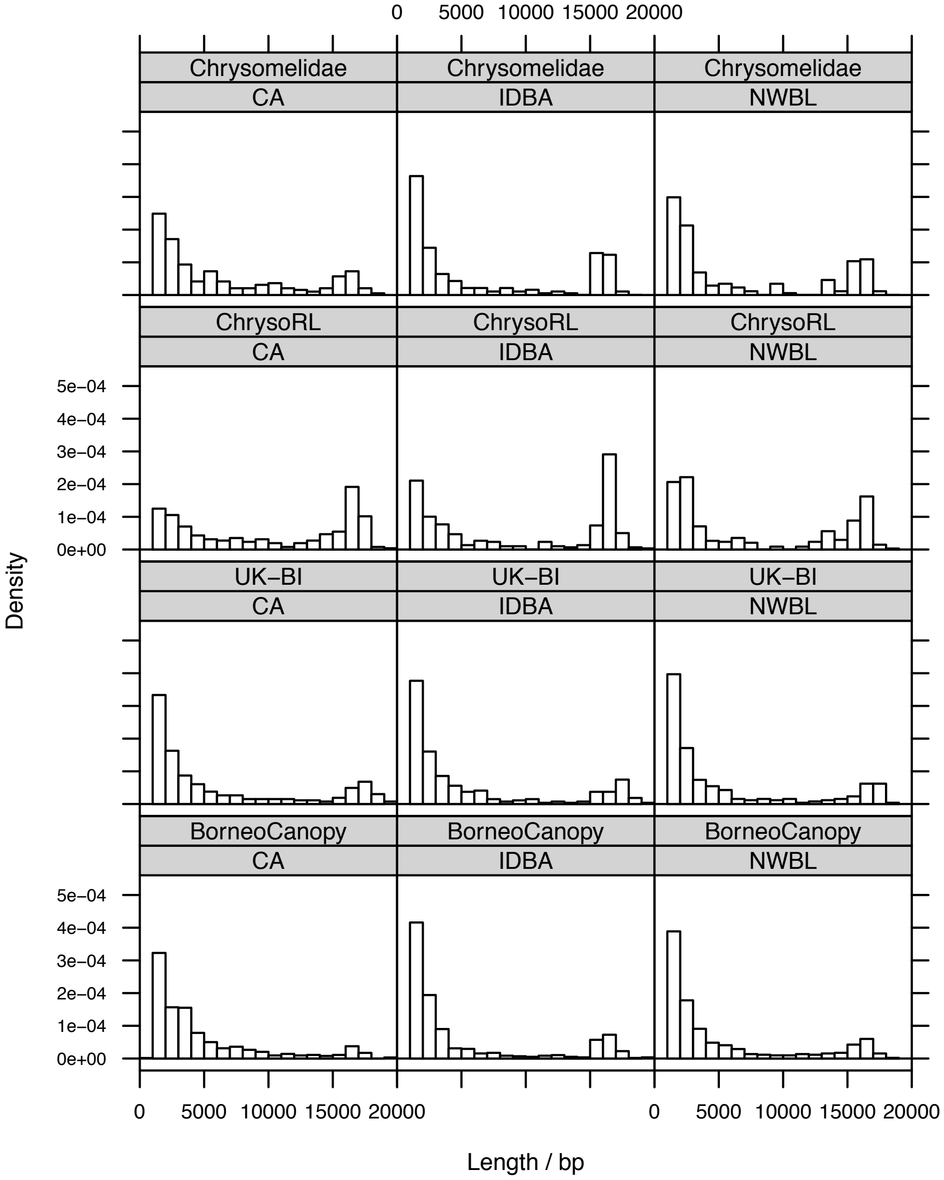
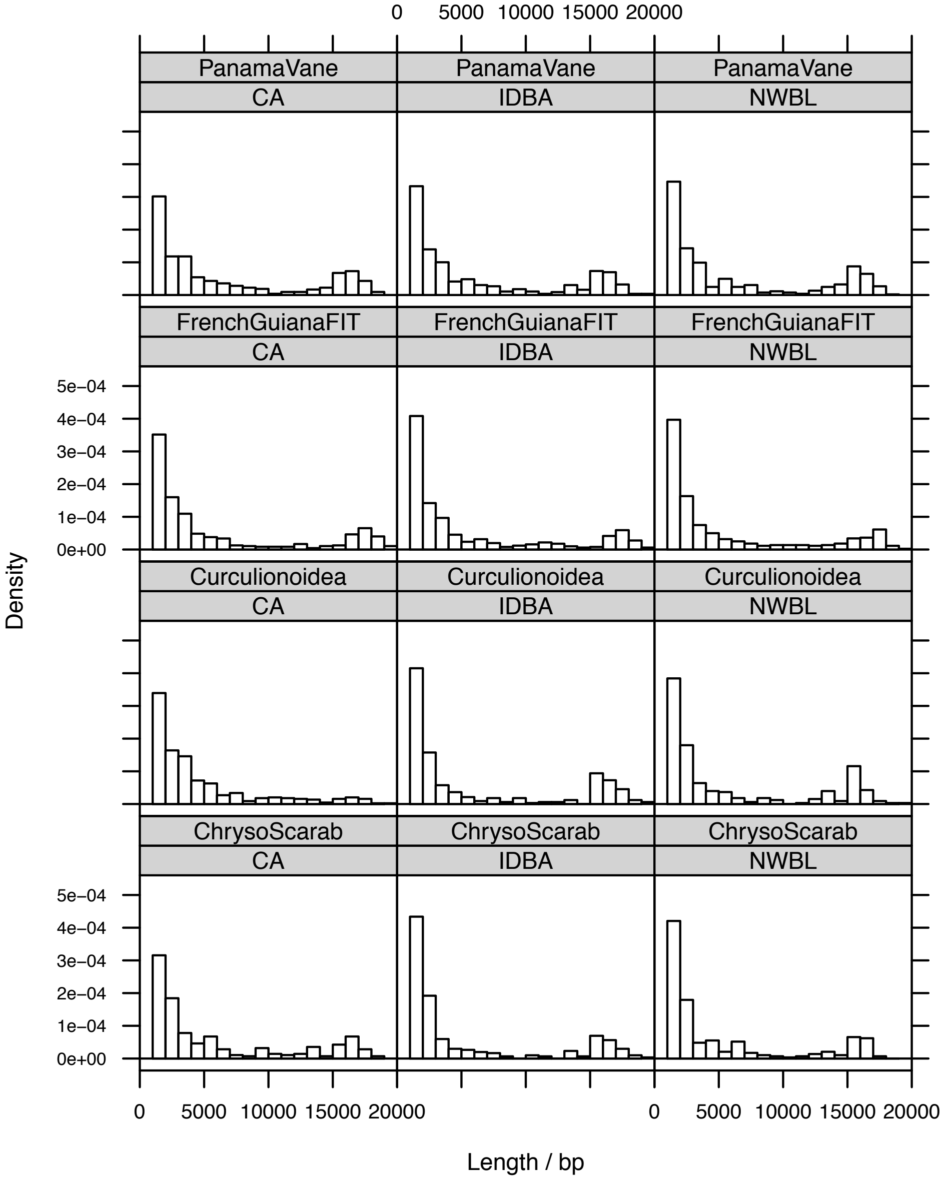


Figure 7.3 Contig length distributions by assembler and dataset, split across four pages.


Chrysomelidae
CA
IDBA
NWBL
ChrysoRL
UK−BI
BorneoCanopy
Density
Length / bp
0
5000
10000
15000
20000
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04


PanamaVane
CA
PanamaVane
IDBA
PanamaVane
NWBL
FrenchGuianaFIT
CA
FrenchGuianaFIT
IDBA
FrenchGuianaFIT
NWBL
Curculionoidea
CA
Curculionoidea
IDBA
Curculionoidea
NWBL
ChrysoScarab
CA
ChrysoScarab
IDBA
ChrysoScarab
NWBL
Density
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
0
5000
10000
15000
20000
Length / bp

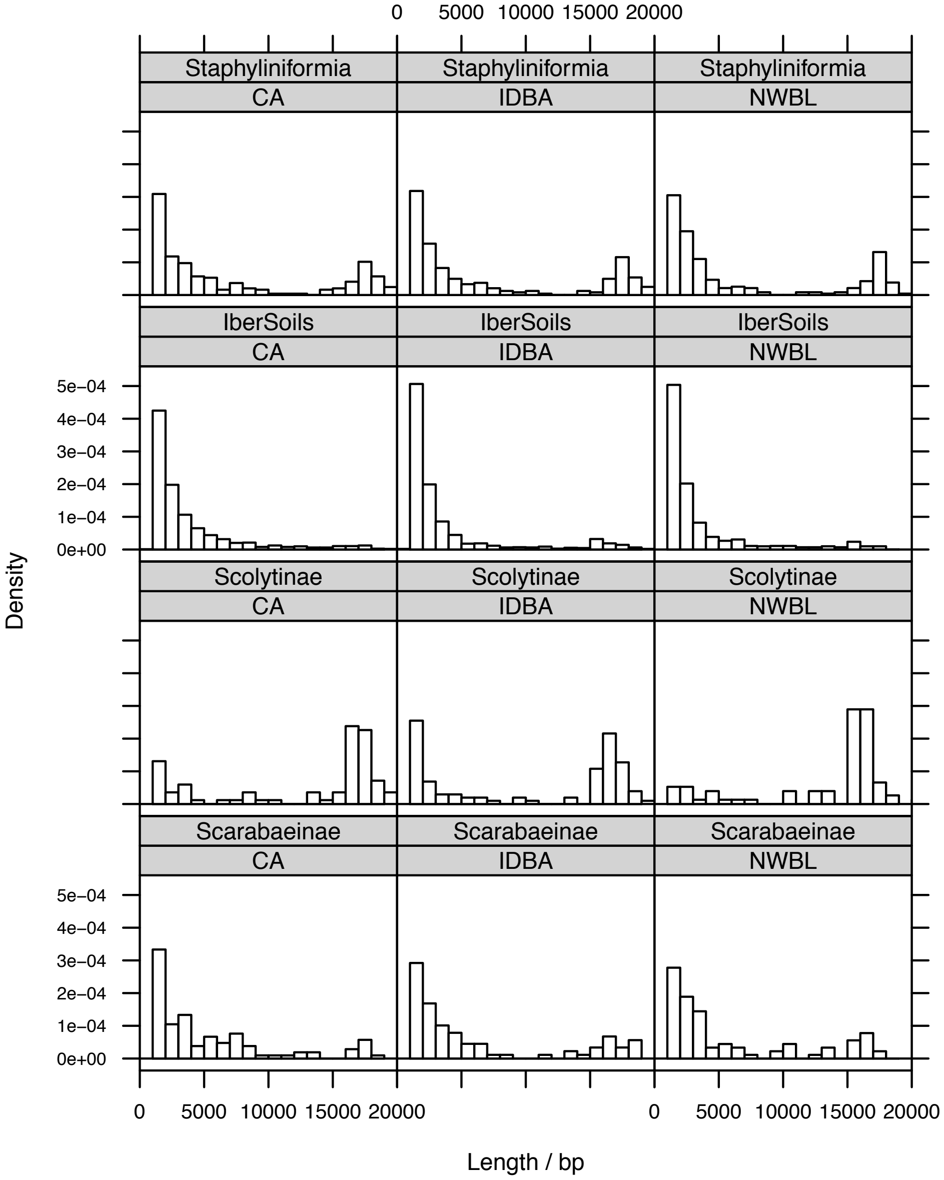

Staphyliniformia
CA
Staphyliniformia
IDBA
Staphyliniformia
NWBL
IberSoils
CA
IberSoils
IDBA
IberSoils
NWBL
Scolytinae
CA
Scolytinae
IDBA
Scolytinae
NWBL
Scarabaeinae
CA
Scarabaeinae
IDBA
Scarabaeinae
NWBL
0
5000
10000
15000
20000
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
Density
Length / bp

Table 7.1 Results of pairwise Kolmogorov-Smirnov tests between assemblies.

Dataset	CA vs IDBA		CA vs NWBL		IDBA vs NWBL	
	D	p	D	p	D	p
BorneoCanopy	0.140	<<0.001	0.108	0.003	0.052	0.456
IberSoils	0.105	<<0.001	0.100	<<0.001	0.031	0.672
ChrysIber (RL)	0.104	0.101	0.220	<<0.001	0.186	<<0.001
ChrysIber (AL)	0.132	0.001	0.084	0.132	0.069	0.204
UK-BI	0.078	0.388	0.089	0.250	0.043	0.969
FrenchGuianaFIT	0.073	0.145	0.076	0.144	0.028	0.993
PanamaVane	0.055	0.380	0.076	0.096	0.023	0.981
RP-Water	0.171	0.612	0.330	0.011	0.263	0.069
RP-Ground	0.33	0.047	0.188	0.682	0.248	0.201
Curculionoidea	0.179	<<0.001	0.155	<0.001	0.102	0.064
Scolytinae	0.204	0.044	0.291	0.002	0.254	0.007
Staphyliniformia	0.061	0.752	0.101	0.173	0.061	0.764
Scarabaeinae	0.122	0.474	0.105	0.662	0.101	0.750
Chrysomelidae	0.180	0.004	0.137	0.065	0.090	0.462
ChrysoScarab	0.135	0.010	0.140	0.007	0.053	0.800
ReferenceSet	0.091	<0.001	0.069	0.021	0.057	0.125

Table 7.2 Results of Hartigan's dip test for unimodality on each assembly.

Dataset	CA		IDBA		NWBL	
	D	p	D	p	D	p
BorneoCanopy	0.016	0.323	0.049	<<0.001	0.036	<<0.001
IberSoils	0.006	0.993	0.018	0.012	0.009	0.931
ChrysIber (RL)	0.093	<<0.001	0.140	<<0.001	0.1	<<0.001
ChrysIber (AL)	0.014	0.939	0.012	0.905	0.014	0.925
UK-BI	0.048	<0.001	0.048	<0.001	0.047	<0.001
FrenchGuianaFIT	0.048	<<0.001	0.037	<<0.001	0.037	<0.001
PanamaVane	0.052	<<0.001	0.052	<<0.001	0.058	<<0.001
RP-Water	0.100	0.003	0.098	0.002	0.061	0.098
RP-Ground	0.104	0.014	0.110	<<0.001	0.132	<<0.001
Curculionoidea	0.010	0.993	0.073	<<0.001	0.057	<<0.001
Scolytinae	0.08	<0.001	0.134	<<0.001	0.043	0.378
Staphyliniformia	0.067	<<0.001	0.079	<<0.001	0.075	<<0.001
Scarabaeinae	0.030	0.751	0.059	0.021	0.053	0.062
Chrysomelidae	0.047	0.003	0.099	<<0.001	0.079	<<0.001
ChrysoScarab	0.040	0.003	0.050	<<0.001	0.047	<<0.001
ReferenceSet	0.007	0.989	0.032	<<0.001	0.021	0.012

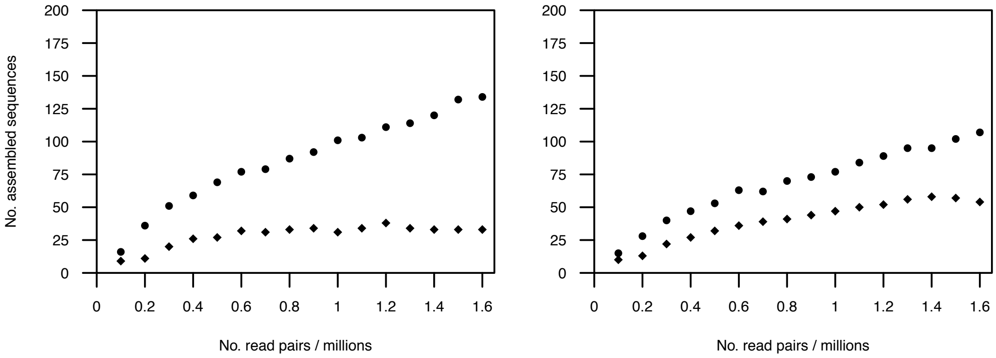


Figure 7.4 Assembled *cox1* (dots) and long contigs (diamonds) in subsampled assemblies of *ChrysIber ChrysoAL*. L: IDBA; R: IDBA-1k.

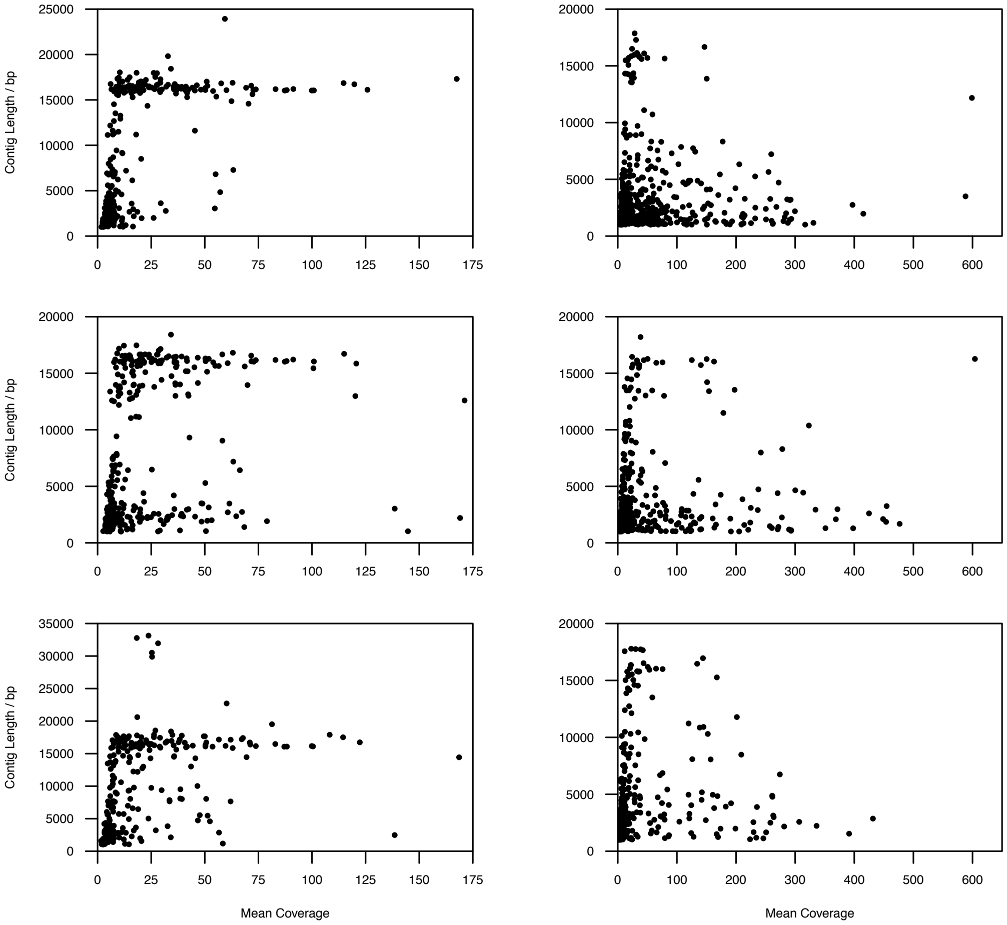


Figure 7.5 Coverage plots for *ChrysIber* assemblies; *ChrysoRL* left, *ChrysoAL* right. Top row IDBA; middle row NWBL; bottom row CA.

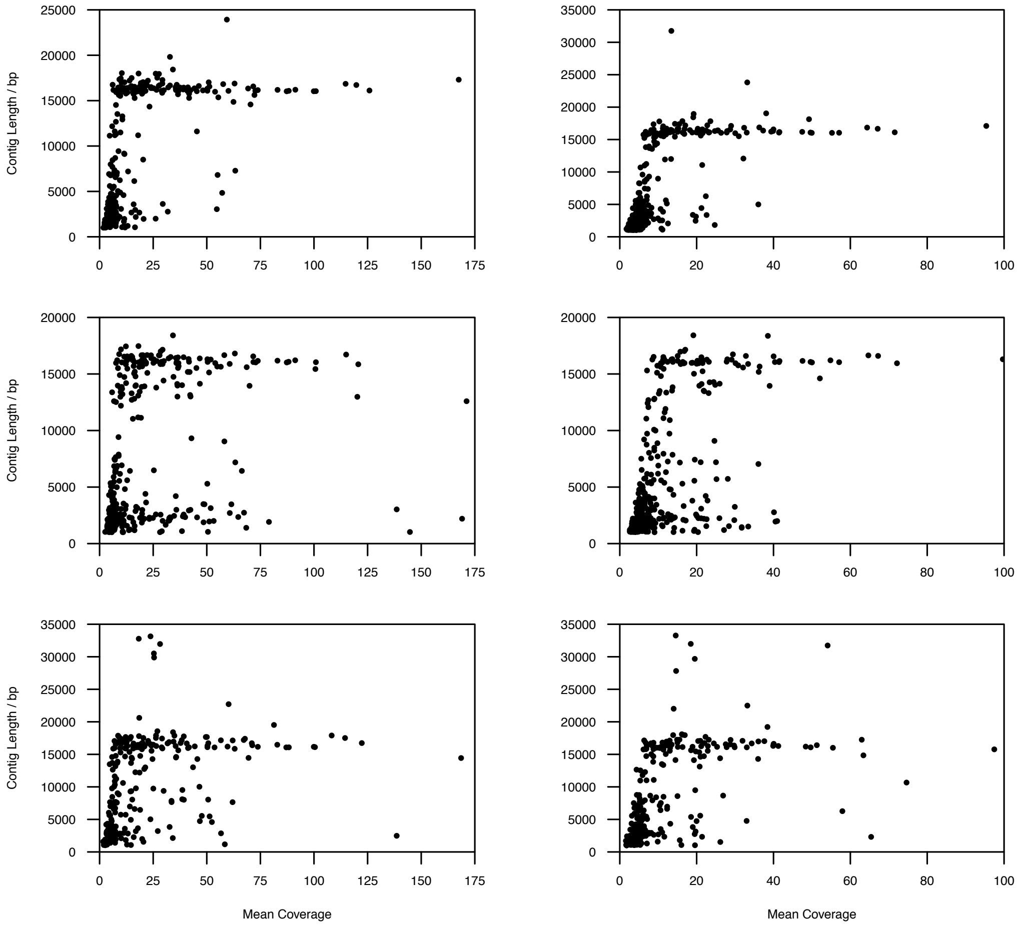


Figure 7.6 Coverage plots for *ChrysoRL* (left) and subsampled *ChrysoRL* (right). Top row IDBA; middle row NWBL; bottom row CA.

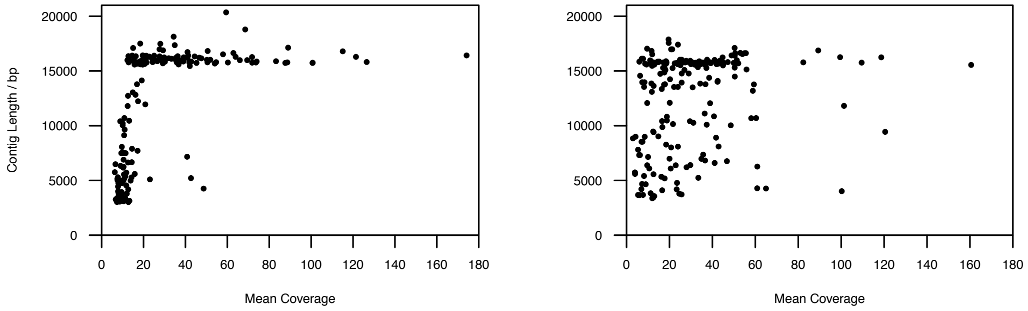


Figure 7.7 Coverage plots for *ChrysIber* assemblies presented in Gomez-Rodriguez et al 2015; *MitoRL*, left; *DeNovoRL*, right.

Chapter 8 Appendix 2

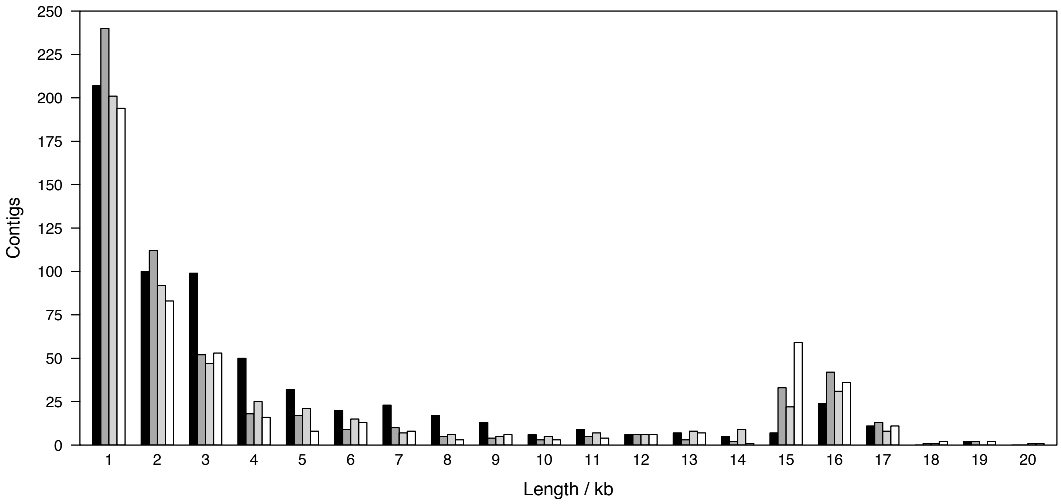


Figure 8.1 Contig length distributions for *BorneoCanopy*. CA: black; IDBA: dark grey; NWBL: light grey; Non-redundant: white.

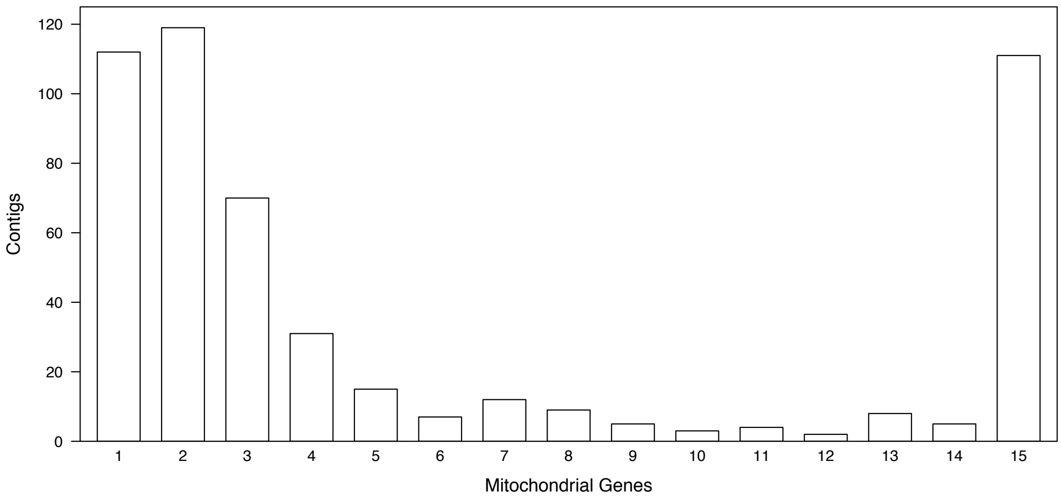


Figure 8.2 Mitochondrial genes per contig in the non-redundant set, *BorneoCanopy*.

Chapter 9 Appendix 3

Table 9.1 Species richness, phylogenetic diversity, and rarefied phylogenetic diversity for the *cox1* and *nad4* datasets.

	Ancient	*cox1*			*nad4*			Inclosure	*cox1*			*nad4*		
		SR	PD	PD_{rare}	SR	PD	PD_{rare}		SR	PD	PD_{rare}	SR	PD	PD_{rare}
Core	BWW	29	13.9	5.4	29	15.5	5.4	SOI	23	11.7	5.5	25	13.7	5.5
	MAW	23	12.4	5.6	25	13.6	5.5	HWI	16	9.4	5.4	16	10.3	5.5
	TTW	37	17.1	5.5	39	20.3	5.4	DLI	17	9.2	5.3	18	10.4	5.2
	WWW	23	12.4	5.7	23	14.2	5.6	NPI	14	8.7	5.6	16	10.5	5.5
Peripheral	ANW	19	12.4	6.3	23	12.3	5.4	SBI	24	16.8	6.6	27	14.7	5.3
	BSW	21	10.0	5.1	21	10.7	5.0	BSI	8	5.7	5.7	7	5.7	5.7
	HLW	20	13.3	6.4	21	11.1	5.2	STI	27	18.7	6.6	31	16.4	5.1
	PHW	25	11.8	5.2	24	13.1	5.3	HLI	32	16.4	5.4	24	14.9	5.5
	RSW	23	11.3	5.4	25	13.8	5.4	GLI	33	14.9	5.4	29	15.2	5.4
	SWW	20	11.1	5.6	20	12.6	5.6	BHI	15	7.9	4.9	13	8.2	5.0

Table 9.2 Multi-site compositional and phylo-beta diversity for the *cox1* and *nad4* datasets. Total beta diversity (β_{SOR}) is decomposed into its turnover (β_{SIM}) and nestedness (β_{SNE}) components.

	Multi-site beta						Multi-site phylo-beta					
	cox1			*nad4*			*cox1*			*nad4*		
	β_{SOR}	β_{SIM}	β_{SNE}	$p\beta_{SOR}$	$p\beta_{SOR}$	$p\beta_{SNE}$	$p\beta_{SOR}$	$p\beta_{SOR}$	$p\beta_{SNE}$	$p\beta_{SOR}$	$p\beta_{SOR}$	$p\beta_{SNE}$
Total	0.88	0.84	0.04	0.88	0.84	0.04	0.84	0.79	0.05	0.82	0.77	0.06
Anc.	0.78	0.73	0.04	0.78	0.74	0.04	0.72	0.66	0.06	0.69	0.62	0.07
Inclos.	0.82	0.75	0.07	0.82	0.75	0.07	0.76	0.67	0.09	0.74	0.65	0.10
Core	0.79	0.74	0.04	0.79	0.74	0.04	0.73	0.67	0.06	0.69	0.63	0.06
Peri.	0.82	0.74	0.07	0.82	0.74	0.08	0.77	0.67	0.09	0.75	0.63	0.11

Table 9.3 Compositional and phylogentic turnover between compartments for *nad4* and *cox1*.

		Compositional			*Phylogenetic*		
		β_{SOR}	β_{SIM}	β_{SNE}	$p\beta_{SOR}$	$p\beta_{SIM}$	$p\beta_{SNE}$
cox1	Habitat	0.2	0.2	0.0	0.2	0.18	0.02
	Position	0.22	0.12	0.10	0.20	0.09	0.11
nad4	Habitat	0.17	0.14	0.03	0.17	0.14	0.02
	Position	0.21	0.15	0.07	0.21	0.15	0.06

Table 9.4 Standardised effect sizes for estimates of phylogenetic structure; PD (phylogenetic diversity); MPD (mean pairwise distance); MNTD (mean nearest taxon distance). Significant positive SES values indicate overdispersion, significant negative SES values indicate clustering. Significant values highlighted in bold.

		PD_{SES}		MPD_{SES}		$MNTD_{SES}$	
		SES	p	SES	p	SES	p
cox1	Ancient	-1.14	0.13	-3.29	**0.005**	-0.50	0.31
	Inclosure	2.41	**1.00**	0.45	0.63	2.76	**1.00**
	Both	-3.24	**0.002**	-1.32	0.11	-2.89	**0.005**
nad4	Ancient	0.88	0.81	0.04	0.45	0.87	0.80
	Inclosure	1.80	**0.97**	0.75	0.77	1.93	**0.98**
	Both	-3.02	**0.002**	-2.10	**0.02**	-2.45	**0.009**

Figure 9.1 Phylogenetic distribution of species exclusively found in ancient (left) of inclosure (right) sites. Top: *nad4*; Middle: *cox1*; Bottom: *cox1*+BOLD.