Evolution in a Marine Gastropod:

Rocks, Clocks, DNA and Diversity

Simon Francis Kahu Hills

2010

A thesis presented in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Evolutionary Biology

Allan Wilson Centre for Molecular Ecology and Evolution
Institute of Natural Resources
Massey University
Palmerston North
New Zealand

ABSTRACT

Comprehensive integration of paleontological and molecular data remains a sought-after goal of evolutionary research. This thesis presents a dataset unlike any previously studied to document changes over time in the evolutionary history of the New Zealand marine mollusc genus *Alcithoe*.

In order to study evolutionary relationships in the *Alcithoe,* DNA sequence of approximately 8Kb of mitochondrial DNA was generated using universal and newly developed PCR primers. The gene composition of the resulting sequences has been thoroughly analysed, using a novel splits-based approach, to gain a clear understanding of the underlying phylogenetic signals in the data. Refinement of the phylogeny was achieved by considering subsets of both the taxa and genes. Taking these analyses into account the combined a robust phylogeny for the *Alcithoe* is presented for use in subsequent analyses.

The *Alcithoe* genus includes species that are exemplars of the problem of correctly identifying species by morphological traits, in both the living and extinct taxa. Taxonomic assignments were explored in a population level analysis of the highly morphologically variable species *A. wilsonae*. Analyses revealed that the various recognised forms of *A. wilsonae* are genetically indistinguishable and that the previously recognised species *A. knoxi* is a synonym of *A. wilsonae*. This result has significant implications for the interpretation of the paleontological data, as *A. knoxi* specimens are known from the Tongaporutuan stage (10.92 – 6.5 Ma) of the New Zealand geological timescale. Therefore, this finding also has implications on the assignment of calibration data in molecular clock analysis.

To ensure accurate estimation of divergence times and rates of molecular evolution, extensive explorations of parameter space in molecular-clock analyses were carried out. These analyses identified the most appropriate models and calibration settings for *Alcithoe* the dataset. The fossil data used to calibrate this analysis is amongst the most robust applied to molecular clock analyses to date. Statistical sampling uncertainty derived from the paleontological data was included in the calculation of prior distributions. Divergence dates inferred for the extant *Alcithoe* are largely consistent with the fossil record. However, the root of the tree was consistently inferred to be younger then expected. Rates of evolution in the species of *Alcithoe* included in this analysis are broadly consistent. However, some small rate differences are observed in some branches, for example, *Alcithoe fusus* appears to

have a faster rate then the rest of the genus. This rate increase is the likely cause of topological inconsistencies observed for four closely related taxa, including *A. fusus*, and indicates that slight rate differences can cause phylogenetic instability when small genetic distances are involved.

Direct comparison of diversification rates between the molecular and paleontological data for the *Alcithoe* illustrated that modern *Alcithoe* species have origins that are around 13 millions years younger then the oldest known *Alcithoe* fossils. The suggestion that *A. fusus* is descended from a series of fossil *Leporemax* species is directly contradicted by the molecular tree. In light of the molecular evidence this result highlights the problem of morphological convergence in the interpretation of fossil *Alcithoe* species. Comparison of the molecular and paleontological datasets was difficult for absolute speciation and extinction rates, as errors inherent to each dataset led to disparate estimates. For example, the fossil record clearly fails to record most recent speciation events observed in the molecular phylogeny, but the molecular data cannot sufficiently account for the amount of extinction evident in the fossil record. It is clear that the assumption of a constant and equal probability of speciation and extinction for all lineages is violated in the *Alcithoe*. However, the general long-term trends estimated for both datasets are concordant, and demonstrate an increase in both speciation and extinction rates over the Cenozoic era.

The research described in this thesis represents significant progress toward the goal of more thorough integration of molecular and paleontological in the study of evolution. I have shown that reconciliation of molecular and paleontological data is not only possible, but can substantially improve the resulting interpretation of evolution. This study is the broadest analysis of the evolution in a single genus using combined molecular and paleontological data that the author is currently aware of. It illustrates the advantage of having quality paleontological data to compare to emerging molecular data, and how the molecular data can further inform the paleontological data. Furthermore, it adds support to the shift in perspective from an adversarial to a complementary approach to the consideration of molecular and paleontological data. This thesis is a comprehensive first step in the synthesis of molecular and paleontological data in the study of evolution of the New Zealand mollusc fauna, and alludes to many promising avenues for future study.

ACKNOWLEDGEMENTS

Firstly to Mary, thank you for your supervision and support throughout the period of my thesis work. I'm sure I have been an 'entertaining' charge.

Many thanks to my co-supervisors Steve, James, and David for your continuing support and inspiration.

Thanks also to my collaborators Bruce, Alan, and Roger, your willingness and expertise have been instrumental in getting me to where I am now.

To the summer students, Melissa and Logan, who had been enlisted into various molluscan related projects, thank for your contributions and enthusiasm.

Klaus, Barbara, Tim, Bennet and Patrick, thanks to you guys for all your help over the years.

Thanks Joy, Susan and Trish, without you none of us would get anything done

Thanks to every one in the AWC, the work environment and camaraderie in the centre is without par.

Mum and Dad, it's been a while, but it looks like I might finally have finished being a student. Thanks for everything you've done to get me here.

Special thanks must be extended to the makers of fine single malt whisky, without whom levels of stress would have been much greater. Thanks also to Warren, for supplying said whisky, amongst other libations.

Finally, Gillian at the risk of being terribly clichéd, there are no words that adequately express my thanks to you. We have shared this arduous journey, and it has been all the richer for sharing it with you.

TABLE OF CONTENTS

4. Chapter Four

Molecular Clock Analysis of Alcithoe

5. Chapter Five

The evolution of New Zealand Volutes: comparison of molecular and paleontological evidence

6. Chapter Six

Reconciling paleontology and molecular biology

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

1 INTRODUCTION

1.1 THE GREAT DIVIDE IN EVOLUTION

Until recently evolution has been analysed from two distinct viewpoints. Paleontological studies observe morphological change through time, and can directly measure differences between extant and extinct morphologies. Molecular evolution, however, captures an instantaneous observation of extant organisms and must infer their phylogenetic relationships. This dichotomy of viewpoints has lead to an entrenched contrast in the way in which these disciplines interpret evolutionary processes, and has led to a situation where there has been little integration of the two disciplines. One point of view is that molecular biology is focused on the mechanisms of evolutionary change, while paleobiology studies the resulting patterns (see Grantham 2004).

An example of the differing viewpoints is seen in the discussion of macroevolution and microevolution (e.g. Reznick and Ricklefs 2009). Microevolution can be defined as the scale small genetic changes within a population that accumulate over time to produce variation, and macroevolution results in the patterns observed in lineages above the species level. The central premise of this dispute is whether processes of microevolution are sufficient to explain patterns of macroevolution. Erwin (2000) suggested that while some aspects of macroevolution can be explained by extended periods of microevolution (eg clade replacement through competition), others could not (punctuated evolution, species or lineage sorting, and mass extinctions). It is argued that these discontinuities impart a structure to

evolution that the processes of microevolution are unable to account for. The counter argument states that current processes can explain paleontological patterns (Leroi 2000; Penny and Phillips 2004), it is simply a process of accumulation of change over time and sorting by natural selection.

In the recent past the primary interface between molecular biology and paleontology has been through molecular clock analysis. A significant source of disagreement has been seen in the disparity in estimation of divergence times between molecular and paleontological data sets. In the study of the divergence of major mammalian lineages, molecular clock analysis suggested that orders diverged quite some time before they appear in the fossil record (Kumar and Hedges 1998). Through quantitative study of the mammalian fossil record Foote *et al.* (1999) concluded that it was unlikely that palaeontologists had missed representatives from the earlier time period indicated by molecular clock studies. Furthermore, they hypothesized that the molecular patterns observed could be a result of punctuated evolution and rapid cladogenesis. Such patterns are seen in phylogenies where several branches radiate from a very nearly common origin with little support for any resolution between them. However, there is little support for the phenomenon of punctuated molecular evolution (but see Pagel 2006). In his review of gastropod phylogenetics Wagner (2001) championed gastropod molluscs as a group that would be informative for testing the hypotheses of Foote et al. (1999).

Patterns of diversity in living and extinct taxa have long been the focus of analysis to elucidate the underlying processes of speciation and extinction (e.g. Raup 1985). Significant work in the last few decades has focused on the identification and understanding of bias in the fossil record, and on developing sophisticated methods of studying the patterns of historical diversity captured therein (e.g. Alroy 2000; Foote 2000; Jablonski 2000; Jackson and Johnson 2001; Crampton et al. 2003; Cooper et al. 2006; Valentine et al. 2006). The transition into methods utilising phylogenies (e.g. Wagner 2000; Ricklefs 2007) allowed molecular based techniques to gain traction on questions regarding diversification patterns. Development of Birth/Death models of diversification has made it possible to calculate speciation and extinction rates from molecular data (Nee 2006), facilitating direct comparisons of independent inferences derived from molecular and paleontological data.

In order to generate a robust evolutionary synthesis these disciplines must be reconciled. There has been a growing call to address the lack of integration between paleontology and molecular biology (e.g. Donoghue and Benton 2007; Wiens 2009). However, opportunities for reconciliation are few. It requires a continuous, well preserved, well studied fossil record with living representatives existing in a system with limited immigration. Fortunately the New Zealand marine mollusc fauna fit all of these requirements.

1.1.1 Molluscs: A Tool

As a tool for the study of evolution in general, molluscs represent a powerful asset. As Lindberg and co-workers outlined in their introduction to the Mollusca for the tree of life (Lindberg et al. 2004), the diversity and history of the clade is perhaps unmatched by any other. With approximately 200,000 extant species, the phylum Mollusca is second only to Arthropoda in size. A huge diversity of physiological, behavioural and ecological adaptations has allowed members of the group to exploit a wide range of habits and habitats, producing an abundant array of variation. Possibly the greatest strength of molluscs for studying evolution is in an unsurpassed fossil record, that has captured the morphological change in molluscs over the last 560 million years. In fact the Mollusca exhibit the greatest diversity of any fossilisable marine phylum (Bouchet et al. 2002). In addition, due to the long interest by malacologists and amateur shell collectors, the quality of knowledge regarding patterns of marine molluscan species richness is amongst the best for any phylum.

1.1.2 Molluscs and molecules

Comparatively little molecular study has been devoted to molluscs when considering their diversity and the wealth of neontological and paleontological research carried out on the group. However, there has been a significant increase in the output of molecular data for molluscs in the last 10 – 15 years (Ponder and Lindberg 2008). Molluscan molecular datasets have increased in size over this time, but generally consist of a set of 'universal' genes with the addition of a small number of more recently developed markers. The core set of genes consists of the mitochondrial genes *cox*1, *cyt*B, 16S and 12S, and the nuclear genes 28S, 18S and the ITS regions. Various combinations of these genes have been used in the vast majority of molecular phylogenetic reconstructions, ranging from very deep inter-

class studies (e.g. Lydeard et al. 2000; Passamaneck et al. 2004) up to within species population analysis (e.g. Nakano and Spencer 2007; Tuan and dos Santos 2007). Several additional nuclear markers have been developed and are becoming more widely utilised such as elongation factor 1-alpha (ef-1a), histone H3 (Colgan et al. 2007) and actin (Adema 2002). Due the great breadth of evolutionary time over which these markers are applied they are, by necessity, obtained from highly conserved parts of the genome. While a very small number of groups have developed markers specific to their study groups (e.g. Strugnell et al. 2004; Imron et al. 2007; Bandyopadhyay et al. 2008), quality assessment of genes used as sequencing markers in molluscs is not often carried out. As the generation of sequence data becomes easier it will become necessary to select markers that have appropriate information content to address the questions posed by researchers.

An emerging area of interest in the molecular analysis of molluscan relationships is the sequencing of complete mitochondrial genomes (see appendix 1). Complete mitochondrial genomes have been found to be critical to the elucidation of the early divergences in birds and mammals (Phillips et al. 2006; Pratt et al. 2009). The amount of phylogenetic signal contained in the mitochondrial genome approaches a sufficient quantity to resolve ancient rapid divergences in these groups. In order to resolve the adaptive radiations in the Mollusca, many of which significantly predate bird and mammal radiations, further development of mitochondrial genome sequencing will be required. To date there have been very few considerations of molluscan mitochondrial genome datasets (e.g. Medina and Collins 2003; Cunha et al. 2009). All these studies have been limited by the small number of molluscan mitochondrial genomes currently available (only 74 representing all Mollusca, on GenBank as of the end of 2009). In addition to the nucleotide sequence of the genome, structural rearrangements are likely to be informative for deep divergences (Boore and Staton 2002; Boore and Fuerstenberg 2008). Indeed, significant rearrangements have been reported in several molluscan mitochondrial genomes published to date (e.g. Hoffmann et al. 1992; Boore and Brown 1994; Hatzoglou et al. 1995; Terrett et al. 1996; Boore et al. 2004; Knudsen et al. 2006).

1.1.3 Comparative Analysis

Although there are few studies reporting combined analysis using both molecular and morphological data, congruence between datasets is uncommon (Ponder and Lindberg 2008). This incongruence reflects the lack of resolution and homplasy observed in both datasets. Morphological data is hampered by issues such as convergence and crypsis. However, the main factor limiting molecular data is that the markers that have been utilised so far are not necessarily capable of clear resolution at the evolutionary depths to which they are applied. Furthermore, it is a non-trivial exercise combining data from these two fields (Grantham 2004). Morphologically based phylogenies pose little problem, except that characters are often under selection and prone to convergence, but it remains unclear as to what is the most appropriate method of integrating stratigraphic data. Grantham makes the very obvious suggestion that stratigraphic data might be relevant in the assessment of phylogenies.

Comparative analysis of molecular and morphological data will continue to play a central role in the study of molluscan evolution. The ultimate goal in the study of a lineage will be to consider a complete picture of the evolutionary history, including both extant and extinct representatives.

1.1.4 Species Delimitation

A key question in comparative analysis is whether species, as recognised by skeletal morphology in the fossil record, are generally the same as those recognised biologically in extant taxa. In molluscs there is some evidence to suggest that such concordance exists, but there are many opposing examples (Jablonski 2000).

Species delimitation is a difficult question that is the source of significant debate. The main issue is that species definitions vary depending on the angle of study (de Queiroz 2005). For example, from a molecular perspective a set of species could be defined by having reciprocal monophyly whereas the biological species concept is defined by reproductive isolation.

Primarily the problem is that some taxonomic characters will not be strongly reflective of true phylogeny. In addition, single-gene analysis only provides a gene tree not necessarily a species tree. This situation will lead to conflict between the

two data sets. A less straightforward problem is when taxonomic characters resolve at a different level to molecular data. At what level of variation should separate species be identified, and what is the difference in variation between species and ecotypes. Another issue is the effect on diversity estimates that an erroneously large number of species would have. This problem arises when morphological variation in the fossil record could be interpreted as many species rather then a single polymorphic species. Molecular analysis of extant taxa can help define the morphological limits of a species to correct for this error.

1.1.5 Molecular Clocks and Rates of Molecular Evolution

Since it was recognised in the early 1960's that the amount of difference between two sequences (initially amino acid sequences, but equally relevant to nucleotide data) could be a function of the time since the divergence of those sequences (Zuckerkandl and Pauling 1962), molecular clock analysis has been used to study evolution. For the past forty years there has been much debate in the literature as to the accuracy and general applicability of molecular clock analysis (e.g. see Bromham and Penny 2003; Kumar 2005). Concerns generally revolved around complications due to the processes of molecular substitutions and uncertainties in the fossil record (Magallon 2004). Specific examples include the generation time effect (where a lineage that progresses through more generations in a given time than another will appear to have a higher rate of molecular change), an inverse relationship between body size and rate, and a relationship between higher metabolic rate and higher rate of change (Martin and Palumbi 1993; Mooers and Harvey 1994; Bromham et al. 1996).

Molecular-clock analysis provides a framework for estimating rates of molecular evolution and inferring divergence times of populations and species. The molecular clock is used as a null hypothesis for testing evolutionary rate. That is, assuming that the accumulation of DNA substitutions over time occurs at a more or less constant rate, when a difference is detected between sequence data sets, or taxa data sets, then it can be inferred that something of evolutionary significance has occurred. Early molecular clock research focused on comparisons between one of a few species from a small number of main groups. These studies also utilised models that assumed a constant rate across all lineages. Often divergence dates inferred from molecular data were older than estimates obtained from fossil data. To some

extent this can be expected, as fossil dates represent minimum divergence estimates (Blair and Hedges 2005). However, the discrepancies observed between molecular and fossil data were often too large to be explained completely by this underestimation, leading to the suggestion that there were gaps in the fossil record (Magallon 2004). Additionally, more recent molecular clock studies have explored relaxed clock models (Drummond et al. 2006). These models avoid the assumption of a global molecular clock and therefore allow for variable rates of change across a studied lineage. Previously molecular-clock analysis required sequence data that passed rate-consistency tests across all taxa involved. Relaxed-clock models now allow studies where rates might vary, and produce more robust dating estimates. In addition, indications of rate differences may be biologically significant.

Molecular-clock analysis is a rapidly developing field of study, and available methods are far from perfect. There is no best method, all have advantages and disadvantages that restrict how and where they should be used. Furthermore, quality molecular-clock results are dependent on appropriate assumptions and good prior data associated with those assumptions. This dependency is particularly true of Bayesian methods where parameter space can be very large.

1.1.6 The New Zealand marine mollusc fauna

Several factors make New Zealand marine molluscs particularly useful for the study evolutionary patterns and processes. Both living and fossil mollusc lineages are well studied (e.g. Powell 1979; Beu and Maxwell 1990; Spencer and Willan 1995). The New Zealand fauna represents a relatively closed system, where immigration has been rare. This isolation has been the result of the long time (60 – 80 million years) since the separation of Zealandia from the Gondwana super-continent. When immigration is evident it has usually been due to the west to east circum-polar current (Fleming 1979). This isolation has led to an increasing number of endemic species observed through the Cenozoic fossil record (Cooper and Millener 1993). In the New Zealand biota endemism is thought to approach 100% for many groups (Daugherty et al. 1993), with molluscs in the order of 85% (Spencer et al. 2009). Therefore it is likely that most species found in the New Zealand evolved locally, and that changes in species diversity should be a result of local processes, not the result of invasive migrations. The fossil record of New Zealand Cenozoic marine molluscs is amongst the best in the world, but to date there have been few

molecular studies carried out on extant New Zealand taxa (e.g. Michaux 1987; Donald et al. 2005; Nakano et al. 2009).

1.1.7 A summary of the New Zealand Cenozoic fossil record

The New Zealand geological timescale is divided in various series and stages based on rock formations found in New Zealand. This timescale is described in detail in Cooper (2004). Different stages are generally characterised by faunas that occur in rocks from a given time. Stage boundaries are recognised where there is clear observable faunal turnover from one stratigraphic sequence to another. In addition the prevalence of warm or cold water taxa, often in combination with oxygen isotope data, can be used to infer the climatic condition at a given time. The general habitat of fossilised organisms can be inferred from the nature of the rock in which they are found. Different environmental conditions lead to different types of sediment, for example deep-water sediments are readily distinguishable from shallow-water sediments. Paleoenvironmental data combined with fossil locality information can be used to infer historical biogeographic patterns, particularly in groups with a rich fossil record.

The New Zealand volute lineage leading to the *Alcithoe* is believed to have originated in the early Cenozoic. Figure 1.1 shows the geological timescale for the New Zealand Cenozoic, and follows Cooper et al. (2004). In order to provide a general environmental context for the evolution of New Zealand volutes I will briefly discuss some relevant features of the New Zealand stages with regards to the mollusc fauna, summarised from Beu and Maxwell (1990), and Cooper (2004). Figure 1.2 accompanies this discussion.

A subtropical climate prevailed during the Eocene, but temperatures appear to be been declining. A sudden increase in molluscan generic diversity in the mid Eocene indicates a possible large influx of taxa (possibly tropical immigrants) during the Bortonian, but many of these taxa do not survive into the Kaiatian. The Bortonian fauna is known from a range of water depths and exhibits tropical elements while the Kaiatian fauna is known from a more limited range of facies, none of which represent shallow water environments.

Geological Timescale			Age of stage base (Ma)
	Epoch	Stage	
Neogene	Holocene	Haweran	0.34
	Pleistocene	Castlecliffian	1.63
	Pliocene	Nukumaruan	2.4
		Mangapanian	3
		Waipipian	3.6
		Opoitian	5.28
	Miocene	Kapitean	6.5
		Tongaporutuan	11
		Waiauan	12.7
		Lillburnian	15.1
		Clifdenian	16
		Altonian	19
		Otaian	21.7
		Waitakian	25.2
Paleogene	Oligocene	Duntroonian	27.3
		Whaingaroan	34.3
	Eocene	Runangan	36
		Kaiatan	37
		Bortonian	43
		Porangan	46.2
		Heretaungan	49.5
		Mangaorapan	53
		Waipawan	55.5
	Paleocene	Teurian	65

FIGURE 1.1—The New Zealand geological timescale (after Cooper 2004). New Zealand geological Stages are shown next to the Epochs of the global geochronogical scale. The currently recognised lower boundaries of each stage are given.

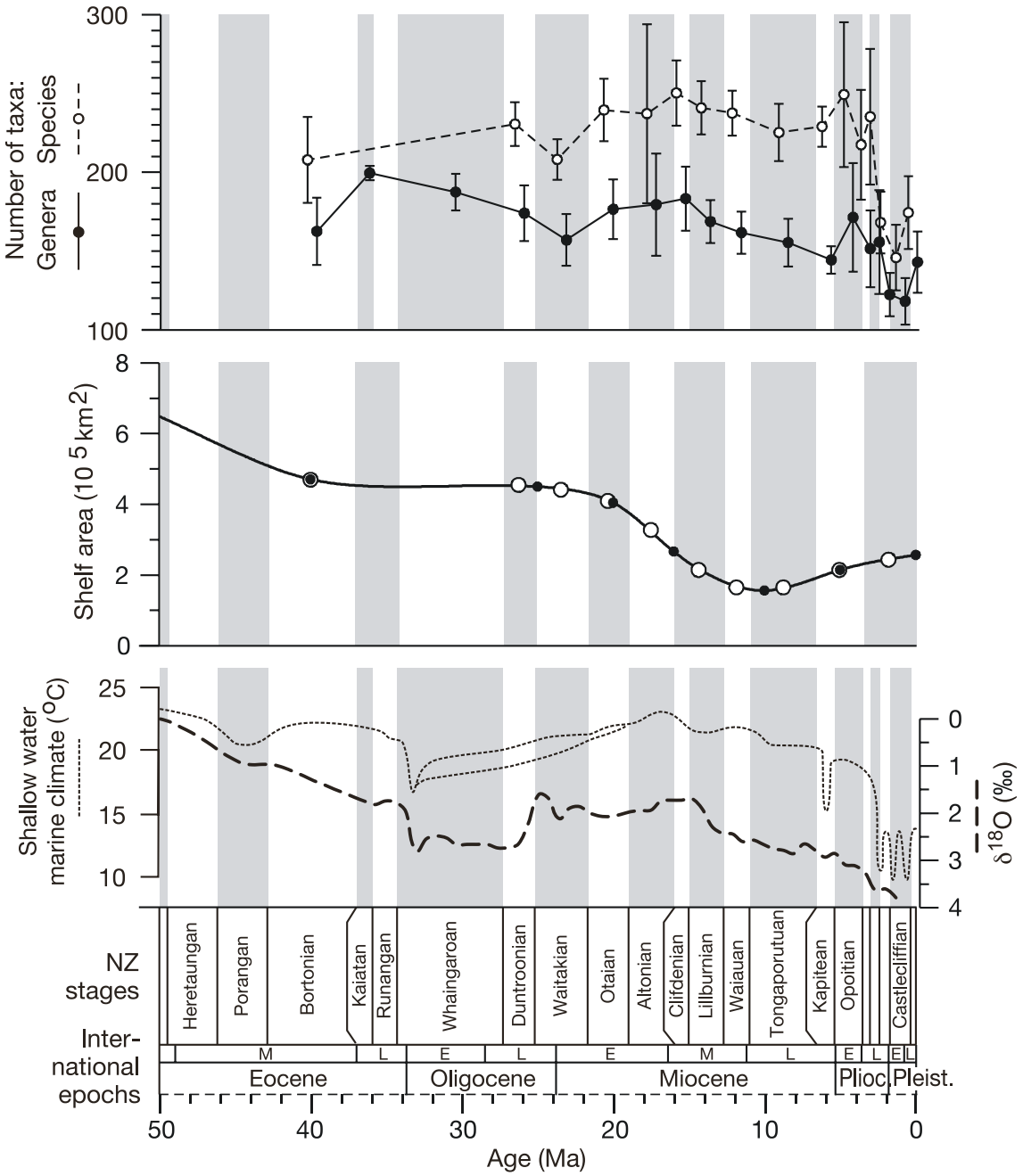

FIGURE 1.2—Summarised trends of diversity and environmental change in New Zealand marine habitat during the Cenozoic (from Figures 12 and 13 of Crampton et al. 2006). Per stage, sampling-standardised diversity curves illustrate the fluctuating species and genus level diversity. The estimated shoreline to shelf-edge area provides a proxy for the changing amount of available habitat. Two climate indicators are shown; global ocean climate, as inferred from foraminifer oxygen isotope data, and a local shallow water marine temperature curve for New Zealand. The two traces shown for the shallow water marine temperature in the Oligocene and early Miocene represent the south and west of the South Island (upper trace) and the east coast of the South Island (lower trace).

Fluctuations in the general climate and a rapid decrease in the temperature of the sea bottom at the beginning of the Whaingaroan marks the Eocene/Oligocene boundary. Few fossils are known from the Whaingaroan stage as a result of little available marine sediment to allow fossilisation, likely due to a significant restriction in landmass at the time. The Duntroonian sees two major events in the geological history of New Zealand. Firstly the peak of the marine transgression occurs during this stage. Secondly, a significant sea-level change is observed, known as the mid-Oligocene unconformity (or Marshall Paraconformity). This event has been linked to a major ocean current shift associated with the onset of the circum-polar current. Following the mid-Oligocene unconformity the greatest increase in Cenozoic molluscan diversity in New Zealand is recorded. However, in more recent work (Crampton et al. 2006) this increase begins in the Waitakian. It is not clear how many of these newly recorded taxa may have originated prior to the Duntroonian, in the poorly sampled Whaingaroan. Duntroonian and Waitakian taxa represent a rich record sampled from a range of habitat types. The climate during this stage is noted as being warm-temperate to subtropical at southern South Island latitudes.

In the early Miocene the onset of movement in the Alpine fault led to a considerable change in sedimentation. For the first time since the Bortonian significant faunas are found in northern North Island localities, and differences in the compositions of the warm northern and cold southern faunas are sufficient to merit the recognition of two distinct provinces. In general this period of time represents one of the warmest in the New Zealand Cenozoic, peaking during the Altonian. In the Otaian warm water was restricted to the northern North Island, but by the late Altonian warm water had spread as far south as Southland latitudes. This warm climate continued into the Clifdenian, which has a less well sampled fauna than earlier stages, and mostly represents deep-water taxa. The climate is then thought to have cooled in the Lilburnian. The faunas of these two stages, and the subsequent Tongaporutuan, are similar, with few discriminating taxa. In addition, the fauna of the Tongaporutuan are generally poorly preserved. The lower Kapitean marks the terminal Miocene glaciation of west Antarctica. While New Zealand did not undergo glaciation at this time, there was a significant increase in the circulation of cool water into the New Zealand marine system. This period low of temperature

coincided with an observed molluscan diversity minimum for the Cenozoic (evident in the genus level curve of Figure 1.2, but not the species curve).

The early Pliocene is represented by shallow, warm water Opoitian faunas. A clear generic turnover marks the boundary of the Waipipian and Mangapanian stages in the late Pliocene. Many of the last remaining Miocene taxa become extinct by the end of the Mangapanian at a time when cold-water taxa characteristic of the Pleistocene abruptly appear. The boundary between the Mangapanian and Nukumaruan stages is not well defined, but Nukumaruan faunas are widespread and represent habits from the near-shore to bathyal depths. During this stage the marine separation of the Northern and Southern landmasses, known as the Manawatu Strait, closed. This change in geography allowed cool water currents to range as far north as the central Hawkes Bay and the Wanganui basin. As a result a high component of warm water taxa are restricted to the north.

The boundary between the Nukumaruan and the Castlecliffian is well defined by extensive generic turnover. The Castlecliffian fauna differs from the Nukumaruan by 17 molluscan genera, the most dramatic faunal turnover in New Zealand during the Cenozoic. An increase in diversity is seen in the Castlecliffian, but to some extent this is due to better preservation of high-energy near shore environments and an influx of planktonic larvae from the East Antarctic and circum-polar currents. Many of the modern taxa originate in the Castlecliffian stage. The major climatic feature of this stage is the period of Pleistocene glaciation. The most recent stage, the Haweran, is mostly represented by near-shore or beach faunas. The composition of these faunas is largely the same as is present in the modern fauna. The main difference observed between the Haweran and the present day faunas is the distribution of taxa, with extensive north/south migration of warm and cold water species ranges in association with glacial cycling.

1.1.8 Volutes in New Zealand

The Volutidae are a group of direct developing, carnivorous neogastropods. A habitat preference of soft sediment means that they are generally well represented in the fossil record. Volutidae are known from throughout the Cenozoic in New Zealand, but the modern fauna is dominated by the genus *Alcithoe* (see Figure 1.3). Only two volute species from different genera are known from the main continental

waters of New Zealand, *Iredalina mirabilis* and *Zygomelon zodion*. A futher two genera (*Lyria* and *Calliotectum*) are known from the more subtropical waters around the Kermadec Islands, which are considered to be included in the New Zealand waters. Reviewing the volute fossil record of New Zealand shows that this has not always been the case. Up to 12 genera and sub-genera are represented in the fossil record. Two of these, *Athleta* and *Lyria* are found in sub-tropical to tropical waters in the present day. The majority of the groups seen in the fossil record are now extinct.

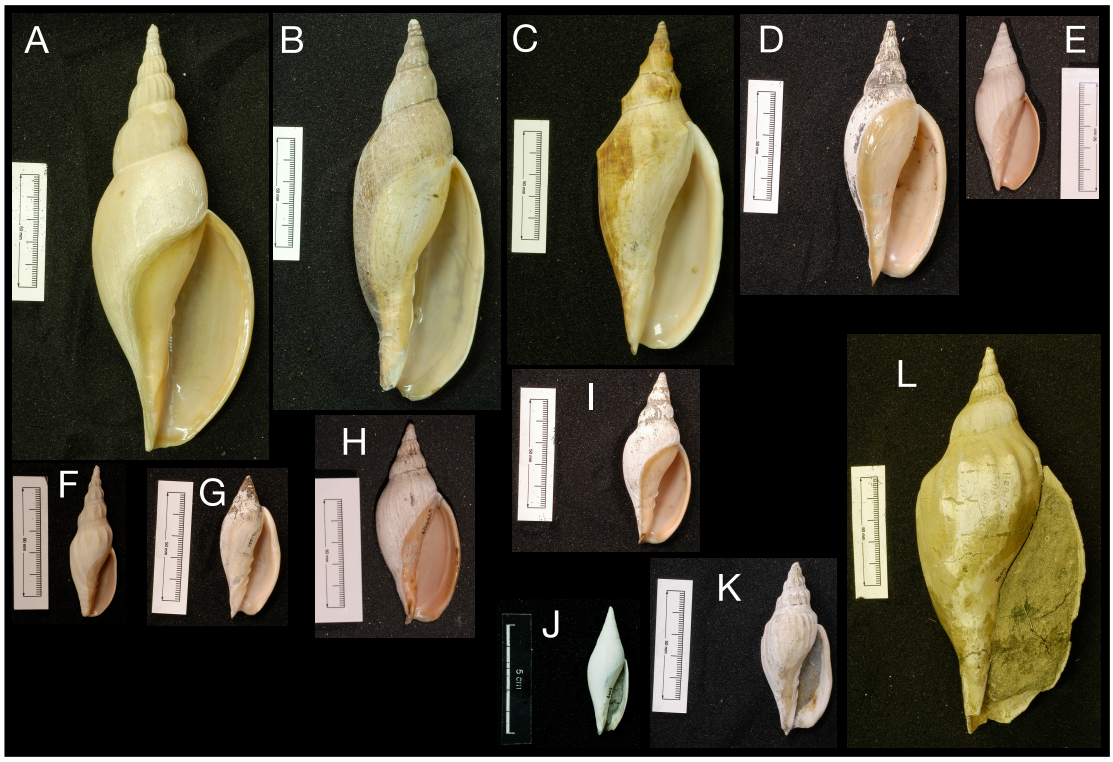

FIGURE 1.3—Examples of the New Zealand Volutidae. (A) *Alcithoe benthicola* [MM90140.A], (B) *Alcithoe fissurata* [M90198.D], (C) *Alcithoe jaculoides* [M110357.B], (D) *Alcithoe larochei* [M88353.B], (E) *Alcithoe wilsonae* [M190079], (F) *Alcithoe flemingi* [M33598.B], (G) *Alcithoe fusus* [M 21661.C], (H) *Alcithoe pseudolutea* [M275009.1], (I) *Alcithoe lutea* [MM95306.B], (J) fossil *Teremelon tumidor* [GS59569], (K) fossil *Alcithoe hurupiensis* [GS5626.A], (L) fossil *Spinomelon parki* [M11598.A]. Shells are shown approximately to scale, and a 5 cm scale is included in each image for reference. These shells are from collections kept at the Museum of New Zealand Te Papa Tongarewa and the Institute of Geological and Nuclear Sciences.

The various species of *Alcithoe* both extant and extinct have been taxonomically categorised (Beu and Maxwell 1990; Spencer and Willan 1995). In addition Bail and Limpus (2005) have produced an iconography of the living New Zealand volutes. This publication presents the group from a conchological point of view, and probably divides species further than is necessary. The morphology of *Alcithoe arabica*, as described by Ponder (1970) provides an excellent basis from which to make comparative studies. There have been no previous molecular studies using this group, and molecular analysis will help clarify the taxonomic status of several currently recognised species.

1.1.9 The taxonomic context of Alcithoe

In addressing the evolutionary history of the *Alcithoe*, it is perhaps best to review the higher level taxonomic grouping of the lineage. The genus *Alcithoe* H. & A. Adams, 1853 is part of the tribe **Alcithoini** Plisbry & Olsson, 1954. This tribe also includes the recently described monotypic genus *Zygomelon* Harasewych & Marshall, 1995 and most of the extinct genera found in the New Zealand fossil record. The tribe **Alcithoini** is included in the subfamily **Zidoninae** H. & A. Adams, 1853. Several southern hemisphere tribes are also included in the **Zidoninae**. The tribes **Adelomelonini** Pilsbry & Olsson, 1954 and **Odontocymbiolini** Clench & Turner, 1964 are found off South America. Members of tribe **Cymbiini** H. & A. Adams, 1853 are found off southern Africa. From southern Australia there is the tribe **Livoniini**. A circum-polar tribe **Zidonini** H. & A. Adams, 1853 includes the New Zealand genus *Iredalina* Finlay, 1926. Darragh (1989) suggested that several Australian genera, *Cymbiola*, *Melo*, *Ericusa* and *Livonia*, which had previously been included in other subfamilies should be included in the **Zidoninae**. Further more Darragh indicated that anatomical features of the members of the subfamily **Amoriinae** suggest a close relationship with the genus *Cymbiola*, and therefore with the **Zidoninae**. It has been proposed that the origins of the **Zidoninae** might have been the result a circum-Antarctic volute radiation during the late Cretaceous (Bondarev 1997).

1.1.10 A history of Volutes in New Zealand

A picture of the current understanding of the fossil record of *Alcithoe* and its relatives and ancestors can be extracted from Beu and Maxwell (1990). The earliest

example of a member of the tribe Alcithoini is an unnamed species attributed to the genus *Teremelon*, from the Paleocene (65 – 55.5 My BP) (Beu and Maxwell 1990). The last record of the *Teremelon* genus *sensu stricto* is from the Clifdenian stage, ending 15.5 My bp. Two genera appear in the Bortonian stage of the middle Eocene, *Waihaoia* and *Mauira*. *Mauira* appears in shallow water faunules in the Bortonian, following which there is an 18 My gap before members of the genus are again seen in the Altonian (19 – 15.9 Mybp). There is then a continuous record of *Mauira* species until the end of the Tongaporutuan (10.92 – 6.5 Mybp). *Waihaoia* has a somewhat patchy record through to the Altonian. The genus *Spinomelon* first appears in the Duntroonian (27.3 – 25.2 Mybp). This genus is characterised by a distinctive apical spike on the protoconch that results from the shedding of the initial whorl (Beu and Maxwell 1990). Fossils attributed to this genus are prevalent through to the Altonian, after which the record becomes poor. The last example of *Spinomelon* was thought to have existed in the Mangapanian (3.0 – 2.4 Mybp), however the structure of the protoconch of the extant *Alcithoe benthicola* has leaves open the possibility that this species could be directly derived from the *Spinomelon* lineage. The first probable occurrence of a specimen attributable to the genus *Alcithoe* is in the Duntroonian. This unnamed species lacks a protoconch, so it is not possible to make a decisive assessment as to whether it is an example of *Alcithoe* or *Spinomelon*, but there is a high degree of similarity to the younger species *Alcithoe turrita*. A 3.5 My gap is observed in the fossil record between the first probable occurrence of *Alcithoe* and the occurrence of *Alcithoe turrita*, in the Otaian (21.7 – 19 My bp). Several species of *Alcithoe* are recognised though to the end of the Altonian, following which there is hiatus until the Waiauan (12.7 – 10.92 Mybp). After this there is a continuous record of *Alcithoe* until recent times. The Otaian stage also sees the first occurrence of *Leporemax*, a subgenus of *Alcithoe*. Species of *Leporemax* are differentiated from *Alcithoe sensu strico* by a smaller size and a more slender shape. Fossil occurrences of *Leporemax* have a similar pattern to those of *Alcithoe*. Several species are named in the Altonian, with the divergence of at least one species into the Clifdenian. A gap in the record follows, until the Tongaporutuan (10.92 – 6.5 Mybp) in which the record is resumed and is continued up to the present, most recently in the form of *Alcithoe (Leporemax) fusus*. The fossil record also includes occurrences of some tropical genera not closely related to *Alcithoe*. *Athleta* and *Lyria* both occur in the early Cenozoic record. While they are not part of the tribe Alcithoini, they are useful indictors of more tropical climatic conditions.

1.2 THESIS STRUCTURE

This first chapter has briefly introduced the thinking into which this thesis fits, and introduces the study system, the New Zealand volute genus *Alcithoe*. The following chapters will present comprehensive analysis of the evolutionary relationships in the *Alcithoe* from a molecular perspective, and compare patterns observed in the molecular data with paleontological data.

Chapter 2 describes an extensive sequencing of *Alcithoe* mitochondrial genome, requiring the development of PCR primers capable of generating long-range PCR products from *Alicthoe* and extraction of DNA samples of adequate quality. The genes comprising the resulting sequence fragments have been assessed for the suitability to robustly infer the phylogeny of the *Alcithoe*. For this assessment a novel splits-based approach was developed to analyse the amounts of signal and noise in the data. These analyses illustrate the importance of exploring the underlying signal in molecular data. The result of this work is a dataset in which the underlying signal is well understood and infers a robust phylogeny of the *Alcithoe*.

In chapter 3 the validity of the species assignments of two key living *Alcithoe* taxa is tested. *A. wilsonae* is known to be a morphologically variable species, and initial sequences showed little genetic differentiation observed between this species and *A. knoxi*. As *A. knoxi* is represented in the fossil record this species assignment required further analysis to ensure correct assignment of a critical calibration node for molecular clock analysis. A population-level analysis proves that these taxa are a single species that diverged approximately 10 million years before present, and therefore are appropriate to calibrate with the Tongaportuan appearance of '*A. knoxi*' in the fossil record. This work demonstrates the use of molecular data to refine species assignments and clarify the interpretation of the fossil record.

The focus of chapter 4 is an extensive examination of parameter space in the Bayesian molecular-clock analysis of the *Alcithoe*. The ultimate goal of this analysis is the inference the most robust divergence date and molecular rate estimates possible with the data at hand, both molecular and paleontological. This study will serve as a demonstration of the technique of evaluating alternative parameter regimes in a Bayesian framework.

Chapter 5 presents a direct comparison of molecular and paleontological data for *Alcithoe* in an analysis of rates of speciation and extinction. The degree of congruence in speciation and extinction rates derived from these datasets is assessed. A high degree of incongruence between absolute rates appears to result from different biases in each dataset, but overall there is a general agreement in increasing long-term rates of speciation and extinction.

Finally, chapter 6 summarises the major findings and relevance of this work to evolutionary biology. In addition, several avenues of further study are introduced.

1.3 REFERENCES

Adema, C. M. 2002. Comparative study of cytoplasmic actin DNA sequences from six species of Planorbidae (Gastropoda : Basommatophora). Journal of Molluscan Studies 68:17-23.

Alroy, J. 2000. New methods for quantifying macroevolutionary patterns and processes. Paleobiology 26:707-733.

Bail, P., Limpus, A. 2005. The recent volutes of New Zealand with a revision of the genus *Alcithoe* H. & A. Adams, 1853 in G. T. Poppe, and K. Groh, eds. A Conchological Iconography. ConchBooks, Hackenheim.

Bandyopadhyay, P. K., Stevenson, B. J., Ownby, J. P., Cady, M. T., Watkins, M., Olivera, B. M. 2008. The mitochondrial genome of *Conus textile*, *cox*I-*cox*II intergenic sequences and Conoidean evolution. Molecular Phylogenetics and Evolution 46:215-223.

Beu, A. G., Maxwell, P. A. 1990. Cenozoic Mollusca of New Zealand. New Zealand Geological Survey Paleontological Bulletin 58:1-518.

Blair, J. E., Hedges, S. B. 2005. Molecular clocks do not support the Cambrian explosion. Molecular Biology and Evolution 22:387-390.

Bondarev, I. 1997. Systematics of the Volutidae. La Conchiglia 282:32-44.

Boore, J. L., Brown, W. M. 1994. Complete DNA-sequence of the mitochondrial genome of the black chiton, *Katharina tunicata*. Genetics 138:423-443.

Boore, J. L., Fuerstenberg, S. I. 2008. Beyond linear sequence comparisons: the use of genome-level characters for phylogenetic reconstruction. Philosophical Transactions of the Royal Society B-Biological Sciences 363:1445-1451.

Boore, J. L., Medina, M., Rosenberg, L. A. 2004. Complete sequences of the highly rearranged molluscan mitochondrial genomes of the scaphopod *Graptacme eborea* and the bivalve *Mytilus edulis*. Molecular Biology and Evolution 21:1492-1503.

Boore, J. L., Staton, J. L. 2002. The mitochondrial genome of the sipunculid *Phascolopsis gouldii* supports its association with Annelida rather than Mollusca. Molecular Biology and Evolution 19:127-137.

Bouchet, P., Lozouet, P., Maestrati, P., Heros, V. 2002. Assessing the magnitude of species richness in tropical marine environments: exceptionally high numbers of molluscs at a New Caledonia site. Biological Journal of the Linnean Society 75:421-436.

Bromham, L., Penny, D. 2003. The modern molecular clock. Nature Reviews Genetics 4:216-224.

Bromham, L., Rambaut, A., Harvey, P. H. 1996. Determinants of rate variation in mammalian DNA sequence evolution. Journal of Molecular Evolution 43:610-621.

Colgan, D. J., Ponder, W. F., Beacham, E., Macaranas, J. 2007. Molecular phylogenetics of Caenogastropoda (Gastropoda : Mollusca). Molecular Phylogenetics and Evolution 42:717-737.

Cooper, R. A. 2004. The New Zealand geological timescale. Pp. 1-284. Institute of Geological and Nuclear Sciences Monograph.

Cooper, R. A., Maxwell, P. A., Crampton, J. S., Beu, A. G., Jones, C. M., Marshall, B. A. 2006. Completeness of the fossil record: Estimating losses due to small body size. Geology 34:241-244.

Cooper, R. A., Millener, P. R. 1993. The New Zealand Biota: Historical background and new research. Trends in Ecology & Evolution - 8:- 433.

Crampton, J. S., Beu, A. G., Cooper, R. A., Jones, C. M., Marshall, B., Maxwell, P. A. 2003. Estimating the rock volume bias in paleobiodiversity studies. Science 301:358-360.

Crampton, J. S., Foote, M., Beu, A. G., Maxwell, P. A., Cooper, R. A., Matcham, L., Marshall, B. A., Jones, C. M. 2006. The ark was full! Constant to declining Cenozoic shallow marine biodiversity on an isolated midlatitude continent. Paleobiology 32:509-532.

Cunha, R. L., Grande, C., Zardoya, R. 2009. Neogastropod phylogenetic relationships based on entire mitochondrial genomes. BMC Evolutionary Biology 9.

Darragh, T. A. 1989. A revsion of the Tertiary Volutidae (Mollusca: Gastropoda) of south-eastern Australia. Memoirs of the Museum of Victoria 49:195-307.

Daugherty, C. H., Gibbs, G. W., Hitchmough, R. A. 1993. Mega-island or micro-continent? New Zealand and its fauna. Trends in Ecology & Evolution - 8:-442.

de Queiroz, K. 2005. Different species problems and their resolution. Bioessays 27:1263-1269.

Donald, K. M., Kennedy, M., Spencer, H. G. 2005. The phylogeny and taxonomy of austral monodontine topshells (Mollusca : Gastropoda : Trochidae), inferred from DNA sequences. Molecular Phylogenetics and Evolution 37:474-483.

Donoghue, P. C. J., Benton, M. J. 2007. Rocks and clocks: calibrating the Tree of Life using fossils and molecules. Trends in Ecology & Evolution 22:424-431.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., Rambaut, A. 2006. Relaxed phylogenetics and dating with confidence. PLOS Biology 4:699-710.

Erwin, D. H. 2000. Macroevolution is more than repeated rounds of microevolution. Evolution & Development 2:78-84.

Fleming, C. A. 1979. The geological history of New Zealand and its Life. University of Auckland and Oxford University Press.

Foote, M. 2000. Origination and extinction components of taxonomic diversity: general problems. Paleobiology 26:74-102.

Foote, M., Hunter, J. P., Janis, C. M., Sepkoski, J. J. 1999. Evolutionary and preservational constraints on origins of biologic groups: Divergence times of eutherian mammals. Science 283:1310-1314.

Grantham, T. 2004. The role of fossils in phylogeny reconstruction: Why is it so difficult to integrate paleobiological and neontological evolutionary biology? Biology & Philosophy 19:687-720.

Hatzoglou, E., Rodakis, G. C., Lecanidou, R. 1995. Complete Sequence and Gene Organization of the Mitochondrial Genome of the Land Snail *Albinaria coerulea*. Genetics 140:1353-1366.

Hoffmann, R. J., Boore, J. L., Brown, W. M. 1992. A Novel Mitochondrial Genome organization for the blue mussel, *Mytilus edulis*. Genetics 131:397-412.

Imron, Jeffrey, B., Hale, P., Degnan, B. M., Degnan, S. M. 2007. Pleistocene isolation and recent gene flow in *Haliotis asinina*, an Indo-Pacific vetigastropod with limited dispersal capacity. Molecular Ecology 16:289-304.

Jablonski, D. 2000. Micro- and macroevolution: scale and hierarchy in evolutionary biology and paleobiology. Paleobiology 26:15-52.

Jackson, J. B. C., Johnson, K. G. 2001. Paleoecology - Measuring past biodiversity. Science 293:2401.

Knudsen, B., Kohn, A. B., Nahir, B., McFadden, C. S., Moroz, L. L. 2006. Complete DNA sequence of the mitochondrial genome of the sea-slug, *Aplysia californica*: Conservation of the gene order in Euthyneura. Molecular Phylogenetics and Evolution 38:459-469.

Kumar, S. 2005. Molecular clocks: four decades of evolution. Nature Reviews Genetics 6:654-662.

Kumar, S., Hedges, S. B. 1998. A molecular timescale for vertebrate evolution. Nature 392:917-920.

Leroi, A. M. 2000. The scale independence of evolution. Evolution & Development 2:67-77.

Lindberg, D. R., Ponder, W. F., Haszprunar, G. 2004. The Mollusca: Relationshios and patterns from their first half-billion years. Pp. 252-278 in J. Cracraft, and M. J. Donoghue, eds. Assembling The Tree Of Life. Oxford University Press, New York.

Lydeard, C., Holznagel, W. E., Schnare, M. N., Gutell, R. R. 2000. Phylogenetic analysis of molluscan mitochondrial LSU rDNA sequences and secondary structures. Molecular Phylogenetics and Evolution 15:83-102.

Magallon, S. A. 2004. Dating lineages: Molecular and paleontological approaches to the temporal framework of clades. International Journal of Plant Sciences 165:S7-S21.

Martin, A. P., Palumbi, S. R. 1993. Body size, metabolic-rate, generation time, and the molecular clock. Proceedings of the National Academy of Sciences of the United States of America 90:4087-4091.

Medina, M., Collins, A. G. 2003. The role of molecules in understanding molluscan evolution. Pp. 14-44 in C. Lydeard, and D. R. Lindberg, eds. Molecular Systematics and Phylogeny of Mollusks. Smithsonian Institution Press, Washington, DC.

Michaux, B. 1987. An analysis of allozymic characters of 4 species of New Zealand *Amalda* (Gastropoda, Olividae, Ancillinae). New Zealand Journal of Zoology 14:359-366.

Mooers, A.O., Harvey, P. H. 1994. Metabolic rate, generation time, and the rate of molecular evolution in birds. Molecular Phylogenetics and Evolution 3: 344-350.

Nakano, T., Marshall, B. A., Kennedy, M., Spencer, H. G. 2009. The phylogeny and taxonomy of New Zealand *Notoacmea* and *Patelloida* species (Mollusca: Patellogastropoda: Lottiidae) inferred from DNA sequences. Molluscan Research 29:33-59.

Nakano, T., Spencer, H. G. 2007. Simultaneous polyphenism and cryptic species in an intertidal limpet from New Zealand. Molecular Phylogenetics and Evolution 45:470-479.

Nee, S. 2006. Birth-death models in macroevolution. Annual Review of Ecology, Evolution, and Systematics 37:1-17.

Pagel, M. 2006. Large punctuational contribution of speciation to evolutionary divergence at the molecular level (vol 314, pg 119, 2006). Science 314:925-925.

Passamaneck, Y. J., Schander, C., Halanych, K. M. 2004. Investigation of molluscan phylogeny using large-subunit and small-subunit nuclear rRNA sequences. Molecular Phylogenetics and Evolution 32:25-38.

Penny, D., Phillips, M. J. 2004. The rise of birds and mammals: are microevolutionary processes sufficient for macroevolution. Trends in Ecology & Evolution 19:516-522.

Phillips, M. J., McLenachan, P. A., Down, C., Gibb, G. C., Penny, D. 2006. Combined mitochondrial and nuclear DNA sequences resolve the interrelations of the major Australasian marsupial radiations. Systematic Biology 55:122-137.

Ponder, W. F. 1970. The morphology of *Alcithoe arabica* (Gastropoda: Volutidae). Malacological Review 3:127-165.

Ponder, W. F., Lindberg, D. R. 2008. Molluscan evolution and phylogeny in W. F. Ponder, and D. R. Lindberg, eds. Phylogeny and Evolution of the Mollusca. University of California Press, Berkeley

Powell, A. W. B. 1979. New Zealand Mollusca. Marine, land and freshwater shells. Collins, Auckland.

Pratt, R. C., Gibb, G. C., Morgan-Richards, M., Phillips, M. J., Hendy, M. D., Penny, D. 2009. Toward Resolving Deep Neoaves Phylogeny: Data, Signal Enhancement, and Priors. Molecular Biology and Evolution 26:313-326.

Raup, D. M. 1985. Mathematical-Models of Cladogenesis. Paleobiology 11:42-52.

Reznick, D. N., Ricklefs, R. E. 2009. Darwin's bridge between microevolution and macroevolution. Nature 457:837-842.

Ricklefs, R. E. 2007. Estimating diversification rates from phylogenetic information. Trends in Ecology & Evolution 22:601-610.

Spencer, H. G., Willan, R. C. 1995. The marine fauna of New Zealand: Index 3: Mollusca. New Zealand Oceanographic Institute, Wellington.

Spencer, H. G., B. A. Marshall, P. A. Maxwell, J. A. Grant-Mackie, J. D. Stilwell, R. C. Willan, H. J. Campbell, J. S. Crampton, R. A. Henderson, M. A. Bradshaw, J. B. Waterhouse, and J. Pojeta. 2009. Phylum Mollusca *in* D. P. Gordon, ed. New Zealand inventory of biodiversity, volume one, Kingdom Animalia. Cantabury University Press, Christchurch.

Strugnell, J., Norman, M., Drummond, A. J., Cooper, A. 2004. Neotenous origins for pelagic octopuses. Current Biology 14:R300-R301.

Terrett, J. A., Miles, S., Thomas, R. H. 1996. Complete DNA sequence of the mitochondrial genome of *Cepaea nemoralis* (Gastropoda: Pulmonata). Journal of Molecular Evolution 42:160-168.

Tuan, R., dos Santos, P. 2007. ITS2 variability of *Biomphalaria* (Mollusca, Planorbidae) species from the Paranapanema Valley (Sao Paulo State, Brazil): Diversity patterns, population structure, and phylogenetic relationships. Genetics and Molecular Biology 30:139-144.

Valentine, J. W., Jablonski, D., Kidwell, S., Roy, K. 2006. Assessing the fidelity of the fossil record by using marine bivalves. Proceedings of the National Academy of Sciences of the United States of America 103:6599-6604.

Wagner, P. J. 2001. Gastropod phylogenetics: Progress, problems, and implications. Journal of Paleontology 75:1128-1140.

Wagner, P. J. 2000. Phylogenetic analyses and the fossil record: tests and inferences, hypotheses and models. Paleobiology 26:341-371.

Wiens, J. J. 2009. Paleontology, genomics, and combined-data phylogenetics: Can molecular data improve phylogeny estimation for fossil taxa? Systematic Biology 58:87-99.

Zuckerkandl, E., Pauling, L. 1962. Molecular disease, evolution, and genic heterogeneity. Pp. 189-225 in M. Kasha, and B. Pullman, eds. Horizons in Biochemistry. Academic Press, New York.

CHAPTER TWO

2 PHYLOGENETIC INFORMATIVENESS OF GENES; ILLUSTRATED WITH MITOCHONDRIAL DATA FROM A GENUS OF VOLUTE MOLLUSC

2.1 INTRODUCTION

In order to confidently build robust phylogenies one needs to critically assess markers to determine molecular datasets most applicable to different levels of divergence.

Comparative approaches allow the phylogenetic utility of markers to be determined (Graybeal 1994). It is also desirable to know whether there is sufficient data for phylogenetic estimation to reflect the evolutionary history of the entire genome, and therefore the organism, rather then the evolutionary history of one, or a small set of genes. For example, given that the genes in the mitochondrial genome are contained as a single linkage group it might be expected that the individual gene trees should agree with each other, but this is not always the case (Cummings et al. 1995).

Phylogenetic information content of genes has been scrutinised from two extreme view points, deep phylogeny (e.g. between orders of vertebrates; Cummings et al. 1995), and within species (e.g. human samples; Non et al. 2007), but many studies that exist in the middle ground, i.e. phylogenies of related species, within genera, or among sister genera, do not usually address the question of phylogenetic information content of the markers they use. The great majority of phylogenetic

studies of animals in this intermediate range have concentrated on mitochondrial *cox1*, 16S, *cytB*, and nuclear 18S, 28S sequencing markers. This has largely come about because of the availability of universal PCR primers, but selection the reliability of DNA extraction to recover mitochondrial sequence has also contributed. The relative information content in mitochondrial genes has been investigated and a range of signals found in different genes (e.g. Corneli and Ward 2000; Mueller 2006, Paton 2006). Some studies separate mitochondrial genes into classes based on the level of phylogenetic usefulness (e.g. Zardoya and Meyer 1996), but the majority of such studies deal with vertebrate lineages, or very broad evolutionary distances (Simon et al. 1994). It is therefore likely that the patterns of gene variability observed are not the same for specific invertebrate lineages such as molluscs. An analysis of the utility and critical selection of the markers to be used to resolve a phylogeny would lend greater confidence to the resulting phylogenetic hypothesis, and provide a foundation from which to assess difficult phylogenetic results. Furthermore, such an analysis would aid marker choice for studies of similar organisms. As the ease and cost effectiveness of DNA sequencing increases, the reliance on universal primers should diminish. Thus targeting genes suitable to a given type of analysis will be a more feasible strategy, rather than marker selection by convenience. An additional benefit of characterising the phylogenetic utility of markers is to provide information as to the most cost-effective regions to sequence from poor quality DNA samples, such as ancient DNA and extractions from poorly preserved museum specimens.

The limitation of available sequence markers for molecular phylogenetic studies is a particular problem for those working with Mollusca. Given the size of the phylum, second only to the Arthropoda in species richness, the Mollusca are significantly under-represented in terms of sequence data. Many “universal” primers do not work for molluscan species, limiting the set of markers available for molecular studies. Investigations of the basal relationships within Gastropoda have sometimes included nuclear genes (EF1-a, Histone H3) in addition to the standard mitochondrial genes (*cox1*, 16S) (Colgan et al. 2007), but in general molecular phylogenetics of snails rely on just three mitochondrial genes; *cox1*, 12S, 16S, and the multi-copy ribosomal nuclear cassette 28S/ITS/18S (for example see Williams and Ozawa 2006; Mejia and Zuniga 2007; Nakano and Ozawa 2007; Klussmann-Kolb et al. 2008; Reid et al. 2008; Turner and Wilson 2008). Due to the need to use sequencing markers based on universal primers that anneal across broad

taxonomic ranges, most of the sequences used for snail phylogenetics are highly conserved fragments. The sequencing markers used in these snail studies are limited to particular taxonomic depths for robust phylogenetic reconstruction. For example the nuclear markers currently used in molluscan studies (18S, 28S) lack resolution even in an intra-genus level study. Conversely, the three mitochondrial genes used for intra-species analysis (*cox*1, 12S, 16S) are more rapidly evolving sequences, but 12S and 16S can be difficult to align for deeper relationships, and *cox*1 becomes saturated at 3rd codon positions and therefore loses resolution (Simon et al. 1994; Roe and Sperling 2007).

Assessing the robustness of molecular datasets is not a trivial problem. Robustness can be judged by both congruence among different tree building methods and by the support for inferred clades. High bootstrap values and Bayesian posterior probabilities are often considered to be indicative of 'true' tree topology. These measures are indicative of accuracy only if the evolutionary model is accurate, but this is rarely the case for biological data. As such, misleading signals can occur, high bootstrap values can be obtained for incorrect topologies (Phillips et al. 2004) and Bayesian support can be inflated and not representative of the probability of correct resolution of clades (Simmons et al. 2004). It is preferable to assess the robustness of a given phylogeny by exploration of the signal in the underlying data independent of a tree. This approach allows one to assess the validity of bootstrap and Bayesian support values, and also to evaluate the signals behind clades with low bootstrap and Bayesian support values. One method of doing this is through the examination of phylogenetic splits, which represent bipartitions of taxa in the DNA dataset (Bandelt and Dress 1992). Any branch in a tree represents a split, were two partitions of a given taxon-set exit, one on either side of the branch. A set of splits is compatible if, when combined, they describe all or part of a fully resolved phylogenetic tree for the taxa involved, if not they are incompatible (Bryant and Moulton 2004). Any given phylogeny derived from a dataset can be described as a set of compatible splits, and any signal in the data that conflicts with that phylogeny can be described by a set of incompatible splits with reference to the compatible split set that describes the tree. For any given split a split-weight (or support) value can derived from an underlying sequence alignment or distance matrix. This support value will reflect the amount of evidence in the underlying data for a given bipartition of the taxa and is analogous to a branch length separating the two sets of taxa. Conflict values for individual splits are calculated from the sum of the support

for splits that contradict a given split, normalised by the ratio of total support for all splits over the total of all conflict values (Lento et al. 1995). Analysis of splits has proven to be a powerful tool for visualising signal and conflict in phylogenetic data (Huson and Bryant 2006, Holland et al. 2004). Spectral analysis allows the visualisation of conflict and support for all signals in a dataset, independently of a tree. When referenced to a tree generated from the same data the spectral analysis can be used to diagnose weaknesses in that tree, and reinforce likely true signals. Identifying genes that provide poor phylogenetic information is done by a comparison of signal and conflict for splits provided by individual genes. Previous studies have shown the potential of spectral analysis for these purposes (Lento et al. 1995, Wagele and Mayer 2007). Further, when large sequence datasets exist it is likely that selection of a subset of the genes which maximise the signal-to-noise ratio will result in phylogenies of greater robustness (Jeffroy et al. 2006).

Here we compare mitochondrial genes using a genus of marine gastropod from New Zealand waters. New Zealand volutes, including the genus *Alcithoe*, are a group of benthic, direct-developing, carnivorous neogastropod molluscs. A prevalence of large intra-specific and low inter-specific morphological variation has made the genus *Alcithoe* difficult to assess phylogenetically using morphological characters, and to date there has been no molecular treatment of the genus. A key feature of the New Zealand volute lineage leading to the *Alcithoe* is an extensive and well-studied fossil record, with the earliest known fossil occurring in 53 – 55 million year old strata (Beu and Maxwell 1990). The taxonomy of both extant and extinct *Alcithoe* is well described although its stability is subject to the vagaries of morphological characters. Based on shell characteristics, Bail and Limpus (2005) recognise 17 living species, three of which are subdivided into eight sub-species. Additionally, they recognised a further nine named “forms” in four of the species. There has been a recent increase in the number of extant taxa recognised as a result of the development of new commercial fisheries and research trips that have yielded new specimens from deeper waters. It is probable, that several of these putative new taxa represent local forms of species. Alternatively, it is possible that wide spread variable taxa may represent a complex of species, but the most recent history of *Alcithoe* taxonomy is dominated by the synonymy of various species as new samples have bridged apparent morphological gaps (Bail and Limpus 2005).

We sequenced more than 7kb of mitochondrial DNA from each species, covering nine genes. In this chapter I aim to:

(1) Infer the most likely phylogeny from the complete nucleotide dataset.

(2) Explore the signal in each of the genes comprising the complete dataset through various summary statistics and tree building methods.

(3) Assess the comparative phylogenetic utility of each gene, using splits to examine the relative contribution of signal and noise in a novel approach using spectral analysis to compare the combined spectra of all genes with a reference tree.

(4) Make recommendations as to the suitability of the genes comprising this dataset for molluscan phylogenetic studies.

2.2 MATERIALS AND METHODS

2.2.1 Taxon Sampling

We sampled 10 of the 17 extant species of *Alcithoe* recognised in Bail and Limpus (2005) from New Zealand waters (Table 2.1). With the exception of *A. larochei tigrina* all putative subspecies have been excluded. *Alcithoe* species not sampled here all have restricted ranges, largely in the far north and far south of New Zealand and have eluded sampling efforts. The only member of the genus not found in New Zealand, *Alcithoe aillaudorum,* has been sourced from New Caledonia. Eight putative out-group species for *Alcithoe* were obtained from New Zealand, Australia, and South America (Table 2.1).

TABLE 2.1—Volute species used to study the phylogenetic information in 9 mitochondrial genes

Genus	Species	Voucher number	Sample Location	
Alcithoe	*aillaudorum*	NB 1024	Isle des Pins	New Caledonia
Alcithoe	*arabica*	M.279684	Wellington	New Zealand
Alcithoe	*benthicola*	M.183806	Coromandel	New Zealand
Alcithoe	*fissurata*	M.183785	Coromandel	New Zealand
Alcithoe	*flemingi*	TAN0408/50	Chatham Rise	New Zealand
Alcithoe	*fusus*	M.279683	Nelson	New Zealand
Alcithoe	*jaculoides*	M.274972	North Island East Coast	New Zealand
Alcithoe	*larochei*	M.274116	North Island East Coast	New Zealand
Alcithoe	*larochei tigrina*	M.183799	Coromandel	New Zealand
Alcithoe	*lutea*	NIWA 30452	Challenger Plateau	New Zealand
Alcithoe	*pseudolutea*	M.183802	Coromandel	New Zealand
Alcithoe	*wilsonae*	M.190062	South Island	New Zealand
Cymbiola	*pulchra subelongata*	M.273459	Queensland	Australia
Odontocymbiola	*simulatrix*	MZSP44320	Cabo Santa Marta	Brasil
Athleta	*studeri*	M.273462	Queensland	Australia
Amoria	*hunteri*	M.273463	Queensland	Australia
Adelomelon	*beckii*		Mar del Plata	Argentina
Adelomelon	*brasiliana*		Mar del Plata	Argentina
Adelomelon	*riosi*	MZSP32971	Cabo Frio	Brasil

2.2.2 DNA Extraction and Amplification

DNA was extracted from foot tissue from both frozen and ethanol preserved specimens using a high-salt buffered extraction method (Norman et al. 1998), modified as follows. Approximately 0.5 mg of tissue was incubated in 300μl of high-salt buffer with 1μl of 10ng/μl ProtK shaking at 60°C for at least 16 hours. 300μl of phenol was added and the solution incubated with shaking at room temperature. Following centrifugation the aqueous phase was removed and mixed with 400μl of chloroform:isoamyl alcohol (24:1). The chloroform wash was repeated, and DNA precipitated with 95% ethanol at -20°C for 8 to 16 hours, before resuspension in 0.1 TE. DNA concentrations were determined using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc.). DNA extractions were diluted to approximately 1ng/μl for amplification.

Initially short-range polymerase chain reaction (SR-PCR) was carried out to amplify fragments of between 300 and 1000 bp of mitochondrial cytochrome oxidase 1 (*cox*1) and 16S ribosomal DNA (16S) using universal primers (Table 2.2). PCR was carried out using Red-Hot Taq (ABgene), following the manufacturer's instructions with a $MgCl_2$ concentration of 2.0mM. Standard thermal cycling conditions were followed, with 50°C annealing temperatures and 30-35 cycles for both primer sets, carried out in a Biometra^{tm} T1 thermocycler. Products generated by SR-PCR were sequenced with both forward and reverse primers using BigDye Terminator v3.1 and an ABI 3730. From these short sequences primers were designed in *cox1* and 16S in order to amplify longer mtDNA fragments (Table 2.2), of approximately 6 kBp. Long-range amplification (LR-PCR) was performed using Extensor Hi-Fidelity PCR enzyme (ABgene). PCR reactions were set up as per the manufacturer's instructions, using 2-5ng of DNA per sample. Thermal cycling conditions were as recommended by the manufacturer, using a 50°C annealing temperature. LR-PCR products were sequenced by primer walking (Kusukawa et al. 1990). From these long sequence fragments, and using complete mitochondrial sequences available on GenBank (*Ilyanassa obsoleta* NC_007781, *Lophiotoma cerithiformis* NC_008098, *Conus textile* NC_008098, *Littorina saxatilis* AJ132137) additional primers were designed for highly conserved regions (Table 2.2), in order to extend the ends of the sequence fragment by up to 1 kB in each direction. The binding sites of primers developed here are shown in Figure 2.1.

Table 2.2—Primers used to amplify mtDNA of volute gastropods

	Primer sequence 5'-3'	Reference
Universal Primers		
cox1		
HCO 2198	TAA ACT TCA GGG TGA CCA AAA AAT CA	Folmer et al. 1994
LCO 1490	GGT CAA CAA ATC ATA AAG ATA TTG G	Folmer et al. 1994
16S		
LR-N-12866	ACA TGA TCT GAG TTC AAA CCG G	Simon et al. 1994
LR-J-12887	CCG GTC TGA ACT CAG ATC ACG T	Simon et al. 1994
Alcithoe primers		
cox1		
AfLR1cox1	CTG GCT CTT AGT TTG CTT ATT CGG G	This study
AfLR3cox1	GTT CAA ATT GCA TAC CAC GTC ATC G	This study
16S		
AfLR216S	GGT ACT CTG ACC GTG CAA AGG TAG C	This study
AfLR416S	TGG TCC AAC ATC GAG GTC ACA AAC C	This study
Upstream of *cox1*		
NGmt_trnI	GAA CGG AAA TCA TTG ATG TTG ATT AWT ATG GG	This study
NGmt_trn2	AAT TAC CCA AAR CAA AGT TAG CAG C	This study
Downstream of 16S		
AfLR14tLEU2	AAG ATG GCA GAT AAA GTG CAT TAG G	This study
SorbND1	GTT CTA AWA GMG TAA AAA AAG CGA CTG C	This study

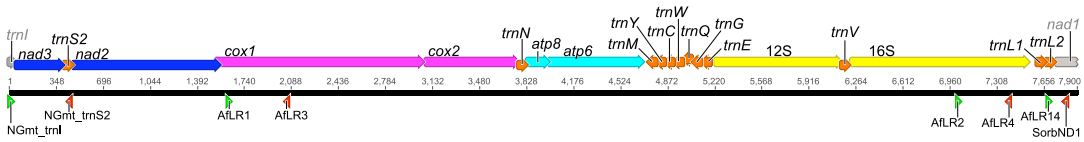

FIGURE 2.1—Mitochondrial gene arrangement in the New Zealand marine mollusc genus *Alcithoe*. Genes that comprise the sequenced DNA fragment from Volutidae are labelled and include the following: NADH dehydrogenase subunit 3 (*nad*3), serine (AGN) tRNA (*trn*S2), NADH dehydrogenase subunit 2 (*nad*2), cytochrome c oxidase subunit 1 (*cox*1), cytochrome c oxidase subunit 2 (*cox*2), asparagine tRNA (*trn*N), ATP synthase F0 subunit 8 (*atp*8), ATP synthase F0 subunit 6 (*atp*6), methionine tRNA (*trn*M), tyrosine tRNA (*trn*Y), cystine tRNA (*trn*C), tryptophan tRNA (*trn*W), glutamine tRNA (*trn*Q), glycine tRNA (*trn*G), glutamic acid tRNA (*trn*E), short subunit rRNA (12S), valine tRNA (*trn*V), long subunit rRNA (16S), leucine (CUN) tRNA (*trn*L1), leucine (UUR) tRNA (*trn*L2). Arrowheads indicate the direction of transcription for each gene. Genes flanking the sequenced region, *trn*I and *nad*1, contain primer sites used to generate the sequenced fragment but do not contribute any nucleotide data to this study. Binding sites for the primers designed to generate this DNA fragment are indicated (see table 2).

2.2.3 Sequence Analysis and Phylogenetic Reconstruction

Sequences were edited using Sequencher (v4.6, Gene Codes Corporation, Ann Arbor, Michigan). Alignments were generated in Sequencher and exported in nexus format. SE-AL v2.0a11 (Rambaut 2002) was used to infer protein sequences from the nucleotide sequences, and to refine alignments as appropriate. Ribosomal DNA genes were aligned based on secondary structure. The ribosomal RNA gene 16S was aligned using the molluscan consensus structure of Lydeard *et al.* (2000), although we found domain 1 was too variable to unambiguously align based on this consensus structure and was aligned based on common secondary structures for volute species returned by Mfold (Zuker et al. 1999). As no consensus structure of 12S is available for molluscs, this alignment was based on similarity to structures on the Comparative RNA Web Site (Cannone et al. 2002) using the secondary structures of *Paracentrotus lividus* (sea urchin) and *Artemia franciscana* (brine shrimp). Due to alignment ambiguity with the 5' and 3' ends of the chosen model sequences, Mfold was used to infer structures of the volute sequences to use as an alignment guide for these regions. Transfer RNA genes were compared to structures reported for *Lophiotoma cerithiformis* (Bandyopadhyay et al. 2006), in order to identify putative stem and loop regions for accurate alignment.

Maximum Parsimony reconstruction, ModelTest v3.7 (Posada and Crandall 1998) and partition homogeneity tests, were implemented using Paup*4.0 (Swofford 1998). Consistency indices and partitioned homogeneity tests were also generated in Paup*. Neighbor-Joining trees were constructed using the Geneious tree builder in the Geneious software package (Drummond 2007). Maximum-Likelihood reconstruction and Bayesian analysis were carried out using PHYML version 2.4.4 (Guindon and Gascuel 2003) and Mr Bayes version 3.1.2 (Huelsenbeck and Ronquist 2001) respectively, using Geneious plug-ins. Maximum-parsimony reconstruction was carried out with default parameters, with the exception that 1000 bootstrap replicates were performed. Alignments that contained gaps were analysed with gaps excluded and with gaps coded as a fifth state, in order to assess the effect of gaps on phylogenetic reconstruction with this dataset. Maximum-likelihood reconstruction was carried out under three sets of model parameters for each dataset. The three sets of parameters represented increasingly complex, and theoretically better fitting models as follows:

1/ HKY with default parameters in PHYML

2/ HKY using the averaged parameters for all models as returned by Modeltest

3/ the specific model and parameters (or as close as possible using PHYML settings) returned by Modeltest by AIC

Bayesian reconstruction utilised default parameters for HKY with 4 heated chains of length 1,000,000, sampling every 1000 generations, with a default 10% burn-in to ensure stationarity of the posterior sampling. Visualisation of conflict in the data through analysis of splits and networks was carried out in SplitsTree4 (Huson and Bryant 2006). Weighed-splits were derived from nucleotide alignments in Splits Tree 4, using the Neighbor-Net method (Bryant and Moulton 2004). These weighted splits were transferred into SpectroNet (Huber 2002) to calculate conflict values. Splits were first calculated in SplitsTree using Neighbor-Net as this generated a smaller number of splits then the methods implemented in Specronet. The reduction in the number of resulting splits is due to Neighbor-Net not considering many trivial splits with very low support values. These values were used to generate Lento-plots (Lento et al. 1995) for individual genes and split support graphs for the collected data.

Spectral analysis of the splits data using Lento-plots allows a detailed examination of the relative signal and noise in a data set. The sum of support and conflict of all the gene partitions for each split, calculated by Neighbor-Net from uncorrected P-distances, illustrates the distribution of information in the dataset. The significance of observed splits was judged not only by the amount of support or conflict that any given gene partition has for a set of splits, but also on the number of genes that supported or conflicted with a given split. Therefore splits that have little support in any given gene partition become significant when most or all considered genes exhibit some support for that split. Phylogenetically problematic genes can be identified where splits are supported or show conflict from only one gene.

2.3 RESULTS

2.3.1 Out-group selection

Prior to the generation of long-range PCR products for outgroup taxa, available samples were tested with short sequences to find which are the most closely related to *Alcithoe*. Initial phylogenetic analyses using 313 bp of 16S and 354 bp of ND3 resolved New Zealand *Alcithoe* as monophyletic but *Alcithoe aillaudorum* from New Caledonia was not part of this clade nor sister to it (Figure 2.2). Constraining the tree topology to include all *Alcithoe* species in a monophyletic clade resulted in a significantly less-likely tree (SH test; $P < 0.05$). Two Australian volute species (*Cymbiola pulchra* and *Amoria hunteri)* were chosen as an out-group for the New Zealand *Alcithoe,* being the most closely related volute taxa in the phylogeny with bootstrap support of 90 for this sister relationship. The two out-group species and 11 New Zealand *Alcithoe* species were then used to generate the full sequence data set.

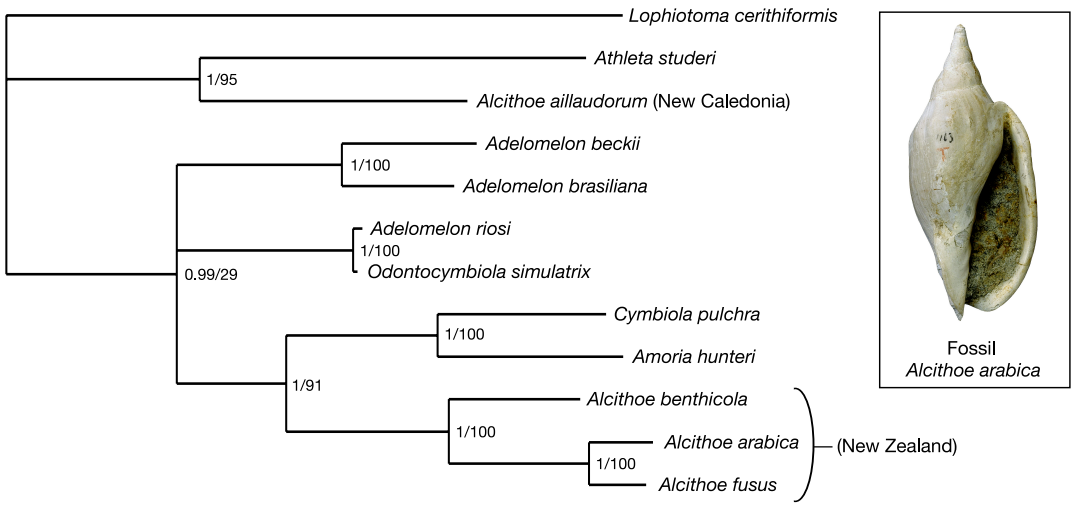

FIGURE 2.2—Phylogeny to establish the molecular context of *Alcithoe* within Volutidae. This phylogeny is derived from a concatenated alignment of 313bp of 16S and 354bp of *nad*3 from 19 volute taxa. *Lophiotoma cerithiformis* (NC_008098), a conoidean gastropod, was used as an out-group for the Volutidae. Bayesian posterior probabilities and Maximum Likelihood bootstrap support are given for each node (B/ML). *Cymbiola* and *Amoria* are shown to be the most closely related volute taxa to *Alcithoe* and therefore represent the most appropriate out-group taxa for phylogenetic reconstruction of *Alcithoe* species. Based on this phylogeny *Alcithoe aillaudorum* from New Caledonia is not sister to the New Zealand *Alcithoe*.

2.3.2 Sequence data

Mitochondrial DNA sequence of between 7681 and 7733 base pairs was generated for each taxon, including all sites from the beginning of the *nad3* gene to the end of the *trnL*(UUR) gene. This DNA fragment represents approximately half the entire neogastropod mitochondrial genome and includes the following genes; *nad3, trnS*(AGN)*, nad2, cox1, cox2, trnD, atp8, atp6, trnM, trnY, trnC, trnW, trnQ, trnG, trnE,* 12S*, trnV,* 16S*, trnL*(CUN)*, trnL*(UUR) (Figure 2.1). The majority of these genes are coded on the heavy strand of the mitochondria. Only the seven tRNA genes *trnM* through to *trnE* are on the light strand. This gene arrangement and order is identical to the four neogastropod mollusc mitochondrial genomes published to date; *Thais clavigera* (NC_010090), *Lophiotoma cerithiformis* (NC_008098, (Bandyopadhyay et al. 2006)), *Conus textile* (NC_008797), and *Ilyanassa obsoleta* (NC_007781, (Simison et al. 2006)). Some gene size variation exists among the taxa considered, and overall GC content ranges from 29.8% to 31.5%, (Table 2.3). Most of the variability in size is seen between the two out-group taxa (*Cymbiola pulchra* and *Amoria hunteri)* and the in-group *Alcithoe* species. The out-group taxa generally have slightly longer ribosomal RNA coding genes than

Alcithoe, and *Alcithoe* have fewer intergenic bases. A maximum size difference of 18 bp is seen among *Alcithoe* species, whereas the difference between any of the *Alcithoe* species and the out-groups ranges from 29 to 52 bp. In contrast, the six protein coding genes are of identical size, with the exception of a single additional amino acid in the *nad2* gene of *Cymbiola* from an insertion of three nucleotides. The difference in sequence lengths is the predominant reason for including gaps to correctly align these sequences. However, additional gaps are required in order to accurately align the structural features of the RNA coding genes.

Table 2.3—Summary of sequenced DNA fragments from 13 marine molluscs of the family Volutidae

Taxa	Total Sequence Length	Protein Coding Regions	16S	12S	tRNAs	Intergenic Regions	Gaps[a]	%GC
Alcithoe arabica	7687	4518	1341	879	806	143	135	31.3
Alcithoe benthicola	7692	4518	1338	878	799	159	130	31.5
Alcithoe fissurata	7690	4518	1345	879	805	143	132	31.4
Alcithoe flemingi	7699	4518	1340	880	800	161	123	31
Alcithoe fusus	7688	4518	1343	878	805	144	134	31.2
Alcithoe jaculoides	7690	4518	1341	880	805	146	132	30.8
Alcithoe larochei	7681	4518	1340	878	804	141	141	31
Alcithoe lutea	7692	4518	1342	878	807	147	130	31
Alcithoe pseudolutea	7688	4518	1340	878	807	145	134	31.3
Alcithoe tigrina	7690	4518	1344	878	805	145	132	31.2
Alcithoe wilsonae	7690	4518	1340	883	802	147	132	31.5
Cymbiola pulchra	7733	4521	1347	885	814	166	89	29.8
Amoria hunteri	7728	4518	1344	888	811	167	94	30.2

[a] Gaps required to accurately align these DNA fragments in an alignment of the 13 taxa listed

Regions of overlap were found for several genes. The terminal base of *nad3* overlaps with *trnS*. *trnW* and *trnQ* overlap by one base. Two terminal bases of *cox2* are shared with *trnD*. The largest overlap is between *nad2* and *cox1*, where the terminal 29 bases of *nad2* are also the initial coding bases of *cox1*. Previous studies have indicated that a reduced TA stop codon exists at the *nad2*/*cox1* junction in neogatropods, however in our data an in-frame full TAA stop codon can be identified for *nad2* at positions 27 to 29 of the *cox1* coding sequence. With the exception of the *nad2*/*cox1*, the overlap of mitochondrial genes observed here is consistent with published results from neogastropods (Bandyopadhyay et al. 2006, Simison et al. 2006).

For the purpose of phylogenetic analysis several partitioned subsets of the sequence data were created. In addition to the complete dataset a concatenated dataset was generated with all intergenic spacers removed, and where an overlap exists the relevant nucleotide positions were included for both genes separately. The six protein coding genes and the two rRNA genes were each given individual partitions, and the 12 tRNA genes were partitioned as a single concatenated set.

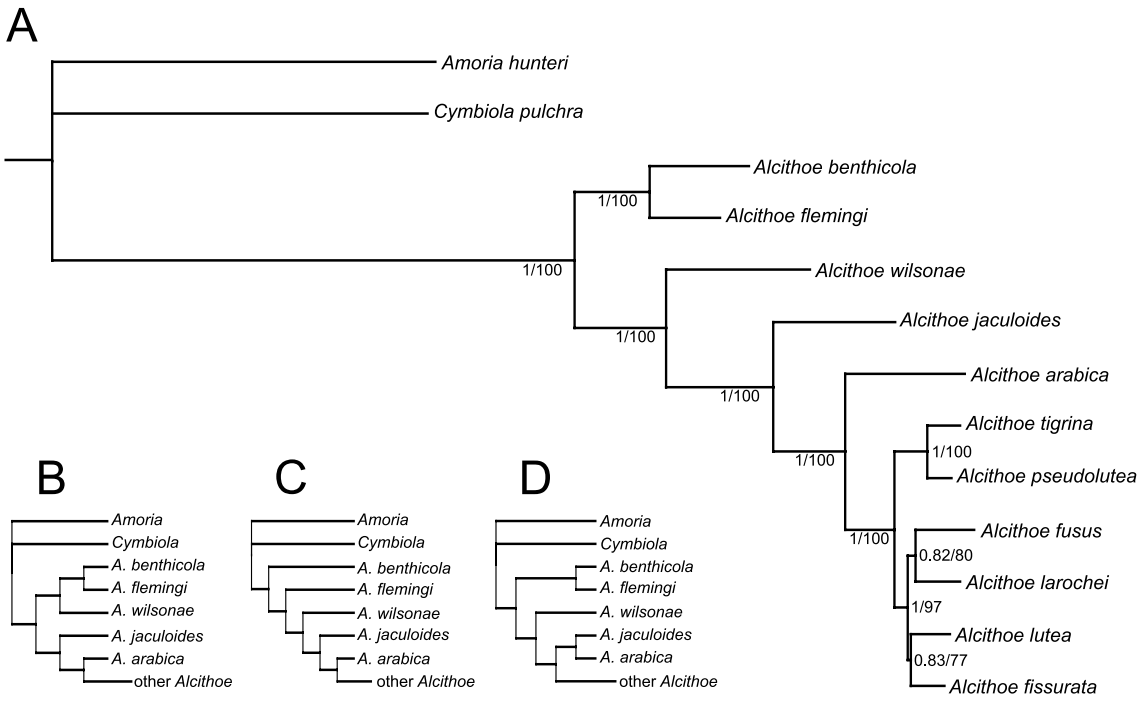

FIGURE 2.3—Molecular phylogeny of 11 species of the New Zealand marine mollusc genus *Alcithoe*. (A) A phylogeny of *Alcithoe* derived from the complete nucleotide dataset (7822 bp), support given as Bayesian posterior probability/ML Bootstrap. Despite absolute Bootstrap and Bayesian support for all deeper divergences conflicting signal can be detected in some genes individually. Three alternative tree topologies are present for the deeper divergences in *Alcithoe*. (B) Alternative topology 1, *A. wilsonae* sister to *A. benthicola*/*A. flemingi* clade. (C) Alternative topology 2, loss of sister relationship of *A. benthicola* and *A. flemingi*. (D) Alternative topology 3, sister relationship of *A. jaculoides* and *A. arabica*.

2.3.3 Phylogeny from the complete nucleotide dataset

Maximum likelihood (using three models), maximum parsimony (including and excluding gaps where appropriate) and Bayesian trees were generated using the complete nucleotide dataset, consisting of 7822 sites (gaps included). These methods recovered identical tree topologies, with bootstrap support of 100 and Bayesian posterior probability of 1 for seven nodes (Figure 2.3A). This consistency, in addition to prior evidence that approximately half the mitochondrial genome can

accurately infer mitochondrial evolutionary history in other taxa (Cummings 1995), is interpreted as evidence that this topology represents a plausible initial phylogenetic hypothesis for *Alcithoe*.

2.3.4 Summary statistics

Alignment length, summaries of variability, consistency index, and Modeltest results for the complete dataset, each of the nine gene partitions and the concatenated dataset are presented in Table 2.4. These statistics provide useful general information about the data, and identify genes that might be problematical for phylogenetic reconstruction. Although relatively short (159bp) *atp*8 exhibits high variability, but 3rd codon position variability is lower than other genes while 2nd position variability is twice that of any other. Overall, low variability is seen in *cox1* and 12S, but the most conserved partition is the set of 12 tRNA genes. The genes *cox*1 and *cox*2 show an accumulation of variability in 3rd codon positions (85% and 77% respectively), but low amino acid variability (3% and 15%) suggesting a high rate of synonymous substitution in these two genes. However, consistency indices for all subsets of the data are very similar and do not indicate high levels of saturation in any genes (Table 2.4). The highest consistency indices are seen where the lowest variation is recorded (tRNAs and 12S), but these genes also have the highest ratio of parsimony-uninformative to parsimony-informative sites.

Four DNA substitution models are recovered for the individual genes using Modeltest. Interestingly, the two most complete subsets of data (Complete and Concatenated) are best modelled by the 9 parameter TVM+I+G model, even though individual genes, such as *cox*2 and *atp*6, recover more complex models.

2.3.5 Partition heterogeneity

Partition Homogeneity tests were carried out on all pair-wise combinations of the nine individual gene partitions, both excluding and including gap information. Only *cox*2/12S and *cox*2/*nad*2 gene combinations showed significant partition heterogeneity. However, when corrected for multiple tests there is no significant partition heterogeneity among the set of nine genes (data not shown).

TABLE 2.4—Summary statistics from alignments of mitochondrial DNA sequence data partitions for 13 volute species.

Gene	Alignment Length	pVAR[a] total	pVAR[a] 1st codon position	pVAR[a] 2nd codon position	pVAR[a] 3rd codon position	Amino Acid pVAR[a]	ci[b]	Proportion of phylogenetically informative variable sites [c]	Modeltest AIC
*nad*3	354	0.395	0.263	0.086	0.679	0.254	0.651	51 / 89	TVM+G
*nad*2	1089	0.41	0.291	0.101	0.609	0.322	0.675	171 / 273	TVM+I+G
*cox*1	1536	0.27	0.138	0.012	0.85	0.031	0.635	154 / 260	HKY+I+G
*cox*2	687	0.322	0.186	0.05	0.765	0.153	0.622	81 / 140	GTR+I+G
*atp*8	159	0.44	0.271	0.2	0.529	0.396	0.661	27 / 43	K81uf+G
*atp*6	696	0.376	0.244	0.103	0.653	0.228	0.644	89 / 173	GTR+I+G
12S	898	0.253	N/A	N/A	N/A	N/A	0.797	101 / 108	HKY+I+G
16S	1375	0.312	N/A	N/A	N/A	N/A	0.756	181 / 208	K81uf+I+G
tRNAs	829	0.211	N/A	N/A	N/A	N/A	0.791	72 / 77	K81uf+G
Concat	7623	0.313	N/A	N/A	N/A	N/A	0.688	927 / 1371	TVM+I+G
Comp	7822	0.328	N/A	N/A	N/A	N/A	0.691	1002 / 1434	TVM+I+G

[a] proprotion of observed variable sites

[b] consistancy index

[c] number of variable sites that are parsimony uninformative / number of variable, parsimony informative sites

2.3.6 Gene trees

As with the complete dataset, Maximum likelihood, maximum parsimony and Bayesian trees were generated for each gene partition and the concatenated dataset. The topology of all trees built from the subset partitions (concatenated dataset and individual gene datasets), were compared to the tree topology derived from the complete dataset, in order to assess the performance of each partition. From a total of 60 combinations of tree-building models and datasets, 25 alternative tree topologies were returned. None of the dataset partitions were as consistent under the alternative tree-building strategies as the complete data. Each of the individual gene partitions produced several tree topologies. Even the concatenated dataset, which only omits 167bp of intergenic spacer, and where nucleotides in overlapping regions appear twice, produces two different tree topologies.

Alternative positions of the six most recently diverged *Alcithoe* species (*A. lutea, A. larochei, A. fusus A. pseudolutea, A. tigrina, A. fissurata*), which differ by between 0.7 to 6.9% pair-wise sequence divergence, are responsible for the majority of the alternative topologies observed. If only the variation in deeper branches is considered, then only four (of a possible 105) topologies are seen. These four tree topologies, characterised by alternative groupings of the basal *Alcithoe* nodes, are as follows:

1/ *A. benthicola* and *A. flemingi* monophyletic and sister to all other New Zealand *Alcithoe*, as in Figure 2.3A

2/ grouping *A. wilsonae* in a clade with *A. benthicola* and *A. flemingi* (Figure 2.3B)

3/ loss of the monophyletic grouping of *A. benthicola* and *A flemingi* (Figure 2.3C)

4/ monophyly of *A. arabica* and *A. jaculoides* (Figure 2.3D)

The degree of topological variability is different across the partitions. The least conflict in tree topology is seen in *cox1* and *atp6,* which each return only one alternative to the tree topology illustrated in Figure 2.3A, while 16S returns a different tree topology for each of the tree estimation methods and parameter sets used. *cox2* is the only gene that returns a sister relationship of *A. jaculoides* and *A. arabica* (Figure 2.3D), although support is low (50-60 bootstrap support, 0.64 Bayesian clade credibility). The topology shown in Figure 2.3BC is returned under some parameter conditions (ML and Bayesian) with *cox2* and *nad3* alignments. A single clade containing *A. wilsonae, A. benthicola* and *A. flemingi* (Figure 2.3B) is consistently recovered from the tRNA dataset, and also seen in MP trees derived from the *cox1* alignment. Additionally, the tRNAs, 12S, *atp*8 and *cox*2 consistently place the divergence the *A. tigrina*/*A. pseudolutea* clade as the most recent in the *Alcithoe*, which is at odds with the complete-dataset phylogeny which places this divergence before the radiation of *A. fusus*, *A. larochei*, *A. lutea* and *A. fissurata*.

2.3.7 Spectral analysis

Phylogenetic reconstruction under different models suggests that there is conflicting signal within the data, a common problem in molecular phylogenetics. A network of the splits, which defines all the bipartitions of taxa in the complete dataset, illustrates the amount and distribution of conflict in the data (Figure 2.4). In order to gain a better understanding of the contribution of signal and noise from each of the gene partitions we explored the phylogenetic information contained in the nucleotide data by visualising support of taxa splits using networks and Lento-plots for individual genes. It is important to reiterate that these splits are a summary of the total signal in an alignment and are not generated assuming any given tree topology. As such they represent a description of the phylogenetic information in the dataset that is independent of any reconstruction method.

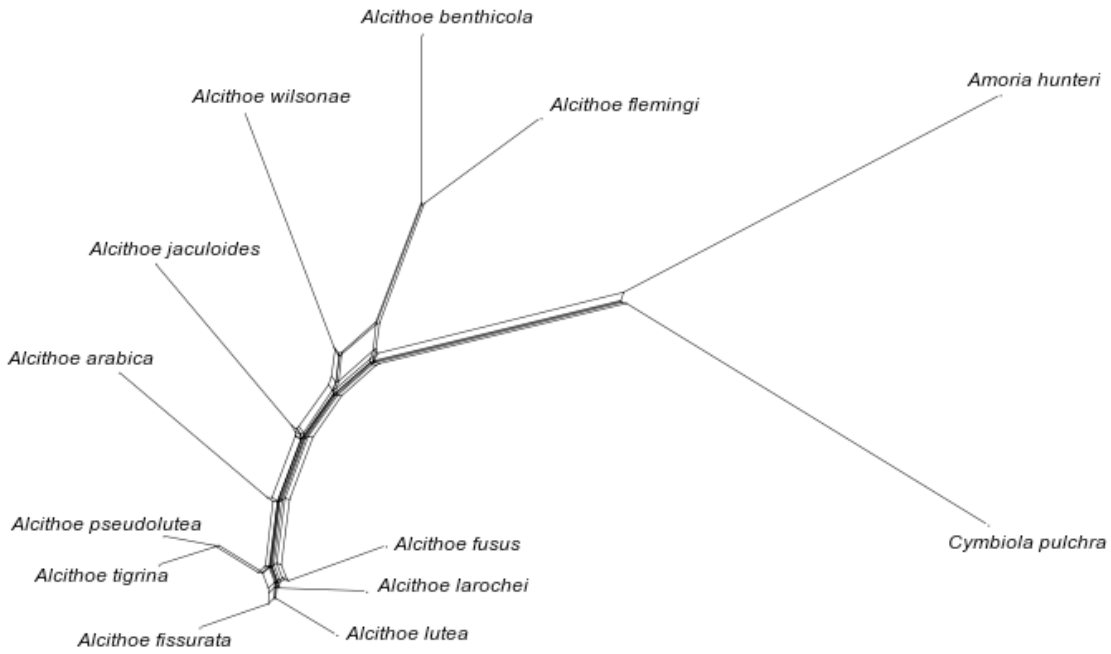

FIGURE 2.4—Spilts Network of *Alcithoe* species based on the complete dataset (7822 bp). Splits were generated using the Neighbor-Net algorithm in SplitsTree 4. The alternative phylogenetic relationships illustrated in Figure 2.3 are apparent in the large box structure at the base of the *A. wilsonae* and *A. benthicola*/*A. flemingi* branches. The phylogenetic noise and short branch lengths of six closely related and recently derived species is clear. This network visualises alternative signal, but only quantifies it in a general way and does not indicate source of the conflicting signals in the data.

The majority of the splits compatible with the complete dataset tree exhibit significant support, in most cases with contributions from all genes (Figure 2.5). Splits compatible with the draft tree in Figure 2.3A will hereafter be referred to by the letters assigned to those splits in the split key in Figure 2.5. Incompatible splits will be referred to by with lower-case letter designations from Figure 2.5. Some splits observed in the reference tree (Figure 2.3A) have relatively large quantities of conflict (e.g. B, C, D, E), although the majority of the conflict for splits C and D comes from *atp8* alone. Many splits representing clades not present in the complete dataset tree (incompatible splits) have very little support and large conflict and tend to be seen in few genes often only one. These are likely to be the result of homoplasy. Five splits seen only in the complete data set are due to signal in the intergenic spacer regions (data not shown), which are gap rich and confidence in site homology is relatively low due to rapid sequence divergence. However, the concordance of the complete and concatenated datasets shows that possible errors in the alignment of these regions do not appear to introduce significant spurious signal.

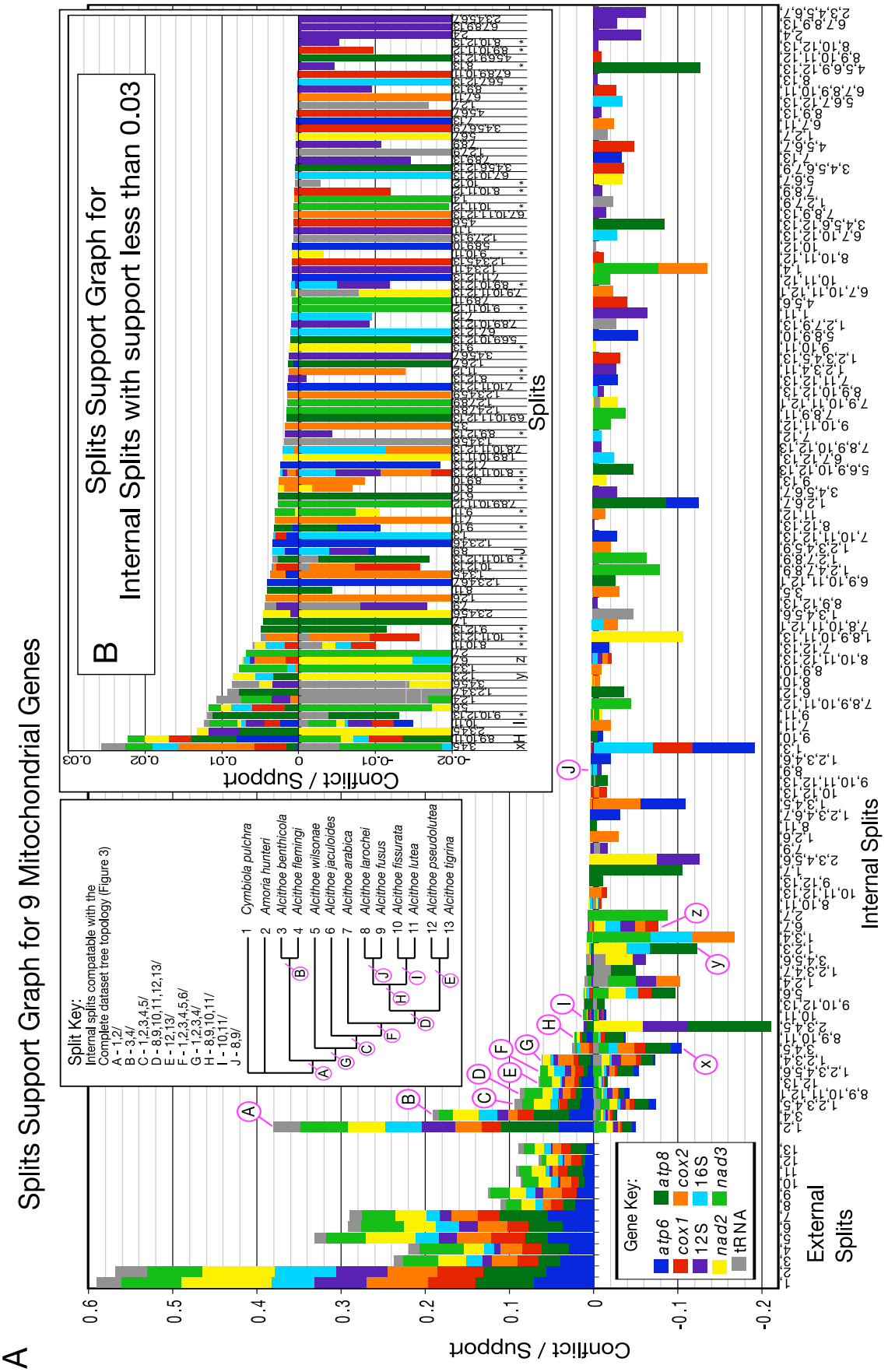
A
Splits Support Graph for 9 Mitochondrial Genes
Conflict / Support
External Splits
Internal Splits
Gene Key:
atp6
atp8
cox1
cox2
12S
16S
nad2
nad3
tRNA
Split Key:
Internal splits compatable with the Complete dataset tree topology (Figure 3)
A - 1,2/
B - 3,4/
C - 1,2,3,4,5/
D - 8,9,10,11,12,13/
E - 12,13/
F - 1,2,3,4,5,6/
G - 1,2,3,4/
H - 8,9,10,11/
I - 10,11/
J - 8,9/
1 Cymbiola pulchra
2 Amoria hunteri
3 Alcithoe benthicola
4 Alcithoe flemingi
5 Alcithoe wilsonae
6 Alcithoe jaculoides
7 Alcithoe arabica
8 Alcithoe larochei
9 Alcithoe fusus
10 Alcithoe fissurata
11 Alcithoe lutea
12 Alcithoe pseudolutea
13 Alcithoe tigrina
B
Splits Support Graph for
Internal Splits with support less than 0.03
Conflict / Support
Splits

FIGURE 2.5 (facing page)—Summed split support from partitioned data reveals the relative amounts of signal and noise contained in each partition of mitochondrial DNA sequence from *Alcithoe* species. (A) Split support and conflict for all gene-based partitions of the data (Table 2.4) are graphed for the total number of splits seen for all partitions. Columns represent the sum of support (above 0) or conflict (below 0) for splits derived from the alignments of each individual gene partition. The contributions of each gene partition are colour coded to illustrate the level of signal and conflict from each gene for each split. Splits referring to a single taxon split from the others (representing external branches in the tree) have no conflict and are clustered on the left. Internal splits are ordered by the total amount of support in the data for each split. In general splits representing branches in the complete dataset have high support and are clustered to the left. However, splits between recently diverged, closely related species are more ambiguous, often with low support and little conflict, and can be distributed further the right of the graph. Splits are listed by the group of taxa represented on one side of the division. The number code for each taxon is given in the split key inset. Internal splits compatible with the complete dataset tree are labelled with capital letters. Lower case letters (x, y, z) mark the splits representing conflicting topologies found for individual genes shown in Figure 2.3. (B) An extracted section of the graph shows the splits with support less than 0.03 in greater detail. Splits marked with an asterisk represent signal relating to the 6 most recently diverged taxa that is inconsistent with the complete dataset tree. Many of these splits exhibit similar degree of support and conflict, indicating a paucity of discriminating signal.

Analysis of splits allows a critical assessment of the signal leading to problems highlighted by tree building using multiple techniques. Tree building demonstrated three major concerns related to the issue of differing signals in genes:

1. conflicting signal around the placement of *A. wilsonae*

2. the position of the *A.tigrina* and *A. pseudolutea* clade as basal or more derived than the four poorly resolved taxa

3. conflicting signal regarding the positions of taxa in the clade containing *A. fusus, A. larochei, A. fissurata,* and *A. lutea*.

1. When the position of *A. wilsonae* is considered with reference to observed high support splits, two mutually exclusive splits are identified (G and x in Figure 2.5). These two splits represent alternative tree topologies returned from different tree building analyses (Figures 2.3A and 2.3B, respectively). While the support for split G is strong, and all genes except *nad3* contain supporting signal, there is also large conflict (Figure 2.5). Support for split x is substantially less, and only a few genes are represented (predominantly *cox2* and *nad3*), while the conflict is much greater and occurs in most genes. This comparison strongly supports the topology for *A. wilsonae* seen for the complete-dataset topology (Figure 2.3A). The support observed for split x is likely due to shared ancestral states in these three taxa in some genes. Genes containing strong conflicting signal of this kind are likely to increase noise in multi-gene datasets and it is likely to be advantageous to identify and remove such genes.

2. The uncertainty of the tree topology of the 6 most recently diverged species (*A. pseudolutea*, *A. tigrina*, *A. larochei*, *A. fusus*, *A. fissurata* and *A. lutea*) is demonstrated by large number of splits involving combinations of these taxa (splits marked with asterisks in Figure 2.5B). The splits D, E and H (Figure 2.5A), which support the topology of the *A. pseudolutea*/*A. tigrina* clade shown in Figure 2.3A, have considerably more support then any conflicting splits (e.g. any splits containing *A. pseudolutea, A. tigrina* and any subset of *A. larochei*, *A. fusus*, *A. fissurata* and *A. lutea*). This result substantiates the high likelihood of the topology seen for the complete dataset.

3. Divisions that are particularly problematic are between *A. larochei*, *A. fusus*, *A. fissurata* and *A. lutea* as there are several splits that have low support, but also relatively low conflict separating two or three of these taxa from the remainder. The

general lack of resolution for these four species in this phylogeny is unsurprising, because pair-wise sequence divergence is only 1% - 5.4%. Homoplasy between these derived taxa and the more basal species, particularly the out-groups, could be confounding signal at this level.

2.3.8 A reduced taxon set clarifies six closely related species

In order to clarify the evolutionary relationships of six closely related *Alcithoe* species, a taxon-reduced dataset was created (*A. fusus*, *A. larochei*, *A. fissurata*, *A. lutea, A. larochei tigrina, A. pseudolutea,* plus *A. arabica* as an outgroup) and the splits spectrum for the nucleotide dataset generated (Figure 2.6A). Using this pruned taxon set reduces the amount of noise produced by homoplasy with more divergent species; this can be seen in the significant reduction in split conflict compared to Figure 2.5. In addition to the nine gene-based data partitions, a partition for the intergenic spacer regions was added as these rapidly evolving regions could be useful in resolving closely related taxa. Low phylogenetic resolution is observed, but, as predicted, a better signal to noise ratio is achieved. This reduction in noise allows for a more clear interpretation of the split support for alternative topologies. Such clarity is important as three different topological solutions are returned from three phylogenetic reconstruction methods, Bayesian (Figure 2.6B), Maximum Likelihood (Figure 2.6C) and Neighbour Joining (Figure 2.6D). The divergence of the *A. pseudolutea*/*A. tigrina* clade prior to the divergences of *A. fusus*, *A. larochei*, *A. fissurata* and *A. lutea* is consistent and well supported in these analyses. This result clarifies an inconsistency seen between the full dataset and individual gene trees (see section 2.3.6).

Lack of concordance in tree building methods for the relationships among the four most closely related taxa is likely a result of low levels of phylogenetic signal within these genes (Figure 2.6A). Additionally, the splits data show that not all genes in the dataset contribute signal to these low-resolution relationships. There are four splits associated with the trees returned, splits C, D, E and F (letters refer to designations in the split key of Figure 2.6A). The most well supported of these, split D, only lacks contribution from *atp*8 and *cox*2, and the majority of the conflict comes from the intergenic spacer. Split D appears consistently in all trees built from this reduced taxon dataset and indicates confidence in the monophyletic grouping of *A. fissurata* and *A. lutea*.

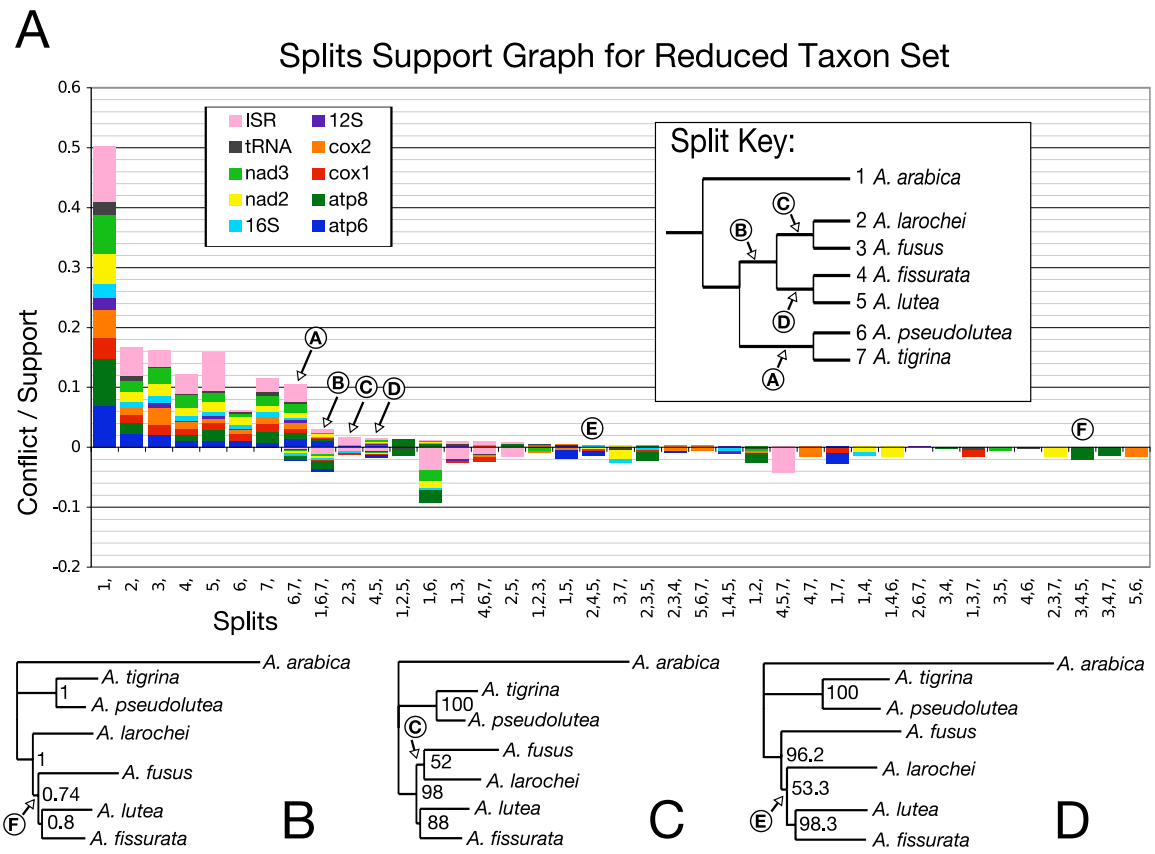

FIGURE 2.6—Refinement of phylogenetic inference for New Zealand Alcithoe *species* by consideration of a sub-tree only. (A) Support and conflict for splits generated from seven *Alcithoe* species are summarised in a splits support graph. Splits representing external branches are clustered on the left and internal splits are then ordered by decreasing split support. In addition to the 9 gene partitions an intergenic spacer region (ISR) sequence partition is included, as these non-coding regions are likely to be more informative at the level of closely related taxa examined here. The splits key shows the taxa included and the expected splits (based on the complete dataset tree shown in Figure 2.3A), these splits are labelled on the graph. The grouping of *A. pseudolutea* and *A. tigrina* is well supported by the split A. The B split supports the basal position of these two taxa and *A. arabica*. Arrangements of the four most closely related species, *A. larochei*, *A. fusus*, *A. fissurata*, and *A. tigrina*, depicted by splits containing two or more of these taxa, all have small amounts of support with near equal or greater conflict. An interesting feature in this dataset is the lack of support for two external branches (3 and 6) in the *atp*8 gene
Three topological solutions for this dataset are returned from different tree-building methods; (B) Bayesian inference, (C) Maximum Likelihood, (D) Neighbor Joining. The differences in each of these can be characterised by one key split, seen in (A), that dictates the positions of *A. fusus* and *A. larochei*; split F for the Bayesian tree (B), split C for the Maximum Likelihood tree (C) and split E for the Neighbor Joining tree (D).

The number of genes contributing signal drops sharply for the remaining splits, which are all in reference to the positions of the two species *A. fusus* and *A. larochei*. The Maximum Likelihood tree (Figure 2.6C) contains split C, which is supported by *atp*6, *cox*1, 16S, and the intergenic spacer, however the majority of the signal comes

from the intergenic spacer. The Bayesian tree has split F, which is supported by only *atp*8, but also has relatively large conflict only from *atp*8. The Neighbour Joining tree includes split E, which has some support from *atp*6, *cox*1, *nad*2 and 16S, but has slightly more conflict from the same genes. This reduced taxa dataset highlights some gene based problems not apparent in the complete dataset. Inclusion of the intergenic spacer, while providing some additional support for some in-tree splits (e.g. C), also introduced conflict for others (e.g. B and D). The splits support-graph shows that *atp*8 and *cox*2 are problematic, providing little support and greater conflict for internal splits. Additionally, in this dataset, *atp*8 shows no support for the external branches for *A. pseudolutea* and *A. fusus*. This means that these two taxa are indistinguishable from other taxa (*A. pseudolutea* from *A tigrina* and *A. fusus* from *A. larochei*), and that data from this gene leads to increased splits conflict.

2.3.9 Refinement of analysis

Taking into account the accumulated information about variability, compatibility, signal and noise now generated for this dataset, an informed decision can be made as to which genes can be included or excluded from an analysis in order to maximise the ratio of signal to noise. In order to reduce noise generated from the high support for conflicting tree topologies we removed the gene that was highlighted by the partition homogeneity test as being potentially problematic; *cox2*. The amount and distribution of variation in *cox2* indicated that the pattern of DNA substitution within this gene is unusual in the context of this data set thus the best substitution model for *cox*2 is unlikely to be a good fit to the rest of the data. The pruned taxon set indicated that *atp8*, the shortest (159 bp) but most variable of the nine genes, is a source of noise. As a result *atp*8 seems to introduce a high level of conflict in the data while not providing a large amount of additional support for any splits. Therefore, in the interest of noise reduction *atp*8 can be trimmed from the data set. Although *nad*3 does carry some noise around some splits (eg split G vs split x in Figure 2.5), it was retained in the data set because it contains consequential signal for several other splits. Finally, 12S and the tRNA set, while exhibiting low resolution also exhibited low conflicts at the taxonomic level of this group.

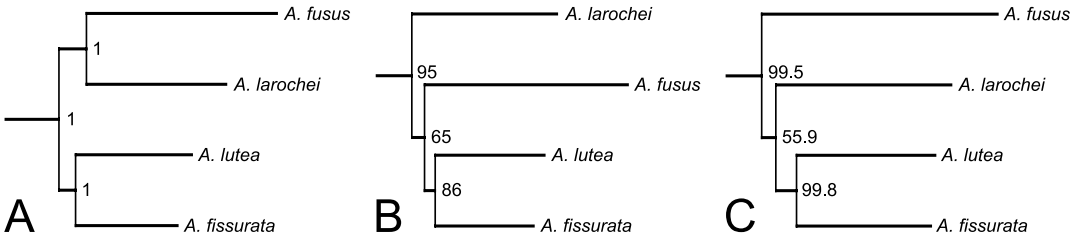

FIGURE 2.7—Alternative phylogenetic topologies in a quality controlled dataset show that the relationships of *A. fusus* and *A. larochei* cannot be resolved with this data. Phylogenetic reconstructions for the 'best gene' dataset (*nad*3, *nad*2, *cox*1, *atp*6, 16S, 12S, and the tRNA set) returned from Bayesian analysis (A), Maximum Likelihood (B), and Neighbour Joining (C), illustrate that the different reconstruction methods interpret low support, low conflict splits in different ways. Only the four most derived taxa are shown here as the topology of the remainder of the tree is identical for all analyses.

The spectrum of splits for the dataset containing the retained seven gene partitions (*nad3*, *nad2*, *cox1*, *atp6*, 16S, 12S, and tRNAs) is similar to that for the complete dataset, but conflicts are reduced. However, trees generated from this reduced dataset are not consistent under different phylogenetic reconstruction methods, with alternative topologies being found for the two closely related species, *A. fusus* and *A. larochei*. Bayesian analysis returns the sister relationship of *A. fusus* and *A. larochei* (Figure 2.7A), whereas ML returns a tree with *A. larochei* as the earliest of the four taxa to diverge (Figure 2.7B), and Neighbor Joining returns a tree with *A. fusus* as first to diverge (Figure 2.7C). For the Bayesian and ML methods these topologies represent an inversion of the reduced taxa dataset result. The split information for this data indicates little signal for any of the possible topologies. SH tests of four possible topologies of *A. fusus* and *A. larochei* (unresolved, sister, *A. fusus* diverging first, and *A. larochei* diverging first) show that these phylogenetic solutions are equally good explanations of the data. Our results show that the branching order of these two taxa is sensitive to the tree building method used. Given that the alternative topologies of these taxa are equally likely the most correct hypothesis of the phylogeny, based on this data, is a three-way polytomy of *A. fusus*, *A. larochei* and the *A. fissurata*/*A. lutea* clade (Figure 2.8), even though no phylogenetic reconstruction method independently returns this result.

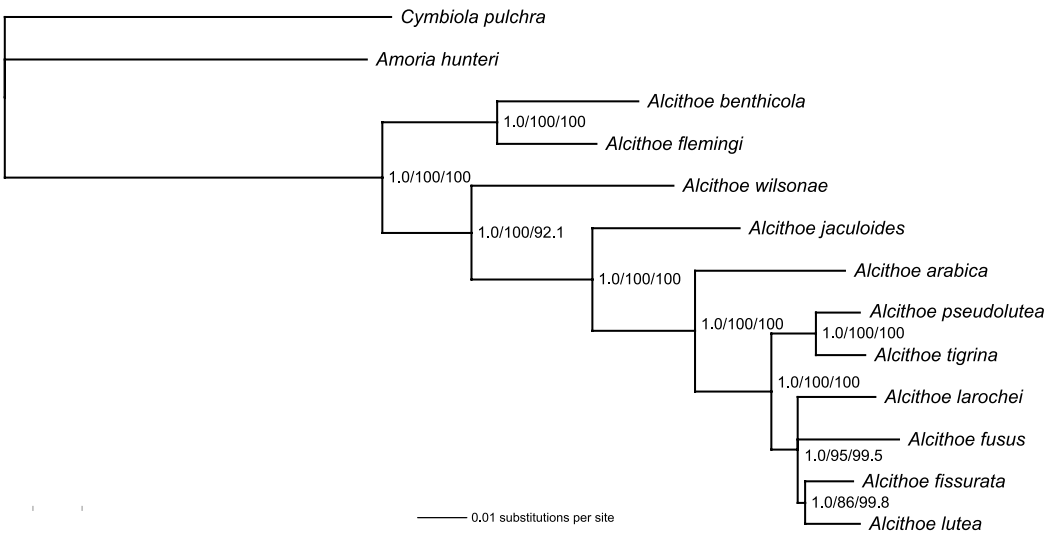

FIGURE 2.8—Molecular Phylogeny of the gastropod genus *Alcithoe* based on a reduced gene dataset with maximised signal and minimised noise. The underlying dataset, once genes with significantly conflicting signal are trimmed, is 6777 bp in length and includes; nad3, nad2, cox1, atp6, 16S, 12S, and the tRNA set. Bayesian posterior probability/Maximum Likelihood bootstrap support/Neighbour Joining bootstrap support is give for each node. The unresolved relationship for *A. larochei* and *A. fusus* has been enforced on this phylogeny. No phylogenetic reconstruction method returned this topology. However, given analysis of the splits data and as different tree building techniques return the different topological alternatives, this is the most accurate depiction of the signal carried by the data.

2.4 DISCUSSION

Large fragments (>7Kb) of molluscan mitochondrial DNA can be obtained by long-range PCR using sequences starting with 'universal' mtDNA markers. Such large-scale sequencing allows the comparison of phylogenetic information content of individual mtDNA genes. Our eleven gene partitions returned a range of DNA substitution models, (from the 6 parameter HKY+I+G to the 10 parameter GTR+I+G) reflecting different complexities in substitution dynamics of the different genes. In combination the partition homogeneity tests and variability data only highlighted one gene as potentially problematic for phylogenetic analysis (*cox2*), but this method has been criticised as an inadequate measure of the combinability of datasets (Barker and Lutzoni 2002). The use of splits support graphs and Lento-plots for presenting phylogenetic signal provide more detailed information about the signal and noise content of individual genes. It is then possible to identify which partitions of a data set bring proportionally more noise than signal to the analysis thus allowing the informed inclusion or exclusion of genes in a multi-gene analysis in order to maximize the robustness of the resulting phylogeny. In this way we have

been able to show that, in this inter-species dataset, the value of the informative signal in the *atp*8 gene is outweighed by the conflicting noise. The finding that *cox*2 and *atp*8 are overly noisy in this dataset is consistent with studies that have demonstrated the limited phylogenetic usefulness of these genes in other taxonomic groups (Mueller 2006, Corneli, 2000, Zardoya, 1996). However it is important to note the phylogenetic utility of a marker is dependent on the distance of evolutionary relatedness and taxonomic context. Additionally, it is likely to be beneficial to select genes that can be modelled by the same or similar substitution models. Until such time as more realistic models of DNA substitution are available, and particularly in the current environment of ever increasing sized nucleotide data sets, it makes sense to identify and analyse sets of genes that better obey the conditions of an existing model rather then attempting to fit a less well suited model or over parameterise by using multiple models.

2.4.1 Taxon subsets exclude unnecessary noise

We recommend analysing subsections of taxa independently to resolve relationships among closely related species and populations, thus reducing noise from homoplasy. By analysing taxa that differ in their degree of divergence separately, models of DNA substitution should more realistically describe DNA evolution in each dataset. Relationships within closely related taxa can then be constrained in larger scale analyses, preventing these relationships from being disrupted by the addition of more divergent taxa (and accompanying homoplasy) and poorly fitting models of DNA evolution.

2.4.2 Marker selectivity

Our refined dataset with fewer genes was not consistent under different phylogenetic reconstruction methods, but generated trees that differed only in the placement of two of the most closely related taxa (*A. larochei* and *A. fusus)*. Unlike the complete dataset, the sensitivity to model of DNA evolution is, we think, a more biologically accurate result. The refined dataset is better because we know, from separate analysis of the recently derived taxa, that relationships can be misled by homoplasy and inappropriate model of DNA evolution. Thus our refined set of genes provided us with a dataset that does not hide the very real difficulty in accurately estimating a phylogeny of this group of volutes. Furthermore, awareness of the specific source of this uncertainty in the data will allow for it to be accounted for in down-stream applications, such as molecular clock analysis.

To date the majority of phylogenetic analyses with molluscan species have included only a few genes. Recent studies of other taxonomic groups have shown that it is of great value to perform multi-gene analyses in order account for idiosyncrasies in individual genes that might otherwise mislead the phylogeny. We go further to suggest that exclusion of genes that can be clearly shown to have anomalous signals or adhere to disparate substitution models is desirable in order to increase the robustness of final evolutionary hypotheses. Ultimately it is likely that the consideration of multi-scale analysis will be appropriate, particularly for large taxon sets. In such an analysis each gene would only be considered up to the level at which the signal to noise ratio for that gene remained acceptable. For example, some genes in a dataset may be informative at the species level but not at the genus level, and the decision as to what level they are informative at will be based on signal to noise ratio. Additionally such genes could be used to resolve some sub-trees but not others, based on their phylogenetic information content.

For the analysis of gastropod taxa separated by between approximately 1 and 50 million years, and not overly species dense, we recommend the combination of the genes in our reduced set; *nad*2, *cox1*, *atp6*, and 16S. The genes *nad*3 and 12S are also suitable, but *nad*3 is preferable for more shallow divergence and 12S for deeper divergence. In neogastropod molluscs an appropriate continuous mitochondrial fragment to target would be an approximately 3 Kb section spaning the genes *nad*3, *nad*2, and *cox1*.

2.4.3 Alcithoe Systematics

In the selection of out-group taxa for this analysis it was clear that *Alcithoe aillaudorum* from New Caledonia is not closely related to the New Zealand *Alcithoe*. If the history of the mitochondria of these volutes is representative of the species history then the genus *Alcithoe* is not monophyletic and we suggest the current placement of the New Caledonian species *Alcithoe aillaudorum* should be re-examined.

Phylogenetic analysis of nine mitochondrial genes from thirteen volute taxa resolved a stable evolutionary hypothesis for the group. Within the New Zealand *Alcithoe* there is one major point of difference in the assignment of species between our data and the work of Bail and Limpus (2005). Bail and Limpus treat *A. tigrina* is a subspecies of *A. larochei* based on shell morphology. This molecular dataset,

however, supports the clear separation of these two species. Our molecular data support the close morphological relationship of *A. larochei*, *A. lutea* and *A. pseudolutea* recognised by Bail and Limpus (2005). The possibility of a common origin of *A. fusus* and *A. jaculoides*, suggested by Bail and Limpus, can be discarded as the molecular data clearly shows that these two species are not closely related. Indeed, the close relationship of *A. fusus*, *A. larochei*, *A. fissurata*, and *A. lutea* is novel and somewhat unexpected. However, this is consistent with fossil evidence for *A. fusus* and *A. larochei*, which indicate a relatively recent origin of both these species around 1.6 million years ago (Beu and Maxwell 1990).

2.5 REFERENCES

Bail, P., Limpus, A. 2005. The recent volutes of New Zealand with a revision of the genus *Alcithoe* H. & A. Adams, 1853 in G. T. Poppe, and K. Groh, eds. A Conchological Iconography. ConchBooks, Hackenheim.

Bandelt, H. J., Dress, A. W. M. 1992. A canonical decomposition-theory for metrics on a finite-set. Advances in Mathematics 92:47-105.

Bandyopadhyay, P. K., Stevenson, B. J., Cady, M. T., Olivera, B. M., Wolstenholme, D. R. 2006. Complete mitochondrial DNA sequence of a Conoidean gastropod, *Lophiotoma (Xenuroturris) cerithiformis*: Gene order and gastropod phylogeny. Toxicon 48:29-43.

Barker, F.K., Lutzoni, F.M. 2002. Utility of the incongruence length difference test. Systematic Biology 51:625-637.

Beu, A. G., Maxwell, P. A. 1990. Cenozoic Mollusca of New Zealand. New Zealand Geological Survey Paleontological Bulletin 58:1-518.

Bryant, D., Moulton, V. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. Molecular Biology and Evolution 21:255-265.

Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y. S., Feng, B., Lin, N., Madabusi, L. V., Muller, K. M., Pande, N., Shang, Z. D., Yu, N., Gutell, R. R. 2002. The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3.

Colgan, D. J., Ponder, W. F., Beacham, E., Macaranas, J. 2007. Molecular phylogenetics of Caenogastropoda (Gastropoda : Mollusca). Molecular Phylogenetics and Evolution 42:717-737.

Corneli, P. S., Ward, R. H. 2000. Mitochondrial genes and mammalian phylogenies: Increasing the reliability of branch length estimation. Molecular Biology and Evolution 17:224-234.

Cummings, M. P., Otto, S. P., Wakeley, J. 1995. Sampling properties of DNA-sequence data in phylogenetic analysis. Molecular Biology and Evolution 12:814-822.

Drummond A.J., Ashton B., Cheung M., Heled J., Kearse M., Moir R., Stones-Havas S., Thierer T., Wilson A. 2007. Geneious v3.8. Available from http://www.geneious.com/

Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology 3:294-299.

Graybeal, A. 1994. Evaluating the phylogenetic utility of genes - a search for genes informative about deep divergences among vertebrates. Systematic Biology 43:174-193.

Guindon, S., Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52:696-704.

Holland, B. R., Huber, K. T., Moulton, V., Lockhart, P. J. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. Molecular Biology and Evolution 21:1459-1461.

Huber, K. T. L., M. Penny, D. Moulton, V. Hendy, M. 2002. Spectronet: A package for computing spectra and median networks. Applied Bioinformatics 1:159-161.

Huelsenbeck, J. P., Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.

Huson, D. H., Bryant, D. 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23:254-267.

Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H. 2006. Phylogenomics: the beginning of incongruence? Trends in Genetics 22:225-231.

Klussmann-Kolb, A., Dinapoli, A., Kuhn, K., Streit, B., Albrecht, C. 2008. From sea to land and beyond - New insights into the evolution of euthyneuran Gastropoda (Mollusca). BMC Evolutionary Biology 8.

Kusukawa, N., Uemori, T., Asada, K., Kato, I. 1990. DNA sequencing report - rapid and reliable protocol for direct sequencing of material amplified by the polymerase chain-reaction. Biotechniques 9:66-&.

Lento, G. M., Hickson, R. E., Chambers, G. K., Penny, D. 1995. Use of spectral-analysis to test hypotheses on the origin of pinnipeds. Molecular Biology and Evolution 12:28-52.

Lydeard, C., Holznagel, W. E., Schnare, M. N., Gutell, R. R. 2000. Phylogenetic analysis of molluscan mitochondrial LSU rDNA sequences and secondary structures. Molecular Phylogenetics and Evolution 15:83-102.

Mejia, O., Zuniga, G. 2007. Phylogeny of the three brown banded land snail genus *Humboldtiana* (Pulmonata : Humboldtianidae). Molecular Phylogenetics and Evolution 45:587-595.

Mueller, R. L. 2006. Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. Systematic Biology 55.

Nakano, T., Ozawa, T. 2007. Worldwide phylogeography of limpets of the order patellogastropoda: Molecular, morphological and palaeontological evidence. Journal of Molluscan Studies 73:79-99.

Non, A. L., Kitchen, A., Mulligan, C. J. 2007. Identification of the most informative regions of the mitochondrial genome for phylogenetic and coalescent analyses. Molecular Phylogenetics and Evolution 44:1164-1171.

Norman, J., Olsen, P., Christidis, L. 1998. Molecular genetics confirms taxonomic affinities of the endangered Norfolk Island Boobook Owl *Ninox novaeseelandiae undulata*. Biological Conservation 86:33-36.

Paton, T. A., Baker, A. J. 2006. Sequences from 14 mitochondrial genes provide a well-supported phylogeny of the Charadriiform birds congruent with the nuclear RAG-1 tree. Molecular Phylogenetics and Evolution 39:657-667.

Phillips, M. J., Delsuc, F., Penny, D. 2004. Genome-scale phylogeny and the detection of systematic biases. Molecular Biology and Evolution 21:1455-1458.

Posada, D., Crandall, K. A. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818.

Rambaut, A. 2002. Se-Al: Sequence Alignment Editor. Available at http://tree.bio.ed.ac.uk/software/seal/.

Reid, D. G., Dyal, P., Lozouet, P., Glaubrecht, M., Williams, S. T. 2008. Mudwhelks and mangroves: The evolutionary history of an ecological association (Gastropoda : Potamididae). Molecular Phylogenetics and Evolution 47:680-699.

Roe, A. D., Sperling, F. A. H. 2007. Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. Molecular Phylogenetics and Evolution 44:325-345.

Simison, W. B., Lindberg, D. R., Boore, J. L. 2006. Rolling circle amplification of metazoan mitochondrial genomes. Molecular Phylogenetics and Evolution 39:562-567.

Simmons, M. P., Pickett, K. M., Miya, M. 2004. How meaningful are Bayesian support values? Molecular Biology and Evolution 21:188-199.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H., Flook, P. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene-sequences and a compilation of conserved polymerase chain-reaction primers. Annals of the Entomological Society of America 87:651-701.

Swofford, D. L. 1998. PAUP*4.0- Phylogentic Analysis Using Parsimony (*and Other Methods). Sinauer Associates Inc., Sunderland, MA.

Turner, L. M., Wilson, N. G. 2008. Polyphyly across oceans: a molecular phylogeny of the Chromodorididae (Mollusca, Nudibranchia). Zoologica Scripta 37:23-42.

Wagele, J. W., Mayer, C. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evolutionary Biology 7.

Williams, S. T., Ozawa, T. 2006. Molecular phylogeny suggests polyphyly of both the turban shells (family Turbinidae) and the superfamily Trochoidea (Mollusca: Vetigastropoda). Molecular Phylogenetics and Evolution 39:33-51.

Zardoya, R., Meyer, A. 1996. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Molecular Biology and Evolution 13:933-942.

Zuker, M., Mathews, D. H., Turner, D. H. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: A Practical Guide in J. Barciszewski, and B. F. C. Clark, eds. NATO ASI Series. Kluwer Academic Publishers.

CHAPTER THREE

3 THE IMPORTANCE OF CORRECT IDENTIFICATION OF FOSSIL SPECIES

3.1 INTRODUCTION

3.1.1 Species delimitation

Discrimination of species in the fossil record is not straight forward. There are many possible sources of uncertainty, not the least of which are the myriad of ways in which species can be delimited. There is no agreement on the definition of a species even when restricting the debate to sexually reproducing living organisms (see debates on species concepts; Mallet 1995; Templeton 2001; Hey 2006, etc). The Biological Species Concept is difficult to apply to extinct species due to the impracticality of demonstrating reproductive isolation, indeed the Biological Species Concept is difficult to apply to the greater part of known life (e.g. prokaryotes, and archaea). Therefore a morphological based approach to delimitation is almost universally applied to fossil species. As morphological discrimination is also the primary basis of the taxonomic description of extant species, the contextualisation of modern and extinct species is done through morphology. However, there can be some significant problems with species defined by morphology alone.

It is argued that morpho-species are an inadequate account of variation. Cryptic species, polymorphic species, ecophenotypic variation and a lack of sampling of

intermediate forms of chronospecies would all lead to an incorrect assignment of morpho-species. Jackson and Cheetham (1990) demonstrated that this not the case in cheilostome Bryzoa, but studies of living molluscs have shown both cryptic species (e.g. Collin 2005; Nakano and Spencer 2007) and polymorphic species (e.g. Palmer 1985; Kartavtsev et al. 2006).

The Phylogenetic Species Concept (PSC), where species are defined as reciprocally monophyletic units, has become more applicable with the advent of molecular methods. The great advantage of the PSC is that it reflects genetic divergence and can be informative where morphology cannot. However, it is hampered in the same way as most methods, by a subjective assignment of the quantity of difference that constitutes separate species. There is ongoing discussion as to the extent of correlation between morphospecies, phylospecies and biological species (eg Samadi and Barberousse 2006; Knowles and Carstens 2007). For the same group of organisms the different methods will often define different species, because the phylogenetic method has a greater power to distinguish cryptic species and to group highly ecophenotypically variable species. Entities defined phylogenetically will represent within-species variability if the variation cut-off for species definition is set too low. The various species concepts can be thought of as a continuum (de Queiroz 2005), where the phylogenetic method can identify entities early in the process of speciation, the biological species concept can identify fully differentiated species, and other methods fit somewhere in between.

3.1.2 Fossil calibration of molecular clock analysis

A major interface between paleontological and molecular data is through molecular clock analysis. When using fossils to calibrate the age of nodes it is critical that the species thought to represent that node is correctly identified. Calibration with an incorrectly identified specimen could have extremely detrimental effects on any conclusions derived from such an analysis. A morpho-species assignment that fails to separate cryptic species could lead to an over estimation of apparent within-species genetic variation and an under estimate of interspecies diversity. Conversely, a highly ecophenotypicaly variable species could give rise to several morpho-species assignments, as is often a problem when intermediate specimens are not sampled. This situation will lead to an over estimation of the species diversity and an inference of low sequence divergence. These taxonomic errors, if

associated with calibration taxa in molecular clock analysis, could lead to significantly erroneously inferred rates of molecular evolution. In the case of unidentified cryptic species an apparently high level of genetic variability will lead to a faster rate estimate. A morphologically variable species that has been classified as several morpho-species will give rise to slower rate. For difficult groups with extant members, molecular and paleontological datasets can be reciprocally illuminative. It is likely that, particularly for difficult groups where extant examples of species with a fossil history exist, the determination of a molecular phylogeny can guide the interpretation of morphological characters that can then be informative for extinct species. (e.g. Michaux 1987; Jackson and Cheetham 1994; Samadi et al. 2000).

3.1.3 Taxonomic history of Alcithoe wilonsae *and* Alcithoe knoxi

Within the marine snail genus *Alcithoe* inter-species morphological convergence, intra-species variation and a paucity of uniquely derived diagnostic characters has made species delimitation difficult (eg Dell 1978; Powell 1979; Bail and Limpus 2005). An exemplar of this problem is the species *Alcithoe wilsonae*. Extensive morphological variability in this species has lead to a variety of generic and species level assignments for different entities that are now recognised as sub-species or forms of *A. wilsonae* (Bail and Limpus 2005) (see Figure 3.1). The variability in generic and specific assignments has been exacerbated by the fact that *A. wilsonae* has a wide distribution (see Figure 3.2) and various morphological forms seem geographically restricted within the overall range. In 1979 Powell recognised three species of *Pachymelon* (*P. wilsonae* Powell 1933, *P. smithi* Powell 1950, and *P. grahami* Powell 1965). Powell also recognised *Alcithoe* (*Leporemax*) *chathamensis* Dell 1956. In the most recent treatment of the *Alcithoe* genus, Bail and Limpus (2005) synonymised these entities into a single subspecies *Alcithoe wilsonae wilsonae*, but maintained the recognition of the forms *smithi, grahami* and *chathamensis*. Furthermore they erected a new sub-species, *Alcithoe wilsonae acuminata*, as the authors considered that there were sufficient constant shell characters to separate this entity from the others. Powell's separation of these various forms resulted from a lack of specimens with intermediate morphologies and because many of these forms appeared to have geographically distinct ranges. The subsequent synonymy was the result of new samples that bridge apparent morphological discontinuities (Bail and Limpus 2005).

FIGURE 3.1—Morphological forms of *Alcithoe wilsonae* and *Alcithoe knoxi*. (A) *A. wilsonae wilsonae* forma *chathamensis*, NMNZ M190111, Chatham Islands, 335m. (B) *A. wilsonae wilsonae*, NMNZ M117117.A, Auckland Islands, 390m. (C) *A. knoxi*, NIWA 30037, Chatham Rise, 516m (D) *A. wilsonae acuminata*, NMNZ M190079, Chatham Rise, 379m. A 5cm scale is shown for each specimen for size reference.

Conversely, the species *Alcithoe knoxi* has had a more stable taxonomic nomenclature. Although it has a similar geographic range, *A. knoxi* does not exhibit the polymorphism of *A. wilsonae*. However, the two species were thought to differ in the ocean depth they occupy; *A. wilsonae* is recorded down to approximately 500 m and *A. knoxi* recorded from approximately 450 to 600 m. Not until the recent work of Bail and Limpus (2005) was *knoxi* included in the *Alcithoe*. Prior to this

knoxi was considered a species of *Teremelon*, the only extant species in this genus of New Zealand volutes. *Alcithoe knoxi* has a long fossil history, the first specimen occurring in the Tongaporutuan stage (10.92 Ma – 6.5 Ma) of the New Zealand geological sequence. This fossil is considered identical in form to extant specimens. As a result *A. knoxi* has been considered a separate and distinct taxonomic unit. Morphologically *A. knoxi* is diagnosed by a small shell with a large protoconch and a steep spire compared to *A. wilsonae*, and exhibits little within-species variation. However, confusion between forms of *A. w. acuminata* and *A. knoxi* has been noted (Bail and Limpus 2005).

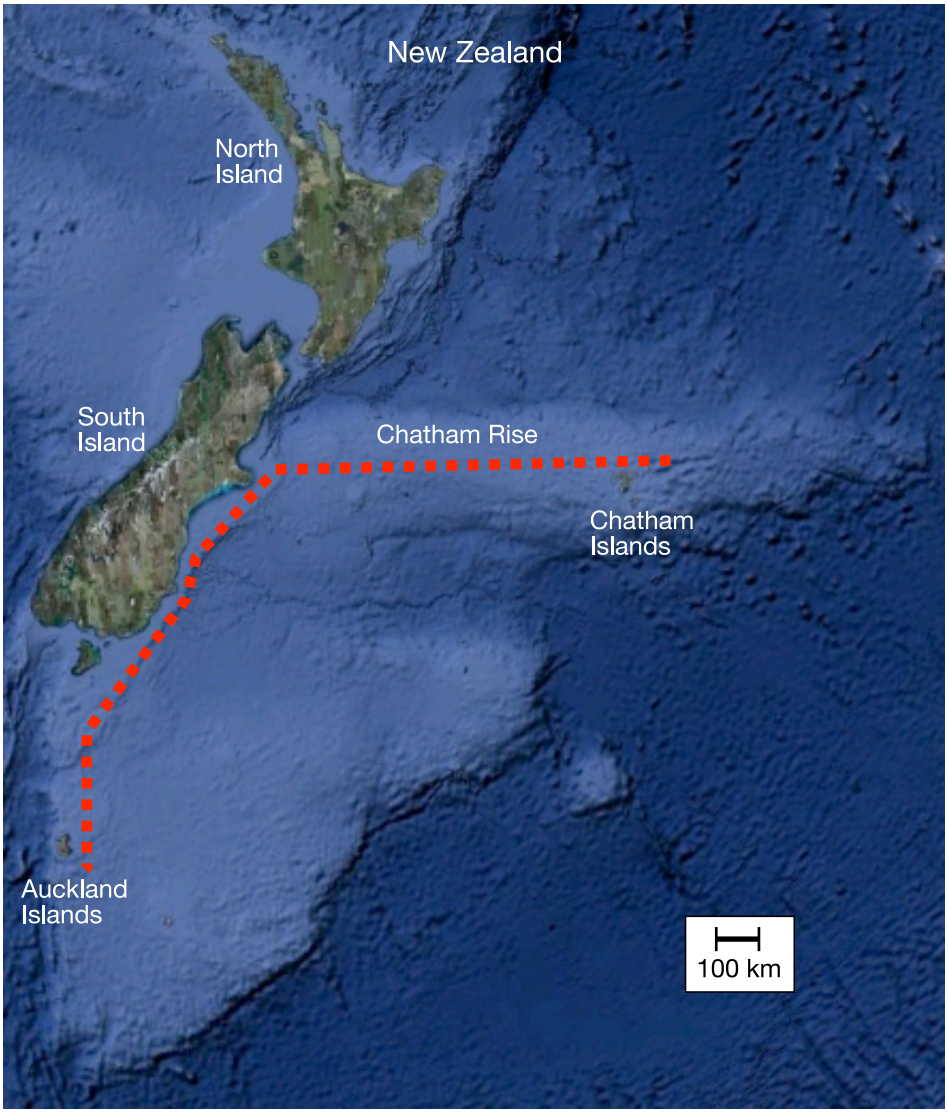

FIGURE 3.2—Approximate geographic range of *A. wilsonae* and *A. knoxi*. The range of these benthic marine gastropods is indicated by the red dotted line. Specimens are known from depths of between approximately 50 and 750 meters.

A lack of clearly defined shell traits is a common problem with accurately identifying extinct species. When extant species are difficult to delimit and large numbers of specimens are required to do so, how confident can we be about the assignment of species level distinctions of fossil taxa? This problem is compounded in extant species found in deeper water, where sampling can be prohibitively expensive and is therefore rather erratic.

Initial molecular phylogenetic reconstructions of the evolutionary relationships of extant species of *Alcithoe* found single specimens of *Alcithoe wilsonae* and *Alcithoe* (*Teremelon*) *knoxi* to be sister taxa with sequence divergence of 0.7%. Between the *A. wilsonae*/*A. knoxi* clade and the most closely related *Alcithoe* species (*A. jaculoides*) the percentage nucleotide difference was 9.9%. Recent sampling trips and observer bycatch specimens have increased the amount of well preserved material available for molecular analysis. Using this new material the sample set of *A. wilsonae* and *A. knoxi* was increased in order to clarify this unexpected result.

In this study I document the genetic variability within the *Alcithoe wilsonae* species complex, and investigate the genetic basis of recognised morphological species. This investigation is of critical importance to the study of the evolutionary history of *Alcithoe,* as the *A. knoxi* fossil would represent the oldest within-clade calibration point in the molecular clock analysis of the genus. Are *A. wilsonae* and *A. knoxi* distinct lineages? And how should fossil *A. knoxi* be treated? The results of this analysis are relevant to the debate about the reality of fossil species, and will illustrate how a greater understanding of the evolutionary history of a lineage can be enhanced through the consideration of both molecular and fossil data.

3.2 MATERIALS AND METHODS

3.2.1 Samples

Specimens were identified by Bruce Marshall at the Museum of New Zealand Te Papa Tongarewa. Identifications were based on shell characters (e.g. Powell 1979; Bail and Limpus 2005), sampling location and depth. Sub-samples of foot tissue were taken for molecular analysis. Sample data is summarised in Table 3.1.

TABLE 3.1—Samples

Voucher number [1]	Taxon ID	General location	Latitude (deg S)	Longitude (deg E)	Depth (m)	Variable sites in alignment of 573 bp of mitochondrail *nad2* [2] 132 136 141 190 243 275 303 309 312 318 321 357 358 378 384 396 403 429 480 487 492 531 561 564 610 639 687
30009/1	acuminata	Western Chatham Rise	43.8434982	176.7050018	479	T T A G A C C A T A A A C G C T C G C T T T A A G C C
30009/2	acuminata	Western Chatham Rise	43.8434982	176.7050018	479	. . . . . G . . . .
30034	knoxi	Western Chatham Rise	44.135334	174.8439941	516-518	. . . . . G . . . . . G A . .
30037	knoxi	Western Chatham Rise	44.135334	174.8439941	516-518	. . . . . T . . . . .
30098/1	acuminata	Central Chatham Rise	44.0183334	178.5234985	767	. . . . . G . . . . .
30098/2	knoxi	Central Chatham Rise	44.0183334	178.5234985	767	. . . . . G . . . . . A . . . . T . . . . .
30098/3	acuminata	Central Chatham Rise	44.0183334	178.5234985	767	. C . . . . .
30113	knoxi	Central Chatham Rise	43.8725014	179.0214996	480-487	. . . . . G . . . . . C . . . . .
30151/1	acuminata	Central Chatham Rise	43.9840012	179.6243286	531-532	. C . . . . .
30151/2	acuminata	Central Chatham Rise	43.9840012	179.6243286	531-532	. . . A . . . . . G . . . . . T . . C . . . . .
30151/3	acuminata	Central Chatham Rise	43.9840012	179.6243286	531-532	. . . . . T . . C . . . . .
30201/1	wilsonae	Western Chatham Rise	43.386658	175.2271729	309-310	. . . . . G . C . . . T . . . . . T . . . . .
30201/2	acuminata	Western Chatham Rise	43.386658	175.2271729	309-310	. . . . . C . . . . .
30227/1	acuminata	Western Chatham Rise	43.4693336	177.14534	251-254	. . . . . T . . C . . . . .
30227/2	acuminata	Western Chatham Rise	43.4693336	177.14534	251-254	. . . . . T . . A . . . . .
30227/3	acuminata	Western Chatham Rise	43.4693336	177.14534	251-254	. . . . . T . . A . . . . .
30227/4	acuminata	Western Chatham Rise	43.4693336	177.14534	251-254	. C . . . . .
30236	knoxi	North East of Chatham Is	43.29083	184.4377	638-644	. . G . . . . . T . . C . . . . .
30265	wilsonae	North East of Chatham Is	43.51117	183.8237	194-218	. . . . . A . . . . G . . . .
30266	wilsonae	North East of Chatham Is	43.51117	183.8237	194-218	. . . . . C . . . . . G . . . . .
30297	acuminata	Central Chatham Rise	43.841	181.4105	460-462	. . . . . G . . . . .
M183116	knoxi	Central Chatham Rise	43 49.975	179 11.34	479-486	. . . . .
M190062	acuminata	Western Chatham Rise	43 30.1	176 14.2	381-385	. . . . . G . . . . . T . . C . . . . .
M190067	acuminata	Western Chatham Rise	43 34.2	176 1.1	350	. . . . . G . . . . .
M190072	acuminata	Western Chatham Rise	43 34.2	176 4.8	360-376	. . . . . G . . . .
M190079	acuminata	Western Chatham Rise	43 30.0	176 6.0	360-376	. . . . G . . . . . T .
M190088	acuminata	Western Chatham Rise	43 36.1	176 14.0	354-365	. . . . . G . G . . . . .
M190092	acuminata	Western Chatham Rise	43 29.0	176 12.0	380	. C . . . . .
M190096	acuminata	Western Chatham Rise	43 38.0	176 16.1	377-380	C . . . . . G . . . . .
M190101	acuminata	Western Chatham Rise	43 28.1	176 7.4	339-375	C . . . . . G . . . . .
M190111	wilsonae	North East of Chatham Is	43 24.0	176 12.1	330-335	. . . . . G . . . . .
M190127	acuminata	Western Chatham Rise	43 36.0	176 8.0	355-364	. . . . . G . G . . . . .
M190129	acuminata	Western Chatham Rise	43 2.0	177 1.0	354-360	. . . . . C . . . . . T
M190345	wilsonae	East of Auckland Islands	50 37.6	167 26.2	412-414	. . . . G . . . . . T . T . . C . . . . .
M274008	wilsonae	East of Auckland Islands	50 48.0	167 1.0	410	. . . . . T . T . . C . . . . .

[1] voucher numbers with an M prefix denote specimens from the Museum of New Zealand Te Papa Tongarewa collection, other specimens are from the collection of the National Institute of Water and Atmospheric Research (NIWA).

[2] positions are numbered from the start codon of the *nad2* gene, the alignment begins at position 129 relative to the start codon

An estimate of the current size of the *Alcithoe wilsonae* population on the Chatham rise was calculated. This was based on the number of specimens collected during the NIWA sampling cruise of 2007. The area sampled was estimated by multiplying the approximate total distance dredged (42 km) by the width of the dredge used (1 m). This sample area was divided by the total number of *A. wilsonae* specimens collected (25) to estimate the number of animals per square kilometer (595) on the Chatham rise. The total amount of habitable area was roughly calculated to be approximately 252,000 km^2, based on the currently known geographic and depth range of *A. wilsonae*. This estimate of the total possible habitable area was multiplied by the estimate of the expected number of *A. wilsonae* animals per square kilometre to give an approximate value of the population size of 150 million.

3.2.2 DNA extraction

DNA was extracted from foot tissue from both frozen and ethanol preserved specimens using a high-salt buffered extraction method (Norman et al. 1998), modified as follows. Approximately 0.5 mg of tissue was incubated in 300µl of high-salt buffer with 1µl of 10ng/µl ProtK shaking at 60°C for at least 16 hours. 300µl of phenol was added and the solution incubated with shaking at room temperature. Following centrifugation the aqueous phase was removed and mixed with 400µl of chloroform:isoamyl alcohol (24:1). The chloroform wash was repeated, and DNA precipitated with 95% ethanol at -20°C for 8 to 16 hours, before resuspension in 0.1 TE. DNA concentrations were determined using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc.). DNA extractions were diluted to approximately 1ng/µl for amplification.

3.2.3 PCR amplification and sequencing

DNA fragments of 823 base pairs were amplified from the mitochondrial NADH 2 (*nad*2) gene. These PCR products where produced from a primer in the preceeding serine tRNA (NG_mtSER2f: 5' - AGA AAA AAC TTG GAG TAA ARC AGG GC) and a primer in the *nad*2 gene (NG_mtND2r781: 5' – CAA AAC CAA GTA AAG GNG GYA ARC C). PCR was carried out using Red-Hot Taq (ABgene), following the manufacturer's instructions with a $MgCl_2$ concentration of 2.0mM. Standard thermal cycling conditions were followed, with 50°C annealing temperatures and 30-35 cycles, carried out in a Biometra^{tm} T1 thermocycler. PCR products were sequenced with both forward and reverse primers using BigDye Terminator v3.1 and sequenced with an ABI 3730. Sequences were edited using Sequencher (v4.6, Gene Codes Corporation, Ann Arbor, Michigan). Alignments were trimed to the start codon of *nad*2 on one end and trimed flush on the other. The resulting alignment was translated in SE-AL v2.0a11 (Rambaut 2002) to confirm nucleotide sequences did not contain errorneus stop codons that would be indicative of nuclear coded mitochondrail pseudogenes. As three samples did not sequence to the *nad*2 start codon the alignment was further trimmed to remove sites containing missing data from these samples.

3.2.4 Phylogenetic reconstruction and population structure

Trees were generated using Neighbor Joining and Maximum Likelihood, as implemented in the Geneious 4.7.5 software package (Drummond et al. 2007). Best-fit nucleotide substitution models were found using Modeltest (Posada and Crandall 1998) as implemented in HYPHY v1 (Pond et al. 2005).

A haplotype network for the *A. wilsonae* and *A. knoxi* samples was constructed manually.

Levels of population structure were estimated by analysis of molecular variance (AMOVA) using GENEALEX 6 (Peakall and Smouse 2006). Four population groupings were tested:

1/ Taxonomy, samples were grouped based on taxonomic identification as it is currently understood (*A. knoxi* – 6, *A. wilsonae wislonae* – 6, *A. w. acuminata* – 23)

2/ Geography, samples were grouped based on the general locations from which they were collected (Western Chatham rise – 20, Central Chatham rise – 9, North East of the Chatham Islands – 4, East of the Auckland Islands – 2)

3/ Depth, samples were grouped based on depth categories from which they where sampled, (less then 450m – 21, 450 to 500m – 7, greater than 500m – 7)

4/ Morphology, samples were grouped based on the ration of the width of the protoconch at the second whorl to the over all length of the shell, (large protoconch relative to shell – 7, small protoconch relative to shell – 23 (5 samples could not be measured))

Bayesian skyline analysis was carried out using BEAST v1.4.8 (Drummond et al. 2005). This analysis was based on the 573 bp *nad*2 alignment for the 35 *A. wilsonae* and *A. knoxi* samples. The root node in this analysis was calibrated using the 95% HPD interval for the divergence between two *A. wilsonae* specimens from dated analysis. A piecewise-constant skyline model was used with 9 groups set. A uniform prior on the population size was set from 100 000 to 1 billion. The MCMC chain was run for 10 million generations, sampling every 1000, and the log file was viewed in Tracer v1.4.1 to ensure the quality of the analysis. Bayesian skyline plots were generated from log files using Tracer v1.4.1.

3.2.5 Dated analysis

Molecular clock analysis was carried out using BEAST v1.4.8 (Drummond and Rambaut 2007), in order to determine the approximate divergence time of the lineage leading to *Alcithoe wilsonae* and to derive an estimate for the age of the origin of the sampled diversity. This analysis used sequence from a subset of the taxa described in chapter 1: *Alcithoe lutea*, *Alcithoe fusus*, *Alcithoe arabica*, *Alcithoe benthicola* and *Amoria hunteri*. This subset of taxa was chosen to maximise the number of nodes that could be calibrated with fossil data, and to minimise topological inconsistencies associated with low phylogenetic resolution at deeper nodes due to the use of the short *nad*2 DNA fragment available for the *A. wilsonae* samples. Two *A. wilsonae* specimens were chosen for this analysis to represent a maximum genetic difference. The samples M.190345 and 30034 were chosen based on a maximum path-length through the haplotype network that did not include any possible reversions.

The root node was calibrated with a normal distribution with a mean of 45 million years and a standard deviation of 5. This calibration was based on an estimate of the divergence of the *Alcithoe* lineage from related Australian volutes (Alan Beu, pers. com.) at between 50 - 55 Ma, but allows for error in fossil identification as the next oldest *Alcithoe* fossils are up to 43 Ma (Beu and Maxwell 1990).

Two internal nodes were calibrated with fossil data, the node representing divergence of *A. arabica* and the node representing the divergence of *A. fusus*. The earliest recognised fossil of *A. arabica* is from the Nukumaruan stage (2.4 – 1.63 Ma) of the New Zealand geological timescale, and the earliest fossil of *A. fusus* is from the Castlecliffian (1.63 - 0.34 Ma) (Beu and Maxwell 1990). These nodes were each calibrated with lognormal distributions where the 2.5% quantiles were set at the younger bound of the relevant time interval, and the majority of the 95% highest probability density interval (HPD) was fit between this and the older bound. A Yule (pure birth) prior was used as this analysis involves predominantly inter-species relationships. An HKY+G nucleotide substitution model was used as was inferred for the dataset by ModelTest. The MCMC chain was run for 10 million generations, sampling every 1000. Input XML files were generated using BEAUTi v1.4.8. Output log files from BEAST were inspected in Tracer v1.4.1 (Rambaut and Drummond 2007) to ensure stationarity of sampling traces and that summary

statistics had adequate effective sample sizes. The maximum credibility tree was generated from the resulting sample using TreeAnnotator v1.4.8, and visualised in FigTree v1.2.3.

3.3 RESULTS

3.3.1 Sequence Data

573 bp of nucleotide sequence data from the mitochondrial *nad*2 gene was generated for 35 *A. wilsoni* and *A. knoxi* samples from 24 distinct locations. An alignment of the 35 sequences contains 27 variable positions, representing 4.71% variable sites across the sample. However, pairwise comparisons amongst *A. wilsonae* and *A. knoxi* individuals ranged from 0.00% to 0.87% difference (average of 0.33%), indicating that a high proportion of the observed variability is unique to individuals. This high degree of individual specific variation is further illustrated by the haplotype diversity in the samples, as there are 25 unique haplotypes. Where more than one individual was sequenced from the same location, multiple haplotypes were detected (e.g. 30151/1, 30151/2, 30151/3). Most haplotypes (18) were found in just one individual, a few (7) were shared among locations, but only one was sampled more then once from a single collection site. The greatest distance separating locations with identical haplotypes was approximately 435 km (between M190067 and 30297).

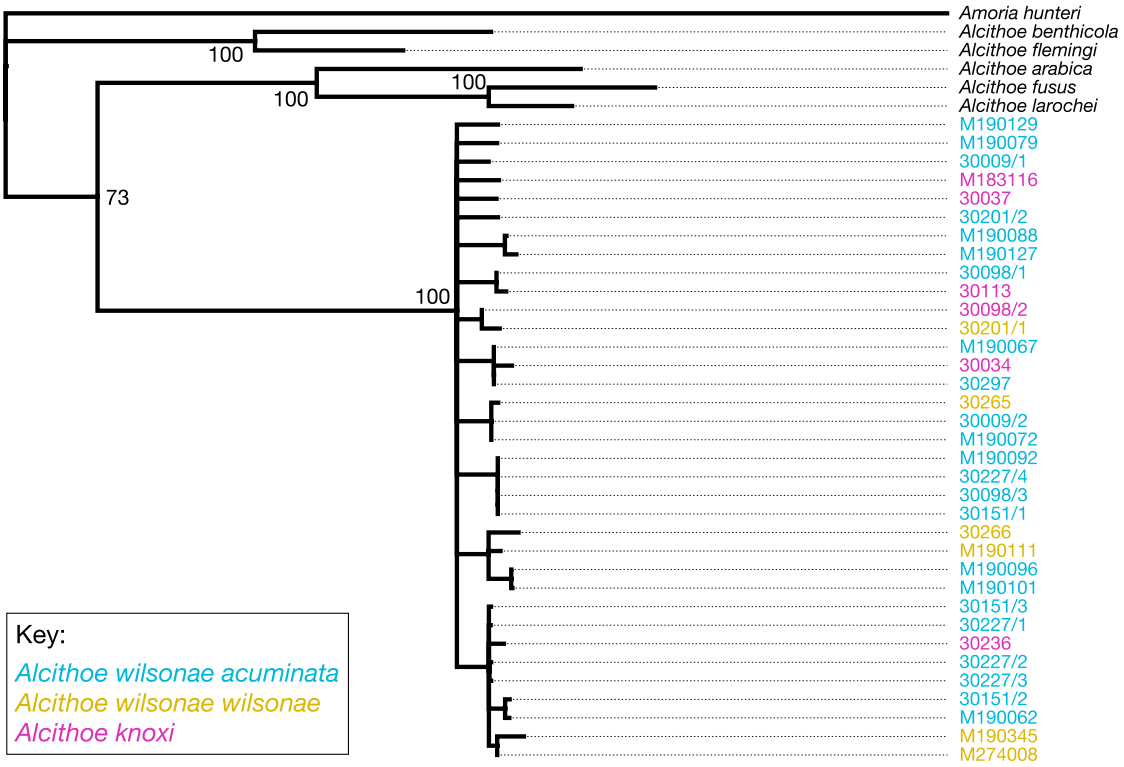

FIGURE 3.3—A Neighbor-Joining consensus tree based 701 bp of mitochondrial *nad2* recovers the monophyly of *A. wilsonae* and *A knoxi*. Consenusus support is indicated for internal nodes, except amongst the *A. wilsonae* and *A. knoxi* samples, where there is little phylogenetic resolution.

Phylogenies including both *Alcithoe wilsonae* and *A. knoxi* do not differentiate the two taxa as reciprocally monophyletic. Putative sub-species represented by specimens with distinct forms (*A. w. acuminata* n=23, *A. w. wilsonae* n=6) are not resolved as clades (Figure 3.3). A haplotype network was constructed using the 25 unique haplotypes of *Alcithoe wilsonae* and *A. knoxi* (Figure 3.4). The haplotype network of the *A. wilsonae*/*A. knoxi* samples shows no structure concordant with morpho-species. Most haplotypes are unique and many are separated by one or more unsampled haplotypes. There are no clear patterns indicative of population subdivision. A central haplotype (30009/1 and M183116, found in both *A. wilsonae* and *A. knoxi*) from which other haplotypes radiate is not more common than other haplotypes as would be expected for recent population growth (Slatkin and Hudson 1991; Posada and Crandall 2001).

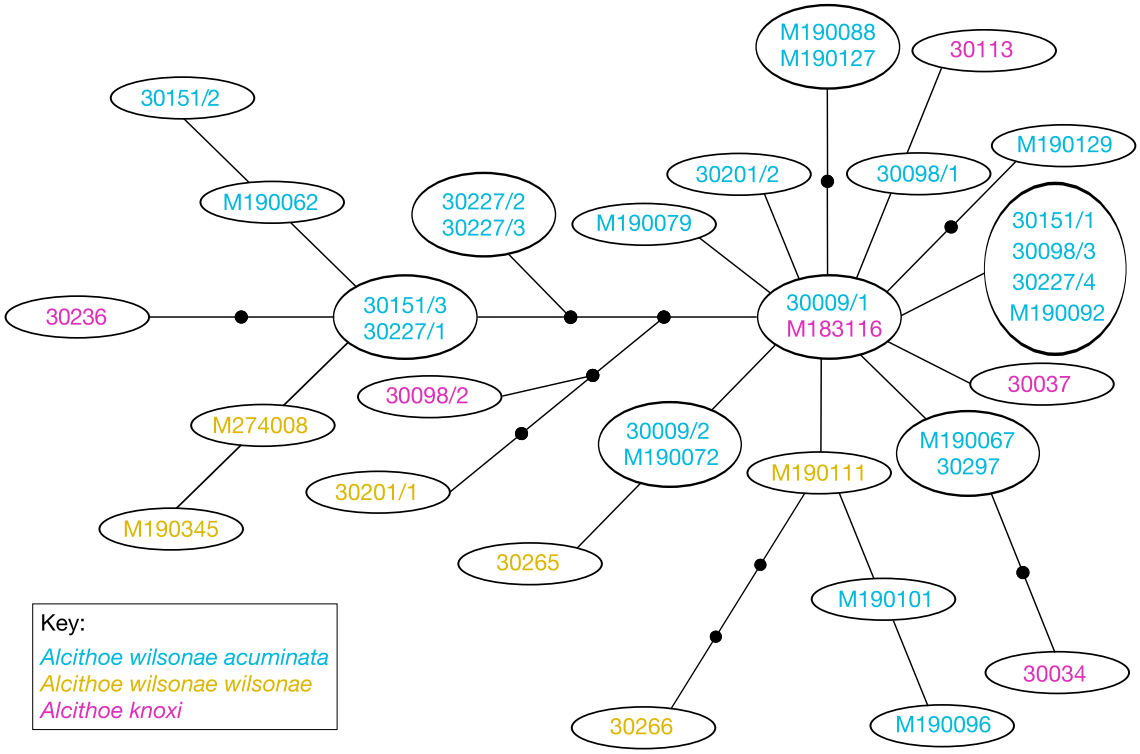

FIGURE 3.4—Haplotype network of 35 *Alcithoe wilsonae* and *Alcithoe knoxi* specimens. The network is based on the 27 variable sites in an alignment of 573 bp of *nad*2 for these 35 specimens. The spread of each of these forms throughout the network demonstrates the lack of any genetic structure related to morphology.

3.3.2 Population Genetic Structure

Snails were assigned to groups based on four attributes, currently recognised taxonomic units, general sampling location, depth, and shell morphology. Separate analyses of molecular variation based on these factors resolved little structure (Table 3.2). Only 1% of the genetic variation is explained by current taxonomic groupings. Broad geographically determined population assignments explain 9% of the genetic variability among populations. For morphology (protoconch size) 2% of the observed genetic variability is explained by among taxa differences. When population assignments are based on 3 depth bands only 5% of the genetic variation attributable to differences between populations. No obvious clustering by geography, depth, morphology, or recognised sub-species identification is apparent. This haplotype network reveals that the *knoxi* specimens are genetically indistinguishable from the *wilsonae* specimens. There is no genetic evidence to differentiate the currently recognised subspecies of *A. wilsonae*. It can also be inferred that the population is (or has been recently) large and that there is no evidence of barriers to gene flow.

TABLE 3.2—Results of AMOVA

Source of variation	df	Sum of squares	Est. Var.	Percentage variation	PhiPT	P
Taxonomic Groupings						
Among Pops	2	4.037	0.027	1%		
Within Pops	32	56.935	1.779	99%		
Total	34	60.971	1.806			
Fixation index					0.015	0.308
Geographic groupings						
Among Pops	3	8.579	0.175	9%		
Within Pops	31	52.393	1.690	91%		
Total	34	60.971	1.866			
Fixation index					0.094	0.131
Grouping by Depth						
Among Pops	2	5.448	0.101	5%		
Within Pops	32	55.524	1.735	95%		
Total	34	60.971	1.836			
Fixation index					0.055	0.099
Morphological groupings						
Among Pops	1	2.254	0.044	2%		
Within Pops	28	50.012	1.786	98%		
Total	29	52.267	1.830			
Fixation index					0.024	0.218

3.3.3 Key dates in the history of Alcithoe wilsonae

Molecular clock analysis of several *Alcithoe* species with good fossil dates indicates that the lineage giving rise to the modern *A. wilsonae* diverged between 3 and 16 million years ago (Figure 3.5). This is consistent with first appearance of the fossil *knoxi*. The inclusion of two maximally genetically divergent *A. wilsonae* samples in this analysis indicates that the modern genetic diversity is only between 50000 and 900000 years old. This estimate is dependent on the assumptions of the Yule prior for the branching process in the tree, which is not an optimal model for within-species divergence (Drummond and Rambaut 2007). However, the divergence time inferred here can be used to construct an informed root prior for further analysis using coalescent models more appropriate to intraspecies inferences.

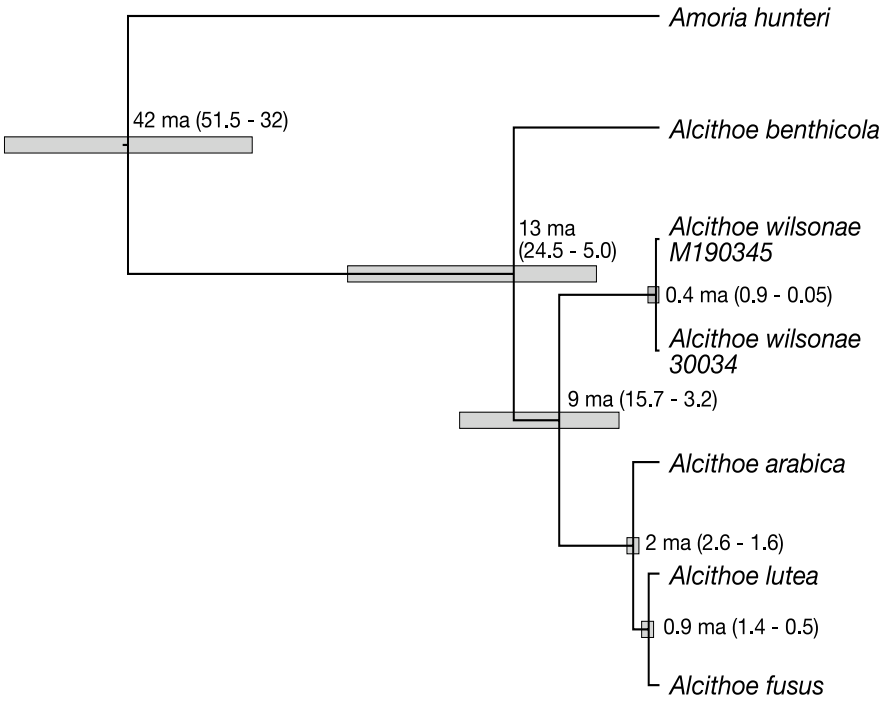

FIGURE 3.5—Molecular clock analysis indicates that *A. wilsonae* diverged around 9 million years ago. The phylogeny is based on 573 bp 701 bp of the mitochondrial *nad*2 gene. All nodes are recovered with a posterior support of 1. For this analysis the *A. arabica* and *A. fusus* nodes were calibrated using fossil data. Based on a maximal nucleotide divergence between *A. wilsonae* M190345 and *A. knoxi* 30034 this analysis suggests a divergence time within the extant population of between 0.9 and 0.05 Ma.

3.3.4 Demographic History of Alcithoe wilsonae

Bayesian skyline analysis of the 35 *A. wilsonae* sequences indicates that the population began expansion around 400 000 - 500 000 years ago, reaching a maximum within the last 50 000 years (Figure 3.6). The effective population size inferred from Bayesian skyline plot expanded from around 300 000 around 550 000 years ago, to approximately 5.5 million today. As the effective population size from the Bayesian skyline plot is the product of the population size and the generation time, to calculate the population size an estimate of the generation time is required. The generation time for *Alcithoe* species is currently unknown, but studies in related volutes from South America have suggested that they reach sexual maturity by 7 – 8 years of age (Bigatti et al. 2008). Assuming that the generation time in *A. wilsonae* is similar to 7 years, we can calculate a population size of around 785 000. An estimation of the current population size, based on sampling data (see materials and methods) indicates a total possible population size of around 150 million. These two population size estimates are considerably different.

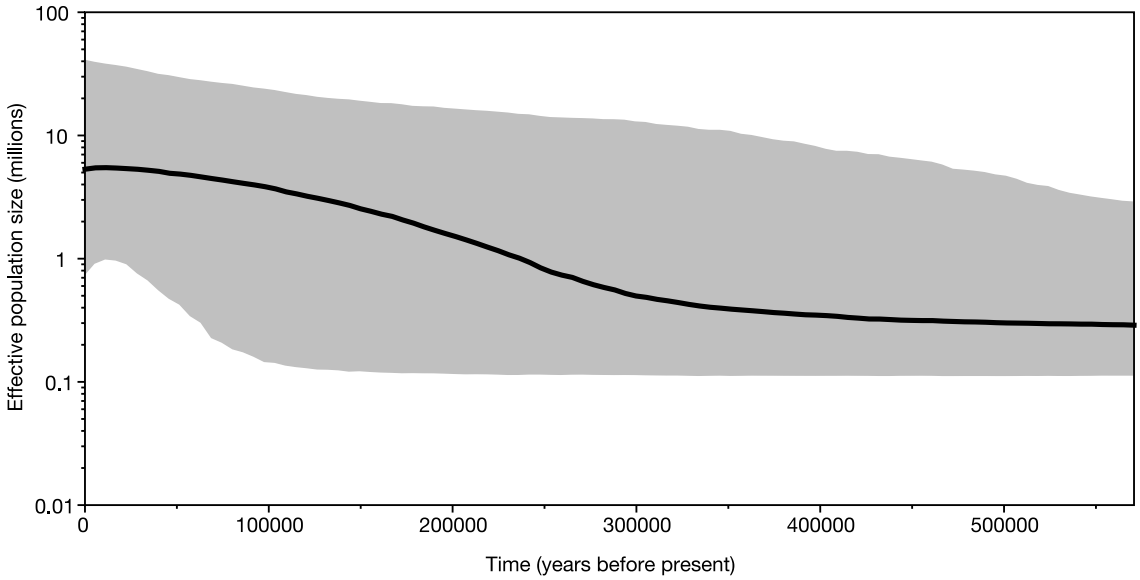

FIGURE 3.6—Bayesian skyline analysis reveals a population expansion in *A. wilsonae*. Based on a 0.9 – 0.05 Ma calibration for the deepest divergence in the sampled population, the skyline plot indicates a population increase from an effective population size of around 400 000 to over 5 million, over the last 200 000 – 300 000 years. The grey shaded area indicates the 95% HPD interval for this estimate.

3.4 DISCUSSION

Population-genetic analysis of *Alcithoe wilsonae* and *Alcithoe knoxi* individuals demonstrates that these two entities reflect morphological variation within one species. Bail and Limpus (2005) noted similarities in general shell shape between some *A. knoxi* specimens and their subspecies *A. w. acuminata*, but maintained a separation of the two species based on specific protoconch and aperture characters. This work shows that these characters are not symptomatic of species separation and are likely to be ecophenotypically plastic. One possible explanation of the large protoconch size, traditionally one of the primary characters for identifying *A. knoxi*, is an ecological preference for embryos or juveniles in deep water habitat to consume siblings while in the egg capsule (Bruce Marshall pers. com.). This behaviour is known as adelphophagy and has been described in other neogastropod species (e.g. Chaparro et al. 1999; Miloslavich and Penchaszadeh 2001). Adelphophagy leads to larger protoconch size in the surviving offspring due to an increased rate of growth as a result of consuming other individuals in the egg capsule. Additionally, there is no evidence of genetic structure concordant with currently recognised sub-species of *A. wilsonae*. The *A. wilsonae* species appears to

be one large population with no apparent barriers to gene flow. These results would seem to indicate the extensive movement of animals at some stage of their life cycle, in contrast to what would be expected for a non-broadcast spawning, direct developing, benthic gastropod.

Approximate molecular clock analysis indicates that the time of divergence of the lineage leading to modern *Alcithoe wilsonae* is consistent with the fossil record for '*Teremelon knoxi*'. Dating the origin of the modern diversity of *A. wilsonae* to around 400 000 years before present reveals a major disparity between the origin of the species and the diversification of the current population. Even allowing for substantial error in the estimation of dates there is a span of several million years between the first observed fossil for *A. knoxi* and the source of the current genetic diversity. This discrepancy could be explained if the modern forms of *A. wilsonae* emerged recently from a *knoxi*-like ancestor. However the genetic structure of the population reveals no such pattern, as *knoxi* forms and *wilsonae* forms occur throughout the haplotype network. The alternative explanation is that the morphological diversity seen in the modern population is a result of phenotypic variability maintained in the species for the majority of its 10 million year history. It also probable that this variability has been retained during past population bottlenecks. The observed morphological diversity may be symptomatic of a species has adapted to be phenotypically responsive to its local environment.

A large discrepancy exists between two population size estimates for the *Alcithoe wilsonae* species. The difference between the two estimates could be reduced if *A. wilsonae* has a shorter generation time than the 7-year estimate extrapolated from a South American volute *Ondontocymbiola magellanica* (Bigatti et al. 2008). However, in order to bring the estimates to values of a similar order of magnitude the generation time of *A. wilsonae* would have to be less than a year. It is unlikely that the biology of *A. wilsonae* is sufficiently different from *O. magellanica* to cause such a large difference, but this question warrants further study. The sampling based estimate (~150 million) is likely to be inflated as it is based on the assumption the whole geographic region is available habitat for *A. wilsonae*. It is possible that suitable habitat for *A. wilsonae* represents only a faction of the area used to estimate the population size. Conversely, the molecular based value (~785 000) is likely to be an under estimate due to the assumptions associated with the effective population size (Wang 2005). It is possible that a small sample size is

responsible for further underestimation of the population size (Wang 2005). Examining a cumulative curve of diversity versus sample size can test this: if this curve plateaus the addition of more samples will not produce significantly better estimates. However, more accurate estimates could be achieved through the addition different loci (particularly nuclear markers) for the existing samples (Felsenstein 2006). It is probable that the molecular-based estimates can be further improved through fine-tuning of the molecular analysis by optimising parameters using additional fossil and molecular data.

Synonymising *A. knoxi* with *A. wilsonae* has significant ramifications for the interpretation of the fossil record. First, the 10 million year old fossil identified as *knoxi*, and considered to be identical in form to modern specimens, represents *wilsonae*. If we extrapolate from the molecular data that similar forms of *wilsonae* were present earlier than the current populations, then we predict that these forms should be present in the fossil record, theoretically up to 10 million years ago. If *A. wilsonae* morphology does exist in the fossil record in sufficient numbers, then the demographic history suggested by Bayesian skyline analysis could be tested. By modelling the historic species occupancy, as done by Foote et al. (2007), and comparing to the predictions from the skyline plot, the accuracy of the inference could be measured. Further-more, if the ecological conditions leading to the various forms of *A. wilsonae* can be identified these could then be extrapolated to infer more detailed ecological information from fossil specimens.

3.4.1 Species identification in the fossil record.

Some molecular studies have found general concordance between snail species identified by shell morphology and genetic identification (e,g. Michaux 1987; Reid et al. 1996; Holford et al. 2009). However, snail shells are famous for convergent evolution (e.g. Booth et al. 1990; Moore and Willmer 1997; Albrecht et al. 2004). Fossil information is routinely used to estimate changes in biodiversity, species longevity and, more recently occupancy, all topics that could be undermined by failure to accurately identify species. For this reason such studies are often done at higher taxonomic levels, but as the processes of speciation and extinction are acting at the species level it is more appropriate to analyse them at this level. Fortunately, where extant representatives of a group exist it is possible to identify traits and methods that are reliable indicators of biologically discrete units. For example,

within *Alcithoe*, rather than single shell traits, morphometric analysis of large samples can produce inferences concordant with genetic data (Crampton et al. 2009).

3.5 REFERENCES

Albrecht, C., Wilke, T., Kuhn, K., Streit, B. 2004. Convergent evolution of shell shape in freshwater limpets: the African genus *Burnupia*. Zoological Journal of the Linnean Society 140:577-586.

Bail, P., Limpus, A. 2005. The recent volutes of New Zealand with a revision of the genus *Alcithoe* H. & A. Adams, 1853 in G. T. Poppe, and K. Groh, eds. A Conchological Iconography. ConchBooks, Hackenheim.

Beu, A. G., Maxwell, P. A. 1990. Cenozoic Mollusca of New Zealand. New Zealand Geological Survey Paleontological Bulletin 58:1-518.

Bigatti, G., Marzinelli, E. M., Penchaszadeh, P. E. 2008. Seasonal reproduction and sexual maturity in *Odontocymbiola magellanica* (Neogastropoda, Volutidae). Invertebrate Biology 127:314-326.

Booth, C. L., Woodruff, D. S., Gould, S. J. 1990. Lack of significant associations between allozyme heterozygosity and phenotypic traits in the land snail *Cerion*. Evolution 44:210-213.

Chaparro, O. R., Oyarzun, R. F., Vergara, A. M., Thompson, R. J. 1999. Energy investment in nurse eggs and egg capsules in Crepidula dilatata Lamarck (Castropoda, Calyptraeidae) and its influence on the hatching size of the juvenile. Journal of Experimental Marine Biology and Ecology 232:261-274.

Collin, R. 2005. Development, phylogeny, and taxonomy of *Bostrycapulus* (Caenogastropoda : Calyptraeidae), an ancient cryptic radiation. Zoological Journal of the Linnean Society 144:75-101.

Crampton, J. S., Hills, S., Fenwick, M., Morgan-Richards, M., Marshall, B., Beu, A., Hendy, A., Buick, D. 2009. Species in the fossil record: hopeless monsters or hopeful messengers? in S. Trewick, N. Hiller, and R. Cooper, eds. Geology & Genes IV. Geological Society of New Zealand, Christchurch.

de Queiroz, K. 2005. Different species problems and their resolution. Bioessays

27:1263-1269.

Dell, R. K. 1978. Additions to the New Zealand Recent molluscan fauna with notes on *Pachymelon (Palomelon)*. National Museum of New Zealand Records 1:161-176.

Drummond AJ, A. B., Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A. 2007. Geneious v3.8. Available from http://www.geneious.com/.

Drummond, A. J., Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7.

Drummond, A. J., Rambaut, A., Shapiro, B., Pybus, O. G. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology and Evolution 22:1185-1192.

Felsenstein, J. 2006. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? Molecular Biology and Evolution 23:691-700.

Foote, M., Crampton, J. S., Beu, A. G., Marshall, B. A., Cooper, R. A., Maxwell, P. A., Matcham, I. 2007. Rise and fall of species occupancy in Cenozoic fossil mollusks. Science 318:1131-1134.

Hey, J. 2006. On the failure of modern species concepts. Trends in Ecology & Evolution 21:447-450.

Holford, M., Puillandre, N., Terryn, Y., Cruaud, C., Olivera, B., Bouchet, P. 2009. Evolution of the Toxoglossa venom apparatus as inferred by molecular phylogeny of the Terebridae. Molecular Biology and Evolution 26:15-25.

Jackson, J. B. C., Cheetham, A. H. 1990. Evolutionary significance of morphospecies - a test with Cheilostome Bryozoa. Science 248:579-583.

Jackson, J. B. C., Cheetham, A. H. 1994. Phylogeny Reconstruction and the Tempo of Speciation in Cheilostome Bryozoa. Paleobiology 20:407-423.

Kartavtsev, Y. P., Zaslavskaya, N. I., Svinyna, O. V., Kijima, A. 2006. Allozyme and morphometric variability in the dogwhelk, Nucella heyseana (Gastropoda : Muricidae) from Russian and Japanese waters: evidence for a single species under different names. Invertebrate Systematics 20:771-782.

Knowles, L. L., Carstens, B. C. 2007. Delimiting species without monophyletic gene

trees. Systematic Biology 56:887-895.

Mallet, J. 1995. A species definition for the modern synthesis. Trends in Ecology & Evolution 10:294-299.

Michaux, B. 1987. An analysis of allozymic characters of 4 species of New-Zealand *Amalda* (Gastropoda, Olividae, Ancillinae). New Zealand Journal of Zoology 14:359-366.

Miloslavich, P., Penchaszadeh, P. E. 2001. Adelphophagy and cannibalism during early development of *Crucibulum auricula* (Gmelin,1791) (Gastropoda : Calyptraeidae) from the Venezuelan Caribbean. Nautilus 115:39-44.

Moore, J., Willmer, P. 1997. Convergent evolution in invertebrates. Biological Reviews 72:1-60.

Nakano, T., Spencer, H. G. 2007. Simultaneous polyphenism and cryptic species in an intertidal limpet from New Zealand. Molecular Phylogenetics and Evolution 45:470-479.

Norman, J., Olsen, P., Christidis, L. 1998. Molecular genetics confirms taxonomic affinities of the endangered Norfolk Island Boobook Owl *Ninox novaeseelandiae undulata*. Biological Conservation 86:33-36.

Palmer, A. R. 1985. Quantum changes in gastropod shell morphology need not reflect speciation. Evolution 39:699-705.

Peakall, R., Smouse, P. E. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes 6:288-295.

Pond, S. L. K., Frost, S. D. W., Muse, S. V. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676-679.

Posada, D., Crandall, K. A. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818.

Posada, D., Crandall, K. A. 2001. Intraspecific gene genealogies: trees grafting into networks. Trends in Ecology & Evolution 16:37-45.

Powell, A. W. B. 1979. New Zealand Mollusca. Marine, land and freshwater shells. Collins, Auckland.

Rambaut, A. 2002. Se-Al: Sequence Alignment Editor. Avilable at

http://tree.bio.ed.ac.uk/software/seal/.

Rambaut, A., Drummond, A. J. 2007. Tracer v1.4. Available from http://beast.bio.ed.ac.uk/Tracer

Reid, D. G., Rumbak, E., Thomas, R. H. 1996. DNA, morphology and fossils: Phylogeny and evolutionary rates of the gastropod genus *Littorina*. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 351:877-895.

Samadi, S., Barberousse, A. 2006. The tree, the network, and the species. Biological Journal of the Linnean Society 89:509-521.

Samadi, S., David, P., Jarne, P. 2000. Variation of shell shape in the clonal snail *Melanoides tuberculata* and its consequences for the interpretation of fossil series. Evolution 54:492-502.

Slatkin, M., Hudson, R. R. 1991. Pairwise comparisons of mitochondrial-DNA sequences in stable and exponentially growing populations. Genetics 129:555-562.

Templeton, A. R. 2001. Using phylogeographic analyses of gene trees to test species status and processes. Molecular Ecology 10:779-791.

Wang, J. L. 2005. Estimation of effective population sizes from data on genetic markers. Philosophical Transactions of the Royal Society B-Biological Sciences 360:1395-1409.

CHAPTER FOUR

4 MOLECULAR-CLOCK ANALYSIS OF *ALCITHOE*

4.1 INTRODUCTION

The merits of inferences made in molecular dating analyses are highly dependent on the quality of the input data and the suitability of assumptions made regarding calibrations and models applied to that data. This dependency is particularly true of Bayesian methods, such as BEAST, in which the researcher is able to place informative prior probability distributions on a very large number of parameters (Drummond and Rambaut 2007). In fact, Drummond and Rambaut (2007) state that in BEAST 'the shear number of possible combinations of models mean that many will be untried and untested'. The great advantage of this freedom is that it allows the application of additional data, such as that from the fossil record, to inform the inference of molecular based phylogenies. However, it is not possible to give general guidelines of use for some parameters, as their application is highly dependent on the data being analysed (Lepage et al. 2007). The onus is therefore on the user of such methods to ensure that the prior probability distributions that they have chosen are the most appropriate for the dataset in question.

The elucidation of a robust molecular phylogeny for the volute genus *Alcithoe* (see Chapter 2) provides and ideal dataset for molecular-clock analysis. *Alcithoe* are a group of predatory, benthic, direct developing, marine gastropods. Species range from the intertidal zone to the edge of the continental shelf. *Alcithoe* are the only currently recognised extant genus of the tribe Alcithoini, which has a long and well-studied fossil record in New Zealand. In order to infer robust times of divergence

and rates of molecular evolution in this group the most appropriate model parameters need to be identified. The molecular dataset for the *Alcithoe* will serve as an example to demonstrate the process of parameter testing in a Bayesian framework.

4.1.1 Calibrating the molecular clock

Accuracy in estimates of divergence times and rates of evolution obtained from molecular clock analysis is dependent on the data used to calibrate the analysis. As a result there has been extensive discussion regarding various aspects of fossil calibration. The discussion has centred on two main themes: the quality of the fossil record, and the adequate allowance for error in the paleontological data. Many of the early molecular divergence dating studies where rightly criticised for using point based calibration methods that removed any uncertainty from the fossil data, leading to the illusion of precision in the inferred date estimates (Graur and Martin 2004; Pulquerio and Nichols 2007). More recently the use of uncertainty distributions has allowed the integration of known error in fossil calibrations (Drummond et al. 2006; Yang and Rannala 2006; Ho 2009). The quality of the fossil record can be problematic as preservation is not consistent over time, nor are currently exposed rock outcrops necessarily representative of the full extent of that time. Furthermore, the biogeographical dynamics of species origination, population growth and extinction are likely to further compound interpretation of the fossil record (Liow and Stenseth 2007). These problems lead to the issue of detectability in the fossil record, where the first known fossil specimen of a taxon may not represent the earliest occurrence of that taxon. To a large extent these errors have been identified in datasets representing taxa that do not have particularly rich fossil records, and such datasets represent the bulk of molecular-clock analysis to date. One solution is to explore data for which there is a much more complete fossil record. A lineage with a well studied fossil history may allow for quantification of preservation biases that lead to errors related to detectability. Additionally it will provide a greater number of more robust calibrations. Together these features will lead to an improvement in the quality of date inferences, as single calibrations can bias results (Rutschmann et al. 2007). However, it has been noted that the effectiveness of multiple calibration points is dependent on correct specification of rate variation across the tree (Pulquerio and Nichols 2007). Various methods of testing the placement of calibration points and/or probability distributions have

been suggested in order to identify calibration data that are inconsistent with the sequence data (Near and Sanderson 2004; Rutschmann et al. 2007; Sanders and Lee 2007). However, the influence of an incorrect prior will be less problematic when multiple priors are included with some estimate of their degree of error (Ho and Phillips 2009) as the combined effect of other calibrations should correct an erroneous one. The main issue with multiple calibrations is that they are not independent. They are connected by branches in the tree, so are able to influence each other in such a way that the joint prior for the analysis might have different distributions than the originally set nodal prior (Ho and Phillips 2009). To diagnose such errors, as well as other prior interactions (such as with the tree prior), it is useful to test the joint prior in the absence of sequence data (Ho and Phillips 2009) to ensure that the prior distributions do not deviate significantly from the distribution set or the calibrations.

4.1.2 Molecular clock models

The accumulation of evidence for widespread violation of the assumption of a strict molecular clock (Welch and Bromham 2005) has led to the development of ever more sophisticated models. The relaxation of the clock was first modelled as an autocorrelated process, then as an uncorrelated process. Autocorrelated models make two assumptions: 1/ that the DNA substitution rate is linked to a variety of heritable traits (Lepage et al. 2006; Welch et al. 2008; Ho 2009), and 2/ rates of mutation (in the individual) are correlated with the substitution rate (in the species) (Bromham and Penny 2003; Ho 2009). However, these assumptions seem to be highly dependent on taxonomic scale (Ho 2009) and sampling (Drummond et al. 2006; Lepage et al. 2007). A high degree of autocorrelation might be expected in an intraspecific analysis due to the close recently shared genetic history of the samples, but over larger evolutionary timescales the relationship between life-history and substitution rates is likely to degrade (Ho 2009). The pitfalls of assuming an autocorrelated clock can be avoided by using an uncorrelated relaxed clock (Drummond et al. 2006). Meaningful inferences under a relaxed clock model are highly dependent on a robust tree, an accurate DNA substitution model and informative calibrations. Several methods for relaxing the molecular clock have been developed, but in all cases the model underling the relaxation of the molecular clock can exert significant influence on estimated dates (Lepage et al. 2007).

4.1.3 Nucleotide substitution models

Branch length inference is sensitive to the way in which nucleotide substitutions are modelled though time. Accurate estimation of node heights and rates of change on branches will then be affected by the substitution model used. Even though more accurate models may not have a large effect on the resulting date estimates, they may nonetheless, have important implications on biological interpretation of the dates (Ware et al. 2008). It has not been established how well suited the models of DNA substitution that are currently used are to the accurate estimation of branch length in a molecular clock framework (Phillips 2009), although the adequacy of current models to approximate natural substitution heterogeneity has been questioned (e.g. Whelan 2008). The calibration regime in BEAST can be setup in a way that can reduce the potential for branch length estimation bias, but not remove it (Phillips 2009).

4.1.4 Speciation or Tree prior

The tree prior is a model for the branching process in a tree. In an interspecies analysis it dictates the tree shape and rate of branching. There are few general guidelines for this parameter, as the optimal setting is highly dependent on the data being analysed (Drummond et al. 2006; Lepage et al. 2007). Three settings for non-demographic data (i.e. interspecific relationships) are currently available in BEAST. A pure birth model, the Yule prior (Yule 1924), models a single parameter, the rate of lineage origination, and tends to generate symmetric trees. More recently a Birth/Death model has been implemented (Gernhard 2008), which models two parameters, the rate of cladogenesis (lineage origination) and the rate of extinction. It is also possible to place a uniform prior on the branching process. In the presence of extensive node calibration data a uniform prior will allow the branching process to be dictated by that data (sequence data and node calibrations). To ensure that the tree prior does not have an anomalous effect on the resulting posterior distributions it is good practice to test the priors in a Bayesian analysis without sequence data and assess the relative influence of the data and the priors (Drummond et al. 2006).

4.1.5 Model discrimination

Model comparison in a Bayesian framework makes use of Bayes factors (Kass and Raftery 1995), which are analogous to likelihood-ratio tests. A Bayes factor is the ratio of the marginal likelihoods of two models on a given dataset and is a measure of which model is a better fit to the data. When under assessment by Bayes-factors, models with a greater number of parameters are inherently penalised (Lepage et al. 2007). Bayes-factor values are often expressed in terms of the Log10 of the Bayes factor (Log(10)BF). In this notation the Bayes factor (BF) for a comparison of model 1 versus model 2, model 1 will be a better fit for the data by a factor of 10^{BF}. A 100 fold better fit is considered decisive evidence that a given model is a better fit to the data, so a Log(10)BF vaule of 2 or more will discriminate a better model (Kass and Raftery 1995). Where Bayes factors are unable to discriminate between models it is sometimes possible to make a qualitative assessment based on the shape of posterior distributions. A posterior distribution with an anomalous shape, such as multimodality, is likely to be an indication of problems in the model used.

4.1.6 Rates of evolution

A well-calibrated dated phylogeny allows the inference of the rate of molecular evolution in a group. Since the early work in the 1960s and 70s to derive rates of evolution from molecular data (see review in Wilson et al. 1977), a range of rates has been reported. These studies have covered a variety of taxa at a range of taxonomic depths. These rates have been derived from an assortment of different types of data (e.g. restriction endonuclease cleave profiles, protein sequence, RFLPs, and nucleotide sequence).

Rates of mtDNA substitution have been reported for molluscs based on DNA sequence data, from *cox*1, *cyt*B, 16S or 12S sequence fragments. In marine gastropods rates range from 0.5% per million years in *Umbonium* (Ozawa and Okamoto 1993) and 0.6% per million years in *Littorina* (Reid et al. 1996), to 2.4% per million years in *Tegula* (Hellberg and Vacquier 1999) and 3-4% per million years in *Nucella* (Collins et al. 1996). Somewhat faster rates have been inferred for land snails, but range from 1% per million years in *Abinaria* (Douris et al. 1998) to 10% per million years in *Mandarina* (Chiba 1999). Rates of between 0.7 – 1.2 % per million years have been inferred for bivalve species separated by the isthmus of Panama (Marko 2002), which are generally concordant with rates inferred for other invertebrate groups (e.g. Lynch and Jarrell 1993; Knowlton and Weigt 1998)

4.1.7 The New Zealand Cenozoic mollusc fossil record

Many of the concerns with fossil calibration have been related to the quality of the fossil data. It is therefore surprising that few molecular-clock studies been based on the fossil record of molluscs, given the rich and well-studied nature of this data. The fossil record of the New Zealand Cenozoic marine mollusc fauna is well studied and well documented (e.g. Beu and Maxwell 1990). The accumulation of much of this paleontological record into an electronic database (www.FRED.org.nz) (Crampton et al. 2003) has allowed the systematic analysis of various elements of the data. Amongst these elements are sources of bias thought to contribute to calibration error in molecular clock analysis such as preservation and collection biases (Crampton et al. 2003; Cooper et al. 2006; Crampton et al. 2006).

The completeness of this fossil record provides multiple calibrations across a broad taxonomic distribution, and deep into evolutionary history. While few specimens have been directly dated by isotope analysis, the stratigraphy in which they occur is largely well understood and described (Cooper 2004). Furthermore, data regarding large scale sampling biases and broad occurrence data have been used to generate sampling probability data for all the stages of the New Zealand Cenozoic series (Crampton et al. 2006). This data is easily adapted for node calibration of molecular phylogenies where fossil species exist.

In order to best assess the quality of divergence date and molecular rate inferences from a molecular-clock analysis it is desirable to study a well resolved phylogeny with associated fossil data for extant species. The phylogeny of the New Zealand volute genus *Alcithoe* provides such a dataset where the tree topology is very stable and for which there is a well-documented paleontological history. The signals in the underlying nucleotide data have been analysed and sources of inconsistent signal identified and minimised (Chapter 2). Here I examine the effect combinations of various prior probability distributions have on a mitochondrial nucleotide dataset in a Bayesian framework, using the BEAST software package (Drummond and Rambaut 2007). Using the most appropriate modelling parameters for this dataset I will elucidate the divergence times of the *Alcithoe*. Finally I report the inferred rate of molecular evolution for this group of gastropod molluscs, based in this mitochondrial DNA dataset.

4.2 MATERIALS AND METHODS

4.2.1 Fossil species calibration priors

Three extant species of *Alcithoe*, *A. arabica*, *A. fusus* and *A. wilsonae*, are well known from the New Zealand Cenozoic fossil record. Verified specimens of *A. arabica* appear in the Nukumaruan stage, between 2.4 and 1.63 millon years before present. *A. fusus* first appears 1.63 – 0.34 million years ago, in the Castlecliffian stage. Previous analysis of population-level data has shown that *Teremelon knoxi* is a synonym of *A. wilsonae*, which means that the fossil record of *A. wilsonae* extends back to the Tongaporutuan stage, 10.92 – 6.5 million years before present. Paleontological data for these species was used to calibrate three nodes in the molecular phylogeny of the *Alcithoe* (Figure 4.1A).

Priors for each of the three calibrations were based on the times of the stages in which fossils of each of the calibrating species first appear, and were modified using per-stage fossil sampling probabilities from Crampton et al. (2006). These values represent the probability that, for a species known from before and after a given stage, that species is found in the stage of interest. This measure provides an estimate of the likelihood that a species, for which a fossil is first found in a given stage, did not occur prior to that stage but has not yet been sampled. In the context of these calibration priors I take these values as proxies for the probability that a fossil species first found in a given stage actually originated in that stage and not before.

Discrete probability distributions were derived for the three calibrations by dividing probabilities, into discrete time bins for each stage, based on the R values for each stage. Firstly the value R for a given stage was assigned to the time-frame of that stage, For example the per-stage fossil sampling probability for the Castlecliffian (in which *A. fusus* is first found) is 0.863, and this stage spans the time from 1.63 mya to 0.34 mya. Next, the probability 1-R (representing the probability that a species originated before that stage and was not sampled) was distributed over the preceding stages. The quantity of the 1-R value assigned to each preceding stage was determined by multiplying the probability $1\text{-}R_1$ of the stage of interest by the $1\text{-}R_2$ value of the preceding stage. This value $(1\text{-}R_1)x(1\text{-}R_2)$ was subtracted from the $1\text{-}R_1$ value, and the remainder was multiplied by $1\text{-}R_3$, the 1-R probability for the next

preceding stage (e.g. $((1-R_1)-(1-R_2))x(1-R_3))$). This calculation was repeated for each preceding stage until the cumulative probability of all stages summed to 1. The probability distributions derived in this way are shown in Figure 4.1B.

These probability distributions are most closely approximated by exponential curves, therefore exponential means and standard deviations were calculated from these distributions and used to create approximate exponential distribution priors, shown in Figure 4.1B. The mean values for each exponential curve used to calibrate each of the selected nodes in BEAST were as follows; 1.30 for the *A. fusus* calibration, 2.28 for the *A. arabica* calibration, and 9,49 for the *A. wilsonae* calibration. Each of these priors had a zero-offset set to the end of the geological stage from which that species is first recorded. The zero-offset on an exponential distribution sets the minimum age for the node that the distribution calibrates.

Greater uncertainty is associated with deep nodes in the *Alcithoe* phylogeny due to a more patchy fossil record for the group. In order to encompass this uncertainty, less stringent priors were used for deep calibrations. The split between the outgroup taxa, *Amoria* and *Cymbiola*, is thought to have occurred during the late Oligocene or very early Miocene (Darragh 1989), approximately 25-30 million years ago. In order to reflect the imprecise nature of this calibration a normal distribution for the prior probability was applied to this node where 95% of the distribution falls between 25 and 30 million years. Calibration of the root of the tree is based on the earliest occurrence of a fossil attributable to the Alcithoini lineage, dated at around 50-55 ma (Beu and Maxwell 1990).

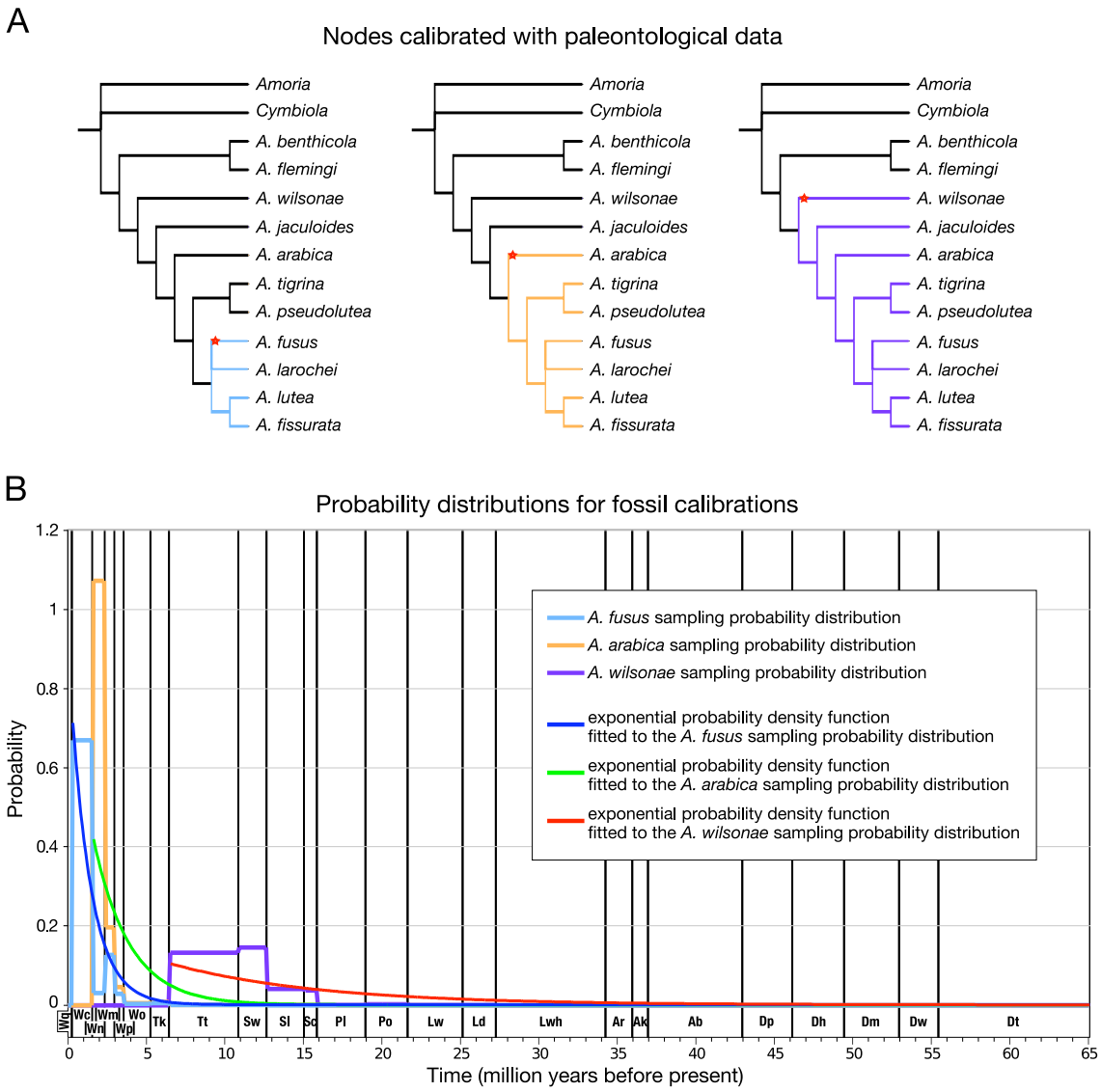

Figure 4.1—Construction of calibration prior distributions for molecular-clock analysis of *Alcithoe*. (A) Taxa sets included in nodes calibrated with fossil data for selected species (*A. fusus*, *A. arabica* and *A. wilsonae*) are colour-coded. Red stars indicate the assumed positions of the oldest known fossils for each of the calibrating species. (B) Probability distributions derived from per-stage sampling probability data Crampton et al. 2006). Sampling probability distributions were calculated for the three calibrating species: *A. wilsonae*, first occurs in the Tongaportuan; *A. arabica*, first occurs in the Nukumaruan, *A. fusus* first occurs in the Castlecliffian. The distributions represent the probability that each species originated in a given stage, and the area enclosed by each distribution sums to a probability of 1. For each sampling probability distribution an exponential probability distribution function was calculated and used as the prior distribution for each node in Bayesian molecular-clock analysis in BEAST. New Zealand geological stages are indicated; Wq – Haweran, Wc – Castlecliffian, Wn – Nukumaruan, Wm – Mangapanian, Wp – Waipipian, Wo – Opoitian, Tk – Kapitean, Tt – Tongaporutuan, Sw – Waiauan, Sl – Lillburnian, Sc – Clifdenian, Pl – Altonian, Po – Otaian, Lw – Waitakian, Ld – Duntroonian, Lwh – Whaingaroan, Ar – Runangan, Ak – Kaiatan, Ab – Bortonian, Dp – Porangan, Dh – Heretaungan, Dm – Mangaorapan, Dw – Waipawan.

Comparisons were made between LogNormal and Uniform prior distributions on the root in order to assess both the extent to which the root prior affects the tree and the effect of the internal tree priors on the root. LogNormal priors were applied such that the modes of the distributions were around 50 -55 ma, and the standard deviation captured the potential uncertainty in the interpretation of the early fossil record and the possibility that the separation of the Alcithioni from *Amoria* and *Cymbioloa* (outgroup taxa) significantly predated the earliest known fossil. In addition, as the next oldest fossils of Alcithoini lineages occur in the Bortonian, starting at 43 ma, the distribution was allowed to significantly predate the 55 ma fossil date in order to allow for misinterpretaion of the fossil record. Three LogNormal distributions were tested for the tree root prior; 1/ mean set to 3.95, standard deviation set to 0.15, 2/ mean set to 4.05, standard deviation set to 0.21, 3/ mean set to 3.26, standard deviation set to 0.2 and a zero offset of 25 million years. A fourth calibration regime consisted of a Uniform distribution with a maximum possible age of the root at 75 ma and a minimum possible age at 25 ma. Analyses based on the LogNormal priors assess the effect of the root prior on the internal nodes of the tree, while the Uniform prior should assess the effect of the internal priors on the root, as the internal calibrations should inform the age of the root.

4.2.2 Sequence data

A refined mitochondrial DNA sequence dataset, described in Chapter 2, was used to reconstruct evolutionary relationships and infer divergence dates and molecular rates. This dataset consists of an alignment of mitochondrial DNA sequences from 13 taxa, is 6777 bp in length, and consists of a concatenated set of the genes *nad3*, *nad2*, *cox1*, *atp6*, 16S, 12S, and 9 tRNAs. The underling signals in this dataset have been thoroughly analysed, and it is considered to generate a robust phylogenetic tree.

4.2.3 Molecular clock analysis

Estimation of node ages and molecular rates was carried out using BEAST v1.4.8 (Drummond and Rambaut 2007). Xml files were generated in BEAUTi v1.4.8. In cases where parameters could not be altered in BEAUTi, such as setting a starting tree, xml files were manually edited. MCMC results were visualised in Tracer v1.4.1. A broad range of parameters were tested in several analyses, alternative models

were compared by Bayes factors calculated using Tracer v1.4.1. Maximum clade credibility trees were generated using TreeAnnotator v1.4.8, and visualised in FigTree v1.2.3. In order to further examine taxa with possible elevated rates of substitution a Perl script was written (extract_node_info.pl, written by Tim White) to extract and collate data from the log file output from BEAST. For every state logged in the file this script identified if a rate inferred on a given branch A was greater then a rate on another given branch B. If A was found to be greater than B a count of 1 was logged, if not the count was 0. The sum was then calculated for every logged state in the BEAST output log file, and when divided by the total number of logged states gave the posterior probability that branch A had a faster rate of substitution then branch B.

4.3 RESULTS

4.3.1 Fossil species calibration priors

The probability distributions for three *Alcithoe* species identified in the New Zealand Cenozoic fossil record were most closely approximated by exponential curves, and were used to create approximate exponential distribution priors. However, in testing the interactions of the calibration priors by MCMC sampling from only the joint-prior for the complete model (i.e. with no DNA data), it was found that the combination of zero-offset exponential and LogNormal priors (used for other calibrations) led to anomalous multi-modal distributions. Posterior distributions of node ages derived from zero-offset exponential prior distributions exhibit additional peaks that are coincident with the distributions of earlier calibrated nodes, but are outside of the prescribed range of the prior for the node in question (Figure 4.2). When these priors are applied to sequence data the extraneous peaks allowed by the joint-prior could lead to inaccurate posterior samples that are skewed toward mean ages that are too old for the relevant node. To avoid these errors the three fossil species calibrations were each converted to LogNormal distributions. These LogNormal prior probability distributions were set so that the mode of the distribution was centred on the mid point of the geological stage in which the species is first observed and the tail approximated the decay of the cumulative distributions calculated for each calibration species. Zero offsets were not applied as the LogNormal distributions could be set to cover the appropriate time frame without the need to apply a hard lower bound, as was necessary for the exponential distributions.

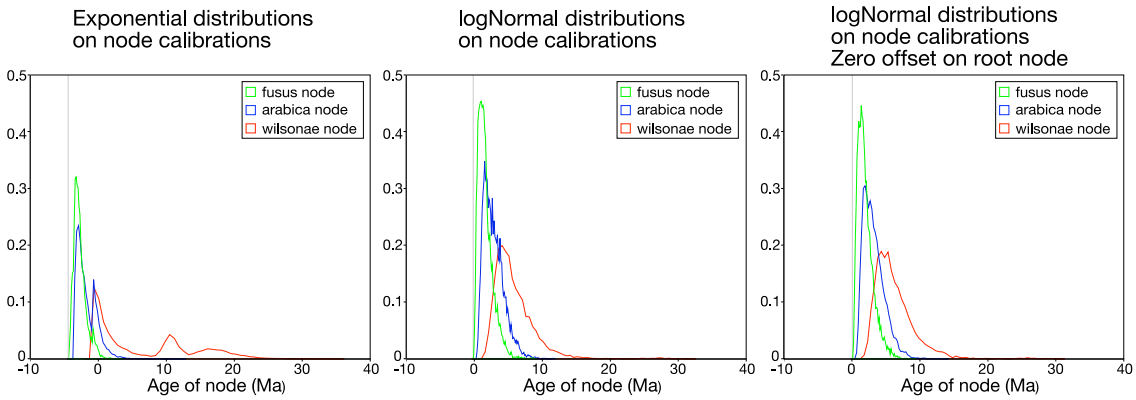

FIGURE 4.2—Posterior distributions for joint-prior tests (no DNA data) of alternative calibration regimes for internal fossil calibrated nodes. Exponential priors exhibit artifactual interactions not seen in the lognormal or zero offset lognormal distribution. Older calibrations appear to co-opt Bayesian sampling from younger priors.

4.3.2 Outgroup calibration

A normal distribution for the prior probability of the split between the outgroup taxa, *Amoria* and *Cymbiola*, was set with 95% of the distribution between 25 and 30 million years. This calibration appears to have little effect on the ingroup, and is not sensitive to alterations of other parameters. However, in combination with the root prior this calibration does determine the length of the branch from the root to the outgroup node. It is therefore important in inferring the rate of evolution in the outgroup.

4.3.3 Tree-root calibration

Calibration of the root of the tree is based on the earliest occurrence of a fossil attributable to the Alcithoini lineage, dated at around 50-55 ma (Beu and Maxwell 1990). Three root-calibration models were examined, a LogNormal distribution, a uniform distribution and a Lognormal distribution with a zero-offset. The joint-prior distributions including these models were tested in the absence of sequence data (Figure 4.3). These tests showed that when the root prior distribution overlaps with the out-group distribution a large peak is seen in the joint prior distribution for the root, and that this peak is correlated with the out-group prior distribution.

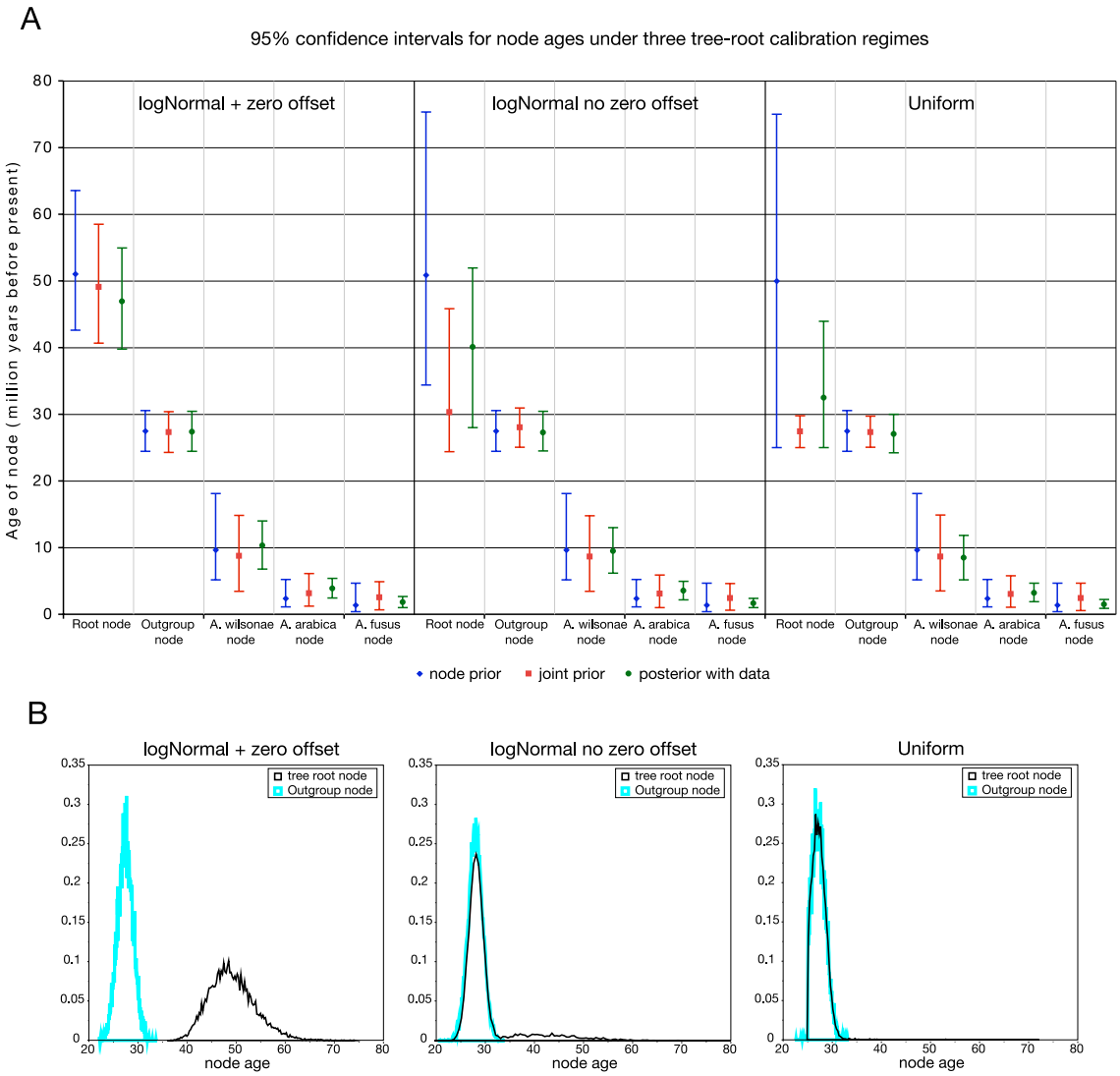

FIGURE 4.3—A zero offset lognormal prior distribution on the root-node calibration is more robust than other root-node calibrations, but the fossil calibrated crown group is not significantly effected the root-node calibration. (A) Interactions between calibrated nodes were tested under three root-node calibration regimes; a zero-offset lognormal distribution, a lognormal distribution with no zero-offset and a uniform distribution. Inferred node ages were inferred from the sequence data, and the posterior distribution of the joint prior was calculated. These distributions were both compared to the individual input prior distributions. A clear interaction between the outgroup prior and the not-offset lognormal and uniform priors that is capable of misleading the data. The zero-offset lognormal prior does not suffer from this interaction, and the fossil calibrated nodes are not significantly effected. (B) Further examination of this effect shows that when the root-node calibration overlaps with the outgroup calibration, MCMC sampling becomes dominated by the outgroup calibration leading to the inference of considerably younger root heights than where the effect is not observed.

This interaction causes a considerable change in the mean for the age of the root node in the joint prior. When a zero-offset was applied to a LogNormal distribution to prevent any overlap this interaction was no longer observed. Setting the younger bound of a uniform prior so that it is older then the upper 95% HPD interval of the out-group calibration prior also avoids erroneous interactions of these priors (data not shown). A comparison of the prior and joint-prior distributions with posterior distributions from analyses of sequence data under these models shows that the skewed joint-prior distributions cause a considerable reduction in the inferred age of the root node. However, different root-calibration priors have little effect on calibrated internal nodes.

An older root node will lead to older internal nodes, but the difference is not significant based on 95% HPD intervals. Even though tests of the joint-priors illustrate an anomalous interaction between two calibration priors the Bayes factors calculated for these analyses indicate that none of the root prior models have a significantly better fit to the data (Table 4.1). In order to avoid interactions between priors misleading subsequent results, further analysis will use a LogNormal prior with a zero-offset to insure the independence of the root prior distribution.

4.3.4 Speciation priors

Although even a cursory study of the fossil record provides evidence that Alcithoini evolution has included extinction, a comparison of the node ages produced using different speciation priors indicates that there is little difference in the results yielded by use of the Yule (pure birth) process versus the Birth/Death process (Figure 4.4). An examination of the joint-prior under each of the models indicates there is no detrimental interaction between these speciation priors and other prior distributions. Application of a uniform prior to the speciation model increases the inferred age of each node, but the difference is not significant based on the 95% HPD intervals. A comparison of Bayes factors indicates that there is no significant difference in fit of any of these priors to the data Table 4.1. The Yule prior is capable of inferring consistent node ages in this dataset, suggesting that extinction has not contributed greatly to the branching process within the lineages represented within the sampled taxa. The Birth/Death process should theoretically be a more real model for this data, as it encompasses extinction that is known to have occurred. Therefore the Birth/Death model was chosen as the most appropriate for this dataset.

TABLE 4.1—Bayes factors for all model comparisons carried out.

Tree root calibration test						
Root calibration distribution	ln P(model | data)	S.E.	log10 Bayes Factors A/	B/	C/	
A/ logNormal + zero-offset	-24663.559	+/- 0.20	-	0.00	0.10	
B/ logNormal (no zero-offset)	-24663.558	+/- 0.17	0.00	-	0.10	
C/ uniform	-24663.798	+/- 0.18	-0.10	-0.10	-	
Speciation Prior Test						
Speciation Prior	ln P(model | data)	S.E.	log10 Bayes Factors A/	B/	C/	
A/ Birth/Death moldel	-24663.559	+/- 0.20	-	0.09	-0.01	
B/ Yule Process	-24663.763	+/- 0.19	-0.09	-	-0.10	
C/ Uniform distribution	-24663.531	+/- 0.20	0.01	0.10	-	
Nucleotide Substitution Model Test 1						
Nucleotide Substitution Model	ln P(model | data)	S.E.	log10 Bayes Factors A/	B/	C/	
A/ TVM+I+G	-24663.434	+/- 0.18	-	0.16	2.24	
B/ GTR+I+G	-24663.798	+/- 0.18	-0.16	-	2.09	
C/ HKY+I+G	-24668.602	+/- 0.21	-2.24	-2.09	-	
Nucleotide Substitution Model Test 2						
Nucleotide Substitution Model	ln P(model | data)	S.E.	log10 Bayes Factors A/	B/	C/	D/
A/ TVM+I+G	-24663.16	+/- 0.12	-	7.41	36.7	592
B/ TVM+I	-24680.225	+/- 0.10	-7.41	-	29.3	584
C/ TVM+G	-24747.715	+/- 0.13	-36.7	-29.3	-	555
D/ TVM	-26025.185	+/- 0.10	-592	-584	-555	-
Molecular Clock Model Test						
Molecular clock model	ln P(model | data)	S.E.	log10 Bayes Factors A/	B/	C/	
A/ Relaxed logNormal	-24663.559	+/- 0.21	-	16.3	0.04	
B/ Strict clock	-24701.175	+/- 0.15	-16.3	-	-16.3	
C/ relaxed exponential	-24663.661	+/- 0.18	-0.04	16.3	-	
Tree Topology Test						
Tree topology	ln P(model | data)	S.E.	log10 Bayes Factors A/	B/	C/	D/
A/ *larochei* diverged first	-24662.886	+/- 0.17	-	2.86	1.23	1.11
B/ *fusus* diverged first	-24669.478	+/- 0.19	-2.86	-	-1.64	-1.75
C/ *larochei* and *fusus* sister	-24665.714	+/- 0.17	-1.23	1.64	-	-0.12
D/ trifurcation	-24665.445	+/- 0.15	-1.11	1.75	0.12	-

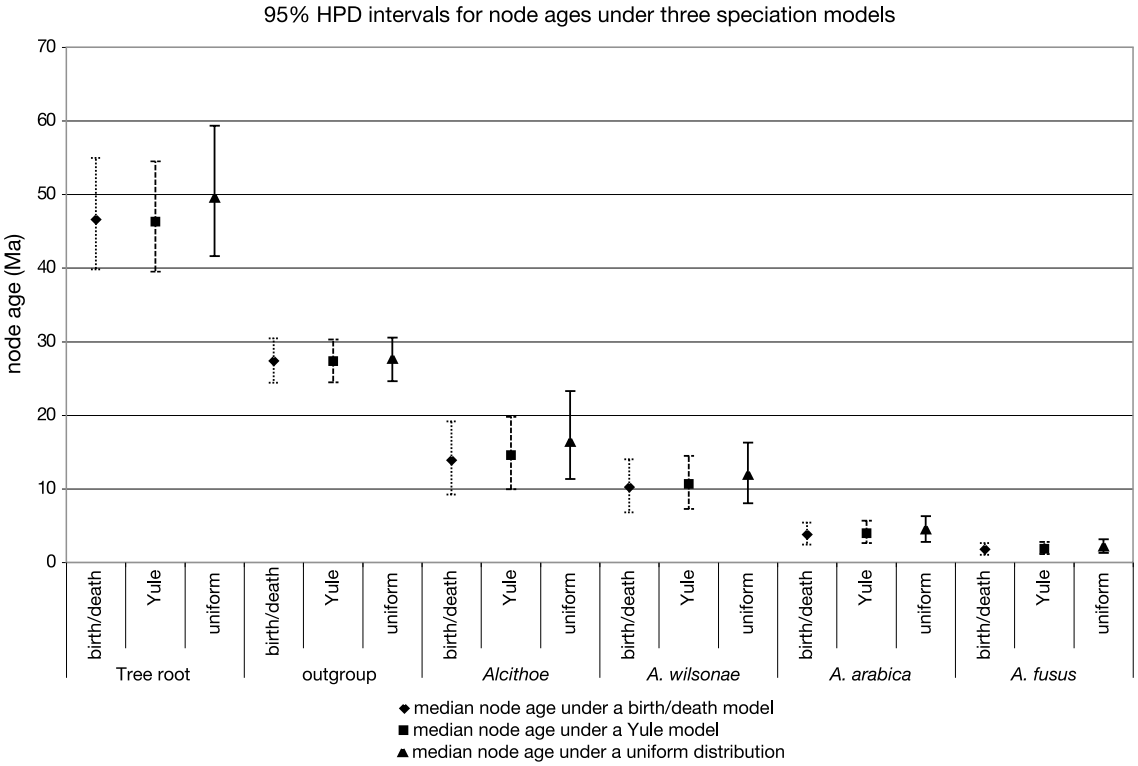

FIGURE 4.4—Divergence time estimation under a Yule prior, a birth/death process and a uniform prior on the branching process generate inferences that are highly consistent with each other. 95% HPD intervals are highly concordant across the three models, and the variances on the distributions are very similar.

4.3.5 Nucleotide substitution models

Previous analysis has shown that DNA evolution in this molecular dataset is best modelled by transversion model TVM+I+G (Chapter 2). It is not known however, how the addition of a time dimension affects the fit of the model. That is, does a different model of DNA substitution fit the data better given that the time frame involved is known? Logically we can assume the inferred node ages would differ as a result of a different DNA substitution model, but what is the effect on the accuracy of node age inference compared to the fossil record? To test this effect the inferred node heights (ages) were compared for otherwise identical analysis, varying the DNA substitution model used. A comparison of three models, GTR+I+G, TVM+I+G and HKY+I+G, shows that the specific substitution model has insignificant effect on the estimation of node heights in this dataset, and that the likelihoods of the results are not very different (Figure 4.5).

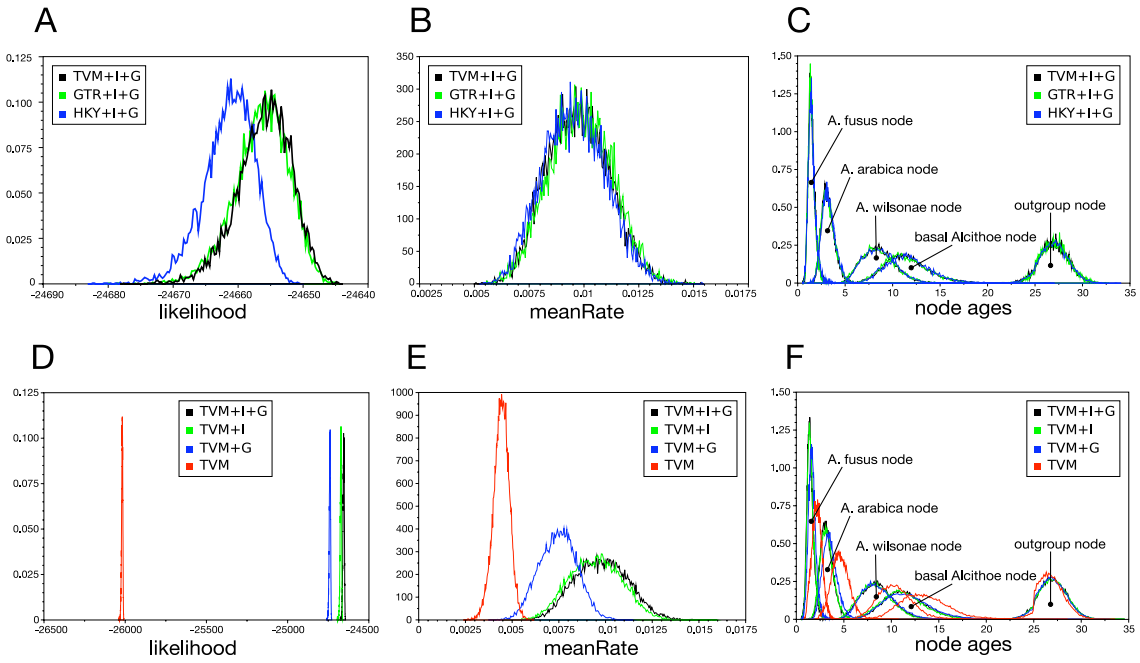

FIGURE 4.5—The proportion of invariable sites is the most important parameter of the nucleotide substitution model for accurate estimation of divergence times for the *Alcithoe*. Insignificant differences are observed between three different substitution models (TVM+I+G, GTR+I+G and HKY+I+G) when comparing: (A) likelihood scores, (B) inferred mean rates, and (C) internal node age estimates. However, a comparison of the elements of the substitution model for the same output parameters; (D) likelihood scores, (E) inferred mean rates, and (F) internal node age estimates; demonstrates a clear hierarchy of importance, with the basic model being the least informative and the full model being the most informative, but the proportion of variable sites parameter accounts for more of the accuracy of the complete model then the gamma parameter.

Bayes factors support a lack of difference between the TVM and GTR models, but both these models are more then 100-fold better in their fit to the data then the HKY model Table 4.1. However, when the parameters within one model are tested (i.e. TVM+I+G, TVM+I, TVM+G and TVM) a significant difference is observed in the resulting distributions of many of the posterior statistics (including the inferred rates and likelihoods) between the TVM+I+G and TVM models (see Figure 4.5). This result is corroborated by the Bayes factors calculated for these analyses Table 4.1. A clear gradient in the fit of these models to the data is observed. Posterior distributions illustrate that the addition of the gamma parameter (G) moves these distributions closer to those of the TVM+I+G, but the addition of the proportion of invariable sites (I) parameter results in distributions that are entirely consistent with the results using the TVM+I+G model. When considering individual nodes in the tree, where posterior distribution differences occur between the different models (i.e. in the TVM and TVM+G models) they are not as marked at the root,

but become more pronounced toward the distal node. This observation implies that the fossil species priors are influencing the posterior sampling. It is clear that, for this dataset, the proportion of invariable sites is the single most influential parameter of nucleotide substitution model for molecular clock analysis.

4.3.6 Molecular clock model

Three molecular clock models, the strict clock, relaxed LogNormal and relaxed exponential, were studied to select the most appropriate clock model to estimate node ages and rates on branches in this dataset. Posterior likelihoods under the LogNormal and exponential relaxed clocks are insignificantly different (mean of -24664 for both), but the likelihood for the strict clock is significantly lower (mean of -24701) based on the non-overlapping 95% highest posterior density intervals. Bayes factors confirm that there is little difference in model fit between the relaxed logNormal and the relaxed exponential models, but they are both significantly better then the strict clock model Table 4.1.

When the mean rates of the trees (or clock rate for the strict clock) are compared, differences between the three clock models are observed, but the 95% HPD intervals overlap. However, the shape of the posterior distribution for the relaxed lognormal clock is more similar to the posterior distribution for the strict clock than either are to the exponential relaxed clock. A consideration of the node ages in the trees shows that slight differences in the posterior distributions of the oldest calibrated nodes (the root and the *A. wilsonae* fossil calibration) are a result of increasing variance of the posterior sampling, strict clock having the least variance and exponential relaxed clock having the most (Figure 4.6). As the outgroup calibration is quite general and largely independent of the ingroup data, the posterior sampling for this node is essentially identical for each of the three clock models. For uncalibrated internal nodes the same pattern of increasing variance is seen as was observed for deep calibrated nodes, but the variance of the posterior sampling for the exponential clock is so much larger that considerably different node age means are estimated under this model. Distal calibrated nodes (*A. arabica* and *A. fusus/A. larochei*) once again demonstrate the difference in the variance of posterior sampling seen for the different clock models. In this case, however, there is a spread in the means of the three clock models.

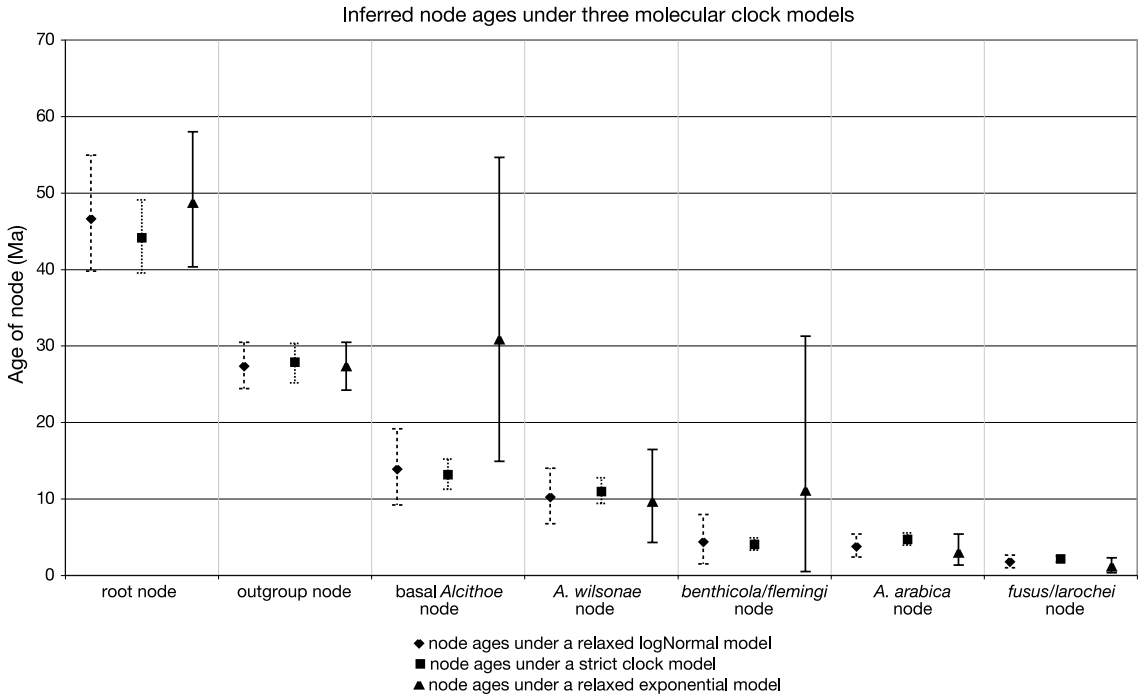

FIGURE 4.6—A relaxed exponential clock model generates a high variance on inferred node ages. Medians and 95% HPD intervals for node-age estimates are shown for 3 molecular-clock models. Results show that the strict-clock and the relaxed-lognormal are consistent, but the relaxed-exponential model recovers posterior distributions with high variance that lead to estimations of significantly different median node ages for uncalibrated nodes.

The increasingly restricted variance in the progressively younger calibrated nodes is indicative of the constraints the calibrated nodes have on the posterior sampling of the Bayesian analysis. Additionally, the distributions for the exponential clock are sampled from a more recent time frame than either the strict or relaxed lognormal clock, in contrast to the basal uncalibrated nodes where the exponential clock samples from deeper in time than the other two models. In general the exponential relaxed clock exhibits less well-sampled posterior distributions then either of the other models, as measured by effective sample sizes of posterior distributions. These results indicate that the exponential relaxed clock model is not as good a fit to the data as the other models. The strict clock shows much less variance in posterior distributions then the relaxed models, as a result of the more restricted parameter space of this model. Posterior distributions obtained using the relaxed lognormal clock tend to have means and standard deviations intermediate to distributions resulting from the other two models. Despite the lack of discriminatory power of the Bayes factors, the posterior distributions of the internal nodes clearly illustrate the poor fit of the relaxed exponential clock. Therefore, as

Bayes factors show the inferiority of the strict clock and the exponential relaxed clock can be discarded due to highly increased variance, the relaxed lognormal clock model is considered most appropriate.

4.3.7 Tree topology

The inferred tree topology remained highly consistent under the various parameter regimes tested, with posterior clade credibility support values of 1 for the majority of nodes. Generally only nodes associated with the four taxa; *A. larochei, A.fusus, A. fissurata* and *A. lutea*, generated posterior probabilities of less than 1. Based on earlier analysis it is known that amounts of signal and noise discriminating these taxa in the data are of similar magnitude, and that this low signal-to-noise ratio will cause different phylogenetic reconstruction algorithms to find different solutions in the same data (Chapter 2). It is therefore appropriate to further examine the behaviour of these taxa in BEAST to ensure that the optimal topological solution for these taxa is inferred. Previous analysis found that the best representation of the evolution of these four taxa is a trifurcation of *A. fusus, A. larochei*, and the *A. fissurata/A. lutea* clade. However, BEAST enforces bifurcation, and, in the case of these four taxa, places *A. larochei* sister to the other three species, and *A. fusus* sister to the *A. fissurata/A. lutea* clade (Figure 4.7). All maximum clade credibility trees generated in BEAST thus far have had this topology, and all trees with this topology imply a substitution rate for the *A. fusus* branch that is elevated compared to nearby branches. This topology was returned irrespective of whether a random starting tree is used or if any of the topological alternatives for the four taxa in question were set as a starting point for MCMC sampling. This reproducibility implies that the result is robust.

In order to further examine the placement of *A. fusus*, analyses were carried out with fixed topologies for each of the topological options for these taxa described in Chapter 2. Figure 4.8 shows the fixed topologies, inferred divergence times and inferred rate ranges for the four taxa in question, the remainder of the tree is concordant in the different tests and is omitted.

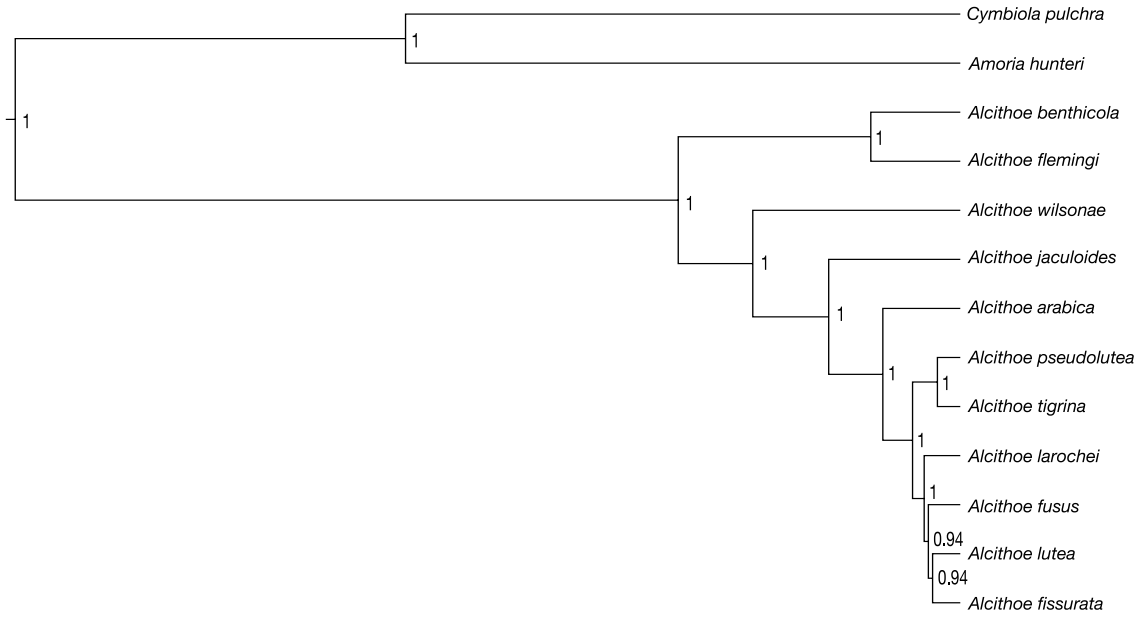

FIGURE 4.7—Bayesian phylogeny of the *Alcithoe* recovers high posterior probability support for all nodes.

Bayes factors (Table 4.1) show that a basal placement of *A. larochei* is the most well supported topology, as would be expected. However, while the evidence suggests that this topology is better then an unresolved relationship of these two taxa (*A. fusus* and *A. larochei*) the Bayes factors are not decisive. The only decisive conclusions that the Bayes factors offer are that a basal placement of *A. larochei* is significantly more likely then a basal placement of *A. fusus*. A trifurcation of *A. fusus, A. larochei*, and the *A. fissurata/A. lutea* clade or sister relationship of *A. larochei* and *A. fusus* cannot be ruled out. A consideration of the rates inferred for these taxa show that, regardless of the topology, *A. fusus* consistently returns faster rates of molecular evolution than surrounding branches. The distributions of inferred rates for the branches in question all tend toward slower rates. Overlapping 95% HPD intervals are indicative of a general concordance with an average rate for at least this part of the tree. However, the rate distribution for the *A. fusus* branch consistently tends towards faster rates than the other branches considered. It is interesting to note that a basal position of *A. fusus* leads to a rate distribution that is the least fast-skewed, even though this topology is rejected by the Bayes factors. When a sister relationship of *A. fusus* and *A. larochei* is enforced the 95% HPD interval on the branch leading to this pair is consistent with the other internal branches, but the range is considerably expanded compared to all other branches. Additionally, the rate range inferred for *A. larochei* is slightly increased. Under a trifurcation model BEAST returns a topology identical to the sister relationship of *A. fusus* and *A. larochei* due to an enforcement of bifurcation. This

result is unsurprising given that BEAST is not able to return a trifurcated topology, but it is interesting that sister relationship is recovered rather than either of the alternative topologies. The rate-ranges are not significantly different to the pattern of rates seen for the sister relationship model. Due to the inability to decisively conclude one best-fit tree topology a polytomy of *A. fusus, A. larochei*, and the *A. fissurata/A. lutea* clade is considered the appropriate topology.

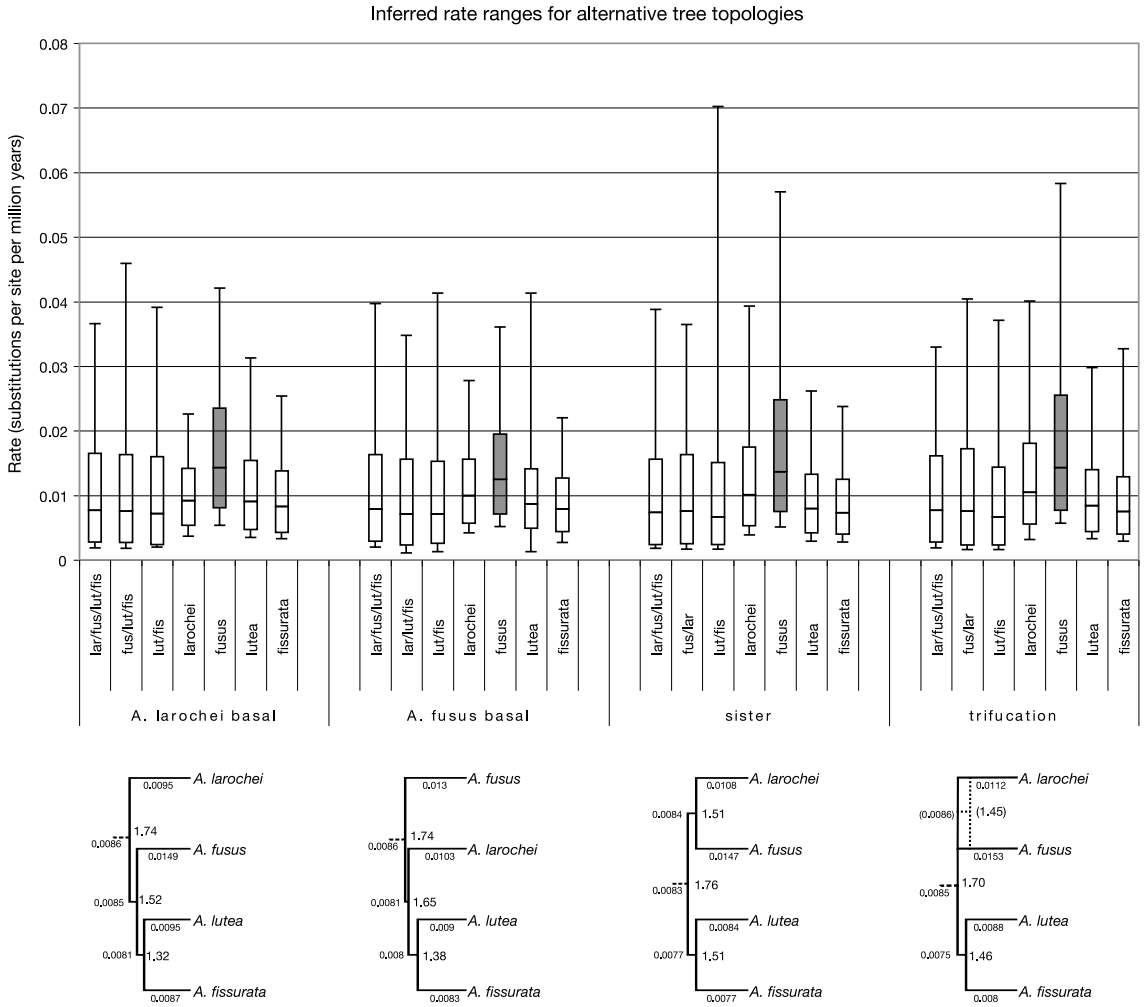

FIGURE 4.8—Inferred rates on branches in alternative topological arrangements of *A. fusus* and *A. larochei*. Box and whisker plots indicate the median, 95% HPD interval and range for rates inferred on branches. The results for the *A. fusus* node are highlighted. The topologies enforced for each analysis are indicated below the graph, these indicate the median inferred rates and divergence dates amongst the four taxa under consideration. The trifurcation topology is shown with solid lines indicating the fixed enforced input state and with dotted lines showing the output topology resulting from BEAST enforcing bifurcation.

4.3.8 Analysis of divergence times in Alcithoe

A final Bayesian dated analysis was carried out using the combination of model parameters found to be the best fit for the sequence data. These parameters were:

1/ the TVM+I+G nucleotide substitution model,

2/ a relaxed uncorrelated lognomal clock model,

3/ the Birth/Death process for the tree model,

4/ a lognormal tree root calibration with a zero offset.

The maximum credibility tree was generated from the posterior sample of this analysis, and is shown in Figure 4.9.

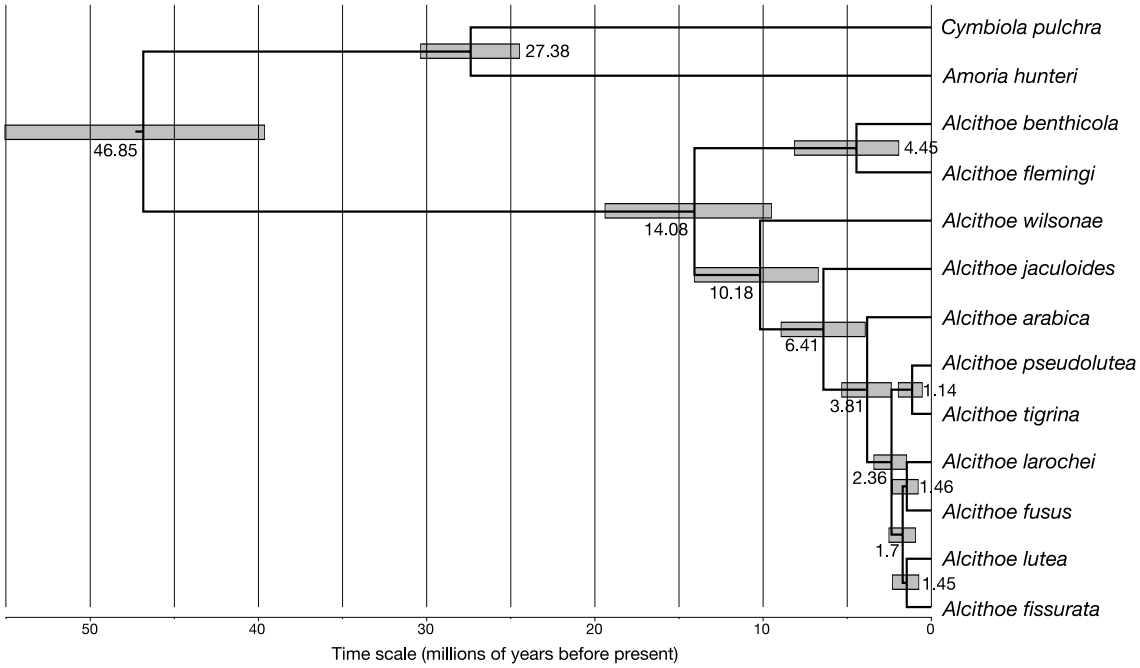

FIGURE 4.9—Time scaled molecular phylogeny of *Alcithoe*. Median node ages are shown and 95% HPD intervals are indicated with grey boxes.

Median inferred ages for the three nodes calibrated with fossil data for a given species (*A. wilsonae*, *A. arabica*, *A. fusus*/*A. larochei*) are all older than prescribed prior medians. The posterior medians for both *A. wilsonae* and *A. fusus* differ from the prior probability distribution medians by only about 500 000 years. The difference between the posterior and prior medians for *A. arabica* is over 1 million years. This difference demonstrates that these deviations are not scaling with depth in the tree, suggesting that these node heights are fundamentally informed by the underlying data. Based on the median node ages the current extant diversity

originated approximately 14 million years ago, but half these taxa diversified in the last 2.5 million years. However, it is more appropriate to consider the 95% highest posterior distribution (HPD) interval, as the true value should exist in this interval but is not necessarily indicated by the mean or median values. This shows that the Alcithoini diverged from *Amoria*/*Cymbiola* between 40 and 55 MYa. The most recent common ancestor of the extant taxa existed between 9.35 and 19.28 million years ago. *Alcithoe wilsonae*, the oldest fossil calibrated species in this dataset, is inferred to have diverged between 6.75 and 13.93 million years ago. This finding is consistent with the fossil record as this places the origin of *A. wilsonae* no more recent then the Tongaportuan stage (10.92 - 6.5 Mya) in which the first fossil specimen of the species is observed. Sampling probabilities, as built into the prior distribution, acknowledge that there is a reasonable chance that *A. wilsonae* could have existed earlier but that no older fossil has yet been identified. The 95% HPD interval for *A. wilsonae* supports this assumption, suggesting that the species could have originated up to 13.93 MY before present. This estimate of node age spans the full Waiauan stage (12.7 - 10.92 mya) and falls within the Lillburnian stage (15.1 - 12.7mya). *Alcithoe fusus* shows a similar pattern, the 95% HPD interval is 0.7 - 2.16 mya. This result places the probable origin of this species during the later part of the Nukumaruan stage (2.4 - 1.63 mya) or in the early to mid Castlecliffian (1.63 - 0.34 mya), an inference that is consistent with the first fossil occurrence of this species in the Castlecliffian. In contrast, molecular evidence suggests that *Alcithoe arabica* is likely to have originated earlier then previously thought (see prior probability distribution Figure 4.1). At present the oldest identified *A. arabica* fossil is from the Nukumaruan stage (Beu and Maxwell 1990). However the 95% HPD interval for the node representing the divergence of *A. arabica* spans 2.97 million years from 2.37 to 5.34 MYA. This interval includes only the earliest 30 000 years of the Nukumaruan. The majority of the interval includes the entirety of the Mangapanian (3.0 - 2.4) Waipipian (3.6 - 3.0) and Opoitian (5.28 - 3.6) stages, and the last 60,000 years of the Kapitean stage (6.5 - 5.28).

4.3.9 Rates of molecular evolution in Alcithoe

The mean rate of molecular evolution inferred from this analysis is 0.0075 substitutions per site per million years. The 95% HPD interval is reasonably small, ranging from 0.0062 to 0.0089. Figure 4.10 shows the inferred rate ranges, 95% HPD intervals and median values for each of the branches in the phylogeny, and illustrates that the range of rates inferred in the tree are generally much greater than mean rate across the tree. However, the 95% HPD intervals are largely overlapping indicating a general concordance. Many branches exhibit rate distributions that have ranges that include outlying values that are considerably larger than the estimated medians, and several branches exhibit interesting, if not statistically significant, deviations from the tree rate. The branch leading to *A. benthicola* and *A. flemingi* has the slowest rate in the tree, with a median below the 95% HPD interval of the mean tree rate and 95% HPD interval below the median value of the mean tree rate. These differences allude to a possible rate reduction in this clade. The rate inferred for *A. flemingi* supports this rate reduction, as it has the next lowest inferred rate, but the 95% HPD interval has less of a tendency toward lower values. *A. benthicola*, however, is at odds with the interpretation of a reduced rate in the clade, exhibiting a 95% HPD interval that includes values that are nearly twice as much as inferred for the mean rate of the tree. Additionally, the distributions for the majority of the internal branches of the tree inculde rate estimates that are considerably faster then the estimated median rates. The median values inferred for most of these branches are well outside of the 95% HPD interval of the mean tree rate.

The most significant apparent departure from the inferred mean tree rate is exhibited by the *A. fusus* branch, which has a 95% HPD interval that is entirely above the median value calculated for the mean tree rate, and has by far the largest rate range of any of the branches. A re-examination of the BEAST data was carried out asking the question; is the rate of substitution inferred for the *A.fusus* branch greater then the rates inferred on other branches? High posterior probabilities, of at least 0.81, were found for pair-wise comparisons of the *A. fusus* branch rate to all other external branches (Table 4.2). When only the three species most closely related to *A. fusus* are considered the posterior probabilities that *A. fusus* has a faster rates are at least 0.99. This result represents strong evidence that *A. fusus* does in fact have a faster rate of substitution, particularly amongst its closest relatives.

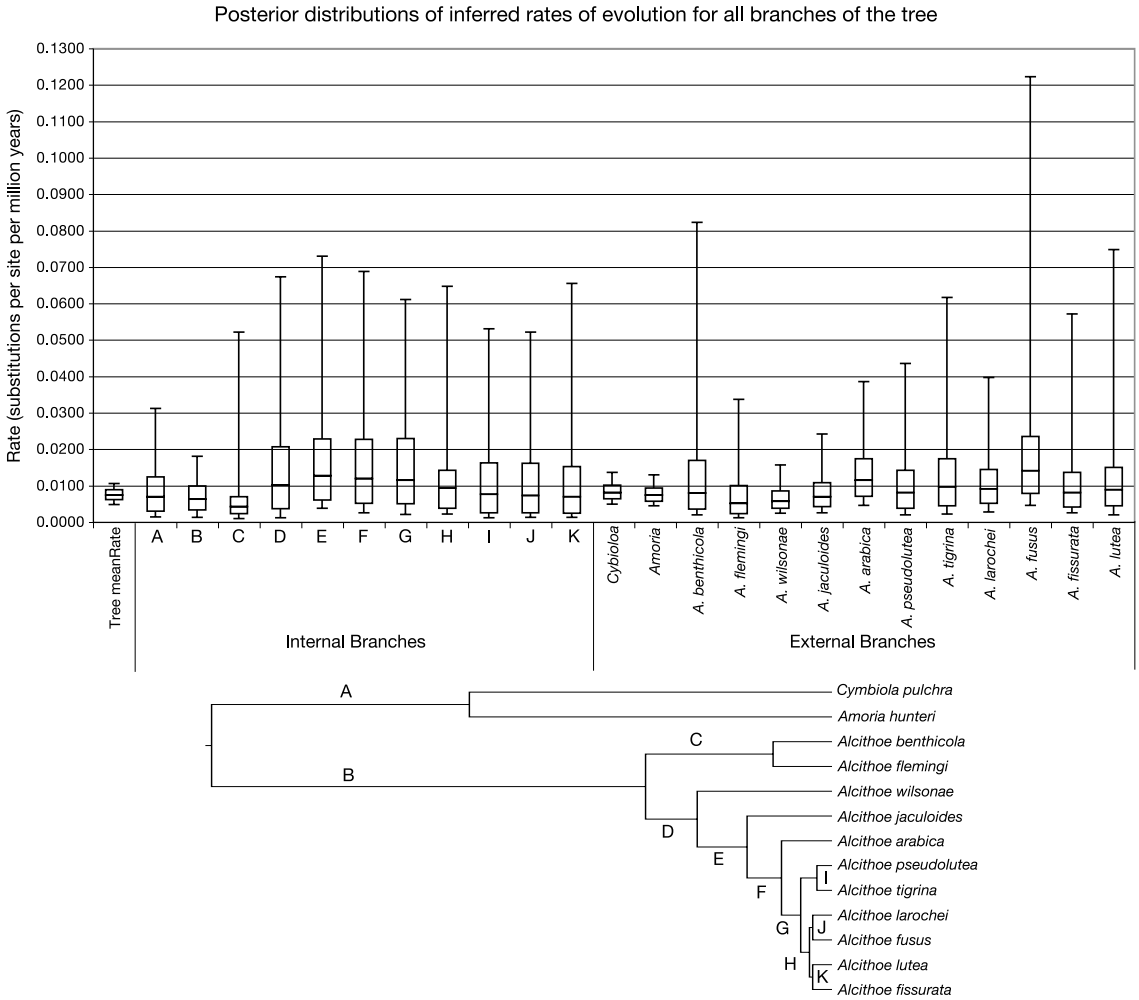

FIGURE 4.10—Inferred rate ranges for branches of the *Alcithoe* tree exhibit considerable variance. Box and whisker plots indicate the median, 95% HPD interval and range for rates inferred on branches in the *Alcithoe* phylogeny. A reference tree is shown below the graph with the internal branches labeled as they appear in the graph.

TABLE 4.2—Posterior probabilities that the substitution rate on the *A. fusus* branch is faster then other branches in the *Alcithoe* phylogeny.

	External branches											
	Cymbiola	*Amoria*	*A. benthicola*	*A. flemingi*	*A. wilsonae*	*A. jaculoidies*	*A. arabica*	*A.pseudolutea*	*A. tigrina*	*A. lutea*	*A. fissurata*	*A. larochei*
A. fusus vs	0.98	0.99	0.91	1.00	1.00	1.00	0.81	0.97	0.89	1.00	1.00	0.99

	Internal branches [1]										
	A	B	C	D	E	F	G	H	I	J	K
A. fusus vs	0.97	0.98	1.00	0.81	0.58	0.64	0.68	0.89	0.88	0.95	0.89

[1] refer to Figure 4.10 for internal branch labels

4.4 DISCUSSION

4.4.1 Parameter testing

Extensive analysis of parameter interactions has elucidated what I infer to be the most appropriate set of model parameters to apply to molecular based divergence date and substitution rate inference for the genus *Alcithoe*. These parameters are:

1/ the TVM+I+G nucleotide substitution model,

2/ the relaxed lognormal clock model,

3/ the Birth/Death speciation prior,

4/ a lognormal distributed tree root model with a mean of 3.26, a standard deviation of 0.2 and a zero-offset of 25 million years.

These specific parameter settings are unlikely to be of general use to other datasets as the parameter space is highly dependent on the dataset in question. However, this analysis also highlights some general principles to consider when performing Bayesian dated molecular analysis.

1. Calibration priors: in addition to being aware of the appropriate application of calibration data in Bayesian evolutionary analysis (Ho and Phillips 2009) it is also necessary to ensure that different types of calibration prior distribution (e.g. exponential or lognormal) do not interact to produce an inappropriate joint prior that is inconsistent with the individual prior probability distributions. An erroneous joint prior resulting from interactions between individual prior distributions can lead to posterior samples that are inconsistent with the data being analysed.

2. Tree Root priors: When the tree root prior distribution and an internal calibration prior distribution overlap the internal prior can dominate the root prior in the joint prior. This interaction can lead to posterior distributions that are biased toward incorrect age estimates, but can be avoided by ensuring that the root node prior distribution does not overlap with any internal calibration priors. This finding reinforces the need to analyse the joint prior distributions of a model configuration to be certain that no anomalous interactions are occurring between individual priors.

3. Nucleotide substitution models: Where the taxa under consideration have relatively consistent patterns of nucleotide substitution, the proportion of variable sites in the data (in other words the distances, or sequence divergence) is the single most important factor in the substitution model. This finding explains why a simple method of taking a genetic distance and applying a rate can provide an informative estimate of dates (e.g. King and Wilson 1975; Wilson et al. 1985). The sensitivity of date estimates to substitution models (i.e. GTR, HKY, etc.) is likely indicative of less consistent patterns of nucleotide substitution in the data. This sensitivity is probably a heterotachy related problem (Lockhart et al. 2006), where the proportion of variable sites is not consistent across the taxa being analysed.
The way in which various parameters interact is dependent on the evolutionary depth and breadth of taxonomic sampling of the dataset in question. This *Alcithoe* dataset has very few taxa when compared to many other molecular-clock based species comparisons, which tend to favour broad taxonomic sampling. One significant advantage of this dataset is the spread of the fossil calibrated species across the tree. Of particular importance is the *A. wilsonae* calibration, which controls the age of the second deepest node inside the diversity of *Alcithoe*. Furthermore, due to its apparent close temporal proximity, this node strongly influences the earlier node as well. This calibration leads to highly stable inferences within the extant diversity of the *Alcithoe* regardless of the parameter regime used. However, this stability is not seen for the root of the tree. The significant temporal gap between the root of the tree and the earliest divergence in the extant taxa means that the internal calibrated nodes have little effect on the root of the tree.

4.4.2 Divergence patterns in Alcithoe

Divergence date estimates for both *A. fusus* and *A. wilsonae* both broadly agree with fossil first occurrence data, while allowing for the possibility of older origins for each species. In both cases more then half the 95% HPD interval of the inferred divergence times are included in the geological time frames for the stages in which they are first found. For *A. arabica*, however, only 1% of the 95% HPD interval coincides with the Nukumaruan stage, in which it is currently thought to have originated. This result is strongly indicative that the true origin of *A. arabica* predates the Nukumaruan. The molecular data suggest that *A. arabica* may have diverged up to nearly 3 million years earlier than is currently accepted. This finding is perhaps unsurprising, as the two stages immediately preceding the Nukumaruan

are short, only around 600 000 years each, and the sampling probabilities are much lower, particularly in the Mangapanian. Some entries in the FRED database list *A. arabica* identifications from stages earlier then the Mangapanian. These are thought to be incorrectly identified specimens (Alan Beu, pers. com.), but the above result suggests that these specimens should be re-examined.

This analysis of the *Alcithoe* phylogeny places the origin of the modern diversity between 9 and 19 million years ago. Furthermore, at least half the species currently recognised have diverged within the last 2.5 million years, and all of the recent species are within a single clade. Yet the root of the tree is still inferred to be in the range of 40 to 55 million years. When combined with the rich fossil history of the tribe Alcithoini these results describe a highly asymmetrical mode of evolution. Most of what we see today has evolved recently and most of the evolutionary history of the Alcithoini is not represented in the modern fauna. It is probable that this asymmetry is the cause of inconsistencies in inference of the root of the tree. These analyses have shown that the estimates for the well-calibrated recent diversity are robust to various parameter-related errors, in that the estimated divergence dates within the extant lineages are not significantly different under most model configurations. However, this well-calibrated recent diversity only represents a fraction of the evolutionary history of the lineage. Nearly 33 million years is not represented by the modern taxa. This time frame is more than double the length of time since the origin of the current diversity, and is seen in the long internal branch in the phylogeny. Variable sites in the extant *Alcithoe* can only inform the substitution rate back to around 19 million years ago. Lineages diverging prior to this are likely to have had a different set of variable sites. As these lineages have gone extinct, and therefore cannot be sampled (in a molecular dataset), the variable sites that originated prior to the diversification of the sampled *Alcithoe* taxa have not been observed. Such unobserved or hidden substitutions will lead to an under-estimation of the substations rates on long internal braches. Furthermore, it is likely that these hidden substitutions will lead to an underestimate of the age of the root of the tree (Phillips 2009). Therefore, in trees that are significantly asymmetric it is possible that branch rates are underestimated. This phenomenon has been called the node density effect, and has been shown to be a possible source of error in rate estimates (Hugall and Lee 2007). However, tree shape can be useful for the analysis of macro-evolutionary patterns (Mooers and Heard 1997).

4.4.3 The tempo of molecular evolution in Alcithoe

Based on the mean tree rates in this analysis a substitution rate of between 0.0062 and 0.0089 substitutions per site per million years is inferred for the *Alcithoe* lineage. While overlapping 95% HPD intervals of the rate estimates of individual branches indicate a general agreement with this, there appear to be some possible departures from this mean rate. A reduced rate in the branch leading to *A. benthicola*/*A. flemingi* clade is implied by the rate distributions. However, an increased rate is apparent in the *A. benthicola* branch. This rate difference could be consistent with biology of *A. benthicola*, which is the largest of the *Alcithoe* species and its divergence may be related to gigantism occurring this lineage (Bruce Marshall, pers. com.) There is a high probability that *A. fusus* has an elevated rated of evolution, compared to the other *Alcithoe* species. This rate difference is likely to be a major contributing factor to problems with the reconstruction of four recently diverged taxa.

A rate range of 0.6 – 0.9 % per million years, inferred for *Alcithoe*, is similar to rates inferred for other marine gastropods at the lower end of the range of estimates (0.5% /ma in *Umbonium* (Ozawa and Okamoto 1993) and 0.6% /ma in *Littorina* (Reid et al. 1996)). Furthermore this rate is generally concordant with the slower end of the range of rates inferred for other invertebrates (e.g. Lynch and Jarrell 1993; Knowlton and Weigt 1998).

4.5 REFERENCES

Beu, A. G., Maxwell, P. A. 1990. Cenozoic Mollusca of New Zealand. New Zealand Geological Survey Paleontological Bulletin 58:1-518.

Bromham, L., Penny, D. 2003. The modern molecular clock. Nature Reviews Genetics 4:216-224.

Chiba, S. 1999. Accelerated evolution of land snails *Mandarina* in the oceanic Bonin Islands: Evidence from mitochondrial DNA sequences. Evolution 53:460-471.

Collins, T. M., Frazer, K., Palmer, A. R., Vermeij, G. J., Brown, W. M. 1996. Evolutionary history of northern hemisphere *Nucella* (Gastropoda, Muricidae): Molecular, morphological, ecological, and paleontological evidence. Evolution 50:2287-2304.

Cooper, R. A. 2004. The New Zealand geological timescale. Pp. 1-284. Institute of Geological and Nuclear Sciences Monograph.

Cooper, R. A., Maxwell, P. A., Crampton, J. S., Beu, A. G., Jones, C. M., Marshall, B. A. 2006. Completeness of the fossil record: Estimating losses due to small body size. Geology 34:241-244.

Crampton, J. S., Beu, A. G., Cooper, R. A., Jones, C. M., Marshall, B., Maxwell, P. A. 2003. Estimating the rock volume bias in paleobiodiversity studies. Science 301:358-360.

Crampton, J. S., Foote, M., Beu, A. G., Cooper, R. A., Matcham, L., Jones, C. M., Maxwell, P. A., Marshall, B. A. 2006. Second-order sequence stratigraphic controls on the quality of the fossil record at an active margin: New Zealand Eocene to recent shelf molluscs. Palaios 21:86-105.

Darragh, T. A. 1989. A revsion of the Tertiary Volutidae (Mollusca: Gastropoda) of south-eastern Australia. Memoirs of the Museum of Victoria 49:195-307.

Douris, V., Cameron, R. A. D., Rodakis, G. C., Lecanidou, R. 1998. Mitochondrial phylogeography of the land snail *Albinaria* in Crete: Long-term geological and short-term vicariance effects. Evolution 52:116-125.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., Rambaut, A. 2006. Relaxed phylogenetics and dating with confidence. PLOS Biology 4:699-710.

Drummond, A. J., Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7.

Gernhard, T. 2008. The conditioned reconstructed process. Journal of Theoretical Biology 253:769-778.

Graur, D., Martin, W. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends in Genetics 20:80-86.

Hellberg, M. E., Vacquier, V. D. 1999. Rapid evolution of fertilization selectivity and lysin cDNA sequences in teguline gastropods. Molecular Biology and Evolution 16:839-848.

Ho, S. Y. W. 2009. An examination of phylogenetic models of substitution rate variation among lineages. Biology Letters 5:421-424.

Ho, S. Y. W., Phillips, M. J. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Systematic Biology 58:367-380.

Hugall, A. F., Lee, M. S. Y. 2007. The likelihood node density effect and consequences for evolutionary studies of molecular rates. Evolution 61:2293-2307.

Kass, R. E., Raftery, A. E. 1995. Bayes Factors. Journal of the American Statistical Association 90:773-795.

King, M. C., Wilson, A. C. 1975. Evolution at 2 levels in humans and chimpanzees. Science 188:107-116.

Knowlton, N., Weigt, L. A. 1998. New dates and new rates for divergence across the Isthmus of Panama. Proceedings of the Royal Society of London Series B-Biological Sciences 265:2257-2263.

Lepage, T., Bryant, D., Philippe, H., Lartillot, N. 2007. A general comparison of relaxed molecular clock models. Molecular Biology and Evolution 24:2669-2680.

Lepage, T., Lawi, S., Tupper, P., Bryant, D. 2006. Continuous and tractable models for the variation of evolutionary rates. Mathematical Biosciences 199:216-233.

Liow, L. H., Stenseth, N. C. 2007. The rise and fall of species: implications for macroevolutionary and macroecological studies. Proceedings of the Royal Society B-Biological Sciences 274:2745-2752.

Lockhart, P., Novis, P., Milligan, B. G., Riden, J., Rambaut, A., Larkum, T. 2006. Heterotachy and tree building: A case study with plastids and eubacteria. Molecular Biology and Evolution 23:40-45.

Lynch, M., Jarrell, P. E. 1993. A method for calibrating molecular clocks and Its application to animal mitochondrial-DNA. Genetics 135:1197-1208.

Marko, P. B. 2002. Fossil calibration of molecular clocks and the divergence times of geminate species pairs separated by the Isthmus of Panama. Molecular Biology and Evolution 19:2005-2021.

Mooers, A. O., Heard, S. B. 1997. Evolutionary process from phylogenetic tree shape. Quarterly Review of Biology 72:31-54.

Near, T. J., Sanderson, M. J. 2004. Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 359:1477-1483.

Ozawa, T., Okamoto, K. 1993. A new movement of phylogeny by synthesis of paleontological approach and moleculer phylogenetical approach, from an example of gastropods *Umbonium*. Gekkan Tikyu 15:589–595.

Phillips, M. J. 2009. Branch-length estimation bias misleads molecular dating for a vertebrate mitochondrial phylogeny. Gene 441:132-140.

Pulquerio, M. J. F., Nichols, R. A. 2007. Dates from the molecular clock: how wrong can we be? Trends in Ecology & Evolution 22:180-184.

Reid, D. G., Rumbak, E., Thomas, R. H. 1996. DNA, morphology and fossils: Phylogeny and evolutionary rates of the gastropod genus *Littorina*. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 351:877-895.

Rutschmann, F., Eriksson, T., Abu Salim, K., Conti, E. 2007. Assessing calibration uncertainty in molecular dating: The assignment of fossils to alternative calibration points. Systematic Biology 56:591-608.

Sanders, K. L., Lee, M. S. Y. 2007. Evaluating molecular clock calibrations using Bayesian analyses with soft and hard bounds. Biology Letters 3:275-279.

Ware, J. L., Ho, S. Y. W., Kjer, K. 2008. Divergence dates of libelluloid dragonflies (Odonata : Anisoptera) estimated from rRNA using paired-site substitution models. Molecular Phylogenetics and Evolution 47:426-432.

Welch, J. J., Bininda-Emonds, O. R. P., Bromham, L. 2008. Correlates of substitution rate variation in mammalian protein-coding sequences. BMC Evolutionary Biology 8.

Welch, J. J., Bromham, L. 2005. Molecular dating when rates vary. Trends in Ecology & Evolution 20:320-327.

Whelan, S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. Molecular Biology and Evolution 25:1683-1694.

Wilson, A. C., Cann, R. L., Carr, S. M., George, M., Gyllensten, U. B., Helmbychowski, K. M., Higuchi, R. G., Palumbi, S. R., Prager, E. M., Sage, R. D., Stoneking, M. 1985. Mitochondrial-DNA and 2 perspectives on evolutionary genetics. Biological Journal of the Linnean Society 26:375-400.

Wilson, A. C., Carlson, S. S., White, T. J. 1977. Biochemical evolution. Annual Review of Biochemistry 46:573-639.

Yang, Z. H., Rannala, B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Molecular Biology and Evolution 23:212-226.

Yule, G. U. 1924. A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, FRS. Philosophical Transactions of the Royal Society of London Series B 213.

CHAPTER FIVE

5 THE EVOLUTION OF NEW ZEALAND VOLUTES: A COMPARISON OF MOLECULAR AND PALEONTOLOGICAL EVIDENCE

5.1 INTRODUCTION

Observed biodiversity is the result of an interaction between rates of species origination and extinction. This interaction produces the myriad tree shapes that we see in species phylogenies. Quantifying the effects of speciation and extinction is a major goal of evolutionary studies, but distinguishing between the two can be difficult particularly as extinction is not independent of speciation. It would be ideal to compare biodiversity trends observed in molecular and paleontological datasets for the same lineage. However, the majority of molecular phylogenetic studies have been concerned with groups that have poor fossil records (Paradis 2004). In many cases robust unequivocal phylogenies are not available for fossil groups (Etienne and Apol 2009) due to the incompleteness of the fossil record and difficulties in species assignments (Kidwell and Flessa 1995; Cooper et al. 2006; Valentine et al. 2006). As a result there have been few direct comparisons between molecular and paleontological studies of evolution for the same group. In the absence of complete, directly comparable datasets, comparisons must be based on some amount of inference from each dataset. An eminently tractable means of comparing the patterns of evolution between paleontological and molecular data is through analysis of the of diversification rates inferred by the two data sets. The fossil record is often used for temporal calibration of molecular phylogenies for

subsequent analysis of the patterns of diversification. However, diversification rates have rarely been compared in equivalent molecular and paleontological datasets. Usually rates are inferred from one of these datasets and compared against a null model.

Methods for calculating origination and extinction rates are well established for both phylogenetic (e.g. Nee 2001; Nee 2006) and fossil (e.g. Foote 2000; Foote 2007) data. Estimates of speciation and extinction rates generated using these datasets form the basis of much research aimed at exploring the generation of biodiversity and are key to our understanding of the interface between micro- and macroevolution (e.g. Nee 2006; Mittelbach et al. 2007; Ricklefs 2007; McPeek 2008; Purvis 2008; Rabosky and Lovette 2008).

For some time speciation rates have been inferable by lineage through time (LTT) analysis of molecular phylogenies. More recently, the advent of modelling a birth/death process in the inference of molecular phylogenetics has enabled the estimation of extinction rates from molecular data (Nee 2001). While speciation rates estimated from fossil data are often more precise, rates inferred from molecular phylogenies can be correctly estimated despite considering only extant species (Paradis 2004). However, when extinction rates are high accurate estimation of rates from molecular data is much less likely (Paradis 2004). A minimum of 15 species has been quoted as being sufficient for accurate estimation of the speciation rate in a clade; fewer than this will lead to an underestimate (Paradis 2004). However, most species in a genus are required for accurate inference (Barraclough and Nee 2001) as missing species will cause a reduction in the inferred rate (Paradis 2004). Extinction rates appear to be well estimated from phylogenies when the clade size ranges from 10 to 20 taxa, but are not as accurate as rates estimated from fossil data and are under-estimated in most cases in simulation studies (Paradis 2004). Random extinction with a constant probability is expected to cause an apparent acceleration in speciation rate towards the present (Barraclough and Nee 2001).

The analysis of origination and extinction rates, sometimes referred to as taxonomic rates, in the fossil record has a long history. Methods to calculate these rates have been reviewed by Raup (1985) and again by Foote (2000). The main problems with inference of rates from paleontological data have been related to the

incomplete nature of the fossil record, difficulties with identification at the species level, and lineages that are found in only a single geological time span (singletons). Significant paleontological research has been devoted to accounting for preservation and sampling bias in the fossil record (e.g. Crampton et al. 2003; Cooper et al. 2006; Crampton et al. 2006a). The problem of species identification has meant that the analysis of taxonomic rates in the fossil record has been based on higher-level groupings, genera and above (Jackson and Johnson 2001). However, it is possible to derive species-level rates from data resolved to the genus level (Foote 2007). Accounting for singletons in a dataset has been shown to result in more accurate estimates of taxonomic rates (Foote and Sepkoski 1999; Alroy 2000). The large body of work devoted to the accurate estimation of taxonomic rates has highlighted some new potential problems. Origination (or speciation) rates may be more significantly distorted by incomplete preservation than extinction rates as extinction events have a greater tendency to coincide with stages with high preservation probabilities (Foote 2001). Additionally, origination and extinction rates are unlikely to be constant over time. Studies show that origination rates vary less then extinction rates over short timescales, but over long time scales origination rates vary at least as much as extinction rates in marine invertebrate families (Kirchner 2002).

5.1.1 Alcithoe

The fossil record of the New Zealand volute lineage (tribe Alcithoini), represented in the modern fauna by the genus *Alcithoe*, has been the subject of study for over 100 years (e.g. Suter 1913; Marwick 1926; Dell 1978; Powell 1979). Volutes are distinctive large snails that that live on soft substrates that are suitable for fossil preservation. Extensive collections of extinct and dry extant specimens allow for an excellent picture of the presence of members of the group through the Cenozoic era (55 Ma until the present). The inference of a robust dated phylogeny for the *Alcithoe* (previous chapter) now allows comparisons to be made between the molecular and paleontological data. While it would be ideal to compare the phylogenetic analyses of the two data sets, as has been discussed (Chapter 1) a paucity of discriminating morphological characters make inference of a complete phylogeny based on the fossil record difficult.

5.1.2 Systematics of the fossil Alcithoini

In an attempt to estimate the phylogenetic relationships of the higher order groupings of the Alcithoini lineages an unofficial review of the generic relationships of many of the New Zealand fossil Volutidae was carried out (Beu, Maxwell, Cooper, Crampton and Marshall, unpublished) see Figure 5.1. This review suggested a revision of several generic relationships within the Alcithoini, and highlighted several affinities that some of the modern species have with other recognised genera. The absence of fossils of *Mauira* through the entire Oligocene suggests that the genus should be divided into two groups. One represents the true *Mauira* from the Eocene, the other encompasses the group of species currently recognised as *Mauira* from the Miocene. It was suggested that this group diverged from *Alcithoe* during the mid Miocene. It was recognised that the genera *Spinomelon*, *Teremelon* and *Waihaoia* share many characteristics and may represent a single lineage. Furthermore, despite the absence of these genera from the fossil record during the Pleistocene, it is possible that *Alcithoe benthicola* is derived from this lineage. This hypothesis is based on the presence of an apical spike on the protoconch of *A. benthicola*, a characteristic feature of *Spinomelon* and *Waihaoia* specimens (Alan Beu and Bruce Marshall, pers com.). *Alcithoe* is presumed to have diverged from the *Spinomelon* lineage during the Whaingaroan stage of the early Oligocene leading up to its first occurrence in the fossil record in the Duntroonian. A separate divergence from *Spinomelon* at a similar time led to the genus *Metamelon*, which is now considered to have a longer fossil history than is indicated in Beu and Maxwell (1990). *Leporemax* is thought to have diverged from *Alcithoe sensu stricto* during the Otaian (21.7 – 19.0 Ma), as the first fossils recognised as *Leporemax* appear in the Altonian. At the time of the revision (Figure 5.1), *Alcithoe fusus* was thought to be a modern example of *Leporemax,* and *A. jaculoides* could be an offshoot of the lineage. However, the taxonomic status of this sub-genus was questioned, as the number of specimens in collection has increased, the number of characters supporting the separation of *Leporemax* has declined. *Mauithoe* is inferred to have been derived from *Alcithoe* in the mid Miocene at a similar time to the Miocene "*Mauira*" group. Finally, *Iredalina mirabilis* appears in the fossil record in the Nukumaruan, possibly immigrating from the region of South America (Figure 5.1).

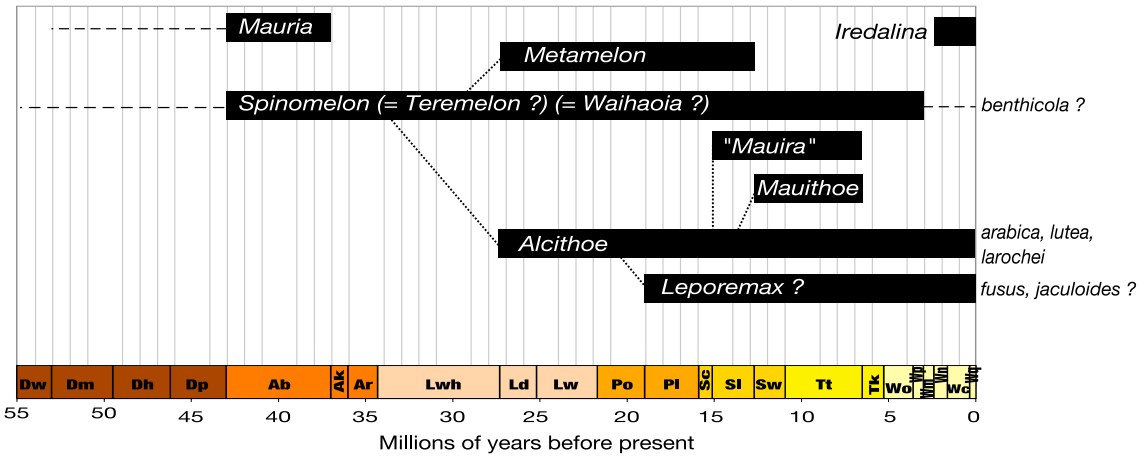

FIGURE 5.1—Revised generic relations for the New Zealand Alcithoini. Dashed lines indicate where unconfirmed fossil data exists. Dotted lines show the theoretical relationships between the genera. Putative generic affinities of some extant species are indicated. The stages of the New Zealand geological timescale are included (Wq - Haweran, Wc - Castlecliffian, Wn - Nukumaruan, Wm - Mangapanian, Wp - Waipipian, Wo - Opoitian, Tk - Kapitean, Tt - Tongaporutuan, Sw - Waiauan, Sl - Lillburnian, Sc - Clifdenian, Pl - Altonian, Po - Otaian, Lw - Waitakian, Ld - Duntroonian, Lwh - Whaingaroan, Ar - Runangan, Ak - Kaiatan, Ab - Bortonian, Dp - Porangan, Dh - Heretaungan, Dm - Mangaorapan, Dw - Waipawan)

5.1.3 Species patterns relevant to modern taxa

There are several species-level patterns in the fossil record that are relevant to the interpretation of the modern taxa. These patterns in particular will be useful to compare with molecular data as they directly involve species that are tractable to molecular analysis. Beu and Maxwell (1990) recognise a morphological grade in the *Leporemax* lineage based on the decreasing prominence of costae in the taxa involved. In this grade *A.*(*L.*) *rugosa* is ancestor to *A.*(*L.*) *gatesi* which is ancestor to *A.*(*L.*) *brevis* which gives rise to *A.*(*L.*) *fusus*. In particular these authors recognise a very gradual intergrade from *A.*(*L.*) *brevis* to *fusus*. However, the close relationship of *A.*(*L.*) *brevis* and *fusus* has been questioned (B. Marshall, pers. com). Acknowledging that this gradient represents a progenitor/progeny species relationship leads to the conclusion that *Alcithoe fusus* belongs to a lineage of morphologically similar species that is at least 10 million years old.

Using the molecular phylogeny for the *Alcithoe* and drawing on data from the paleontological record of the New Zealand volutes I address the question; in the study of evolution are molecular and paleontological data comparable? More specifically, can a data from one discipline be used to test hypotheses arising from

data from the other? Or, do they provide complementary but distinct information? By comparing patterns and rate estimates of clade evolution derived from molecular and paleontological data for *Alcithoe*, I will examine where the two datasets are concordant and where they are not. Finally I will show that it is appropriate to consider these datasets in parallel.

5.2 METHODS

Using values of B-D and D/B, estimated from molecular data, the approximate rates of speciation (B) and extinction (D) can be calculated. Bayesian phylogenetic inference under a birth-death model of clade expansion, or speciation, yields posterior statistics for the values diversification rate B-D and the extinction to speciation ratio D/B, which is a measure of the degree that by which the diversification rate differs from a pure birth process. Larger D/B values indicate a greater deviance (Nee 2006). Values of B-D and D/B for the *Alcithoe* molecular phylogeny were obtained from the final analysis described in Chapter 4. As described in Chapter 4, three internal nodes were calibrated using the fossil record of the species *A. wilsonae*, *A. arabica* and *A. fusus*. To investigate estimates of these values over shorter timeframes Bayesian analysis was carried out using nested subsets of the *Alcithoe* phylogeny. These analyses where carried out using BEAST v1.4.8 using model parameters identified in Chapter 4, with the exception of the tree root-height calibration. This calibration varied for each analysis as the oldest divergence in the taxa under examination decreased. Root node calibrations for each of the taxon subsets were set with lognormal priors that best approximated the mean and 95% HPD for the relevant node in the complete dataset. A lineage through time (LTT) plot was generated using the complete molecular phylogeny, but only considering the *Alcithoe* species. The number of branches was counted every two million years, beginning at 20 million years ago (Ma). The median node heights and 95% highest probability density (HPD) interval were considered in order to plot an estimated confidence interval for the LTT plot. Fossil occurrence data for the New Zealand volute tribe Alicthoini was extracted from Beu and Maxwell (1990). This data is based on first and last occurrence in the fossil record. Rates were derived from the fossil data using calculations for dynamic survivorship analysis and per stage rate estimates described in Foote and Miller (2007).

Due to different analyses providing estimates in different units the ratio of extinction to speciation may be an appropriate way to compare values generated by different methods as used by Ricklefs (2007).

5.3 RESULTS

5.3.1 Patterns in molecular and paleontological datasets

Comparing the molecular phylogeny to the revised generic relationships (Figure 5.2) reveals several inconsistencies. However, the root of the molecular tree, representing the divergence of the Alcithoini tribe, is broadly consistent with the fossil data, based on the overlapping 95% highest probability density (HPD) interval. However, given that only the upper third of the 95% HPD interval overlaps with the expected 50 - 55 Ma age of *Alcithoe*, there is a persuasive suggestion that the molecular data support a younger origin than the oldest currently recognised fossil for the group. The earliest divergence in the extant taxa shown in the molecular phylogeny significantly post-dates the earliest recognised *Alcithoe* fossil. In addition, the diversification of the modern species, including *A. fusus*, is later than the implied origin of the *Leporemax* subgenus. As *A. fusus* is recognised as the type species for *Leporemax*, and given the close relationship of *A. fusus* to *Alcithoe sensu stricto,* it appears clear that *Leporemax* is a synonym of *Alcithoe* and fossil species assigned to *Leporemax* cannot be placed there. It is not yet clear if the fossil species currently recognised fossil Leporemax species truly represent a separate clade or are members of the *Alcithoe s. str.*, additional morphological analyses will be required to clarify the affinities of these extinct species. Furthermore, the inclusion of *Alcithoe benthicola* in the *Spinomelon* genus, as this genus is currently understood, is not supported by the molecular phylogeny. An interesting coincidence of divergence times exists between the molecular based inference of the origin of the extant *Alcithoe* lineage and the fossil based inferences of divergence times of the *Mauithoe* and "*Mauira*" groups, all in the middle Miocene.

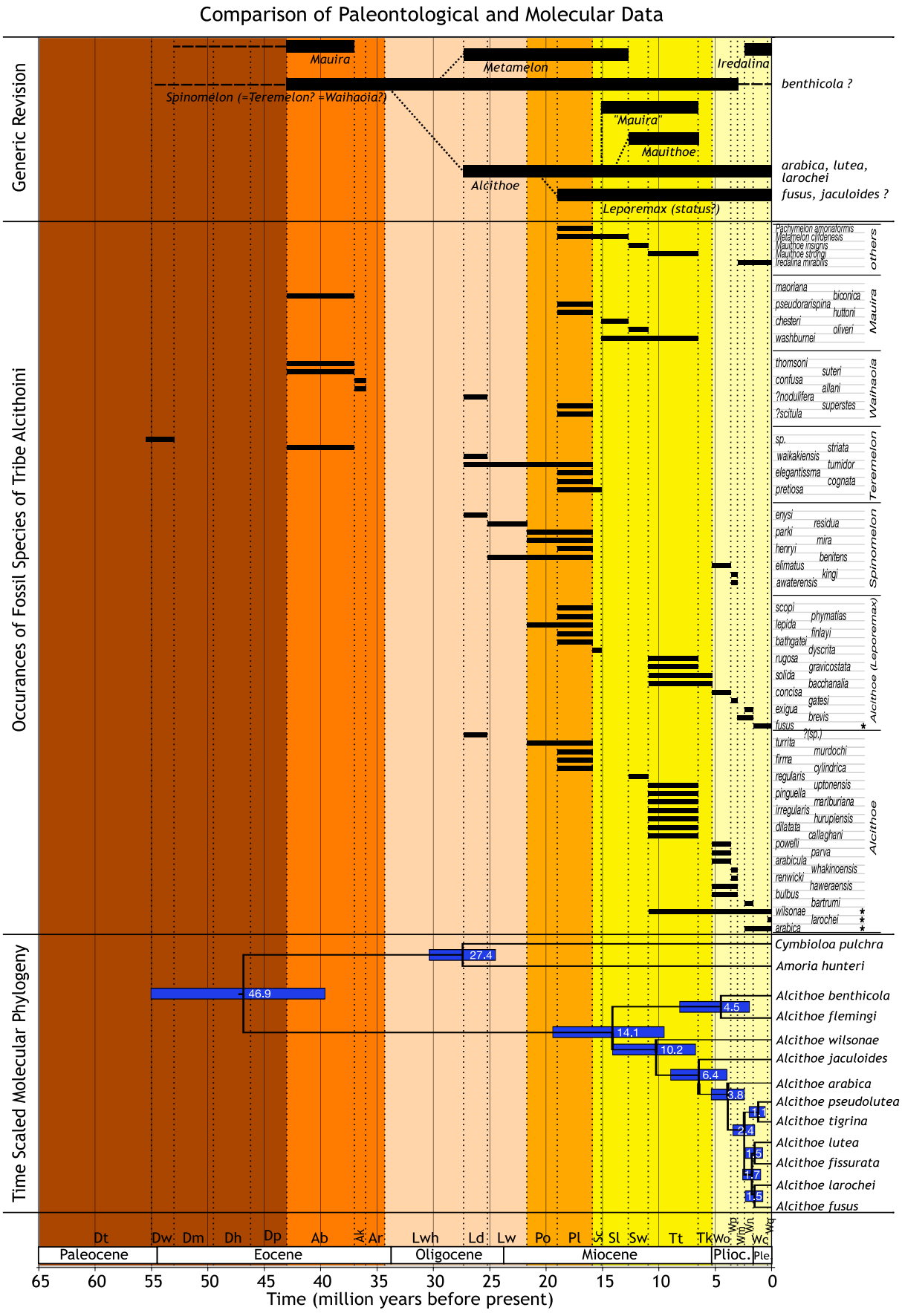
Comparison of Paleontological and Molecular Data
Generic Revision
Occurances of Fossil Species of Tribe Alcithoini
Time Scaled Molecular Phylogeny
Mauira
Metamelon
Iredalina
benthicola ?
Spinomelon (=Teremelon? =Waihaoia?)
"Mauira"
Mauithoe
arabica, lutea, larochei
Alcithoe
fusus, jaculoides ?
Leporemax (status?)
Pachymelon amoriaformis
Metamelon cliftonensis
Mauithoe insignis
Mauithoe strongi
Iredalina mirabilis
others
maoriana
biconica
pseudorarispina
huttoni
chesteri
oliveri
washburnei
Mauira
thomsoni
suteri
confusa
allani
?nodulifera
superstes
?scitula
Waihaoia
sp.
striata
waikakiensis
tumidor
elegantissma
cognata
pretiosa
Teremelon
enysi
residua
parki
mira
henryi
benitens
elimatus
kingi
awaterensis
Spinomelon
scopi
phymatias
lepida
finlayi
bathgatei
dyscrita
rugosa
gravicostata
solida
bacchanalia
concisa
gatesi
exigua
brevis
fusus
?(sp.)
Alcithoe (Leporemax)
turrita
murdochi
firma
cylindrica
regularis
uptonensis
pinguella
marlburiana
irregularis
hurupiensis
dilatata
callaghani
powelli
parva
arabicula
whakinoensis
renwicki
haweraensis
bulbus
bartrumi
wilsonae
larochei
arabica
Alcithoe
Cymbioloa pulchra
Amoria hunteri
Alcithoe benthicola
Alcithoe flemingi
Alcithoe wilsonae
Alcithoe jaculoides
Alcithoe arabica
Alcithoe pseudolutea
Alcithoe tigrina
Alcithoe lutea
Alcithoe fissurata
Alcithoe larochei
Alcithoe fusus
27.4
46.9
4.5
14.1
10.2
6.4
3.8
2.4
1.8
1.7
1.5
Dt Dw Dm Dh Dp Ab Ak Ar Lwh Ld Lw Po Pl Sc Sl Sw Tt Tk Wo Wp Wn Wq Wc
Paleocene
Eocene
Oligocene
Miocene
Plioc.
Ple
65 60 55 50 45 40 35 30 25 20 15 10 5 0
Time (million years before present)

FIGURE 5.2— (facing page) Direct comparison of the molecular and paleontological data for the tribe Alcithoini. Putative affinities of a selection of extant species to the recognised genera and the revised generic relationships are shown for comparison to the molecular data. Fossil occurrence for recognised Alcithoini species are shown, based on first and last occurrence. Living species of *Alcithoe* that are represented in the fossil record are marked with an asterisk. The time scaled dated molecular phylogeny is shown with the median node ages and 95% HPD intervals. The globally recognised Epochs are indicated and the stages of the New Zealand geological timescale are included (Wq – Haweran, Wc – Castlecliffian, Wn – Nukumaruan, Wm – Mangapanian, Wp – Waipipian, Wo – Opoitian, Tk – Kapitean, Tt – Tongaporutuan, Sw – Waiauan, Sl – Lillburnian, Sc – Clifdenian, Pl – Altonian, Po – Otaian, Lw – Waitakian, Ld – Duntroonian, Lwh – Whaingaroan, Ar – Runangan, Ak – Kaiatan, Ab – Bortonian, Dp – Porangan, Dh – Heretaungan, Dm – Mangaorapan, Dw – Waipawan)

A side-by-side comparison of the molecular phylogeny and a species level summary of the fossil record of the New Zealand Volutidae (Figure 5. 2) allows a consideration of the stage-based, discontinuous paleontological data with continuous relationships inferred by the molecular data. The continuous picture shows that diversification of a single lineage of the Alcithoini tribe, beginning in the early to middle Miocene, has led to all of the modern diversity. The fossil data shows that diversity in the Alcithoini lineage was relatively high in the early Miocene. However, by the middle Miocene the Alcithoini fauna is poorly represented in the fossil record. Species diversity is again relatively high in the late Miocene, and extant taxa potentially originating at this time represent between 18 and 45% of the modern diversity. The majority of diversification leading to extant *Alcithoe* species, as inferred by the molecular data, has occurred in the Pliocene. Indeed, as much as 45% of the modern diversity originated in the last 2.5 million years. By contrast, no extinct species are seen in the fossil record from the same timeframe, likely due to the lack of fossil deposits representing deeper water faunas from this time.

When considering only the fossil record of *Alcithoe* species, the data shows that most extinct species occur in only one stage (Figure 5.2). Similar levels of diversity, as are seen in the modern taxa, are observed in the Altonain and Tongaporutuan stages. However, a conspicuous fossil gap in the middle Miocene, particularly the Lilburnian stage (15.1 – 12.7 Ma), masks the faunal turnover occurring around the time of the earliest divergence in the modern taxa. As has been discussed earlier (Chapters 2 and 3) fossil and molecular data are highly concordant in the inference of the time of origin of the species *Alcithoe wilsonae*. Much of the diversification of the extant *Alcithoe* has occurred in the last 2 million years.

5.3.2 Rates from different datasets

Bayesian estimation inferred a median diversification rate (B-D) for the complete molecular phylogeny of 0.0496. The median D/B value for the same analysis is 0.5671. This value infers a considerable departure from a pure birth process (where D = 0) indicating that, as expected, extinction has played a significant role in the evolutionary history of the New Zealand Volutidae. Using these values the calculated speciation rate is 0.1146 speciations Myr^{-1} and the extinction rate is 0.0650 speciations Myr^{-1}, for this phylogeny. As the extinction to speciation (D/B) ratio indicates, the extinction rate accounts for 57% of the speciation rate in the molecular data.

Proportional rates of speciation and extinction were calculated from the fossil record to compare with the molecular based estimates. These rates were based on the numbers of originations and extinctions observed in the species level summary of the Alcithoini lineage (Figure 5.2) over the time interval roughly equivalent to the molecular phylogeny (approximately 50 Ma and the present). The per-taxon extinction rate is around 0.018 per million years, and the origination rate is approximately 0.02 per million years. Therefore in the fossil record the extinction rate is 90% of the speciation rate.

In order to incorporate the time frame between the origination of the Alcithoini tribe and the divergence of the modern *Alcithoe* species, the molecular analysis included out-group taxa. This time frame includes the *Alcithoe* stem lineage, and has clearly been dominated by extinction. However, the inclusion of out-group taxa in the molecular phylogeny is probably a source of significant error as they significantly under-represent the taxonomic diversity that exists between the ingroup and out-group. This error will have the effect of misleading the estimates of the B-D and D/B values, but its magnitude is unknown. To be able to include more of the timeframe that is effectively missing in the molecular phylogeny, sampling a taxon that diverged much closer to the root of the Alcithoini tribe would be necessary. As such a sampling has not yet been achieved, we can only make meaningful comparisons between the molecular data and fossil record of the *Alcithoe* species.

When the molecular analysis is repeated, considering only the timeframe beginning with the oldest divergence of the extant *Alcithoe*, B-D and D/B are estimated to be 0.1702 and 0.368 respectively. This is a significant increase in the diversification rate and a reduction in the deviation from a pure birth rate (D = 0). The calculated values for speciation rate B and the extinction rate D are; [B = 0.2693 and D = 0.0991]. Over this shorter time frame the extinction rate now only 37% of the speciation rate.

In order to compare an approximately equivalent time scale, the origin of the *Alcithoe* was considered to be around 15 ma, and it was assumed that a single unobserved lineage gave rise to the currently recognised genus. The taxonomic rates estimated from the fossil record under these assumptions are a speciation rate of 0.064 and an extinction rate of 0.057. In this instance the extinction rate is 89%, very similar to the ratio calculated for the complete timeframe, and still considerably higher then the values derived from the molecular data.

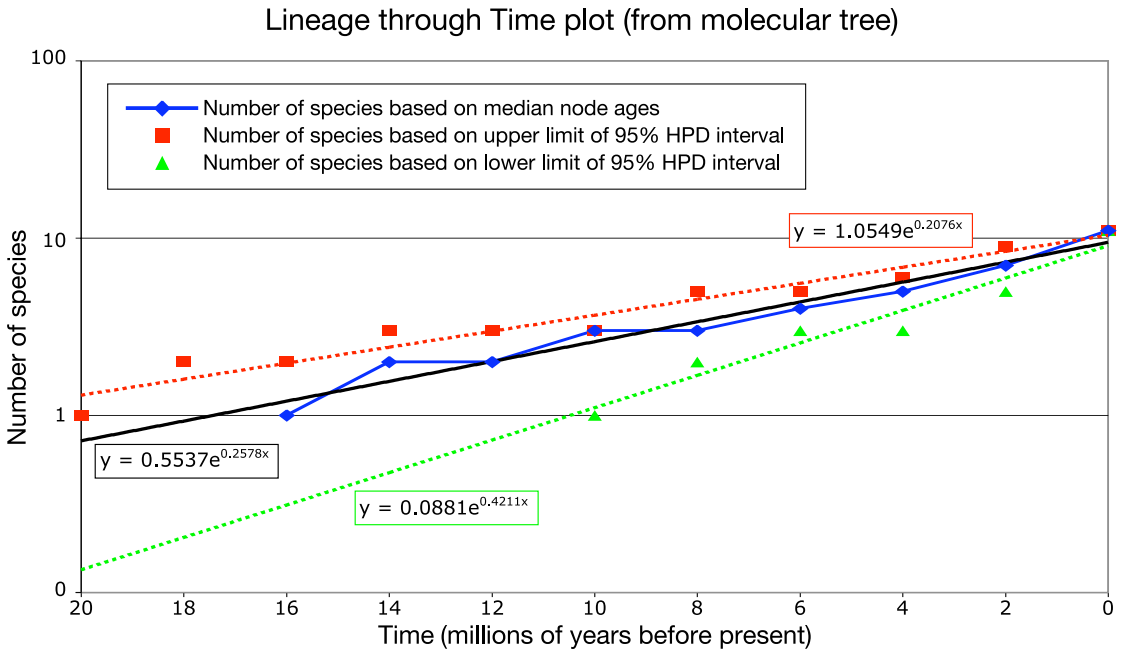

FIGURE 5.3—A lineage through time plot reveals an increasing speciation rate from the dated molecular phylogeny of the living *Alcithoe*. The number of lineages in tree was counted at 2 million year intervals, diamonds indicate the count based on the inferred median node ages, squares indicate the count based on the upper limit of the 95% HPD interval, triangles indicate the counts based on the lower limit of the 95% HPD interval. A solid trend line is given for the median values, and the slope of this line give the speciation rate (0.2076). Dotted lines indicate the trend lines based on the 95% HPD interval and show the level of uncertainty based on the distributions of node ages in the molecular analysis.

Alternative methods of estimating these rates were applied to generate additional values for comparison. Using the molecular tree a lineage through time (LTT) plot was generated (Figure 5.3). This plot does not indicate a significant departure from an expected linear best-fit trendline and infers a speciation rate of between 0.2 and 0.4 (median of 0.26) based on the 95% HPD intervals of the Bayesian inferred node heights in the tree. This rate is similar to the rate derived from the diversification rate data obtained from the Bayesian phylogenetic reconstruction, but is considerably different to the fossil record data. A dynamic survivorship analysis of the fossil data (Figure 5.4) estimated an extinction rate in the Alcithoini tribe of approximately 0.4. This rate is considerably higher then any of the extinction rate estimates thus far.

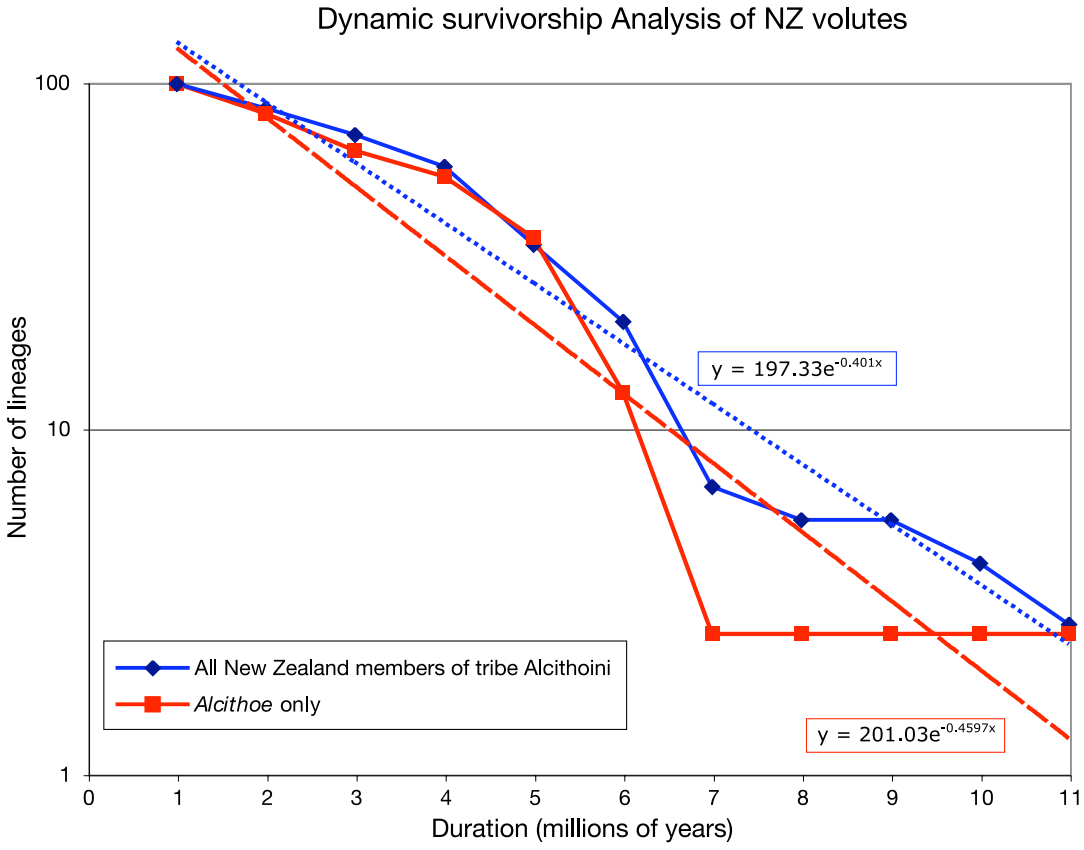

FIGURE 5.4—A dynamic survivorship analysis indicates the extinction rates for the fossil Alcithoini and the fossil *Alcithoe*. Rates of extinction derived from slopes of the trend lines (dotted for the Alcithoini, dashed for the *Alcithoe*) show that the estimated extinction rate in the *Alcithoe* is similar to the Alcithoini, in which the genus is nested.

5.3.3 Trends

In order to examine the consistency of rate estimates in the molecular dataset, B-D and D/B values were estimated for a series of subsets of the taxa representing decreasing ages at the root of each subset. This approach effectively simulates sampling the taxonomic rates at a series of time points. Comparison of molecular inferred rate estimates at different times shows that the diversification rate increases dramatically toward the present while the D/B value decreases (Figure 5.5). The observation of a decreasing departure from a pure birth process, as indicated by the reducing D/B values, is consistent with the expectation that extinction will have a decreasing role over time as the time in which extinction can occur is reduced. However, the inference of a dramatically increasing extinction rate is at odds with this expectation. Interestingly the rates inferred from the complete phylogeny, including out-groups, appear to be consistent with the patterns seen for within *Alcithoe* rates.

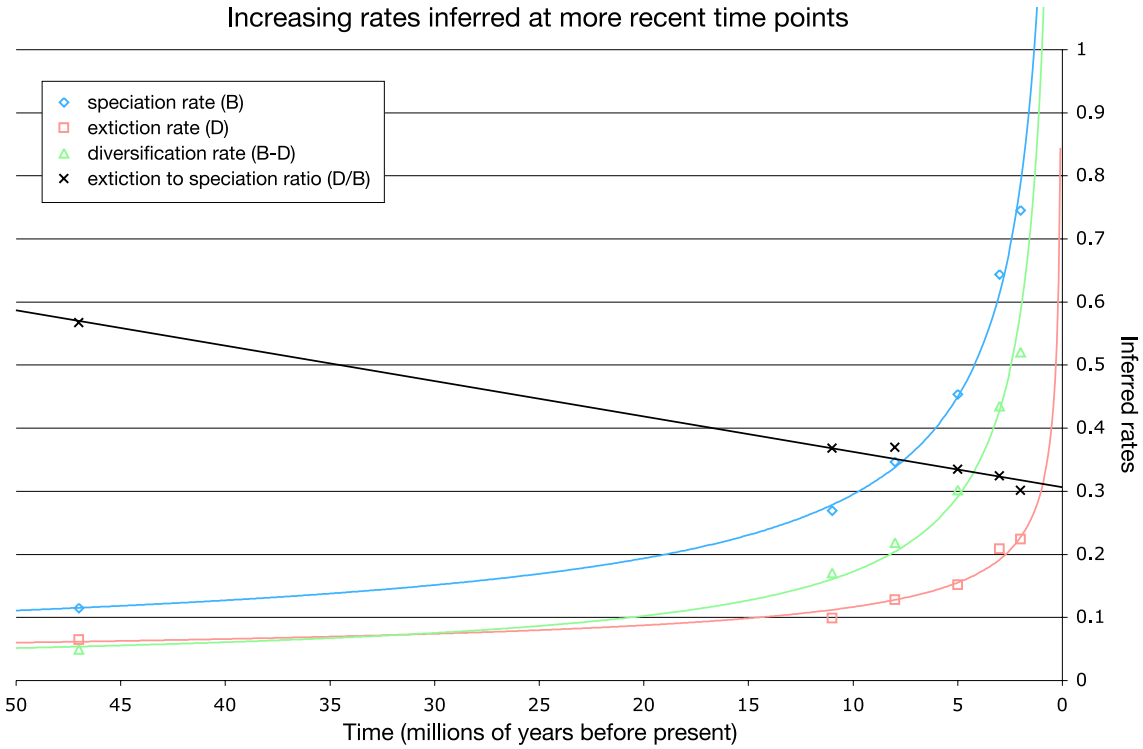

Figure 5.5—Rates derived from a sliding scale molecular analysis show that both extinction and speciation rates in Alcithoe have increased during the Cenozoic. Estimates of the diversification rate (D) and the extinction to speciation ratio (D/B) were obtained for a series of molecular clock analyses of increasingly younger *Alcithoe* species. Speciation (B) and extinction (D) rates were then calculated. These values were plotted against the inferred age of the root node for taxon subset in question. Trend lines were fitted to each set of rate estimates.

An evaluation of within stage rate estimates inferred from the fossil record illustrates long term trends comparable to those inferred from the molecular data. The long-term trends seen in the fossil record are of both increasing origination and extinction rates during the Cenozoic (Figure 5.6). In contrast to the molecular data, the paleontological data depicts linear trends of increasing lineage origination and extinction.

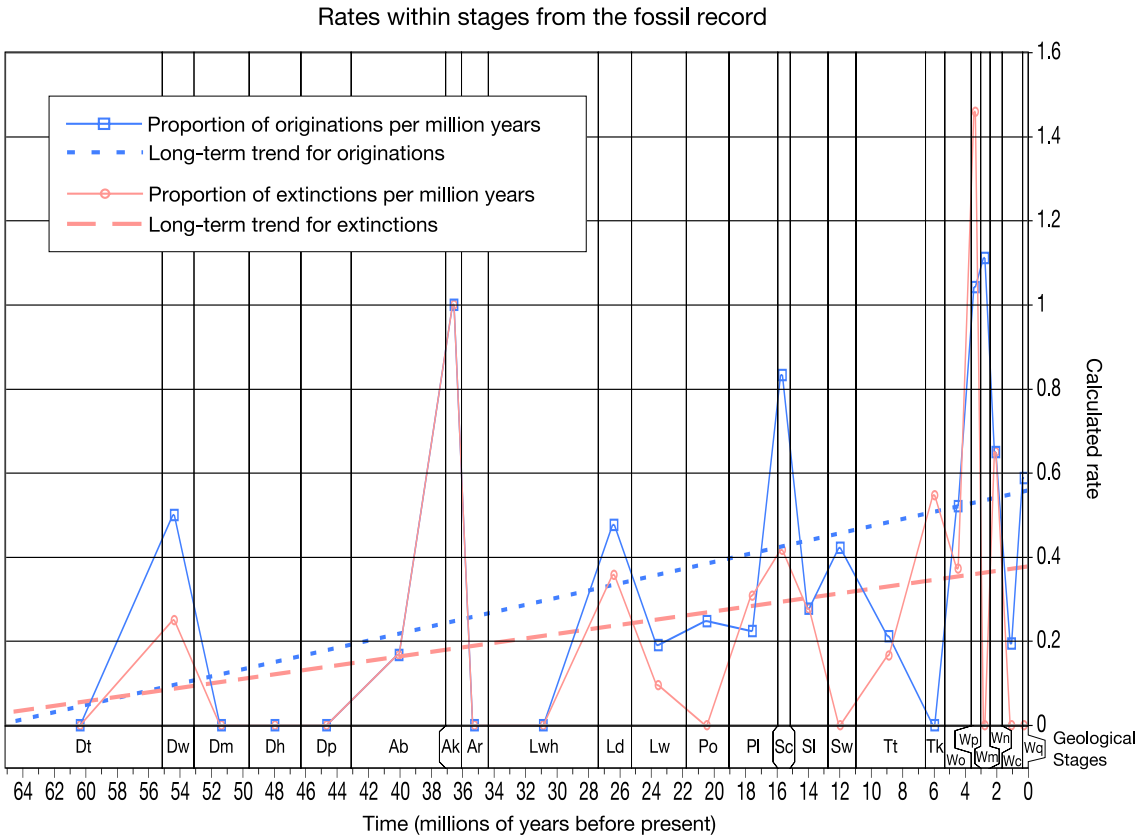

Figure 5.6—Origination and extinction rate trends derived from within stage rates indicates increasing rates during the Cenozoic. Within stage rates where calculated based on stage-boundary crossing taxa. Long-term trends were derived by fitting a line of best-fit to the within-stage rates. The stages of the New Zealand geological timescale are indicated (Wq – Haweran, Wc – Castlecliffian, Wn – Nukumaruan, Wm – Mangapanian, Wp – Waipipian, Wo – Opoitian, Tk – Kapitean, Tt – Tongaporutuan, Sw – Waiauan, Sl – Lillburnian, Sc – Clifdenian, Pl – Altonian, Po – Otaian, Lw – Waitakian, Ld – Duntroonian, Lwh – Whaingaroan, Ar – Runangan, Ak – Kaiatan, Ab – Bortonian, Dp – Porangan, Dh – Heretaungan, Dm – Mangaorapan, Dw – Waipawan).

5.4 DISCUSSION

5.4.1 Patterns in the Paleontological and Molecular Data

There are clearly several aspects requiring further investigation in the comparison of the molecular and current paleontological interpretations of the evolution of the *Alcithoe* lineage. The earliest spilt in the extant taxa occurs in the middle Miocene and yet the fossil record of *Alcithoe* species extends back into the late Oligocene. The earliest divergence separates the ancestor of the species *A. benthicola* and *A. flemingi* from the remainder of the *Alcithoe* species. This clade has morphological affinities with the *Spinomelon* genus, but the relative origination times of these two genera in the fossil record is inconsistent with this interpretation. The middle Miocene appears to be have been an important time in the evolution of the modern *Alcithoe*, and the hiatus of specimens in the fossil record from this time is a major limiting factor in the interpretation of the group. Low sampling probabilities through middle Miocene stages (Crampton et al. 2003) indicate that the lack of fossils is a result of poor preservation rather than there being few taxa at that time. However, the molecular inference of the origin of the two clades of modern *Alcithoe* during a time of poor fossil preservation is consistent with observations by (Foote 2001) of Phanerozoic taxa. The existence of a modern clade representing the *Leporemax* subgenus based on morphology is not supported by the molecular data. Inference of a direct ancestor/descendent relationship of the *A.*(*L.*) *rugosa* - *A.*(*L.*) *gatesi* - *A.*(*L.*) *brevis* - *A.*(*L.*) *fusus* lineage is refuted by the molecular data. The origin of the *A. fusus* species is much too recent, and is part of a radiation from an ancestor that is likely to have been more similar to *A. arabica* rather than *Leporemax*-like. I suggest that the current interpretation of the evolutionary affinities of *A. fusus* is a result of convergent morphological evolution.

Despite specific discrepancies highlighted between the two datasets, it is clear that when considered together they allow for a more complete assessment of evolutionary patterns observed in the *Alcithoe*. The combined data suggest that the mode of evolution in this group has been one of succession and replacement with subsequent diversification in the succeeding lineage.

5.4.2 Disparate absolute rate estimates

Rates inferred from the molecular data are consistently higher for both origination and extinction than those inferred from the fossil record. However, extinction rates derived from the fossil record consistently account for a much larger proportion of the fossil origination rates (~90%) then the molecular rates (~37 – 57%). Over the timeframes in question these two rates would result in substantially different tree diversities. These differences probably result from compounding bias in each dataset, as opposed to an overall deficiency in one.

It is likely that the origination rates derived from the fossil record are lower than the actual rates. Less then half of the extant taxa are represented in the fossil record leading to the lack of any evidence for most of the speciation events leading to the modern species. This observation, coupled with the finding that origination events tend to be distorted by variable preservation (Foote 2001), leads to an expectation of unobserved speciation.

The expectation of speciation rate underestimation in the molecular data, as a result of a small number of taxa and missing taxa (Paradis 2004), seems not to hold for the *Alcithoe*. Rates derived from the molecular data are consistently higher then those inferred from the fossil data. The greater problem in the molecular data seems to be an inability to adequately capture the extinction process. This discrepancy may be indicative of the inability of the molecular data accurately measure extinction rate when the relative extinction rate (D/B) is high in the lineage under consideration (Rabosky and Lovette 2008).

The molecular data has shown that the morphologically variable species *A. wilsonae* has been recognised as multiple separate species and subspecies (Chapter 3). One of the forms is known in the fossil record in the Tongaporutuan stage (10.92 – 6.5 ma) (Beu and Maxwell 1990) and it is probable that other forms exist in the fossil record but are currently known as different species. It is therefore likely that a reconsideration of the fossil record will alter the number of recognised species and this will impact on the analysis of diversity in the lineage.

5.4.3 Long-term trends

Despite the inconsistent absolute rate estimates, both datasets reveal a general trend of increasing origination and extinction rates in the *Alcithoe* lineage. This rate increase is at odds with the large-scale patterns of declining biodiversity in the New Zealand Cenozoic shown by Crampton et al. (2003). The rate of increase seen in the molecular data is of a considerably different scale to that of the fossil data. The fossil data depict a linear increase, while the molecular data imply a power function increase. This molecularly derived increase is reminiscent of the J-shaped curve reported by Ho et al. (2005) for molecular substitution rate time dependency. It seems unlikely that the two are related as no indication of this time dependency was seen in the molecular clock analysis of the *Alcithoe* (Chapter 4). A more likely explanation is that the birth/death model used to approximate the branching process in the *Alcithoe* molecular phylogeny under Bayesian reconstruction is not an adequate model for the data. Primarily, the assumption that rates of speciation and extinction are the same for all lineages is unlikely to hold for this group. Two key features seen in a comparison of the two datasets support this hypothesis. Firstly, all of the species that have diverged in the last 2 million years effectively represent a radiation of a single modern lineage. Secondly, the modern diversity represents a single surviving lineage from what appears to have been a rich early Miocene group.

Additionally, the power function increase seen in the rate inferred from the molecular data appears to coincide with a logarithmic decrease in the number of informative observed patterns in the data. This observation may indicate that these rates are connected to the underlying site pattern in the nucleotide data, and not just the branching pattern. This finding alludes to a correlation between the speciation rate and the substitution rate, a situation that has rarely been demonstrated (Webster et al. 2003).

5.4.4 Concluding remarks

The analysis of diversification rates presented here is predominantly of an explorative and demonstrative nature. A complete analysis with greater statistical rigour needs to be carried out to establish accurate measures of taxonomic rates in the *Alcithoe*. However, these approximate inferences serve to compare the general similarities and differences between the paleontological and molecular datasets.

Individually both data sets appear to contain shortcomings that hinder the calculation of absolute rates of speciation and extinction. The molecular dataset appears to depart from the basic models of Birth/Death diversification processes. This problem could be overcome with the application of recently developed conditional Birth/Death models (e.g. Etienne and Apol 2009). The paleontological dataset seems to be more robust than the molecular data for the inference of taxonomic rate, but is still hampered by significant gaps in the fossil record.

Regardless of the drawbacks inherent to each dataset, they are concordant in the inference of increasing rates of extinction and origination in the *Alcithoe*. It is important to reiterate that this finding is counter to the trends shown for the general Cenozoic molluscan fauna (Crampton et al. 2006b). In addition, the combined analysis of the molecular and paleontological data is both possible and appropriate. The consideration of both datasets is required to infer the pattern of succession and replacement apparent in the *Alcithoe*. Furthermore, the side-by-side comparison of both datasets highlights specific areas in each dataset that can be further developed to allow more robust analysis.

5.5 REFERENCES

Alroy, J. 2000. New methods for quantifying macroevolutionary patterns and processes. Paleobiology 26:707-733.

Barraclough, T. G., Nee, S. 2001. Phylogenetics and speciation. Trends in Ecology & Evolution 16:391-399.

Beu, A. G., Maxwell, P. A. 1990. Cenozoic Mollusca of New Zealand. New Zealand Geological Survey Paleontological Bulletin 58:1-518.

Cooper, R. A., Maxwell, P. A., Crampton, J. S., Beu, A. G., Jones, C. M., Marshall, B. A. 2006. Completeness of the fossil record: Estimating losses due to small body size. Geology 34:241-244.

Crampton, J. S., Beu, A. G., Cooper, R. A., Jones, C. M., Marshall, B., Maxwell, P. A. 2003. Estimating the rock volume bias in paleobiodiversity studies. Science 301:358-360.

Crampton, J. S., Foote, M., Beu, A. G., Cooper, R. A., Matcham, L., Jones, C. M., Maxwell, P. A., Marshall, B. A. 2006a. Second-order sequence stratigraphic

controls on the quality of the fossil record at an active margin: New Zealand Eocene to recent shelf molluscs. Palaios 21:86-105.

Crampton, J. S., Foote, M., Beu, A. G., Maxwell, P. A., Cooper, R. A., Matcham, L., Marshall, B. A., Jones, C. M. 2006b. The ark was full! Constant to declining Cenozoic shallow marine biodiversity on an isolated midlatitude continent. Paleobiology 32:509-532.

Dell, R. K. 1978. Additions to the New Zealand Recent molluscan fauna with notes on *Pachymelon (Palomelon)*. National Museum of New Zealand Records 1:161-176.

Etienne, R. S., Apol, M. E. F. 2009. Estimating speciation and extinction rates from diversity data and the fossil record. Evolution 63:244-255.

Foote, M. 2007. Symmetric waxing and waning of marine invertebrate genera. Paleobiology 33:517-529.

Foote, M. 2000. Origination and extinction components of taxonomic diversity: general problems. Paleobiology 26:74-102.

Foote, M. 2001. Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis. Paleobiology 27:602-630.

Foote, M., Miller, A. I. 2007. Principles of Paleontology. W.H. Freeman and Company, New York.

Foote, M., Sepkoski, J. J. 1999. Absolute measures of the completeness of the fossil record. Nature 398:415-417.

Ho, S. Y. W., Phillips, M. J., Cooper, A., Drummond, A. J. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Molecular Biology and Evolution 22:1561-1568.

Jackson, J. B. C., Johnson, K. G. 2001. Paleoecology - Measuring past biodiversity. Science 293:2401-2404.

Kidwell, S. M., Flessa, K. W. 1995. The quality of the fossil record - Populations, species, and communities. Annual Review of Ecology and Systematics 26:269-299.

Kirchner, J. W. 2002. Evolutionary speed limits inferred from the fossil record. Nature 415:65-68.

Marwick, J. 1926. Teriary and Recent Volutidae of New Zealand. Transactions of the New Zealand Institute 56:259-303.

McPeek, M. A. 2008. The ecological dynamics of clade diversification and community assembly. American Naturalist 172:E270-E284.

Mittelbach, G. G., Schemske, D. W., Cornell, H. V., Allen, A. P., Brown, J. M., Bush, M. B., Harrison, S. P., Hurlbert, A. H., Knowlton, N., Lessios, H. A., McCain, C. M., McCune, A. R., McDade, L. A., McPeek, M. A., Near, T. J., Price, T. D., Ricklefs, R. E., Roy, K., Sax, D. F., Schluter, D., Sobel, J. M., Turelli, M. 2007. Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. Ecology Letters 10:315-331.

Nee, S. 2006. Birth-Death models in macroevolution. Annual Review of Ecology, Evolution, and Systematics 37:1-17.

Nee, S. 2001. Inferring speciation rates from phylogenies. Evolution 55:661-668.

Paradis, E. 2004. Can extinction rates be estimated without fossils? Journal of Theoretical Biology 229:19-30.

Powell, A. W. B. 1979. New Zealand Mollusca. Marine, land and freshwater shells. Collins, Auckland.

Purvis, A. 2008. Phylogenetic approaches to the study of extinction. Annual Review of Ecology Evolution and Systematics 39:301-319.

Rabosky, D. L., Lovette, I. J. 2008. Explosive evolutionary radiations: Decreasing speciation or increasing extinction through time? Evolution 62:1866-1875.

Raup, D. M. 1985. Mathematical-models of cladogenesis. Paleobiology 11:42-52.

Ricklefs, R. E. 2007. Estimating diversification rates from phylogenetic information. Trends in Ecology & Evolution 22:601-610.

Suter, H. 1913. Manual of the New Zealand Mollusca: with an atlas of quarto plates. Government Printer, Wellington.

Valentine, J. W., Jablonski, D., Kidwell, S., Roy, K. 2006. Assessing the fidelity of the fossil record by using marine bivalves. Proceedings of the National Academy of Sciences of the United States of America 103:6599-6604.

Webster, A. J., Payne, R. J. H., Pagel, M. 2003. Molecular Phylogenies link rates of evolution and speciation. Science 301:478-478.

CHAPTER SIX

6 RECONCILING PALEONTOLOGY AND MOLECULAR BIOLOGY

6.1 HOW HAVE THE *ALCITHOE* HELPED?

6.1.1 Species identification

Identification of character differences in the fossil record that accurately define biological species remains a serious problem (Forey et al. 2004), not only for phylogenetic reconstruction but also studies of biodiversity change, historical biogeography and occupancy, and speciation processes. Species identification in extant taxa can also suffer from this problem, and the solutions are usually specific to a single lineage or clade of related species.

> Hence, in determining whether a form should be ranked as a species or a variety, the opinion of naturalists having sound judgment and wide experience seems the only guide to follow (Darwin 1859).

Alcithoe is no exception and the extant fauna has, as expected, provided challenges that genetics can overcome. While many species identified by shell morphology were concordant with mtDNA results, two particular examples illustrate the problems associated with the lack of informative shell characters in *Alcithoe*. The morphological diversity within *A. wilsonae* was considered sufficient to recognise at least two separate lineages (*A. wilsonae* and *A. knoxi*), and to place them in separate genera at one time (Powell 1979). However, the genetic diversity of the

clade, measured with mitochondrial DNA sequence, clearly identifies a single species (Chapter 3). Furthermore, it has been demonstrated that two morphologically very similar species (*A. larochei* and *A. tigrina*) are clearly distinct and quite separate, contrary to the original taxonomic placement of *A. tigrina* as a subspecies of *A. larochei*. Thus this thesis has identified that units of evolution are not necessarily well represented by morphological characters. Although not a new finding, the results allow development of tools to better understand phenotypic evolution in this gastropod lineage, such as morphometrics (Crampton et al. 2009).

6.1.2 Phylogenetic inference

Phylogenetic relationships inferred from fossil material rely on characters that are likely to be under selection (i.e. shell shape), and it is recognised that convergence to similar forms will mislead the inference of relatedness (Moore and Willmer 1997). The large scale of DNA data sets and characters that evolve in a more neutral way, lead to phylogenetic hypotheses we consider closer to the truth than morphological-based trees. The extensive fossil material for *Alcithoe* provided a high level of confidence in phylogenetic hypotheses based on paleontological data, but this thesis has found that confidence to be misplaced. Evidence provided by analysis of almost 8Kb of mtDNA (chapter 1) suggested conflict with a number of relationships inferred by the fossil record. One example is the *Leporemax* subgenus, recognised in the fossil record as species similar to *Alcithoe* but are considerably smaller and more slender (Beu and Maxwell 1990), and for which *A. (L.) fusus* is the type species, and from which *A. jaculoides* was suggested to have diverged (Bail and Limpus 2005). The molecular data shows that *A. fusus* is clearly recently diverged from the *Alcithoe sensu stricto*, not from a long series of "*Leporemax"* species. Furthermore, *A. jaculoides* is clearly nested within the *Alcithoe sensu stricto* clade. This thesis therefore illustrates that even an excellent fossil record may not provide accurate phylogenetic information, although the problem of morphological convergence may be more extreme in gastropod evolution, any tree based on characters under strong selection can be misleading. Where the fossil record is less dense (such as for birds, mammals) interpretation of relationships must be taken with even greater care.

6.1.3 Rocks and clocks

Molecular clock analyses are dependent on paleontological data to accurately calibrate the time dimension (Ho and Phillips 2009). Calibrations do not just rely on correctly dating fossil specimens; there is also a requirement that the fossil data is applied to the correct node in the tree. Therefore species identification is extremely important (chapter 3), to ensure accurate estimation of divergence times. This thesis illustrates how easily one can be mislead by convergence of morphology, but also shows how population-genetic analysis can provide a powerful tool to ensure the correct identification and placement of calibrating taxa.

High quality and extensive fossil data provides very robust calibration for most nodes. However, distant nodes such as the root node of the *Alcithoe* tree, appear to still be susceptible to error. The *Alcithoe* dataset appears to be well suited to quantify effect the that branch contraction (Phillips 2009) might have on the accurate estimation of divergence dates. In general though, when all the priors are set to minimise interaction and artifact, then DNA dating and fossil dating are in excellent agreement with each other.

6.1.4 Biodiversity patterns

This thesis presents a dataset unlike any previously studied to document changes over time in biodiversity. Although molecular phylogenetics and paleontology provide quite distinct estimates of speciation and extinction rates both methods identify an increase in overall species diversity resulting from higher speciation rates rather than lower extinction rates. The prevailing thought on global biodiversity has been that it has been increasing over time, peaking in the late Cenozoic (e.g. Jablonski et al. 2003), however it has been suggested that this is not real and is the result of sample bias (e.g. Foote 2000). Indeed, Crampton et al. (2006) have shown that in the New Zealand Cenozoic fossil record there has been a relatively constant diversity, with a decline in the last 5 million years. The findings of this thesis are at odds with this result, particularly for the recent decline as the *Alcithoe* data suggests an increase in diversity in the last 5 million years, with around half the modern species of *Alcithoe* have originated in that time. If these opposing trends are real, then it would be very interesting to explore the factors associated with the generation of biodiversity (Ricklefs 2007; McPeek 2008) by

identifying the ecological factors that have promoted the recent diversification of the *Alcithoe* while the majority of New Zealand marine taxa have been in decline.

6.2 WHERE TO FROM HERE?

6.2.1 Expanding the Alcithoe dataset

6.2.1.1 Complete sampling of Alcithoe

The taxonomy of the *Alcithoe* will be greatly enhanced by completing the sampling of species represented in the molecular dataset. A number of named species and subspecies were not sampled in the scope of the work descried here, but it is likely that most of those represent forms of species present in this study and that only three distinct species are missing (Bruce Marshall, pers. com.). What will be greatly beneficial to further study of the *Alcithoe* is broader sampling within each species, similar to the dataset described for *A. wilsonae* in Chapter 3. A complete sample of extant species is desirable for the study of diversification rates (Ricklefs 2007), as greater accuracy can be achieved when the extent of the modern diversity is known. More extensive sampling within each species will enable an assessment of the morphological and genetic limits of the known species. Comparison of these levels of variation and the ecology of these species will be highly informative. It is likely that such a comparison will allow an objective analysis of the niche space that *Alcithoe* species evolved to exploit. This will then facilitate the consideration of the ecological factors that have generated selective pressure and the biological adaptations that have led to the success of the modern species (e.g. Chiba 1998).

6.2.1.2 Refinement of morphological analysis

Now that a robust molecular phylogeny exists for the *Alcithoe* there can be a clarification of the morphological characterisation of the genus. A high level of within-species variation and convergence between species continues to confound morphological analysis of the New Zealand volutes. The molecular phylogeny allows a reassessment of shell morphology characters in light of a stable hypothesis of the relationships amongst the extant taxa. Morphological characters can be selected that are concordant with the molecular phylogeny, and these characters can then be used to assess the fossil taxa. However, a paucity of discriminating gross morphological features confound this solution somewhat, by offering little

resolution. In order to circumvent this lack of information morphometric data has obtained from large numbers of *Alcithoe* specimens (Crampton et al. 2009). An analysis of the principle coordinates of the morphometric data recovers clouds of data points that represent collections of identified species. Furthermore the means of these clouds are statistically significantly different. However, phylogenetic reconstruction using continuous morphometric data has is not well developed yet. The robust molecular phylogeny will be instrumental as a reference in the development of phylogenetic reconstruction using the morphometric data.

6.2.1.3 From population rates to species rates

The incongruence of population-level mutation rates and species-level substitution rates is a known issue (Ho and Larson 2006) that requires further analysis. Such a study should be carried out in a group of species for which both types of rate can be calculated and compared. With the assembly of population-level datasets for many of the species of *Alcithoe* the genealogical rates for each of the species can be determined. Within-population rates can then be compared directly to the rates on individual branches and overall mean rate inferred for the *Alcithoe* in this thesis. From such a comparison observed differences in the population and species rates can be quantified and further examined. In addition, a detailed analysis of the underlying nucleotide variation can be carried out in order to determine if it is appropriate to consider all sites equally at all levels of analysis. For example, what is the most appropriate way to consider sites that are variable within the population as well as between species? Current methods of reconstructing species phylogenies do not consider this problem, as each species is often represented by only one sequence. Therefore a site that maybe saturated in a population sample would be represented by only one base in a species sample, and if in this species sample it is observed to be informative will lead to inflated confidence in a conclusion. Furthermore, expanded population level sampling will facilitate the investigation of population size effects on rates of evolution. Theoretically population size should have an effect on evolutionary rates (e.g. Bromham and Woolfit 2004), but some studies have detected no such effect (Bazin et al. 2006; Nabholz et al. 2009). Such effects should be associated with different founder-population size of species and differing timeframes between species origination and population expansion.

6.2.2 Markers

6.2.2.1 Development of nuclear markers

It is of critical importance to develop new nuclear DNA markers for use in molluscs. Several were tested on the *Alcithoe*, but were found to have insufficient variability (28S, 18S), excessive heteroplasmy (Histone H3), multiple unalignable copies (ITS1, ITS2) or simply did not amplify (hemocyanin, elongation factor 1-alpha, engrailed, arginine kinase, beta-catenin). Second-generation sequencing provides a method for cost effective non-directed (i.e. specific primers are not required) generation of nuclear DNA sequence that can then be examined to develop informative sequencing markers. Development of a suite of nuclear markers will facilitate multiple locus 'tube tree' analysis, as described by (Heled and Drummond 2008), which is capable of inferring more accurate population-size estimates and coalescence times for both genes and species.

6.2.2.2 Complete mitochondrial genome sequences

The development of methods to sequence and analyse complete mitochondrial genomes will prove invaluable for the elucidation of mollucsan relationships (Boore and Brown 1994; Simison and Boore 2008). At the population level such data will allow researchers to specifically target more variable elements of the mitochondrial genome that are more suitable for inference of close relationships. At deeper phylogenetic levels structural elements of the genome, such as gene order and non-coding secondary structures (Boore 2006), will provide data to discriminate relationships. Ongoing research on the early diversification of birds and mammals has made extensive use of complete mitochondrial genomes to gain enhanced resolution of rapid diversifications (e.g. Phillips et al. 2006; Pratt et al. 2009). The timeframes of the bird and mammal diversifications (e.g. Brown et al. 2008) are approximately similar to that of the neogastropod radiation (Cunha et al. 2009). Furthermore, phylogenetic inference of the relationships during the neogastropod diversification suffer from the same lack of resolution in most markers, that led to the development of complete mitochondrial genome sequencing for birds and mammals. It is likely that whole mitochondrial genome data would also provide traction on earlier rapid diversifications seen in the molluscan lineage, such as the diversification of the caenogastropods and the earliest diversification of the major molluscan orders.

Improved resolution of the diversifications will allow us to investigate different macroevolutionary patterns observed at different levels of evolutionary depth. For example, analysis of basal relationships in neogastropods shows that there is significant branching close to the root of the neogastropod phylogeny (Harasewych et al. 1997; Colgan et al. 2007; Ponder and Lindberg 2008). However, for within genus level patterns seen in younger lineages, such as *Alcithoe*, the majority of the diversification is toward the tips of the tree. Are these opposed patterns indicative of different mechanisms or the same mechanism over different time scales?

6.2.2.3 Splits to assess markers

Analysis of splits data is a highly informative method to assess the signal in molecular data. By considering partitions in the data it is possible to diagnose sources of conflict in an extensive dataset. As the number of genes in datasets increase, and particularly where relatively small genetic distances or short internal branches are seen, the assessment of signal and noise will become more important. The development of an index of partition compatibility based on splits data could be of benefit, particularly as the utility of traditional methods such as the consistency index have been called into question (Barker and Lutzoni 2002). A splits based index could make use of a measure of the split weight and conflict data for a given partition and compare this to all other partitions in a dataset. This measure would assess the degree to which the signal in that partition deviates from the rest of the dataset. Such quality control of the data will prevent erroneous signals, which might deviate from a more or less neutral pattern, from introducing confounding noise to an analysis.

6.2.2.4 Multiscale analysis

As the ease by which the amount of molecular data increases, the size and complexity of phylogenetic datasets increases dramatically. Even though our computational capabilities are currently sufficient for the scale of data we analyse now, this could soon become the bottleneck in the study of evolutionary biology. In the near future, as second- and third-generation sequencing systems continue to promise more data in shorter time frames, the way we analysis this data will have to evolve. It is not yet clear if a 'total-evidence' approach will be appropriate for analysis of this data. Growing evidence suggests that not all data is useful at all levels of analysis. For example highly conserved markers are generally not

informative at the population level. One solution to this problem will be multiscale analysis, where markers are applied only at levels where they are informative, and excluded where they are not. For example structural elements of the genome (such as gene order) can be applied to the inference of very old divergences while mitochondrial control regions or hypervariable nuclear sequences are used to differentiate the very youngest divergences in individual sub-trees of the same analysis.

6.2.3 More illuminative answers through more refined questions

6.2.3.1 What effect do extinction patterns have on divergence date estimates?

An explanation for the consistent inference of younger than expected dates for the root of the *Alcithoe* tree is branch contraction (Phillips 2009). This contraction appears to result from an asymmetry in the pattern of extinction leading to lack of living representatives of old divergences. As there is an absence of older lineages an accurate measure of the molecular divergence from this time is impossible. This hypothesis can be tested through simulation of divergence and extinction. By simulating approximately 50 million years of evolution under substitution parameters inferred for the *Alcithoe*, and a pure birth speciation model with parameters informed from the fossil record, a dataset can be generated that approximates a complete sample of possible species to evolve in the time period. Extinction can then be simulated by trimming the taxon set to equal the number known extant *Alcithoe* species. Different patterns of extinction can be tested by selective removal of simulated taxa. Molecular divergence estimates can then be tested under different extinction patterns and compared to known simulated divergence times. In this way the difference in divergence estimates can be directly compared for balanced and unbalanced trees. The effect of specific tree shapes on divergence date estimates can be tested.

6.2.3.2 What exactly is a different rate of evolution?

At present it is not clear at what scale rate estimates should be differentiated. Currently rates must differ by orders of magnitude to be considered significantly different. This uncertainty is in large part due to wide margins of error associated with existing rate estimates (Bromham and Penny 2003). However, even small differences in rate can lead to significantly different amounts of accumulated

change when considered over geological timescales. Furthermore, within relatively closely related taxa much smaller differences in substitution rate are likely to be meaningful.

The degree to which rates of molecular evolution are directly comparable is, in my opinion, highly questionable at present. Absolute rates of evolution reported in the literature to date have been derived from a multitude of sources, such as; proteins (e.g. Sarich and Wilson 1967; King and Wilson 1975), RFLP data (e.g. Brown et al. 1979; Brower 1994), and sequence data from a number of markers (e.g. Wilson et al. 1985; Kumar and Hedges 1998; Smith et al. 2006). Mutation rates are unlikely to be invariable across all species and the rate at which mutations are fixed in populations to become substitutions between species will not be the same. Functional and structural constraints that allow variation to occur will vary between markers in addition to variability between species. As a result the range of rates reported in the literature is unsurprising. In order to properly compare rates inferred for different taxon sets the data used to derive the rates must be equivalent, otherwise the source of any observed differences could be attributed to several causes.

6.2.3.3 Getting more information with more specific questions

The quantity and detail of data resulting from Bayesian inference of molecular divergence times allows much more specific questions than are currently asked to be addressed. For example, in examining the rates of substitution in *Alcithoe* it is not until we ask the specific question "is the rate in *A. fusus* faster than the rates in other species?" that we find a significant rate difference in the group. To extend this example, by asking a series slightly different questions we could construct a posterior probability distribution on how much faster the rate in *A. fusus* is than in each other species.

Current examination of node heights addresses the rather general question of how old a given node is. A more specific question that can then be asked is what is the posterior probability that a given node coincides with a given geological stage. Or, what is the posterior probability of there being a given length of time between two (consecutive) nodes.

6.2.4 The way forward in the synthesis of molecular evolution and paleontology

6.2.4.1 More than just age calibrations

Additional information can be obtained from the fossil record to inform parameters in molecular clock analysis. Paleontological data, such as exists for the New Zealand marine mollusc fauna, can be used to derive rates of extinction and speciation that could be used as prior information. An informed choice of branching process in a molecular analysis could be made based on the patterns of speciation and extinction observed in the fossil record. It is likely that comparison to real patterns in the fossil record will be an informative approach to testing new conditional birth/death models (e.g. Etienne and Apol 2009). In addition, incorporating fossil species occupancy data (Foote et al. 2007), will likely inform more robust molecular inferences.

6.2.4.2 Use fossil data a test of the quality of molecular clock data

The current practice of using fossil data to calibrate the molecular clock under-utilises the wealth of data available for the New Zealand Cenozoic marine mollusc fossil record. One avenue of making more thorough use of the exceptional fossil record offered by groups such as the *Alcithoe* is to use the high quality fossil-record data as a test of the results of a molecular analysis (Magallon 2004). For example divergence times inferred by molecular analysis could be measured by how much they differ from origination times in the fossil record (Hills et al. 2009). Summing these across a tree will give a measure of how consistent the phylogeny is with the fossil record and allow objective comparison between trees constructed under different parameters. The problem that independent data are required for calibration of the clock can be avoided by using a subset of the fossil data (such as the deeper divergences), and the fossil record for specific species with very good fossil data could be used to assess results. In addition, it might be informative to consider species occupancy data from the fossil record in parallel with branch lengths in the molecular tree. Comparisons between the inferred age of the node and the occupancy curve for the taxon represented by the branch will inform objective limits on the length of time the species may have existed before leaving sampled fossil remains (e.g. Foote et al. 1999).

6.3 REFERENCES

Bail, P., Limpus, A. 2005. The recent volutes of New Zealand with a revision of the genus *Alcithoe* H. & A. Adams, 1853 in G. T. Poppe, and K. Groh, eds. A Conchological Iconography. ConchBooks, Hackenheim.

Barker, F. K., Lutzoni, F. M. 2002. The utility of the incongruence length difference test. Systematic Biology 51:625-637.

Bazin, E., Glemin, S., Galtier, N. 2006. Population size does not influence mitochondrial genetic diversity in animals. Science 312:570-572.

Beu, A. G., Maxwell, P. A. 1990. Cenozoic Mollusca of New Zealand. New Zealand Geological Survey Paleontological Bulletin 58:1-518.

Boore, J. L. 2006. The use of genome-level characters for phylogenetic reconstruction. Trends in Ecology & Evolution 21:439-446.

Boore, J. L., Brown, W. M. 1994. Mitochondrial genomes and the phylogeny of mollusks. Nautilus 108:61-78.

Bromham, L., Penny, D. 2003. The modern molecular clock. Nature Reviews Genetics 4:216-224.

Bromham, L., Woolfit, M. 2004. Explosive radiations and the reliability of molecular clocks: Island endemic radiations as a test case. Systematic Biology 53:758-766.

Brower, A. V. Z. 1994. Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial-DNA evolution. Proceedings of the National Academy of Sciences of the United States of America 91:6491-6495.

Brown, J. W., J. S. Rest, J. Garcia-Moreno, M. D. Sorenson, and D. P. Mindell. 2008. Strong mitochondrial DNA support for a Cretaceous origin of modern avian lineages. BMC Biology **6**.

Brown, W. M., George, M., Wilson, A. C. 1979. Rapid evolution of animal mitochondrial-DNA. Proceedings of the National Academy of Sciences of the United States of America 76:1967-1971.

Chiba, S. 1998. A mathematical model for long-term patterns of evolution: effects of environmental stability and instability on macroevolutionary patterns and mass extinctions. Paleobiology 24:336-348.

Colgan, D. J., Ponder, W. F., Beacham, E., Macaranas, J. 2007. Molecular phylogenetics of Caenogastropoda (Gastropoda : Mollusca). Molecular Phylogenetics and Evolution 42:717-737.

Crampton, J. S., Foote, M., Beu, A. G., Maxwell, P. A., Cooper, R. A., Matcham, L., Marshall, B. A., Jones, C. M. 2006. The ark was full! Constant to declining Cenozoic shallow marine biodiversity on an isolated midlatitude continent. Paleobiology 32:509-532.

Crampton, J. S., Hills, S., Fenwick, M., Morgan-Richards, M., Marshall, B., Beu, A., Hendy, A., Buick, D. 2009. Species in the fossil record: hopeless monsters or hopeful messengers? in S. Trewick, N. Hiller, and R. Cooper, eds. Geology & Genes IV. Geological Society of New Zealand, Christchurch.

Cunha, R. L., C. Grande, and R. Zardoya. 2009. Neogastropod phylogenetic relationships based on entire mitochondrial genomes. BMC Evolutionary Biology **9**.

Darwin, C. 1859. On the origin of species by means of natural selection. John Murray, London.

Etienne, R. S., Apol, M. E. F. 2009. Estimating speciation and extinction rates from diversity data and the fossil record. Evolution 63:244-255.

Foote, M. 2000. Origination and extinction components of taxonomic diversity: Paleozoic and post-Paleozoic dynamics. Paleobiology 26:578-605.

Foote, M., Crampton, J. S., Beu, A. G., Marshall, B. A., Cooper, R. A., Maxwell, P. A., Matcham, I. 2007. Rise and fall of species occupancy in Cenozoic fossil mollusks. Science 318:1131-1134.

Foote, M., Hunter, J. P., Janis, C. M., Sepkoski, J. J. 1999. Evolutionary and preservational constraints on origins of biologic groups: Divergence times of eutherian mammals. Science 283:1310-1314.

Forey, P. L., Fortey, R. A., Kenrick, P., Smith, A. B. 2004. Taxonomy and fossils: a critical appraisal. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 359:639-653.

Harasewych, M. G., Adamkewicz, S. L., Blake, J. A., Saudek, D., Spriggs, T., Bult, C. J. 1997. Neogastropod phylogeny: A molecular perspective. Journal of Molluscan Studies 63:327-351.

Heled, J., Drummond, A. J. 2008. Bayesian inference of population size history from multiple loci. BMC Evolutionary Biology 8.

Hills, S., Morgan-Richards, M., Trewick, S., Crampton, J. 2009. Enhancing molecular analysis with rigorous paleontological sampling probabilities in S. Trewick, N. Hiller, and R. Cooper, eds. Geology & Genes IV. Geological Society of New Zealand, Christchurch.

Ho, S. Y. W., Larson, G. 2006. Molecular clocks: when times are a-changin'. Trends in Genetics 22:79-83.

Ho, S. Y. W., Phillips, M. J. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Systematic Biology 58:367-380.

Jablonski, D., Roy, K., Valentine, J. W., Price, R. M., Anderson, P. S. 2003. The impact of the pull of the recent on the history of marine diversity. Science 300:1133-1135.

King, M. C., Wilson, A. C. 1975. Evolution at 2 levels in humans and chimpanzees. Science 188:107-116.

Kumar, S., Hedges, S. B. 1998. A molecular timescale for vertebrate evolution. Nature 392:917-920.

Magallon, S. A. 2004. Dating lineages: Molecular and paleontological approaches to the temporal framework of clades. International Journal of Plant Sciences 165:S7-S21.

McPeek, M. A. 2008. The ecological dynamics of clade diversification and community assembly. American Naturalist 172:E270-E284.

Moore, J., Willmer, P. 1997. Convergent evolution in invertebrates. Biological Reviews 72:1-60.

Nabholz, B., Glemin, S., Galtier, N. 2009. The erratic mitochondrial clock: variations of mutation rate, not population size, affect mtDNA diversity across birds and mammals. BMC Evolutionary Biology 9.

Phillips, M. J. 2009. Branch-length estimation bias misleads molecular dating for a vertebrate mitochondrial phylogeny. Gene 441:132-140.

Phillips, M. J., McLenachan, P. A., Down, C., Gibb, G. C., Penny, D. 2006. Combined mitochondrial and nuclear DNA sequences resolve the interrelations of the major Australasian marsupial radiations. Systematic Biology 55:122-137.

Ponder, W. F., Lindberg, D. R. 2008. Molluscan evolution and phylogeny in W. F. Ponder, and D. R. Lindberg, eds. Phylogeny and Evolution of the Mollusca. University of California Press, Berkeley

Powell, A. W. B. 1979. New Zealand Mollusca. Marine, land and freshwater shells. Collins, Auckland.

Pratt, R. C., Gibb, G. C., Morgan-Richards, M., Phillips, M. J., Hendy, M. D., Penny, D. 2009. Toward resolving deep Neoaves phylogeny: Data, Signal Enhancement, and Priors. Molecular Biology and Evolution 26:313-326.

Ricklefs, R. E. 2007. Estimating diversification rates from phylogenetic information. Trends in Ecology & Evolution 22:601-610.

Sarich, V. M., Wilson, A. C. 1967. Rates of albumin evolution in primates. Proceedings of the National Academy of Sciences of the United States of America 58:142-&.

Simison, W. B., Boore, J. L. 2008. Molluscan evolutionary genomics in W. F. Ponder, and D. R. Lindberg, eds. Phylogeny and Evolution of the Mollusca. University of California Press, Berkeley

Smith, A. B., Pisani, D., Mackenzie-Dodds, J. A., Stockley, B., Webster, B. L., Littlewood, T. J. 2006. Testing the molecular clock: Molecular and paleontological estimates of divergence times in the Echinoidea (Echinodermata). Molecular Biology and Evolution 23:1832-1851.

Wilson, A. C., Cann, R. L., Carr, S. M., George, M., Gyllensten, U. B., Helmbychowski, K. M., Higuchi, R. G., Palumbi, S. R., Prager, E. M., Sage, R. D., Stoneking, M. 1985. Mitochondrial-DNA and 2 perspectives on evolutionary genetics. Biological Journal of the Linnean Society 26:375-400.

INDEX-FREE *DE NOVO* ASSEMBLY AND DECONVOLUTION OF MIXED MITOCHONDRIAL GENOMES[1]

Bennet J McComish[1,2,*,†], Simon F K Hills[1,3,†], Patrick J Biggs[1,4,5] and David Penny[1,2]

[1]Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand.

[2]Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand.

[3]Institute of Natural Resources, Massey University, Palmerston North, New Zealand.

[4]Massey Genome Service, Massey University, Palmerston North, New Zealand.

[5]Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand.

[†]These authors contributed equally to this work.

[*]Author for correspondence: Bennet McComish, Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, PO Box 11-222, Palmerston North, New Zealand.

Tel: +44 (6) 356 9099 ext 7626.

Fax: +44 (6) 350 5626.

Email: b.mccomish@massey.ac.nz

[1] This manuscript has been accepted by Genome Biology and Evolution.

ABSTRACT

Second-generation sequencing technology has allowed a very large increase in sequencing throughput. In order to make use of this high throughput, we have developed a pipeline for sequencing and *de novo* assembly of multiple mitochondrial genomes without the costs of indexing. Simulation studies on a mixture of diverse animal mitochondrial genomes showed that mitochondrial genomes could be re-assembled from a high coverage of short (35 nucleotide) reads such as those generated by a second-generation Illumina Genome Analyser. We then assessed this experimentally with long-range PCR products from mitochondria of a human, a rat, a bird, a frog, an insect and a mollusc. Comparison to reference genomes was used for deconvolution of the assembled contigs, rather than for mapping of sequence reads. As proof of concept, we report the complete mollusc mitochondrial genome of an olive shell (*Amalda northlandica*). It has a very unusual putative control region that contains an unusual structure which would probably only be detectable by next-generation sequencing. The general approach has considerable potential, especially when combined with indexed sequencing of different groups of genomes.

Keywords: multiplex sequencing, informatic deconvolution, control region, non-complementary, molluscs

INTRODUCTION

DNA sequence information is fundamental to our understanding of genome structure, function and evolution. A major advance in sequencing methodology was introduced by the Sanger group in the 1970s, with the development of the chain-termination DNA sequencing reaction (Sanger et al. 1977). Sequencing has subsequently undergone increasing degrees of industrialization, with the introduction of fluorescent radiolabeled terminators, and capillary electrophoresis, allowing the sequencing of entire genomes. In the last few years however, so-called 'second-generation' sequencing technologies have been developed using strategies such as pyrosequencing (Margulies et al. 2005) and sequencing by synthesis (Bentley 2006); strategies that are radically different from the Sanger dideoxy methodology.

Four commercial second-generation DNA sequencing platforms are now available: Roche's (454) Genome Sequencer FLX System, Illumina's Genome Analyzer (GA), Applied Biosystems' SOLiD System, and Helicos' HeliScope Single Molecule Sequencer. These all use a massively parallel approach, producing hundreds of thousands to tens of millions of sequence reads at a time, however they are much shorter than Sanger dideoxy reads. Instead of creating a clone library (which could have ethics and/or genetic modification issues), the sample DNA is fragmented and the fragments ligated to adapters, eliminating library-construction and cloning-host biases. At the time these experiments were carried out, a single run on the 454 system produced 400,000 reads of around 250 nucleotides (nt), a GA run produced over 40 million 36 nt reads, and a SOLiD run promised 86 to 114 million 35 nt reads. However, these output figures are all increasing rapidly as the technologies from each company are developed further. For example, a single GA run can currently produce 12 to 15 GB of sequence data (i.e. more than 10 million 75-bp paired-end reads per lane).

For robust phylogenetic reconstruction it is highly advantageous to demonstrate concordance between independent datasets. In molecular datasets this is often achieved by comparing results from nuclear data and mitochondrial and/or chloroplast data (e.g. Pratt et al. 2009). These datasets have often not been concordant due to the limited amount of sequence data being more indicative of aberrant histories of the gene involved rather than the evolutionary history of the

genome (Nichols 2001). With the advent of second-generation sequencing it has become increasingly possible to generate large quantities of data. Large multi-gene datasets are significantly less likely to be dominated by aberrant individual gene histories. It is therefore desirable to sequence both nuclear and organelle genomes. Due to issues such as nuclear copies of mitochondrial genomes, it is necessary to segregate organelle genomes from the nuclear sequence. However, the size of these genomes is such that much of the sequence will be wasted in many times more coverage than is needed.

If even a single lane of a GA flow cell is used to sequence something as small as a typical animal mitochondrial genome, there is a high degree of redundancy. For the 16.5 kbp human mitochondrial genome, for example, raw coverage could be over 90,000 ×, and each read would be present in 300 copies. Current *de novo* sequence assembly algorithms perform well with much lower coverage. For example, Hernandez et al. (2008) successfully assembled a *Staphylococcus aureus* genome from 35-bp reads with a raw coverage of 48 ×.

A solution to this problem is sequencing a mixture of many organelle genomes; however this leads to the difficulty of separating the individual genomes from the resulting short sequence reads. Clearly a method is required to informatically allocate *de novo* contigs to a given genome, maybe via a pooling or an indexing strategy. There are many examples of pooling and indexing strategies in the literature, though none of them do exactly the same as the strategy we are proposing. Prior to next generation sequencing there were a variety of methodologies to look at pooling and/or indexing (see for example Cai et al. 2001; Ng et al. 2006; Fullwood et al. 2009), however these kinds of approach rely on finding segments in a genome for subsequent mapping and analyses, but not for sequencing whole genomes. Illumina have developed and marketed their own indexing technology that allows up to 12 samples to be mixed in one lane of a GA flow cell. Using current protocols, each sample must be prepared individually, resulting in a linear cost increase for the number of samples under investigation. There is some cost reduction with the mixing of samples for running on the machine, but overall, this is still an expensive procedure. At the other end of the indexing continuum are new “hyperindexing” methods, such as DNA Sudoku (Erlich et al. 2009) and BARCRAWL (Frank, 2009). However, again economies of scale mean that these approaches are useful for large numbers (thousands) of short

sequences sometimes using multiple lanes and/or pooling, and so the sequencing of organellar genomes would not be appropriate with this approach either.

Our aim here is to test the hypothesis that for distantly related species (that is, for highly divergent sequences) assembly should be straightforward and unambiguous. Where there is a high degree of similarity between two sequences, however, it becomes more difficult to assemble short reads unambiguously, as there will be longer overlaps between reads from the different genomes. For these more similar genomes, we expect that indexing would be more appropriate, but we need to develop a method that could combine both approaches, index-free multiplexing and indexing. Ultimately we would like to get the cost of a mitochondrial genome to under $100, but that is beyond the scope of our present work.

We first used combined simulated reads from a set of several animal mitochondrial genomes to explore the ability of sequence assembly algorithms to separate and assemble sequences from a mixture of reads from different sources. Once optimised, the same methods were successfully applied to reads from a single lane of a GA flow cell containing a mixture of six different mitochondrial genomes.

Mitochondrial sequences from four species were successfully assembled, thus establishing that it is possible to disambiguate and assemble a complete organellar genome from a mixture of sequence reads from more distantly related species. The complete mitochondrial genome of the neogastropod mollusc *Amalda northlandica* is reported in more detail, and we identify a novel putative regulatory element, most likely a reduced control region. This structural feature can, under certain assembly conditions, interfere with complete assembly of the genome, and this control feature is unlikely to be detected by classical sequencing techniques.

This approach is complementary to the indexing strategies mentioned above. Indexed sequencing will allow our approach to be used for several mixtures in a single run, with each mixture assigned a single index. This will enable us to sequence a large number of samples with a fraction of the sample preparation that would be required if we were to assign an index to each sample. The combination of index-free multiplexing and indexing should reduce costs considerably. In the application reported here we use a disparate mixture of mitochondrial genomes (from humans to molluscs), but other combinations can certainly be used.

METHODS

SIMULATIONS

Simulations were carried out using known animal mitochondrial genome sequences which were downloaded and stored in a MySQL database. Custom Perl scripts (available from http://awcmee.massey.ac.nz/downloads.htm) were used to simulate 35-bp reads at random positions in the sequence, and to introduce errors in these reads based on observed error profiles from previous GA sequencing experiments. Reads were then extracted from the database to simulate mixtures of different genomes in predefined ratios, and written to files in FASTA format. A total of four million reads were extracted for each simulation, a conservative approximation to the number of usable reads produced on a single lane of a GA flow cell at the time of these experiments.

The simulated reads were assembled using Velvet version 0.7.26 (Zerbino and Birney 2008) and Edena version 2.1.1 (Hernandez et al. 2008), with a range of values for the hash length *k* (Velvet) or the minimum overlap between reads (Edena). The assembled contigs were aligned to the original genomes using the assembly tool of the Geneious package (v4.5.3, Drummond et al. 2008). Since the reference sequences were those used to generate the reads, stringent parameters were used for the alignment (minimum overlap 50, overlap identity 98%). The contigs were also aligned to related reference sequences using less stringent parameters (minimum overlap 40, overlap identity 60%) to test how closely related the reference needed to be to separate the contigs unambiguously.

The statistics package R (version 2.8.1, R Development Core Team 2008) was used to examine the distribution of coverages for each set of contigs. If the coverage distribution showed discrete peaks corresponding to the five different genomes, the contigs were grouped according to their coverage. Each group was then assembled into supercontigs using Geneious. No reference was used for the supercontig assembly—separating the contigs into groups corresponding to the different mitochondria should eliminate the ambiguous overlaps that broke up the initial assembly (except in the case of repeats), so that each group of contigs will assemble into a small number of supercontigs.

Another approach used to separate contigs from different genomes was to align the contigs to a set of reference sequences using the Exonerate sequence

alignment package (v2.2.0, Slater and Birney 2005). Exonerate was set to report the five best alignments for each contig, and to output a table showing, for each alignment, the names of the contig and the reference, the beginning and end of the aligned region in each, and the score and percent identity. As with the Geneious alignments, this was performed using the source genomes, and using genomes with differing degrees of relatedness. The resulting table was used to group the contigs according to which reference produced the highest scoring alignment, and Geneious was used to assemble each group into supercontigs.

SEQUENCING

Long-range PCR products were generated from a diverse set of templates in order to create a mixture of templates to sequence using an Illumina Genome Analyser. The organisms used were a human, a rat (bush rat, *Rattus fuscipes*), a bird (tawny frogmouth, *Podargus strigoides*), a frog (Hamilton's frog, *Leiopelma hamiltoni*), an insect (ground weta, *Hemiandrus pallitarsis*) and a mollusc (Northland olive, *Amalda northlandica*). PCR products of between approximately 1 kb and 8 kb were generated using primers specific to, and thermal cycling conditions optimised for, each DNA template (available from the authors). PCR products were processed by SAP/EXO digestion to remove unincorporated oligonucleotides and then quantified using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc.). Aliquots were taken in order to have an approximately even relative molarity for all DNA fragments in the final mix. All samples were then pooled and processed for sequencing in one lane using the genomic DNA sample preparation kit from Illumina (part # 1003806).

A 50-bp single read run was performed on an Illumina Genome Analyser GA2 (Illumina, Inc.) according to the manufacturer's instructions. Unfortunately, there was an instrument problem at cycle 33, which meant that only 32 nucleotides were usable. Due to the availability of the raw material for sequencing, the run was continued to completion. After sequencing, the resultant images were analysed with the proprietary Illumina pipeline (version 1.0) using default parameters. This resulted in approximately 238Mb of sequence, with 63% of the clusters passing the initial filtering step.

Additional assessment of an anomalous section of the *Amalda northlandica* mitochondrial genome was performed by traditional Sanger sequencing of a 300 bp PCR product spanning a region between *nad*5 and *cox*3. This PCR product was

generated from a total genomic DNA sample using specifically designed primers (Anor_nad5_f1618: 5'-ATGTCACAAGCAAACCAAAAGATCC-3'; Anor_cox3_r100: 5'-TTACTGTAATATACCCATATCCGTG-3') and using Taq DNA polymerase (Roche Applied Science) under the manufacturer's recommended conditions. The PCR product was processed by SAP/EXO digestion, and sequenced on an ABI3730 automated sequencer (Applied Biosystems) in both the forward and reverse directions using the specific designed PCR primers. The resulting sequences and electropherograms were visualised using Geneious.

DE NOVO ASSEMBLY

Due to high error rates observed for bases 1 to 5 and 33 onwards in the control lane of the Illumina flowcell, the reads were trimmed before assembly, removing the first five bases and the last 18 to leave 27-bp reads consisting of bases 6 to 32 of the original reads.

Perl scripts were used to run Velvet with a range of values for the hash length *k* and the coverage cutoff, and to extract the number of nodes, maximum contig length and N50 (median length-weighted contig length—half of all bases assembled are in contigs of this size or longer) values reported by Velvet.

Because of the large numbers of contigs produced, a Perl script (available from http://awcmee.massey.ac.nz/downloads.htm) was used to automate the procedure of aligning contigs against the reference sequences and separating them to produce a FASTA file of contigs aligning to each of the references, along with a file containing those contigs that fail to align to any of the references. The same script also converted the de Bruijn graph of contigs for each assembly produced by Velvet to DOT format, so that the graph could be visualised using GraphViz (Gansner and North 2000).

Identification of coding regions of the sequenced portions of the mitochondrial genomes was achieved through comparison to published complete mitochondrial genome sequences available through GenBank.

RESULTS

SIMULATIONS

35-bp reads were extracted for a human mitochondrial genome (GenBank accession number J01415; see Table 1 for a list of the mitochondrial sequences used in this study), the white-faced heron, the dark-spotted frog, the oriental mole cricket and the eastern mudsnail. These organisms were chosen as they represent a mixture similar to that used in our experimental run. To test the effect of having the genomes present at different concentrations, the reads were extracted in a ratio of 10:15:20:25:30, in several permutations. These simulated reads were then assembled using Velvet version 0.7.26 (Zerbino and Birney 2008) and Edena version 2.1.1 (Hernandez et al. 2008). The largest possible overlap (the largest hash length in Velvet) gave the highest N50 in all cases.

The two sets of contigs produced by Velvet and Edena were aligned to each of the five genomes in turn. Each contig mapped perfectly to one of the five genomes, indicating that there were no misassemblies and that all sequencing errors were eliminated by the high coverage.

Coverage distributions for both sets of contigs for a single permutation are shown in Figure 1A. All permutations that were tested gave similar results, with a single peak corresponding to each of the five genomes. This meant that, for simulated reads, the coverage values could easily be used to separate the contigs into five groups, one for each genome. In practice, however, it has been reported that for GA reads, coverage is not uniform, but is correlated to GC content, perhaps due to AT-rich fragments being more prone to denaturation than GC-rich fragments (Dohm et al. 2008; Hillier et al. 2008). This highlights the need for caution when using simulations to test new methods.

To separate the contigs produced from simulated reads without using coverage information, Exonerate (Slater and Birney 2005) was used to align each contig against a set of reference sequences related to the mitochondrial genomes used to generate the reads. For the human, rat, bird, snail and squid mitochondria listed above, the references used were another human mitochondrial genome (accession NC_001807), another Norway rat (NC_001665), the Australian pelican, the turrid snail and Bleeker's squid respectively. Because the relatedness between the reference and the original sequence was different for each genome (the same

species for the human and rat, different orders for the bird) and the degree of sequence conservation varies across the genome, only the relative values of the alignment scores for each contig could be used to separate the mixture of contigs into its component genomes. For each contig of our Edena assembly, the best alignment identified corresponded to the correct reference, except for two short contigs from the control region of the bird which failed to align to any of the reference sequences.

Once separated, each group of contigs was assembled to give one or more supercontigs for each genome. For the Velvet assembly of the permutation described above, the cricket and snail mitochondrial genomes each gave two contigs, which overlapped to form a single supercontig covering the whole of each genome. The human mitochondrial genome gave seven contigs, which formed a single supercontig covering the whole genome, although one of the overlaps was very short (seven bases). The bird and frog mitochondrial genomes both contain tandem repeat regions which could not be assembled from short reads. This would be the case regardless of whether they are sequenced separately or as part of a mixture (see Chaisson *et al.* 2004 and Kingsford *et al.* 2010 for analysis of the limitations of short reads for repeat resolution). However, the remainder of each genome was successfully assembled into two supercontigs. We obtained similar results for the other permutations we examined, and for the Edena assemblies—there were small differences in the numbers of contigs produced, but these did not affect the assembly into supercontigs.

To test whether more closely related mitochondrial genomes could be separated in the same way, the exercise was repeated using the same human, bird and snail mitochondrial genomes, together with a Norway rat (accession AJ428514) and reef squid. The relatively closely related human and rat mitochondrial genomes were each broken up into a larger number of contigs (twelve each), but these could still be separated by their different coverage levels, and each set then assembled into a single supercontig. Two short contigs (length 52 bp and 54 bp) had coverage equal to the sum of the expected coverages for the human and rat genomes, and aligned with 100% sequence identity to both the human and rat references. These represent regions of the 16S ribosomal RNA gene that are conserved between the two species, and were included in both sets of contigs.

The squid mitochondrial genome contained a duplicated region, which gave a 505 bp contig with twice the expected coverage. The double coverage made it

possible to identify the contig as a repeat, and to include it twice when assembling the contigs into a supercontig.

Our simulations thus confirmed that it is possible to assemble short reads from this type of mixture of mitochondrial genomes, and to separate the assembled contigs into the individual components of the mixture.

BIOLOGICAL DATA

For some of the organisms chosen, closely related reference genomes were available. We also chose some more difficult examples, for which the closest available reference was much more distant, for example in a different taxonomic order in the case of the bird.

Trimmed 27-bp GA reads were assembled using Velvet. As expected, the best results were obtained with the longest possible hash length (25, giving 536 contigs with an N50, or median length-weighted contig length, of 598). The coverage cutoff parameter of Velvet was used to eliminate short, low-coverage nodes (which are likely to be errors), giving considerably higher N50 values. It is likely that the six samples were present at different concentrations, so we expected that different values of the coverage cutoff would be optimal for each genome. The number of nodes, maximum contig length and N50 values reported by Velvet with coverage cutoff values up to 150 are shown in Figure 2. Assemblies with coverage cutoff set to 12, 26, 35, 45 and 58 were examined.

Probably because of the differences in GC content within a genome, coverage was not sufficiently uniform to separate the contigs belonging to the different genomes (see Figure 1B). Consequently, they were separated by aligning them to a set of reference genomes. The references used were the mitochondrial genomes of: a human (accession J01415); the spiny rat; the common swift; the common midwife toad; the cave cricket; and the eastern mudsnail. The degree of relatedness between the target and reference sequences was thus different in each case: for the human, target and reference were two members of the same species; for the rat, different species of the same genus; and for the bird, frog, cricket and mollusc, target and reference were in different families or even higher-order taxa.

Of a total of 964 contigs for the five assemblies examined, 762 were correctly grouped into species in this first step. However, because the single best alignment for each contig was used regardless of the relative scores of alignments to the other

references, 64 contigs were initially assigned to incorrect species. Where sequences are highly conserved (or highly divergent), contigs may align with similarly high (or low) scores to several references, thus it was expected that not all contigs would be assigned correctly by this method. A further 113 contigs failed to produce any alignments with scores above Exonerate's default threshold. However, all contigs belonging to the human and rat sequences were assigned correctly, presumably as a consequence of having closely related reference sequences for these organisms.

A second round of separation was carried out using the assembly graphs produced by Velvet. An example of an assembly graph, with each node coloured according to the reference to which it aligned, is shown in Figure 3. We used the graph to identify contigs that appeared to have been assigned to the wrong genome, and to ascertain the origin of those contigs that failed to align using Exonerate. These were checked against the GenBank (Benson et al. 2009) nucleotide database using the web-based BLASTn algorithm (Altschul et al. 1997). BLASTn found closer alignments for most of these contigs than those to our reference genomes, as we expected, since GenBank contains many shorter sequences in addition to the relatively small number of whole mitochondrial genomes known. Such comparisons are therefore very useful in aiding assembly.

Any contigs that were connected in the assembly graph to contigs which aligned to different references were checked against GenBank. Node 206 of the assembly in Figure 3, for example, aligned to the spiny rat, while the two neighbouring contigs aligned to the common swift, but when checked against GenBank, the best alignment found for node 206 was to *Gallirallus okinawae* (Okinawa rail) mitochondrial DNA, so it was reassigned to the pool of bird contigs. Another example is node 75, which was found to match fragments of mitochondrial 16S sequence from the frogs *Leiopelma archeyi* and *L. hochstetteri* in the GenBank database with 100% identity, despite aligning more closely to our bird reference than to our frog reference. No useable alignments were found for 17 unmatched nodes, all of which grouped in the graphs with contigs which aligned to the insect reference, and these were assigned to the insect pool on the basis of their position in the graph. This general problem will certainly decrease as more complete genomes become available, but it still requires care at present.

Once separated, the contigs aligning to each reference were imported into Geneious (Drummond et al. 2008). Each set of contigs was assembled into

supercontigs, and these supercontigs, along with any contigs not included in the supercontigs, were aligned against the reference.

The human sequence was a single long range PCR product from a human Melanesian sample, the remainder of this mitochondrial genome having been sequenced in a previous experiment. The best results were obtained with a coverage cutoff of 12, giving three overlapping contigs with a total length of 10,485 nucleotides spanning from *cox*1 to 12S rRNA as expected. Higher coverage cutoff values still gave the same three overlapping contigs, except that the longest contig (and hence the overall length) was slightly shorter. This human Q2 haplotype is reported separately in Corser et al. (2009), and has the GenBank accession number GQ214521.

The best assembly for the mollusc sequence was obtained with the higher coverage cutoff values. Coverage cutoffs of 45 and 58 both produced seven contigs that overlapped to form a single supercontig 15,361 bp in length, whose ends overlap by 7 bp. Although this overlap is short, it is within the *trn*H gene, and is part of a short (18 bp) overlap between two long-range PCR products. All other overlaps between contigs were 19 bp or longer. The supercontig appears therefore to constitute the entire mitochondrial genome of *Amalda northlandica*. Lower coverage cutoffs gave six contigs, covering the whole genome except for a gap of 11 nucleotides in the non-coding region between *trn*F and *cox*3. We discuss this genome in more detail below.

All coverage cutoff values gave similar results for the frog, with five contigs forming three supercontigs of 616 bp, 1,385 bp and 6,361 bp at coverage cutoffs of 45 and 58. This represents almost all of the frog template loaded (long-range PCR was only able to generate one fragment representing approximately half of the frog mitochondrial genome). At the lower coverage cutoff values six contigs were produced, but they still formed the same three supercontigs, although the longest was slightly shorter, at 6,350 bp.

The rat assembly was also largely unaffected by the coverage cutoff. However, there were two regions where polymorphisms were observed. These can be seen as criss-cross patterns in the graph in Figure 3—where the two sequences have diverged, they form a pair of parallel contigs both of which overlap with contigs on either side where the sequences are identical. These regions are in the 12S and 16S rRNA genes, and in *cox*1. The two sequences observed in each of these

regions were highly similar, and open reading frames were preserved. These might indicate the presence of nuclear DNA sequences of mitochondrial origin (numts, see Lopez et al. 1994, Richly and Leister 2004).

The contigs where the sequence was unambiguous were used in conjunction with further sequencing experiments to determine the complete mitochondrial genome sequence of *Rattus fuscipes*, extending the work of Robins et al. (2008). This sequence has the GenBank accession GU570664, and will be published separately, along with the mitochondrial genome sequences of several other *Rattus* species.

The bird sequences show a more complicated pattern again, as can be seen in Figure 3. It appears that, as well as containing the intended tawny frogmouth DNA, the sequencing reaction was contaminated with DNA from a common moorhen, a sandhill crane, and a red-fronted coot. Unfortunately no reference sequences are available at present that can be used to distinguish these birds across the whole mitochondrial genome, and the problem appears to have arisen through tissue contamination (see later).

Aligning the contigs to the common swift reference genome showed a single sequence stretching from the middle of the 12S rRNA gene to *trn*M, with a small gap in 12S rRNA. From *nad*2 to *atp*6 there were two parallel sequences, and from the end of *cox*3 to the middle of *cytB* there were three. Comparing contigs to the GenBank nucleotide database using MegaBLAST showed that the sequence covering 12S rRNA to *trn*M matched tawny frogmouth sequence fragments: one partial 12S rRNA sequence, and one sequence covering *trn*L, *nad*1, *trn*I and *trn*Q. Of the two parallel sequences from *nad*2 to *atp*6, one gave an exact match to existing partial *cox*1 and *atp*8 sequences for the Southern American common moorhen *Gallinula chloropus galeata*, and the other gave an exact match to an existing partial *cox*1 sequence for the red-fronted coot *Fulica rufifrons*. At the *cytB* locus, where there were three parallel sequences, one was found to match tawny frogmouth, the second matched common moorhen, and the third matched the sandhill crane *Grus canadensis*. One of the three sequences also matched an existing tawny frogmouth fragment covering part of *nad*1 and *trn*H, *trn*S and *trn*L, and another matched an existing sandhill crane fragment covering part of *cox*3, *nad*3 and *trn*G.

Many of the bird contigs had relatively low coverage values (since the presence of contaminants meant that the overall sequence length was much longer than expected), so that when assembly was carried out with a higher coverage cutoff they were eliminated, or two parallel contigs were merged to form a single contig.

DNA was extracted from a sandhill crane sample in our laboratory alongside the tawny frogmouth sample. However, neither common moorhen nor red-fronted coot have ever been studied in this laboratory (nor are the species present in this country), so it is likely that either the tawny frogmouth or the sandhill crane tissue sample was contaminated with DNA from these two species before our laboratory received them. Using the same scalpel for dissecting different birds is a possible explanation. This highlights the need for good laboratory practice—the high dynamic range of these DNA sequencing techniques means that minute traces of DNA will be amplified and sequenced.

As with the bird, the insect sequences show a rather convoluted assembly, with regions where two or three sequences align in parallel to the same region of the reference. The insect sequences, however, appear to be nuclear DNA sequences of mitochondrial origin, as we were unable to identify open reading frames corresponding to the genes to which the sequences align. A possible solution to this problem would be the isolation of whole mitochondria, followed by DNA extraction from these mitochondria. This would exclude nuclear DNA, thereby eliminating the contribution of any nuclear copies of mitochondrial genes to the resulting sequence reads.

AMALDA NORTHLANDICA MITOCHONDRIAL GENOME

The mitochondrial genome of *Amalda northlandica* is 15,354 bp in length and contains 13 protein coding genes, two ribosomal RNA genes and 22 tRNA genes (Figure 4), and has the GenBank accession number GU196685. All protein coding genes begin with the standard ATG start codon with the exception of *nad*6, which starts with an ATA codon. All the protein coding genes terminate with standard TAA or TAG codons. The gene composition and order is consistent with neogastropod complete mitochondrial sequences currently available in GenBank (*Ilyanassa obsoleta*, *Lophiotoma cerithiformis*, *Conus textile*, *Thais clavigera*, *Rapana venosa*, *Terebra dimidiata*, *Cancellaria cancellata*, *Fusiturris similis*, *Conus borgesi*, *Cymbium olla*, *Nassarius reticulatus*, *Bolinus brandaris*). In addition, a novel structural element (outlined below) was identified during

assembly of the *A. northlandica* mitochondrial genome sequence. This structure may represent a reduced mitochondrial control region (which has not yet been identified in neogastropod molluscs).

AN UNUSUAL CONTROL REGION?

A very unusual feature of the assembly was that under certain coverage cutoff regimes a fragment of the mitochondrial sequence was omitted. This 11 bp fragment was found to be in an intergenic region between *trn*F and *cox*3 and it is surprising that such an apparently short region should disrupt assembly. In order to identify possible causes of incomplete assembly the non-coding intergenic spacers were analysed for secondary structure formation. This could also elucidate structural features, such as the origin of replication and control region, which have not yet been identified in neogastropod molluscs. The highly variable 3′ and 5′ domains of the rRNA genes mean that the precise boundaries of 12S rRNA and 16S rRNA are not yet known. Due to this uncertainty the regions flanking the rRNA genes were not considered.

The longest intergenic spacer in *Amalda northlandica* is located between the genes *trn*F and *cox*3. It is 56 bp in length and contains two predicted secondary structural elements, a strong stem-loop element and a second small stem-loop element (Figure 5A). Of the remaining intergenic sequences in the *Amalda* mitochondrial genome, only seven are longer than 10 bp. All of these seven exhibit some secondary structure (as predicted by the program MFold, Zuker et al. 1999). Including sequence of *trn*F showed that the initial stem in the intergenic spacer overlaps with the 3′ end of the acceptor stem of the tRNA by 5 bp. This initial stem of 14 bases is by far the strongest secondary structure in the intergenic regions (-20.03 kcal/mol). The presence of the short second possible stem-loop reduces the stability of the combined structure to -19.85 kcal/mol.

The incomplete assembly observed for lower coverage cutoff regimes (see earlier) was identified to be the result of a loss of 11 nucleotides representing the complete loop of the structure shown in Figure 5A. This appeared to suggest that palindromic sequence of sufficient size may cause the loss of sequence during assembly under specific cutoff regimes. However, further analysis of the sequence coverage of this region revealed that identical (not complementary), but reversed sequence existed in both the forward and reverse directions of the loop region of this structure (Figure 6). Although we are able to confirm the sequence of the loop

region, we are unable to show in which of two possible orientations this sequence exists naturally in the *Amalda* mitochondrial genome. Re-examination of this region with Sanger sequencing confirmed the presence of ambiguous base calls within the expected 11 bp section. The Sanger sequence also confirms that this anomalous region is not the result of an artefact introduced in the Illumina sequencing or short-read assembly (Figure 6C).

It is not yet possible to confirm whether this structure represents either the control region or an origin of replication. There are no clear homologies with known structures or known conserved sequence blocks associated with either structure. However, this region can be identified in six published neogastropods (*Ilyanassa obsoleta*, *Thais clavigera*, *Rapana venosa*, *Fusiturris similis*, *Bolinus brandaris* and *Nassarius reticulatus*), where the size is nearly identical (56-58 bp). The predicted secondary structures are very similar (data not shown) with well conserved sequences for the stem structure (see Figure 5), but the nucleotide sequences for the remainder of the region are quite divergent in these species. The mitochondrial genomes of *Cancellaria cancellata*, *Cymbium olla*, *Lophiotoma cerithiformis, Conus textile* and *Conus borgesi* are all longer, have more complex predicted secondary structures and, with the exception of *Cancellaria*, have no significant sequence homology to the previously mentioned neogastropods. The remaining published neogastropod (*Terebra dimidiata*) has a considerably larger intergenic region in this position that exhibits no clear homology with the other known neogastropod mitochondrial genome sequences.

In addition, the positions of other structure bearing intergenic regions are not conserved across the known neogastropod mitochondrial genomes. For example, an intergenic region of 25 bp is observed in *Amalda* between *nad*1 and *trn*P; while most of the known neogastropod sequences have some intergenic sequence at this position only five have a region that is larger than 10 bp. Furthermore, there is no unambiguously homologous sequence in these variable intergenic regions. It remains uncertain whether homologous structures exist at different positions in the other mitochondrial genomes.

DISCUSSION

These results show that, given an appropriate reference sequence for each genome under consideration, it is possible to assemble short reads from a mixture of mitochondrial genomes and deconvolute the resulting contigs without the need to index the reads. The reference sequence for each genome must be considerably closer to that genome than to any of the others, but it is not necessary for the references to separate the sequences perfectly, as the assembly graph can be used to identify spurious alignments, as well as to reallocate contigs that fail to align to any of the references.

In principle, the same approach could be applied to other mixtures of sequences, for example chloroplast genomes. We have successfully assembled chloroplast genomes from short read data (data not shown), although not yet from a mixture.

The main difficulties encountered in assembling the genomes in this study were not due to problems in separating the contigs, but to problems with sample preparation, namely the presence of numts and contamination. These same issues would have arisen if the six genomes had been sequenced separately. It is clear that it is important to have high quality DNA samples for *de novo* assembly. Any contamination can lead to ambiguities which make it difficult to distinguish between the sample and the contamination. This issue is significantly compounded if the contamination is closely related to the target sequence, relative to the reference sequence used (e.g. two birds), with varying degrees of sequence incompleteness or incorrect contigs generated depending on the level of relatedness. However, contamination will normally only affect assembly of the most closely related sequence, leaving the other samples unaffected. In the absence of contamination and numts, we would expect fewer contigs to be produced, making the process of deconvolution considerably simpler.

In generating the complete sequence of the mollusc *Amalda northlandica* we have characterised a novel structural element in a mitochondrial genome. The identification of apparently identical DNA sequence in both the heavy and light strands of this structure leads to two possible explanations (see Figure 5B):

1. that separate mitochondrial genome molecules exist in an individual, differing only in alternative orientations of the sequence of this loop; or

2. that the sequence on both strands of the DNA molecule is identical in this loop and therefore non-complementary in double-stranded DNA.

It is difficult to envisage a functional explanation for the first hypothesis. However, extrapolating from the second hypothesis, it could be suggested that this non-complementary sequence enforces the formation of a functionally important structural element in double-stranded DNA (Figure 5C). One difficulty with this hypothesis is how such a non-complementary region would be replicated. RNA mediation is a possible solution, and could be involved in an initiation process. Furthermore, given that the identical loop sequences are in opposite directions on each DNA strand, this might impart a directionality to each strand (e.g. for replication). Similar stem structures have been proposed for bidirectional transcriptional promoters in vertebrate mitochondrial genomes (L'Abbé et al. 1991; Ray and Densmore 2002), but the suggestion of non-complementary DNA in the double stranded mitochondrial genome is, as far as we are aware, unprecedented. Such an arrangement could be a result of the contraction of the mitochondrial genome in neogastropod molluscs, and the structure we have identified may represent a highly reduced control region. It is extremely unlikely that traditional Sanger sequencing is capable of characterising this novel sequence feature, although it might be detectable as a region of poor quality sequence. Indeed, several reported neogastropod mitochondrial genomes share sequence and structural homology with the stem structure shown here for *Amalda*, but there is very little sequence homology seen for the loop. Furthermore, the sequence of the mitochondrial genome of *Ilyanassa obsoleta* is reported with ambiguous bases in the region homologous to the *Amalda* loop, alluding to the presence of ambiguous sequence that we predict would be observed in Sanger sequence of this region. It is probable that the case reported here is not limited to *Amalda*. A detailed characterisation of the structure and evolutionary significance of the genomic region that we have identified here will be reported elsewhere.

The unusual arrangement of sequence in this structure was detectable in short-read sequencing as it led to an apparently structure-mediated loss of sequence during contig generation. The extent to which this prevails is unknown, as such an arrangement has never been described. However, clearly the development of new DNA sequencing technologies might allow the discovery of features that were intractable with earlier techniques.

The utility of complete mitochondrial genome sequences to the analysis of molluscan phylogenetic relationships is reinforced with the addition of the *Amalda northlandica* sequence. Neogastropoda represent a lineage that appears to have undergone a rapid diversification. Standard analysis of nucleotide sequence is often insufficient to resolve deep relationships in such cases (e.g. birds, Pratt et al. 2009). It is thought that structural organization of mitochondrial genomes ("rare genomic changes") could be used to resolve uncertainties in deep relationships in molluscs (Boore, 2006). As the gene content and order of known neogastropod mitochondrial genomes is identical, positional data for genes will not be informative. However positional information for intergenic spacer regions can provide important additional data. When the *Amalda* sequence is compared to the twelve known neogastropod sequences, a tantalising picture of lineage-specific arrangements of structure-bearing intergenic spacers emerges. However, very little can be concluded from such a small sample of molluscs. Fortunately, as methods are developed to enable the deconvolution of mixed samples from second-generation sequencing runs, large numbers of mitochondrial genomes or other short genomic regions can now be quickly and cost-effectively generated. Through sufficient sampling of maximally informative taxa, inference of phylogenetic relationships of molluscan lineages will then be robust and free of the bias associated with insufficient taxon sampling and inadequate sequence coverage to achieve resolution.

The mixture strategy that we have developed can readily be combined with an indexing approach. For example, if we wish to sequence mitochondrial genomes from, say, twelve birds, twelve molluscs, twelve insects and twelve human individuals, rather than using 48 index tags, we could use twelve, each with a mixture consisting of one bird, one mollusc, one insect and one human. A single set of four reference sequences could then be used to separate all twelve mixtures.

It should be noted that the approach developed here is very general in that it can be applied to a wide range of mixtures of DNA sequences. One that we have simulated is a mixture with a chloroplast and several mitochondria (data not shown), but in principle any mixture could be used, provided that for each sample we have a reference sufficiently close to separate that sample from the other components of the mixture. However, whatever mixture is tried, we would strongly advocate that the simulation approach be used to test that the software can successfully separate the mixture before committing to the cost of an actual run.

FUNDING

This work was supported by the Allan Wilson Centre for Molecular Ecology and Evolution.

ACKNOWLEDGEMENTS

The authors wish to thank Trish McLenachan, Chris Corser, Gillian Gibb, Judith Robins, Renae Pratt and Jan Binnie for sample preparation; and Lorraine Berry and Maurice Collins of the Allan Wilson Centre Genome Service for sequencing. We would also like to thank Bruce Marshall of the Museum of New Zealand Te Papa Tongarewa for formally identifying the *Amalda northlandica* specimen and lodging it in the national collection (voucher number NMNZ M.289187).

REFERENCES

Akasaki T, Nikaido M, Tsuchiya K, Segawa S, Hasegawa M, Okada N. 2006. Extensive mitochondrial gene arrangements in coleoid Cephalopoda and their phylogenetic implications. Mol Phylogenet Evol. 38:648-658.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Anderson S et al. 1981. Sequence and organization of the human mitochondrial genome. Nature. 290:457-465.

Bandyopadhyay PK, Stevenson BJ, Cady MT, Olivera BM, Wolstenholme DR. 2006. Complete mitochondrial DNA sequence of a Conoidean gastropod, *Lophiotoma* (*Xenuroturris*) *cerithiformis*: gene order and gastropod phylogeny. Toxicon. 48:29-43.

Bandyopadhyay PK, Stevenson BJ, Ownby J-P, Cady MT, Watkins M, Olivera BM. 2008. The mitochondrial genome of *Conus textile*, *coxI-coxII* intergenic sequences and Conoidean evolution. Mol Phylogenet Evol. 46:215-223.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2009. GenBank. Nucleic Acids Res. 37:D26-31.

Bentley DR. 2006. Whole-genome re-sequencing. Curr Opin Genet Dev. 16:545-552.

Boore JL. 2006. The use of genome-level characters for phylogenetic reconstruction. Trends Ecol Evol. 21:439-446.

Cai W-W, Chen R, Gibbs RA, Bradley A. 2001. A Clone-Array Pooled Shotgun Strategy for Sequencing Large Genomes. Genome Res. 11:1619-23.

Chaisson M, Pevzner P, Tang H. 2004. Fragment assembly with short reads. Bioinformatics. 20:2067-2074.

Corser CA, McLenachan PA, Pierson MJ, Harrison G, Penny D. 2009. The Q2 mitochondrial haplotype in Oceania. J Polynesian Soc. (in prep.).

Cunha R, Grande C, Zardoya R. 2009. Neogastropod phylogenetic relationships based on entire mitochondrial genomes. BMC Evol Biol. 9:210.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36:e105.

Drummond AJ et al. 2008. *Geneious v4.0*. Available from http://www.geneious.com/.

Erlich Y, Chang K, Gordon A, Ronen R, Navon O, Rooks M, Hannon GJ. 2009. DNA Sudoku-harnessing high-throughput sequencing for multiplexed specimen analysis. Genome Res. 19:1243-53.

Fenn JD, Song H, Cameron SL, Whiting MF. 2008. A preliminary mitochondrial genome phylogeny of Orthoptera (Insecta) and approaches to maximizing phylogenetic signal found within mitochondrial genome data. Mol Phylogenet Evol. 49:59-68.

Frank, DN. 2009. BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. BMC Bioinformatics. 10:362.

Fullwood MJ, Wei C-L, Liu ET, Ruan Y. 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res. 19:521-532.

Gansner ER, North SC. 2000. An open graph visualization system and its applications to software engineering. Software Pract Exper. 30:1203-1233.

Gibb GC, Kardailsky O, Kimball RT, Braun EL, Penny D. 2007. Mitochondrial genomes and avian phylogeny: complex characters and resolvability without explosive radiations. Mol Biol Evol. 24:269-280.

Hernandez D, Francois P, Farinelli L, Osterås M, Schrenzel J. 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res. 18:802-809.

Hillier LW et al. 2008. Whole-genome sequencing and variant discovery in *C. elegans*. Nat Methods. 5:183-188.

Hills S, Trewick S, Morgan-Richards M. 2009. Phylogenetic information content of mitochondrial genes in Volutidae (Gastropoda). (in prep.).

Ingman M, Kaessmann H, Pääbo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. Nature. 408:708-713.

Kim I, Cha SY, Yoon MH, Hwang JS, Lee SM, Sohn HD, Jin BR. 2005. The complete nucleotide sequence and gene organization of the mitochondrial genome of the oriental mole cricket, *Gryllotalpa orientalis* (Orthoptera: Gryllotalpidae). Gene. 353:155-168.

Kingsford C, Schatz MC, Pop M. 2010. Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics. 11:21.

L'Abbé D, Duhaime JF, Lang BF, Morais R. 1991. The transcription of DNA in chicken mitochondria initiates from one major bidirectional promoter. J Biol Chem. 266:10844-10850.

Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. 1994. *Numt*, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J Mol Evol. 39:174-190.

Margulies M et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 437:376-380.

Ng P, Tan JJS, Ooi HS, Lee YL, Chiu KP, Fullwood MJ, Srinivasan KG, Perbost C, Du L, Sung W-K, Wei C-L, Ruan Y. 2006. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. Nucl Acids Res. 34:e84.

Nichols R. 2001. Gene trees and species trees are not the same. Trends Ecol Evol. 16:358-364.

Nilsson MA, Gullberg A, Spotorno AE, Arnason U, Janke A. 2003. Radiation of extant marsupials after the K/T boundary: evidence from complete mitochondrial genomes. J Mol Evol. 57:S3-12.

Pratt RC, Gibb GC, Morgan-Richards M, Phillips MJ, Hendy MD, Penny D. 2009. Toward resolving deep Neoaves phylogeny: data, signal enhancement, and priors. Mol Biol Evol. 26:313-326.

R Development Core Team. 2009. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ray DA, Densmore L. 2002. The crocodilian mitochondrial control region: general structure, conserved sequences, and evolutionary implications. J Exp Zool. 294:334-345.

Richly E, Leister D. 2004. NUMTs in sequenced eukaryotic genomes. Mol Biol Evol. 21:1081-1084.

Robins JH, McLenachan PA, Phillips MJ, Craig L, Ross HA, Matisoo-Smith E. 2008. Dating of divergences within the *Rattus* genus phylogeny using whole mitochondrial genomes. Mol Phylogenet Evol. 49:460-466.

San Mauro D, García-París M, Zardoya R. 2004. Phylogenetic relationships of discoglossid frogs (Amphibia:Anura:Discoglossidae) based on complete mitochondrial genomes and nuclear genes. Gene. 343:357-366.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 74:5463-5467.

Sasuga J, Yokobori S, Kaifu M, Ueda T, Nishikawa K, Watanabe K. 1999. Gene contents and organization of a mitochondrial DNA segment of the squid *Loligo bleekeri*. J Mol Evol. 48:692-702.

Simison WB, Lindberg DR, Boore JL. 2006. Rolling circle amplification of metazoan mitochondrial genomes. Mol Phylogenet Evol. 39:562-567.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 6:31.

Sumida M, Kanamori Y, Kaneda H, Kato Y, Nishioka M, Hasegawa M, Yonekawa H. 2001. Complete nucleotide sequence and gene rearrangement of the mitochondrial genome of the Japanese pond frog *Rana nigromaculata*. Genes Genet Syst. 76:311-325.

Tomita K, Ueda T, Watanabe K. 1998. 7-Methylguanosine at the anticodon wobble position of squid mitochondrial $tRNA^{Ser}GCU$: molecular basis for assignment of AGA/AGG codons as serine in invertebrate mitochondria. Biochim Biophys Acta. 1399:78-82.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821-829.

Zuker M et al. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. NATO ASI SERIES 3 HIGH TECHNOLOGY. 70:11-44.

FIGURE LEGENDS

Figure 1: Coverage distributions.

A. Coverage, weighted by contig length, for contigs assembled from simulated reads by Velvet with k = 31. The sequences used in this simulation were human (25%), bird (30%), frog (10%), cricket (20%) and snail (15%). It is clear from these distributions that the contigs from each genome have tightly clustered coverage values, with the coverage for each genome directly proportional to the percentage of reads from that genome.

B. Coverage, weighted by contig length, for contigs assembled from biological data by Velvet with k = 25. Coverage for each genome is clearly not sufficiently uniform to be useful as a means of separating contigs.

Coverages are given as k-mer coverage (see Zerbino and Birney 2008).

Figure 2: Assembly statistics for biological data.

Median length-weighted contig length (N50, solid line), maximum contig length (dotted line), and number of nodes (dashed line) plotted against coverage cutoff for Velvet assemblies with hash length k = 25. Contig lengths are in k-mers (length in base-pairs can be obtained by adding k - 1). Increasing the coverage cutoff eliminates low-coverage nodes, removing some branching in the graph and allowing some of the higher-coverage nodes to merge. The distinct steps in the N50 plot may reflect different coverages for the different DNA fragments sequenced. The longest contig is stable, with a length of 8231 for all coverage cutoffs up to 129, except that for coverage cutoffs between 45 and 64, 10 nucleotides are added to one end of the contig to give a length of 8241.

Figure 3: Assembly graph.

The assembly graph for sequences assembled by Velvet with k = 25 and *cov_cutoff* = 26. Nodes are coloured according to the reference sequence to which the corresponding contigs align: green for human, purple for rat, blue for bird, red for frog, yellow for mollusc and orange for insect. Grey nodes failed to align to any of the references, and white nodes are shorter than 2k - 1 (Velvet does not output contigs for these nodes). The area of each node is proportional to the length of the

sequence it represents, and the width of an edge between two nodes is proportional to the number of reads that connect those nodes. The human, mollusc and frog sequences are assembled into relatively small clusters of long contigs, while the insect, bird and rat show more complex chains of shorter contigs. The reasons for these patterns are discussed in the text.

Figure 4: The *Amalda northlandica* complete mitochondrial genome.

Arrowheads depict the direction of transcription. Genes with offset annotations (*cox*1, *trn*C, *trn*Q, and *nad*4) overlap with genes preceding them. Binding sites for the primers used to generate the long-range PCR products are indicated in green.

Figure 5: The *trnF-cox3* intergenic region of the *Amalda northlandica* mitochondrial genome.

A. The position and inferred structure of stem-loop elements in this region; the positions of the *trn*F gene and the initial bases of the *cox*3 gene are also indicated. The smaller predicted stem-loop reduces the overall stability of both structures.

B. Two hypotheses could explain the sequence data:

   I. there is a mixture of two mitochondrial genome copies that differ in the orientation of the loop sequence, or

   II. there is a single genome that contains a non-complementary region, which could exist in either of two possible orientations.

C. Hypothesis II suggests the formation of a double stem structure in double-stranded DNA.

Figure 6: Gbrowse visualisations of short reads from the *Amalda northlandica* mitochondrial control region showing reads present in either orientation, and electropherograms confirming the sequence.

Parts A and B show a representative sample of 27-bp sequence reads across each orientation. The loop sequence between the stems is shown in magenta in the 'Annotation' track. Short reads are shown in the forward and reverse strands (blue and green respectively). The reads that give directionality to the loop sequences (i.e.

that cross the boundary of either the 5′ stem or 3′ stem into identifiable sequence) are shown in the forward (yellow) and reverse (pink) strands.

Part C shows Sanger sequence confirmation of ambiguous nucleotide sites at the positions predicted by the short-read mapping in A and B above. Electropherograms show the base calls for the nucleotide sequence reads in both the forward and reverse directions. Sequence quality scores are indicated for each site as a histogram in parallel with the electropherograms. Scores range from 55 for high quality base calls to 12 for the lowest quality call of the ambiguous nucleotide positions. The sequence shown includes only the 100 bases that align with the short-read assemblies shown in A and B, and comes from a sequence fragment of length 300 bp.

Table 1: Mitochondrial sequences referred to in this study.

Accession no.	Species	Common name	Reference
J01415[a,b]	*Homo sapiens*	human	Anderson et al. (1981)
NC_001807[b]	*Homo sapiens*	human	Ingman et al. (2000)
AJ428514[a]	*Rattus norvegicus*	Norway rat	Nilsson et al. (2003)
NC_001665[b]	*Rattus norvegicus*	Norway rat	
EU273708[b]	*Rattus praetor*	spiny rat	Robins et al. (2008)
NC_008551[a]	*Ardea novaehollandiae*	white-faced heron	Gibb et al. (2007)
DQ780883[b]	*Pelecanus conspicillatus*	Australian pelican	Gibb et al. (2007)
NC_008540[b]	*Apus apus*	common swift	
AB043889[a]	*Rana nigromaculata*	dark-spotted frog	Sumida et al. (2001)
NC_006688[b]	*Alytes obstetricians*	common midwife toad	San Mauro et al. (2004)
AY660929[a]	*Gryllotalpa orientalis*	oriental mole cricket	Kim et al. (2005)
EU938374[b]	*Troglophilus neglectus*	cave cricket	Fenn et al. (2008)
NC_007894[a]	*Sepioteuthis lessoniana*	reef squid	Akasaki et al. (2006)
AB029616[b]	*Loligo bleekeri*	Bleeker's squid	Tomita et al. (1998), Sasuga et al. (1999)
DQ238598[a,b]	*Ilyanassa obsolete*	eastern mudsnail	Simison et al. (2006)
NC_008098[b]	*Lophiotoma cerithiformis*	turrid snail	Bandyopadhyay et al. (2006)
NC_008797	*Conus textile*	cloth-of gold cone	Bandyopadhyay et al. (2008)
NC_010090	*Thais clavigera*	rock shell	
NC_011193	*Rapana venosa*	veined rapa whelk	
NC_013239	*Terebra dimidiate*	dimidiate auger shell	Cunha et al. (2009)
NC_013241	*Cancellaria cancellata*	cancellate nutmeg	Cunha et al. (2009)
NC_013242	*Fusiturris similis*		Cunha et al. (2009)
NC_013243	*Conus borgesi*		Cunha et al. (2009)
NC_013245	*Cymbium olla*	pata-del-burro	Cunha et al. (2009)
NC_013248	*Nassarius reticulates*	reticulate nassa	Cunha et al. (2009)
NC_013250	*Bolinus brandaris*	purple dye murex	Cunha et al. (2009)
GU196685	*Amalda northlandica*	Northland olive	This study

[a]Genomes from which simulated reads were extracted.

[b]Genomes used as references.

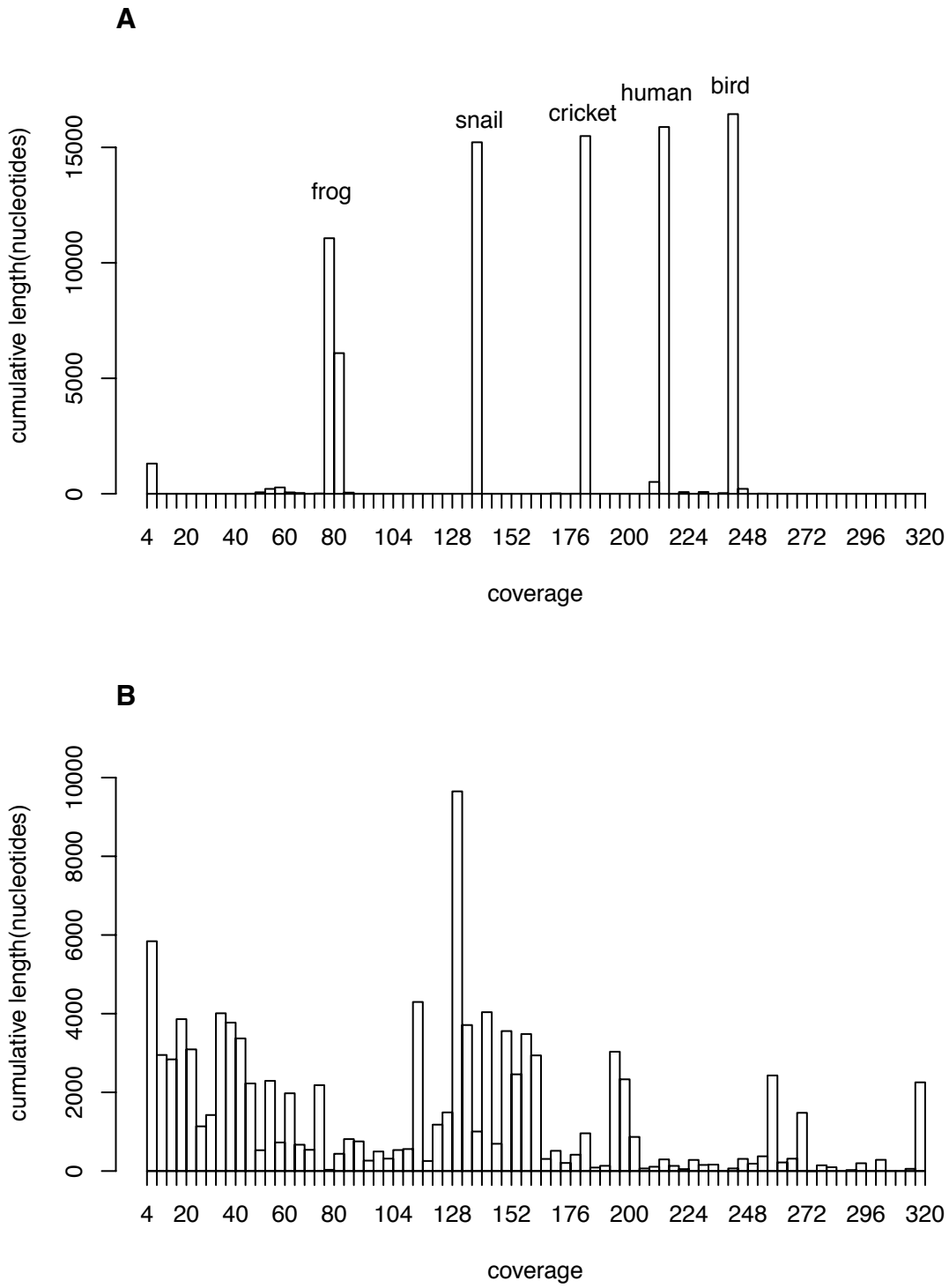

Figure 1: Coverage distributions.

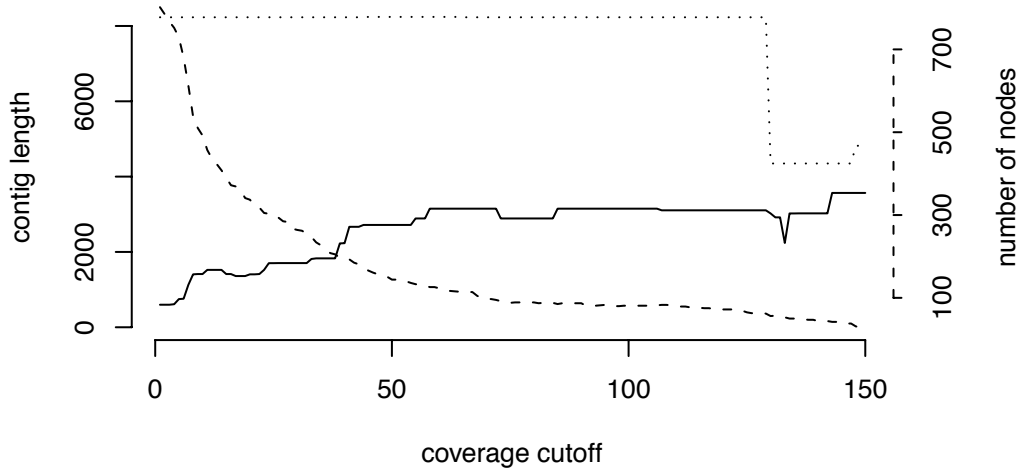

Figure 2: Assembly statistics for biological data.

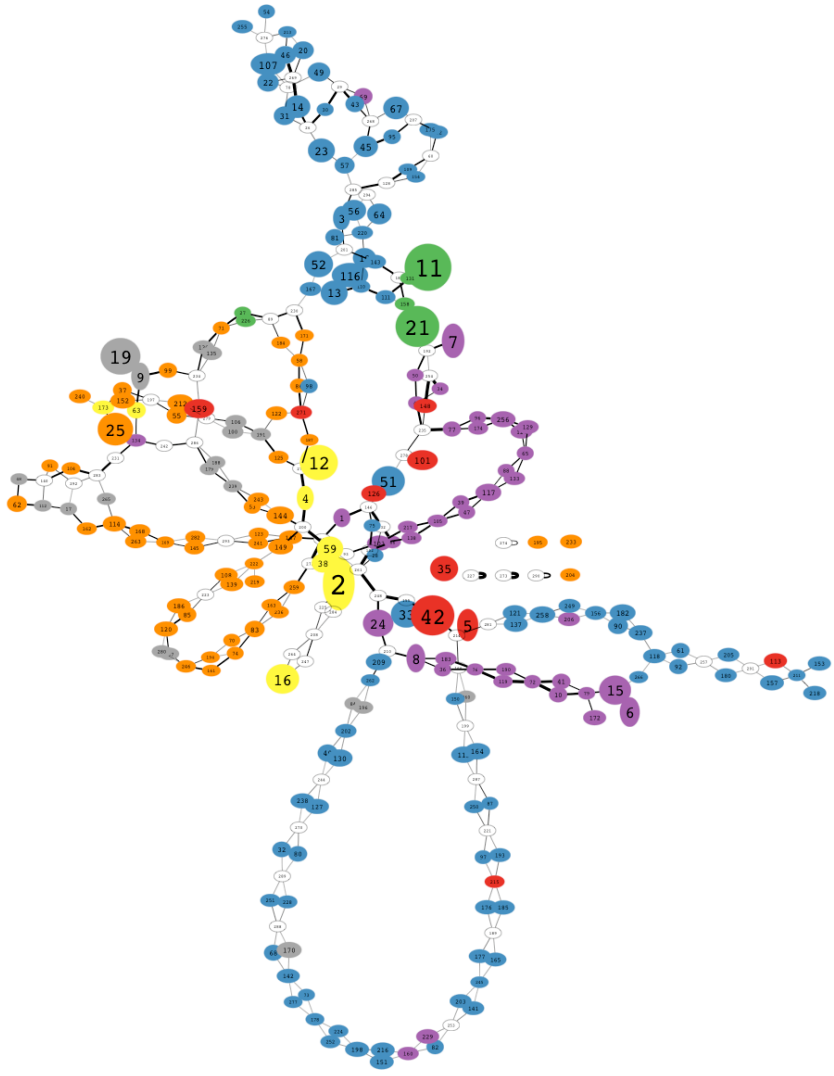

Figure 3: Assembly graph.

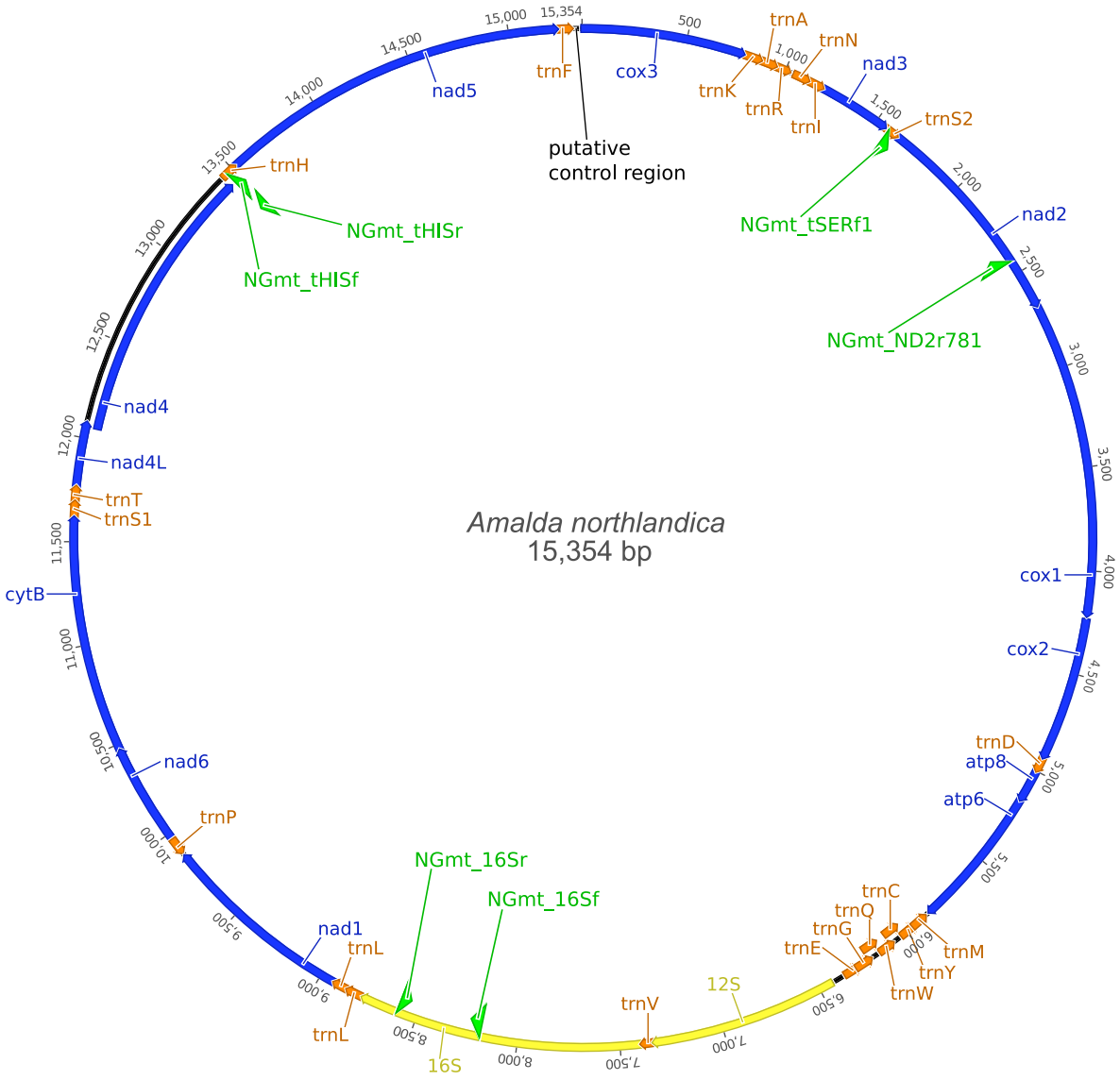

Figure 4: The *Amalda northlandica* complete mitochondrial genome.

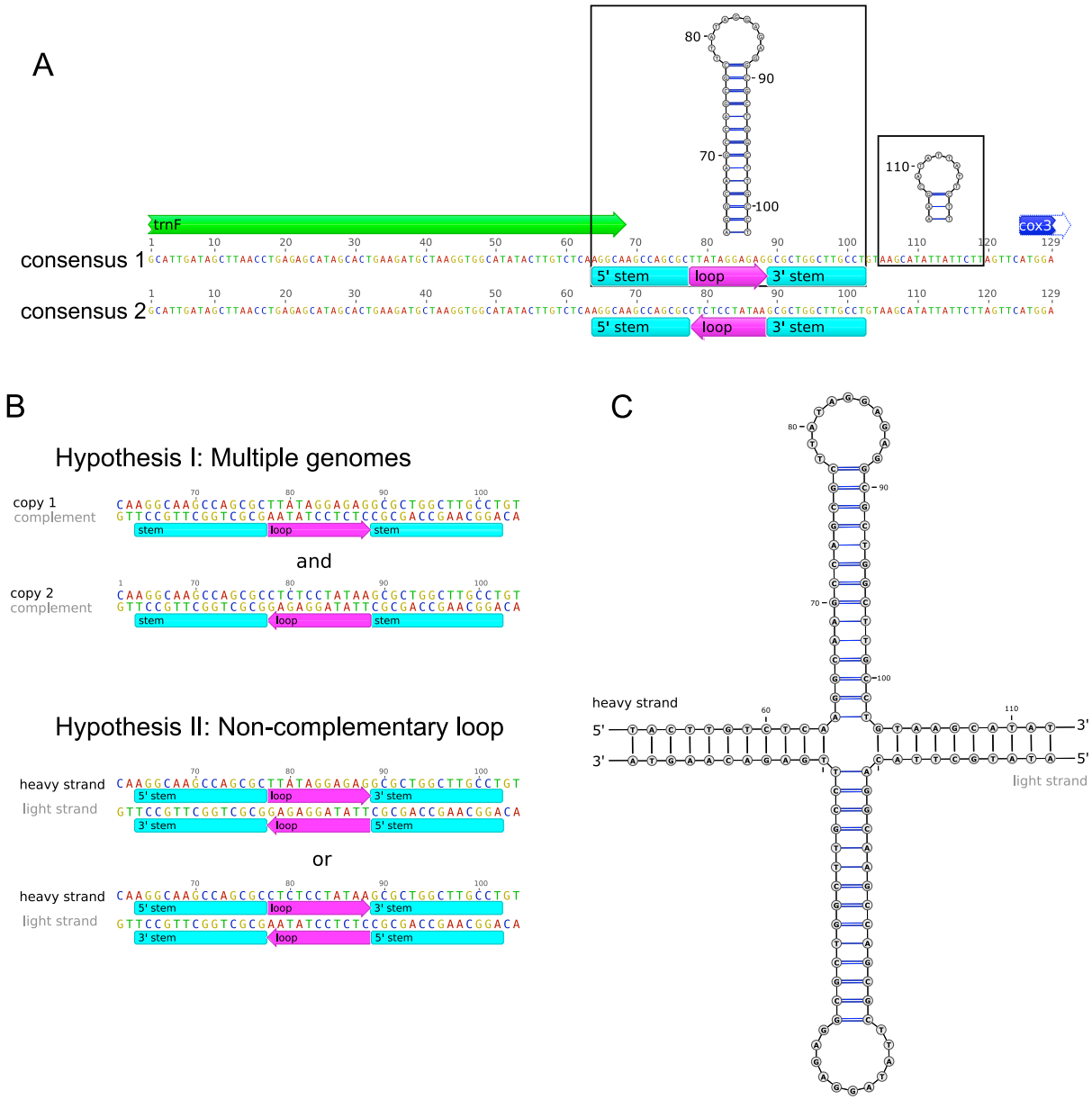

Figure 5: The *trnF-cox3* intergenic region of the *Amalda northlandica* mitochondrial genome.

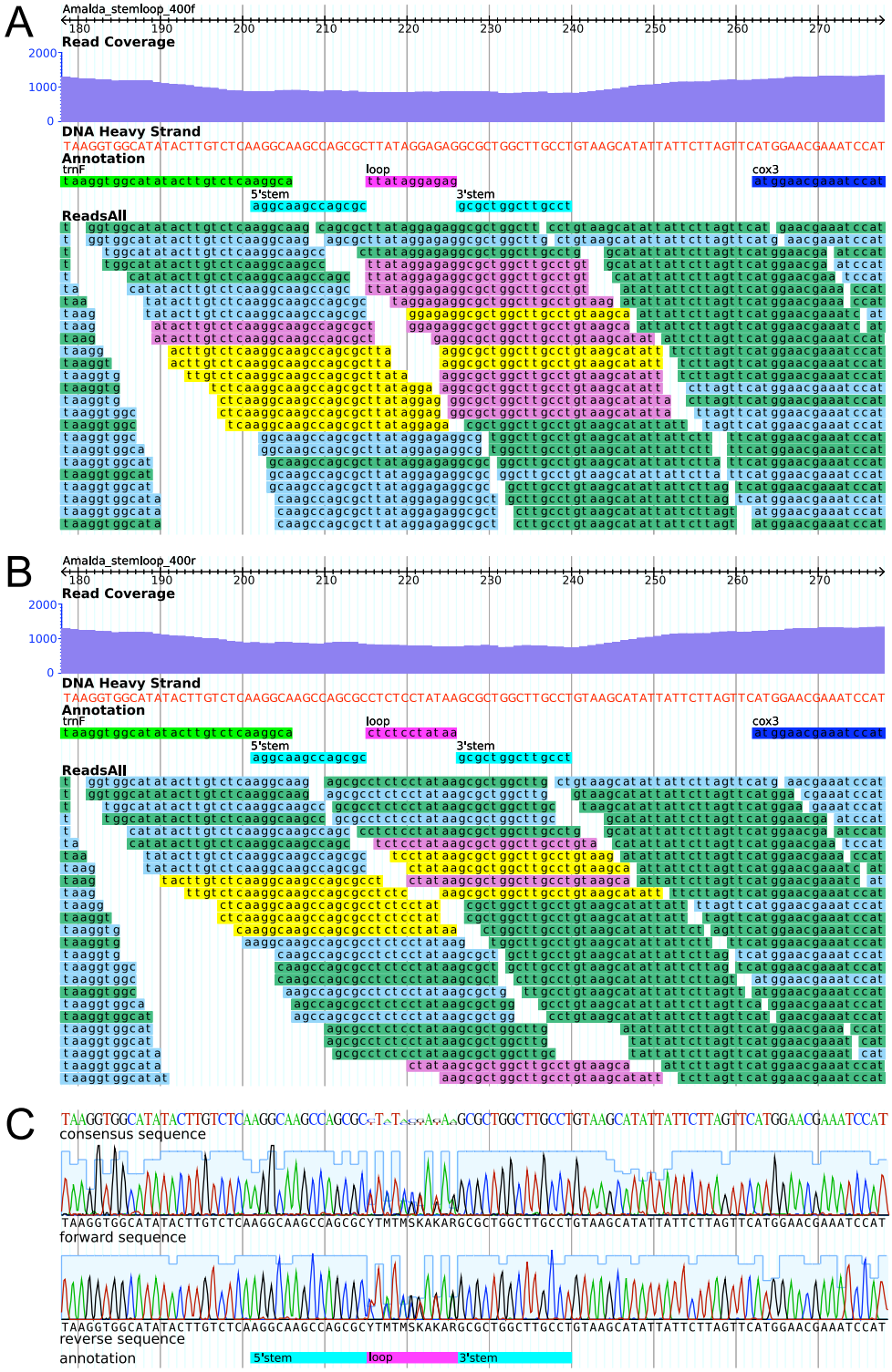

Figure 6: Gbrowse visualisations of short reads from the *Amalda northlandica* mitochondrial control region showing reads present in either orientation, and electropherograms confirming the sequence.