Exploring Biological Sequence Space

Selected problems in sequence analysis and phylogenetics

A thesis presented in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

in Computational Biology

at Massey University

Bennet James McComish

2012

Abstract

As the volume and complexity of available sequence data continues to grow at an exponential rate, the need for new sequence analysis techniques becomes more urgent, as does the need to test and to extend the existing techniques. These include, among others, techniques for assembling raw sequence data into usable genomic sequences; for using these sequences to investigate the evolutionary history of genes and species; and for examining the mechanisms by which sequences change over evolutionary time scales. This thesis comprises three projects within the field of sequence analysis.

- It is shown that organelle genome DNA sequences can be assembled *de novo* using short Illumina reads from a mixture of samples, and deconvoluted bioinformatically, without the added cost of indexing the individual samples. In the course of this work, a novel sequence element is described, that probably could not have been detected with traditional sequencing techniques.

- The problem of multiple optima of likelihood on phylogenetic trees is examined using biological data. While the prevalence of multiple optima varies widely with real data, trees with multiple optima occur less often among the best trees. Overall, the results provide reassurance that the value of maximum likelihood as a tree selection criterion is not often compromised by the presence of multiple local optima on a single tree.

- Fundamental mechanisms of mutation are investigated by estimating nucleotide substitution rate matrices for edges of phylogenetic trees. Several large alignments are examined, and the results suggest that the situation may be more complex than we had anticipated. It is likely that genome scale alignments will have to be used to further elucidate this question.

Acknowledgements

First and foremost I would like to thank my supervisor, David Penny, for encouraging me to do a PhD after many years away from science. Along with my co-supervisor, Mike Hendy, and despite the best efforts of the University, David has spent many years building a world-class research group conveniently located in my home town. My other co-supervisor, Lesley Collins, has helped knock this thesis into shape, and fought valiantly against my inclination toward understatement.

Thank you to all those who helped with code, data and other technical stuff: Patrick Biggs, Klaus Schliep, Trish McLenachan, Robin Atherton, Simon Hills, Judith Robins, Abby Harrison, Bojian Zhong, Pete Lockhart, Eric Bapteste and others. And to Tim White for the LaTeXtemplate I used to typeset this thesis. Also to the many colleagues who asked questions and made useful suggestions at talks and conferences.

I must thank the Allan Wilson Centre and the Institute of Molecular BioSciences for financial and logistical support, and especially Katrina Ross for having, as David puts it, "the terrible failing of always being on the side of the students."

Thanks to all my friends, old and new, for making the PhD experience a social one as well as a scientific one. In particular, those who do crosswords at lunchtime (Gillian, Simon, Nick, Barbara) and those who play board games on Saturday evenings (some of the above plus Rogerio, Paul, Tim, Sylvia, Robin, Klaus, Atheer, Rick, Sam, Elsa and more). Another thank you goes to Klaus for putting me up and showing me the sights on my visits to Paris. Also to Paul and Carolyn, David and Helen, the Tappers and the Slees, for their generous help and hospitality during my visits to Melbourne.

Finally, I would like to thank my family: both my parents for always encouraging my scientific tendencies; my stepmother Heather for convincing me that doing a PhD would be a good idea; my Belgian extended family, especially Charles and Nanou, for their hospitality during my visits to Europe; and my daughter Petra for providing a good excuse to take a break every few months and for distracting me from my work. And lastly Barbara, for your love and support.

Contents

Chapter 1

Introduction

The analysis of biological sequence data is a relatively young field of research. DNA sequencing techniques were first developed in the mid 1970s (most notably by Sanger et al., 1977), but it was not until the early 1990s that both the polymerase chain reaction (PCR) and capillary electrophoresis allowed considerable automation of the process. The volume of available sequence data has grown at an ever-increasing rate since, and has exploded in recent years with the advent of high-throughput sequencing, so that data generation is no longer the major challenge. The consequent availability of sequence data has enabled biologists to tackle a wide range of new questions, particularly about evolutionary history and about mechanisms of evolution.

Unfortunately, the increase in the breadth and depth of data available does not automatically entail an increase in biological knowledge—the data must first be analysed. While this glut of new data is opening up ever more opportunities for analysis of said data, many of these new opportunities require new methods, and existing methods must be tested to verify that they perform as expected with larger datasets. This thesis represents a modest contribution to the field, developing a new method for assembling and deconvoluting sequences, as well as testing and extending existing methods of phylogenetic analysis.

We will begin by outlining some of the terminology used in this thesis.

1.1 Background

1.1.1 Sequence data

The type of data we consider here are biological sequences, usually the nucleotide sequences of DNA, although we also use the amino acid sequences of proteins for some of the analyses carried out in Chapter 3. The fact that biological macromolecules exist as *ordered sequences* of a small number of different subunits or *characters* (four nucleotides, twenty amino acids) is something we often take for granted, but it is of huge importance in that it makes them easily amenable to powerful statistical analyses that might otherwise be intractable.

The first step in converting raw sequence data into knowledge is usually the *assembly* of sequencing reads into longer contiguous sequences (*contigs*) representing the DNA in the sample. This can be done either by *aligning* (or *mapping*) the reads to a known *reference* sequence, or by *de novo assembly*, that is, relying entirely on the information in the reads themselves without reference to any known sequence.

While traditional Sanger sequencing produces sequences up to about 1,000 base pairs in length, the recently developed *high-throughput sequencing* (sometimes referred to as *next-generation* or *second-generation* sequencing) techniques generally produce much shorter sequence reads, but in much greater volume. Early versions of the Illumina technology produced tens of millions of 35-base-pair reads per run, and both the read length and throughput have increased steadily, so that an Illumina run can now produce several billion 150-bp paired-end reads. This shorter read length and higher data volume has necessitated the development of a new generation of tools for assembly and mapping.

The high throughput of the new sequencing technology has also led to widespread use of *indexing*, which entails the ligation of a specific short sequence (the *index* or *barcode*) to one end of each sample so that multiple samples can be sequenced in a single run and the barcodes can then be used to determine which sample any given read came from.

If we wish to make any inferences concerning sequence evolution, we need an *alignment* of *homologous* sequences (i.e. sequences that have evolved from a common ancestor) from a number of *taxa* (taxonomic units such as species, populations or individual organisms). The differences between aligned sequences allow us to infer *substitutions* of one nucleotide or amino acid for another. The problem of aligning sequences so that each column (or *site*) contains only homologous characters is a difficult and interesting one, and is an area of active research, but this is not one of the questions that we address in this thesis.

1.1.2 Phylogenetic trees

A *phylogenetic tree* is a representation of a hypothesis concerning the evolutionary history of a set of taxa. It comprises a connected acyclic graph (a collection of *nodes*, or *vertices*, connected by *edges*) whose *tips* (or *leaves*) are labelled with the names of the taxa. The term *branch* is sometimes used to refer to either a single edge or to a *subtree*—we shall avoid this ambiguity by not using the term in this thesis. A tree can be *rooted*, in which case all edges are directed away from a node called the *root*, which represents the common ancestor of all the taxa at the tips. The edges leading to the tips are referred to as *pendant* edges.

A *weighted* tree has weights or lengths assigned to each edge. These *edge lengths* usually represent either numbers of substitutions (observed or inferred), or elapsed time. The edge structure of a labelled tree, in other words its branching pattern, ignoring any edge weights, is often referred to as its *topology*, although in graph theory this term is used for the branching pattern of an unlabelled tree (i.e., ignoring the positions of any taxon names or other labels), which we refer to as its *shape*.

Choosing the 'best' tree is obviously a crucial step in phylogenetic analysis, and one way to do this is to use an *optimality criterion*, by which we can assign a numeric score to any given tree. We then search for the tree or trees that maximise this score. The most widely used optimality criteria are maximum likelihood (Felsenstein, 1981), Bayesian maximum *a posteriori* probability (Rannala and Yang, 1996), and maximum parsimony (Fitch, 1971). Alternatively, we can use a purely algorithmic approach to build a tree—the most widely-used example of this approach is the neighbour-joining method (Saitou and Nei, 1987).

Often the evolutionary process is not strictly treelike. Hybridisation, recombination, lateral gene transfer and gene duplication can all lead to situations where the sites in a sequence evolve on different trees. This can be problematic for phylogenetic inference, especially when sequences from different parts of the genome are concatenated and analysed under a single model. In such cases phylogenetic *networks* may be used as an alternative to trees—a good overview of these methods can found in Huson et al. (2010).

1.1.3 Substitution models

A model of molecular evolution has three components: a tree, initial conditions (the distribution of bases at the root), and a mechanism of change. This last component, the mechanism of change, is usually modelled as a *continuous-time Markov process*, that is, a stochastic process where the probability distribution of future states depends only on the present state, and not on the site's history. The *states* of this process are the elements of the sequence, i.e. nucleotides or amino acids.

The Markov process is characterised by *transition rates* q_{ij} between states i and j. These transition rates are given as the off-diagonal elements of the *instantaneous rate matrix* $\mathbf{Q}$, and the diagonal elements are defined as $q_{ii} = -\sum_{j \neq i} q_{ij}$, so that each row sums to zero. The probabilities of transition from one state to another over time t, $p_{ij}(t)$, are given as the elements of the *transition matrix* $\mathbf{P}(t)$. These matrices are related by the equation $\mathbf{P}(t) = e^{\mathbf{Q}t}$.

We often model sites in the sequence as *independent and identically distributed* (i.i.d.), implying both that each site evolves by the same process, and that it is unaffected by any other sites. This i.i.d. assumption can be relaxed by allowing some sites to remain invariable, and by allowing the variable sites to evolve independently under several different Markov models. Further common assumptions are that the process at each site is *stationary* (the marginal probability of each nucleotide remains the same, so that nucleotide frequencies are expected to be constant over time), *reversible* (so that the process is not directed in time—this is particularly useful if we wish to consider unrooted trees), and *homogeneous* ($\mathbf{Q}$ remains constant). These assumptions may apply *locally* on a single edge, or *globally* across the whole tree.

More detailed discussions of phylogenetic trees and substitution models can be found in Penny et al. (1992), Liò and Goldman (1998), and Jermiin et al. (2008).

1.2 Thesis outline

This thesis covers three sequential aspects of sequence space exploration:

1. *de novo* assembly of genomic DNA sequences using short Illumina reads from a mixture of samples;

2. investigating whether multiple optima of likelihood on a given topology are likely to cause problems for the reconstruction of evolutionary trees from biological data; and

3. investigating the mechanisms of mutation by estimating nucleotide substitution rate matrices for edges of a tree.

The structure of the thesis is sequential in the sense that each project depended on analysis of the type carried out in the preceding project(s). A basic prerequisite of most sequence analysis projects is a DNA sequence of some length, which must first be assembled from raw sequence reads. This sequence can then be analysed in a variety of ways, one of which is the construction of a phylogenetic tree (which requires an alignment of several homologous sequences). Once we have both sequences and a reconstruction of their evolutionary history in the form of a phylogeny, we can use these to infer some properties of the evolutionary processes that took place along the branches of the phylogeny.

Each of the three projects is presented in the form of a paper: one published, one submitted, and one draft manuscript. This thesis is therefore modular in nature, in the sense that, since the projects are written up as papers, each one can, of course, be read alone.

Two of the papers included in the thesis have been written in collaboration with others. The draft manuscript presented in Chapter 4 is currently solely authored by myself, but other authors may be added as their contributions are included. This notwithstanding, the work presented here is my own. An outline of the relative contributions of the various authors of these papers can be found in Appendix A.

1.2.1 Assembly of mixed mitochondrial genomes

For any study involving sequence analysis, the most obvious requirement is to have sequences, and these sequences must be of sufficient length so that we can draw statistically meaningful inferences. Unfortunately such long sequences do not come ready-made—many individual sequence reads must be combined by identifying overlaps between them. Due to the prevalence of sequencing errors and the possibility of misleading overlaps between reads, sequence assembly is a far-from-trivial problem. High-throughput (or 'next-generation') sequencing methods in particular generally produce relatively short sequence reads, making the problem more difficult.

Here we do not focus on the problem of assembly as such—several assembly programmes are available that can assemble short reads with a high degree of accuracy. Instead, in **Chapter 2** we tackle the question of whether relatively short reads from a mixture of sources can be successfully assembled without the costly step of indexing each sample. In particular, we look at mixtures of mitochondrial genomes, and assemble them using the program Velvet (Zerbino and Birney, 2008). We show that provided we have a suitable set of reference genomes, it is possible to assemble contigs from such a mixture, separate these contigs by alignment to the reference genomes, and assemble each resulting subset of contigs into a complete mitochondrial genome.

In the course of assembling these mitochondrial genomes, we identified a highly unusual structural feature of one of them, that of the mollusc *Amalda northlandica*. This consisted of an 11 bp stretch of non-complementary double-stranded DNA (in this case the sequence on one strand is the reverse, instead of the reverse complement, of the sequence on the other strand) flanked by an inverted repeat. As far as we are aware, this is the first time non-complementary DNA has been identified in a mitochondrial genome, and this raises the interesting question of how the non-complementary strands could be synthesised during DNA replication.

A published manuscript (McComish et al., 2010) is the primary basis of this chapter, but I also describe further sequence assembly projects that I have undertaken, which have demonstrated that the methods work even with relatively low coverage of the target genomes and in the presence of large quantities of contaminating nuclear DNA. A paper describing one of these projects assembling a chloroplast genome (Atherton et al., 2010), of which I am a co-author, is provided as Appendix B. The approach outlined in Chapter 2 has now been used fairly extensively.

1.2.2 Multiple optima of likelihood

A common use for sequences is to reconstruct the evolutionary history of a group of organisms, in the form of a phylogenetic tree. This can be carried out using various algorithms and optimality criteria. One of the most popular optimality criteria used in phylogenetic analysis is maximum likelihood (ML).

In **Chapter 3** we investigate a potentially serious problem with ML optimisation—the existence of multiple local optima of likelihood on a given tree topology. It has been proven analytically that topologies can have multiple optima of likelihood (Steel, 1994; Chor et al., 2000), but simulation studies (Rogers and Swofford, 1999) suggest that the prevalence of multiple optima is not high enough in practice to compromise the efficacy of ML. However, the simulations used by Rogers and Swofford were carried out using sequences simulated on a single tree. Real biological data, in contrast, often contains conflicting signals as a result of processes such as recombination, introgression and lateral transfer of genes between species. It is therefore important to verify that the results obtained by simulation reflect the situation with real sequences. Here we explore the prevalence of multiple optima of likelihood on trees constructed with biological sequence data, and our results tend to support those of the simulations.

The basis of this chapter is a manuscript that has been submitted to *Systematic Biology* and accepted with revisions.

1.2.3 Investigating mutational mechanisms

Armed with a sequence alignment (or alignments) and a tree on which the sequences are hypothesised to have evolved, we can begin to draw inferences about the evolutionary process that has given rise to the observed sequences (or, to be more precise, to the observed differences between the sequences). One of the most fundamental aspects of that process is the underlying mutational mechanism, that is, the mechanism by which changes occur in the sequence. These changes can be either damage to the DNA, or errors in DNA replication, that are not repaired. We assume that mutations occur randomly (although repair of the mutations may not be random), but this does not imply that all types of mutation occur at the same rate; for example transitions usually occur more frequently than transversions (Wakeley, 1996). If we model molecular evolution as a Markov process, the mutational mechanism is reflected in its instantaneous rate matrix $\mathbf{Q}$.

In **Chapter 4**, substitution rate matrices are estimated in an attempt to detect differences in the underlying mutational mechanisms between different lineages. Several large alignments are examined, consisting of both coding and non-coding, nuclear and mitochondrial sequences. Our results tend to suggest that underlying mutational mechanisms cannot yet be untangled from other causes of nucleotide substitution using the alignments considered here, and further work will be needed, perhaps using genome-scale alignments. So far the availability of genome-scale alignments covering a broad taxonomic range (as opposed to alignments of individuals within a single species or genus) has lagged behind the surge in genome sequencing, but we expect this to change as more resources are shifted from data generation to analysis.

1.2.4 Future directions

Finally, in **Chapter 5**, I outline some directions for further research that could follow on from the three projects described here. While each of these projects has led to useful and important results, these results open the field for a number of other interesting questions, as well as providing material that will be of use in answering those questions.

Bibliography

R. A. Atherton, B. J. McComish, L. D. Shepherd, L. A. Berry, N. W. Albert, and P. J. Lockhart. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. *Plant Methods*, 6:22, 2010.

B. Chor, M. D. Hendy, B. R. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Mol Biol Evol*, 17(10):1529–1541, 2000.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

W. M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst Biol*, 20(4):406–416, 1971.

D. H. Huson, R. Rupp, and C. Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, 2010.

L. S. Jermiin, V. Jayaswal, F. Ababneh, and J. Robinson. Phylogenetic Model Evaluation. In J. M. Keith, editor, *Bioinformatics, Volume I: Data, Sequence Analysis, and Evolution*, volume 452 of *Methods in Molecular Biology*, pages 331–364. Humana Press, Totowa, NJ, 2008.

P. Liò and N. Goldman. Models of molecular evolution and phylogeny. *Genome Res*, 8:1233–1244, 1998.

B. J. McComish, S. F. K. Hills, P. J. Biggs, and D. Penny. Index-free de novo assembly and deconvolution of mixed mitochondrial genomes. *Genome Biol Evol*, 2:410–424, 2010.

D. Penny, M. D. Hendy, and M. A. Steel. Progress with methods for constructing evolutionary trees. *Trends Ecol Evol*, 7(3):73–79, 1992.

B. Rannala and Z. Yang. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol*, 43(3):304–311, 1996.

J. S. Rogers and D. L. Swofford. Multiple local maxima for likelihoods of phylogenetic trees: a simulation study. *Mol Biol Evol*, 16(8):1079–1085, 1999.

N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, 1987.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *P Natl Acad Sci USA*, 74(12):5463–5467, 1977.

M. Steel. The maximum likelihood point for a phylogenetic tree is not unique. *Syst Biol*, 43(4):560–564, 1994.

J. Wakeley. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends Ecol Evol*, 11(4): 158–162, 1996.

D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 18(5):821–9, 2008.

Chapter 2

Assembly of mixed mitochondrial genomes

2.1 Preamble

This chapter presents the paper "Index-free de novo assembly and deconvolution of mixed mitochondrial genomes", which I published in the journal *Genome Biology and Evolution* in 2010 (McComish et al., 2010). Next-generation sequencing methods now produce much longer reads than the 50 bp reads used here, but the principles behind the method are general, and the method becomes more powerful with longer reads. Three of the mitochondrial genomes sequenced in this study have been deposited in the GenBank database: the mollusc *Amalda northlandica* (accession number GU196685) described in the paper; the rat *Rattus fuscipes* (accession number GU570664), published by Robins et al. (2010); and the human haplotype FA064 (accession number GQ214521), published by Corser et al. (2012). The mollusc and rat mitochondrial genomes have since been added to NCBI's Reference Sequence collection with the accession numbers NC_014403 and NC_014867 respectively.

The Perl scripts used in this paper are now freely available for download from `http://sourceforge.net/p/mixed-assembly`.

As a result of this work, similar approaches have been applied in several additional projects. These are described in Section 2.3, after the *Genome Biology and Evolution* paper which is given as Section 2.2.

Index-Free De Novo Assembly and Deconvolution of Mixed Mitochondrial Genomes

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.

†Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, PO Box 11-222, Palmerston North, New Zealand.

‡These authors contributed equally to this work.

*Corresponding author: Email: b.mccomish@massey.ac.nz

Accepted: 19 May 2010

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 reassembled from a high coverage of short (35 nt) reads, such as those generated by a second-generation Illumina Genome Analyzer. We then assessed this experimentally with long-range polymerase chain reaction products from mitochondria of a human, a rat, a bird, a frog, an insect, and a mollusc. Comparison with 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, which contains a structure that 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.

Key words: multiplex sequencing, informatic deconvolution, control region, noncomplementary, 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 are 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 nt, a GA run produced over 40 million 36 nt reads, and a SOLiD run promised 86–114 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 12–15 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 data sets. In molecular data sets, this is often achieved by comparing results from nuclear data and mitochondrial and/or chloroplast data (e.g., Pratt et al. 2009). These data sets 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 multigene data sets 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-kb 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, although 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, e.g., 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 1 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 (i.e., 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 optimized, the same methods were successfully applied to reads from a single lane of a GA flow cell containing a mixture of 6 different mitochondrial genomes.

Mitochondrial sequences from 4 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 4 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). Because the reference sequences were those used to generate the reads, stringent parameters were used for the alignment (minimum overlap 50 and overlap identity 98%). The contigs were also aligned to related reference sequences using less stringent parameters (minimum overlap 40 and 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 2009) was used to examine the distribution of coverages for each set of contigs. If the coverage distribution showed discrete peaks corresponding to the 5 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 polymerase chain reaction (PCR) products were generated from a diverse set of templates in order to create a mixture of templates to sequence using an Illumina GA. 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, *A. northlandica*). PCR products of between ∼1 and 8 kb were generated using primers specific to, and thermal cycling conditions optimized 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 GA GA2 (Illumina, Inc.) according to the manufacturer's instructions. Unfortunately, there was an instrument problem at cycle 33, which meant that only 32 nt were usable. Due to the availability of the raw material for sequencing, the run was continued to completion. After sequencing, the resultant images were analyzed with the proprietary Illumina pipeline (version 1.0) using default parameters. This resulted in ∼238 Mb of sequence, with 63% of the clusters passing the initial filtering step.

Additional assessment of an anomalous section of the *A. 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′ and 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 specifically designed PCR primers. The resulting sequences and electropherograms were visualized using Geneious.

De Novo Assembly

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

Table 1
Mitochondrial Sequences Referred to in this Study

Accession Number	Species	Common name	Reference
J01415[a,b]	*Homo sapiens*	Human	Anderson et al. (1981)
NC_001807[b]	*H. sapiens*	Human	Ingman et al. (2000)
AJ428514[a]	*Rattus norvegicus*	Norway rat	Nilsson et al. (2003)
NC_001665[b]	R. *norvegicus*	Norway rat	unpublished
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	unpublished
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 obsoleta*	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	unpublished
NC_011193	*Rapana venosa*	Veined rapa whelk	unpublished
NC_013239	*Terebra dimidiata*	Dimidiate auger shell	Cunha et al. (2009)
NC_013241	*Cancellaria cancellata*	Cancelate 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.

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 visualized 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

Thirty-five base-pair 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 5 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.

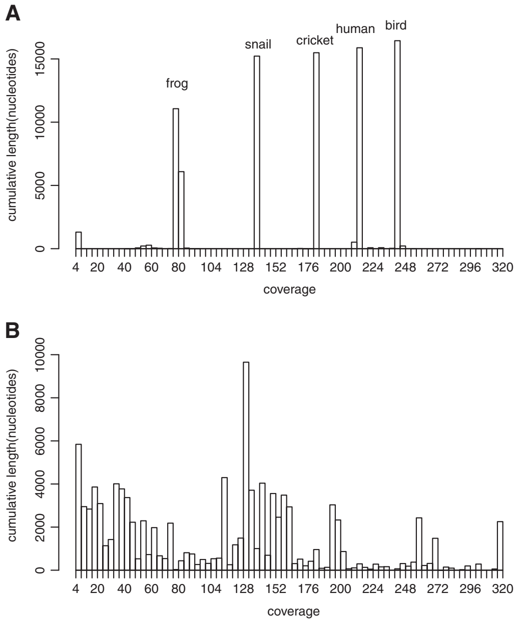


Fig. 1.—Coverage distributions. (*A*) Coverage, weighted by contig length, for contigs assembled from simulated reads by Velvet with $k = 31$. The sequences used in this simulation were human (25%), bird (30%), frog (10%), cricket (20%), and snail (15%). It is clear from these distributions that the 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).

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 number NC_001807), another Norway rat (accession number 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 and 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 contain tandem repeat regions, which could not be assembled from short reads. This would be the case regardless of whether they were 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 number AJ428514) and reef squid. The relatively closely related human and rat mitochondrial genomes were each broken up into a larger number of contigs (12 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 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

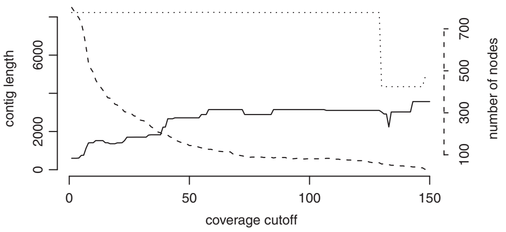


Fig. 2.—Assembly statistics for biological data. Median length–weighted contig length (N50, solid line), maximum contig length (dotted line), and number of nodes (dashed line) plotted against coverage cutoff for Velvet assemblies with hash length $k = 25$. Contig lengths are in *k*-mers (length in base 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 8,231 for all coverage cutoffs up to 129, except that for coverage cutoffs between 45 and 64, 10 nt are added to one end of the contig to give a length of 8,241.

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 fig. 1*B*). Consequently, they were separated by aligning them to a set of reference genomes. The references used were the mitochondrial genomes of a human (accession number 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 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.

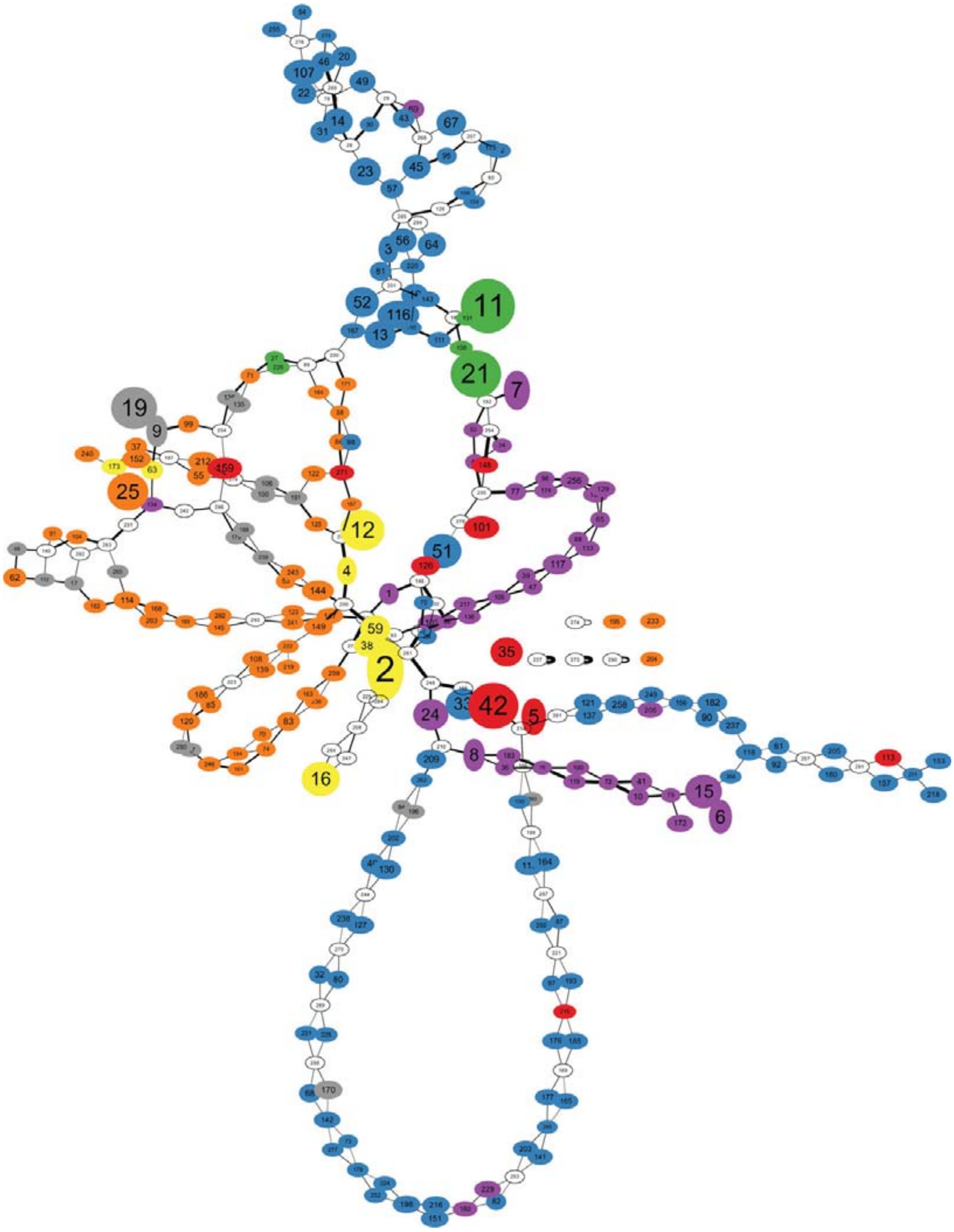
Of a total of 964 contigs for the 5 assemblies examined, 762 were correctly grouped into species in this first step. However, because the single best alignment for each contig was used regardless of the relative scores of alignments to the other references, 64 contigs were initially assigned to incorrect 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 colored 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, because 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 that aligned to different references were checked against GenBank. Node 206 of the assembly in figure 3, for example, aligned to the spiny rat, whereas the 2 neighboring 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 *Leiopelma 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 that 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


107
46
20
49
22
14
31
43
67
23
45
57
3
56
64
52
116
13
11
21
7
19
9
25
37
152
212
55
159
63
12
4
51
126
144
149
1
59
38
2
101
35
42
5
33
24
8
209
117
15
6
16
114
83
120
62
130
238
127
32
80
170
151
216
229
121
258
182
90
113
164

experiment. The best results were obtained with a coverage cutoff of 12, giving 3 overlapping contigs with a total length of 10,485 nt 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 will be reported separately 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 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 *A. northlandica*. Lower coverage cutoffs gave six contigs, covering the whole genome except for a gap of 11 nt in the noncoding 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, 1,385, 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 crisscross 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 *R. fuscipes*, extending the work of Robins et al. (2008). This sequence has the GenBank accession number 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 *cyt*B, there were three. Comparing contigs to the GenBank nucleotide database using BlastN showed that the sequence covering 12S rRNA to *trn*M matched tawny frogmouth sequence fragments: 1 partial 12S rRNA sequence and 1 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 *cyt*B 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 3 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 (because 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

← **Fig. 3.**—Assembly graph. The assembly graph for sequences assembled by Velvet with $k = 25$ and cov_cutoff = 26. Nodes are colored 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. Gray 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 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, whereas the insect, bird, and rat show more complex chains of shorter contigs. The reasons for these patterns are discussed in the text.

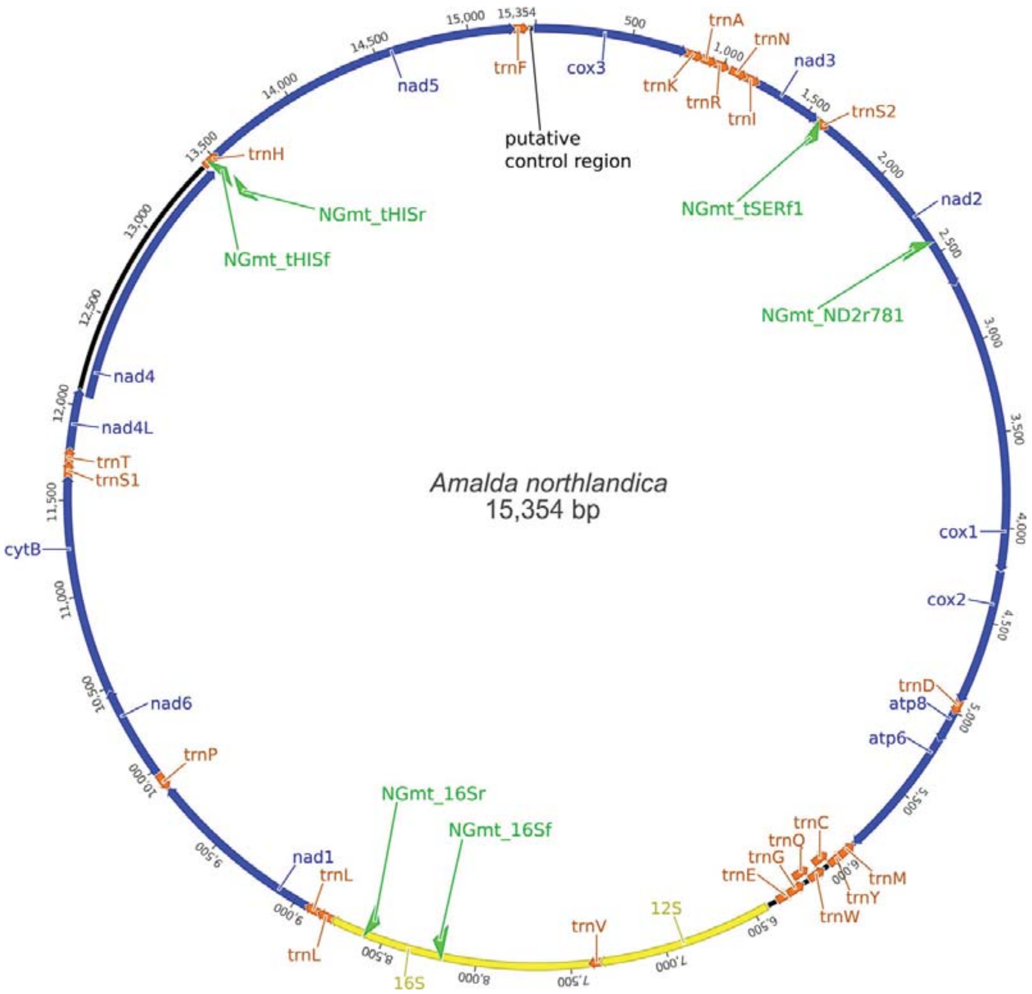


Fig. 4.—The *A. northlandica* complete mitochondrial genome. Arrowheads depict the direction of transcription. Genes with offset annotations (*trn*C, *trn*Q, and *nad*4) overlap with genes preceding them. Binding sites for the primers used to generate the long-range PCR products are indicated in green.

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 *A. northlandica* is 15,354 bp in length and contains 13 protein-coding genes, 2 ribosomal RNA genes, and 22 tRNA genes (figure 4) and has

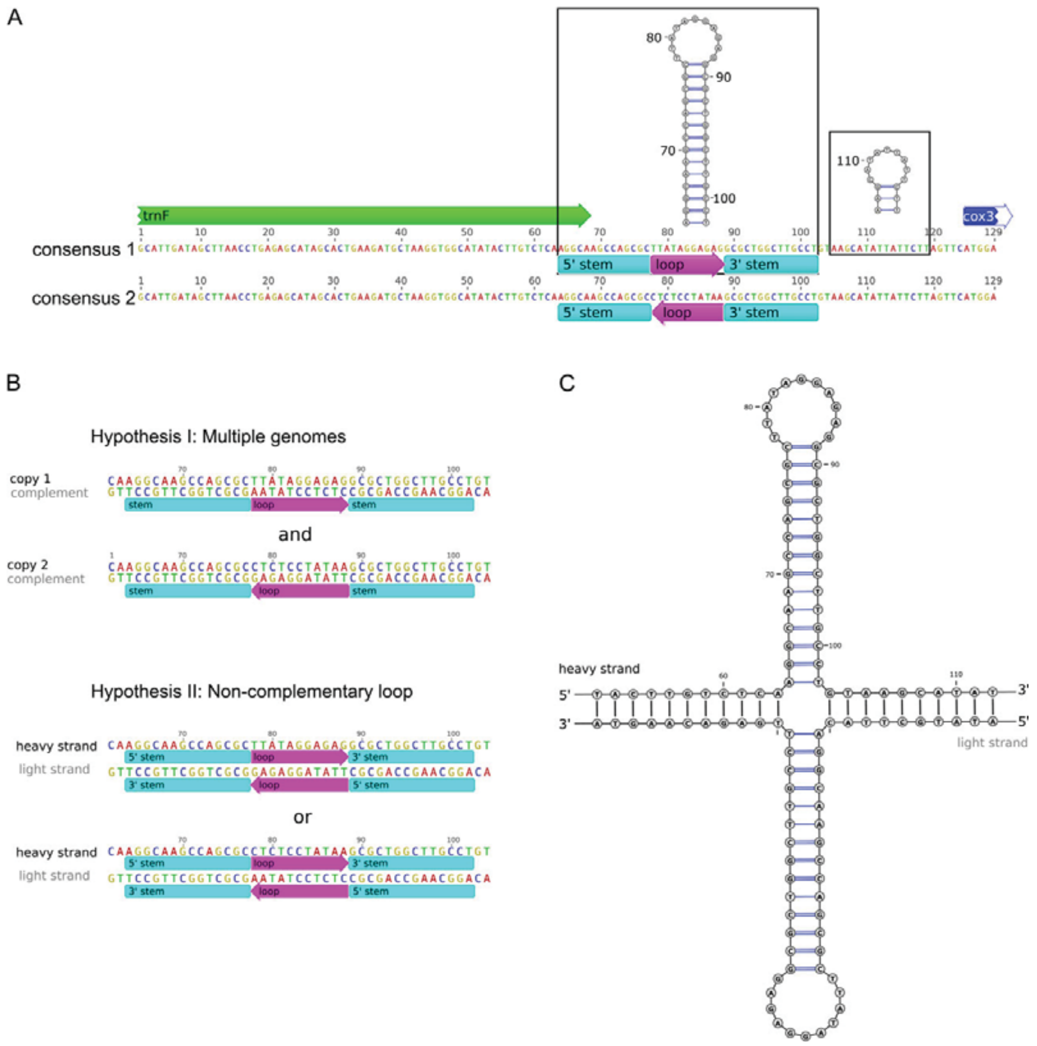
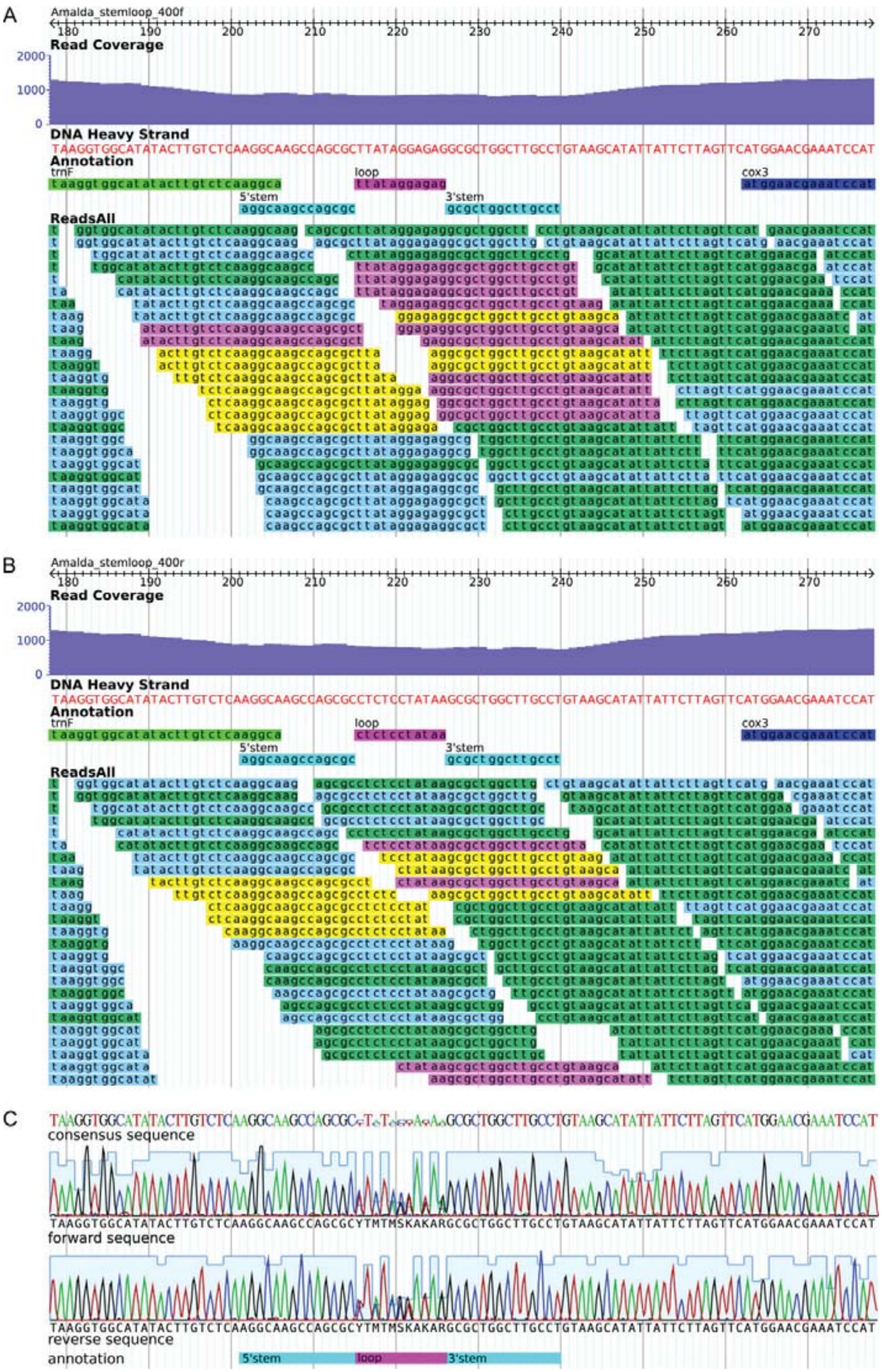

A
80
90
70
100
110
trnF
cox3
1 10 20 30 40 50 60 70 80 90 100 110 120 129
consensus 1 GCATTGATAGCTTAACCTGAGAGCATAGCACTGAAGATGCTAAGGTGGCATATACTTGTCTCAAGGCAAGCCAGCGCTTATAGGAGAGGCGCTGGCTTGCCTGTAAGCATATTATTCTTAGTTCATGGA
5' stem
loop
3' stem
1 10 20 30 40 50 60 70 80 90 100 110 120 129
consensus 2 GCATTGATAGCTTAACCTGAGAGCATAGCACTGAAGATGCTAAGGTGGCATATACTTGTCTCAAGGCAAGCCAGCGCCTCTCCTATAAGCGCTGGCTTGCCTGTAAGCATATTATTCTTAGTTCATGGA
5' stem
loop
3' stem
B
Hypothesis I: Multiple genomes
70 80 90 100
copy 1 CAAGGCAAGCCAGCGCTTATAGGAGAGGCGCTGGCTTGCCTGT
complement GTTCCGTTCGGTCGCGAATATCCTCTCCGCGACCGAACGGACA
stem
loop
stem
and
1 70 80 90 100
copy 2 CAAGGCAAGCCAGCGCCTCTCCTATAAGCGCTGGCTTGCCTGT
complement GTTCCGTTCGGTCGCGGAGAGGATATTCGCGACCGAACGGACA
stem
loop
stem
Hypothesis II: Non-complementary loop
70 80 90 100
heavy strand CAAGGCAAGCCAGCGCTTATAGGAGAGGCGCTGGCTTGCCTGT
5' stem
loop
3' stem
light strand GTTCCGTTCGGTCGCGGAGAGGATATTCGCGACCGAACGGACA
3' stem
loop
5' stem
or
70 80 90 100
heavy strand CAAGGCAAGCCAGCGCCTCTCCTATAAGCGCTGGCTTGCCTGT
5' stem
loop
3' stem
light strand GTTCCGTTCGGTCGCGAATATCCTCTCCGCGACCGAACGGACA
3' stem
loop
5' stem
C
80
90
70
100
heavy strand
60
110
5'
3'
3'
5'
light strand

Fig. 5.—The *trn*F–*cox*3 intergenic region of the *A. 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: hypothesis I: there is a mixture of 2 mitochondrial genome copies that differ in the orientation of the loop sequence; or hypothesis II: there is a single genome that contains a noncomplementary region, which could exist in either of 2 possible orientations. (*C*) Hypothesis II suggests the formation of a double-stem structure in double-stranded DNA.

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 codon. 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*, and *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).


A
Amalda_stemloop_400f
180 190 200 210 220 230 240 250 260 270
Read Coverage
2000
1000
0
DNA Heavy Strand
TAAGGTGGCATATACTTGTCTCAAGGCAAGCCAGCGCTTATAGGAGAGGCGCTGGCTTGCCTGTAAGCATATTATTCTTAGTTCATGGAACGAAATCCAT
Annotation
trnF
taaggtggcatatacttgtctcaaggca
loop
ttataggagag
cox3
atggaacgaaatccat
5'stem
aggcaagccagcgc
3'stem
gcgctggcttgcct
ReadsAll
B
Amalda_stemloop_400r
180 190 200 210 220 230 240 250 260 270
Read Coverage
2000
1000
0
DNA Heavy Strand
TAAGGTGGCATATACTTGTCTCAAGGCAAGCCAGCGCCTCTCCTATAAGCGCTGGCTTGCCTGTAAGCATATTATTCTTAGTTCATGGAACGAAATCCAT
Annotation
trnF
taaggtggcatatacttgtctcaaggca
loop
ctctcctataa
cox3
atggaacgaaatccat
5'stem
aggcaagccagcgc
3'stem
gcgctggcttgcct
ReadsAll
C
consensus sequence
TAAGGTGGCATATACTTGTCTCAAGGCAAGCCAGCGCYTMTMSKAKARGCGCTGGCTTGCCTGTAAGCATATTATTCTTAGTTCATGGAACGAAATCCAT
forward sequence
TAAGGTGGCATATACTTGTCTCAAGGCAAGCCAGCGCYTMTMSKAKARGCGCTGGCTTGCCTGTAAGCATATTATTCTTAGTTCATGGAACGAAATCCAT
reverse sequence
annotation
5'stem
loop
3'stem

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 noncoding intergenic spacers were analyzed 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 *A. northlandica* is located between the genes *trn*F and *cox*3. It is 56 bp in length and contains 2 predicted secondary structural elements, a strong stem–loop element and a second small stem–loop element (fig. 5A). Of the remaining intergenic sequences in the *Amalda* mitochondrial genome, only 7 are longer than 10 bp. All of these 7 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 nt 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 (fig. 6). Although we are able to confirm the sequence of the loop region, we are unable to show in which of 2 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 artifact introduced in the Illumina sequencing or short-read assembly (fig. 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 6 published neogastropods (*I. obsoleta*, *T. clavigera*, *R. venosa*, *F. similis*, *B. brandaris*, and *N. 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 fig. 5), but the nucleotide sequences for the remainder of the region are quite divergent in these species. The mitochondrial genomes of *C. cancellata*, *C. olla*, *L. cerithiformis*, *C. textile*, and *C. 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 (*T. 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, whereas most of the known neogastropod sequences have some intergenic sequence at this position only 5 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

⟵

Fig. 6.—Gbrowse visualizations of short reads from the *A. 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.

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 due 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., 2 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 mitochondrial genome sequence of the mollusc *A. northlandica*, we have characterized 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 fig. 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 noncomplementary 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 noncomplementary sequence enforces the formation of a functionally important structural element in double-stranded DNA (fig. 5C). One difficulty with this hypothesis is how such a noncomplementary 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 noncomplementary 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 characterizing 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 *I. 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 characterization 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 *A. 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 with the 12 known neogastropod sequences, a tantalizing 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, 12 birds, 12 molluscs, 12 insects, and 12 human individuals, rather than using 48 index tags, we could use 12, each with a mixture consisting of 1 bird, 1 mollusc, 1 insect, and 1 human. A single set of 4 reference sequences could then be used to separate all 12 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.

Acknowledgments

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

Literature Cited

Akasaki T, et al. 2006. Extensive mitochondrial gene arrangements in coleoid Cephalopoda and their phylogenetic implications. Mol Phylogenet Evol. 38:648–658.

Altschul SF, et al. 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, et al. 2008. The mitochondrial genome of *Conus textile, cox*I- *cox*II 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–D31.

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. 11:1619–1623.

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

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 [cited 2008 December 22]. Available from http://www.geneious.com/.

Erlich Y, et al. 2009. DNA Sudoku—harnessing high-throughput sequencing for multiplexed specimen analysis. Genome Res. 19:1243–1253.

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.

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

Kim I, et al. 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, et al. 2006. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. Nucleic 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–S12.

Pratt RC, et al. 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. Vienna, Austria: R Foundation for Statistical Computing.

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, et al. 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 U S A. 74:5463–5467.

Sasuga J, et al. 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, et al. 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. In Barciszewski J, Clark BFC, editors. RNA biochemistry and biotechnology. NATO ASI Series, Dordrecht (NL): Kluwer Academic Publishers. 11–43.

Associate editor: B. Venkatesh

2.3 Further sequence assembly work

The work referred to in the *Genome Biology and Evolution* paper has proved useful to a wide range of other researchers. A similar approach has also been developed independently using the longer reads of the Roche/454 sequencing platform to sequence long-range PCR fragments and assemble partial mitochondrial genomes of 30 species of beetles (Timmermans et al., 2010). Those projects for which I have carried out assemblies are described below.

2.3.1 Rats

As mentioned in the paper above, some of the contigs from this experiment were used, along with further sequencing experiments, to determine the complete mitochondrial sequence of *Rattus fuscipes*. This was published, along with six other rat mitochondrial genomes, which were sequenced separately and which I assembled concurrently, in Robins et al. (2010), of which I am a co-author. Assembly of these genomes was complicated by the apparent presence of numts (nuclear sequences of mitochondrial origin). Ambiguous regions were resolved using Sanger sequencing, and these sequences were used to identify the correct mitochondrial contigs for each of the rat genomes.

2.3.2 Chloroplasts

An additional sequencing project, of a single chloroplast genome, was carried out in early 2009. The complete chloroplast genome of karaka (*Corynocarpus laevigatus*) was sequenced from chloroplast-enriched genomic DNA. Sequencing was carried out on the Illumina GAII platform to give 75 bp paired-end reads, and *de novo* assembly was carried out using Velvet (v. 0.7, Zerbino and Birney, 2008). The best assembly gave four contigs of chloroplast origin, along with nine nuclear contigs. The four chloroplast contigs were aligned to the cucumber (*Cucumis sativus*, which is in the same order, Cucurbitales, as karaka) chloroplast genome in order to identify overlaps, and combined to give the complete chloroplast genome sequence. The paper describing this project, Atherton et al. (2010), is provided as Appendix B, and I am a co-author of that paper.

2.3.3 More mixtures

In late 2011, a modified version of the pipeline described in the paper above was applied to four more mixtures of mitochondrial genomes, comprising a total of 18 samples. For each of the four mixtures, a series of preliminary *de novo* assemblies was performed using Velvet (v. 1.1), and these were used to determine appropriate parameter settings for further assemblies. The mitochondrial genomes included in these mixtures are listed in Table 2.1, along with some parameters and statistics of the final Velvet assemblies used.

In contrast to the mixture described in the paper above (where the samples were long range PCR products) all but three of the 18 samples in this case were total genomic extracts. This meant that large amounts of nuclear DNA were present along with the mitochondrial DNA, so that overall coverage of the mitochondrial genome was highly variable and often relatively low. Low coverage meant that in some cases we were working at the limits of what can be achieved with *de novo* assembly.

One sample, the flax snail *Placostylus fibratus*, could not be assembled, and part of its mitochondrial genome was later sequenced separately. The sequence reads from the mixture containing *Placostylus* were mapped against the partially assembled genome using BWA (Li and Durbin, 2009), and only 14 reads could be aligned, which suggests that little or no *Placostylus* mitochondrial DNA was present in the mixed sample.

Because of the presence of nuclear DNA, Velvet assemblies, as expected, produced much larger numbers of contigs than the assembly described in the paper above. This meant that the assembly graphs could not be used to help separate contigs from the different samples in each mixture, because the graphs consisted largely of isolated short contigs of nuclear origin. However, it was possible to identify almost all of the mitochondrial contigs by simply altering the script used to separate contigs so that it called Exonerate (v. 2.2.0, Slater and Birney, 2005) with a score threshold of 150. This was sufficient to filter out all of the nuclear contigs, as these gave scores lower

Table 2.1: **Mitochondrial genomes sequenced as mixtures in September 2011.**

Mix	Sample	Length	Gaps	k	k-mer cov.[a]	Nuc. cov.[a]	
18	Fiji tree frog (*Platymantis vitiensis*)	16,819	4	45	4.7	8.3	frog
18	Limpet (*Notoacmea* sp.)	18,529	2	45	12.2	21.9	mollusc
18	Chestnut rail (*Eulabeornis castaneoventris*)	17,439	1	61	54.9	137.3	bird
18	Hapuku (*Polyprion oxygeneios*)	16,509	1	45	7.1	12.6	fish
19	Tree weta (*Hemideina crassidens*)[b]	17,164	0	61	30.1	75.3	insect
19	Flax snail (*Placostylus fibratus*)[c]	-	-	-	-	-	mollusc
19	Coot (*Fulica atra*)	17,029	0	61	19.3	48.3	bird
19	Eastern grey kangaroo (*Macropus giganteus*)	16,887	0	63	314.7	828.3	mammal
19	Blue cod (*Parapercis colias*) ACH-1[d]	16,630	4	-	-	-	fish
19	Moss-forest rat (*Rattus niobe*)	16,298	0	61	19.9	49.7	mammal
20	San Cristoval frog (*Hylarana kreffti*)	17,185	3	37	5.9	9.3	frog
20	Fernbird (*Megalurus punctatus*)	17,991	1	63	14.4	37.8	bird
20	Top snail (*Calliostoma simulans*)	15,993	6	37	3.9	6.1	mollusc
20	Blue cod (*Parapercis colias*) ACH-3	16,631	0	37	7.7	12.1	fish
21	Raukumara tusked weta (*Motuweta riparia*)[b]	16,371	0	63	19.9	52.5	insect
21	Ostrich foot snail (*Struthiolaria papulosa*)	16,001	5	25	5.0	6.6	mollusc
21	Oriental bay owl (*Phodilus badius*)	17,045	9	25	4.8	6.3	bird
21	Snapper (*Pagrus auratus*)	16,728	0	51	12.7	25.5	fish

[a] The k-mer coverage is the length-weighted mean of the k-mer coverage of all Velvet contigs used in the assembly; nucleotide coverage is calculated from k-mer coverage using the formula given in the Velvet manual.

[b] For both weta samples, mitochondrial DNA extractions were used rather than total genomic extracts.

[c] No *Placostylus* DNA was detected.

[d] The blue cod ACH-1 sequence was determined by mapping reads to the successfully assembled ACH-3 sample of the same species (in mixture 20), so no Velvet parameters are given.

than 150 when aligned to the mitochondrial reference genomes.

Geneious (Biomatters, 2011) was used to align assembled contigs against reference genomes and to identify overlaps and combine contigs into preliminary whole mitochondrial genome assemblies. In some instances the coverage of the mitochondrial genome was so low that there were gaps in the Velvet assemblies. Additional gaps were caused by regions of repetitive sequence in a few of the genomes. In order to fill these gaps, the original sequence reads were mapped to the preliminary assembly using BWA (Li and Durbin, 2009), and these mappings were visualised using Tablet (Milne et al., 2010) in order to identify reads that overlapped into the gaps. This was repeated until either all gaps were filled or no more overlapping reads could be found, which required up to four iterations in some cases. Some gaps could not be closed in this way, and the numbers of remaining gaps are shown in Table 2.1. For most of these gaps, the assembled sequences have been used to design PCR primers, which have then been used to generate transcripts across the gaps. These transcripts have been sequenced, enabling complete mitochondrial genomes to be finished.

These mitochondrial genomes will be used in several PhD and MSc theses, as well as in at least seven publications that are currently in preparation. To date, one has been deposited with GenBank: *Rattus niobe*, with accession number KC152486.

2.4 Conclusions

Overall, the methods reported here have been useful to a significant number of other researchers, both here at Massey and at Victoria University, the University of Auckland, and Queensland University of Technology. Additionally, carrying out assembly work on a variety of projects in collaboration with others has given me a better understanding of how *de novo* assembly (and Velvet in particular) performs with varying coverage levels, and in the presence of non-target sequence.

The length of the sequence reads from next-generation sequencing and the availability of reference genomes both continue to increase, making the method increasingly powerful. Longer reads mean that the elements of a mixture will be assembled into fewer larger contigs, because there are fewer stretches of identical sequence long enough to break up the assembly. Longer reads also make it possible to separate more closely related genomes, as they become less likely to have stretches of identical sequence longer than the read length. The increasing number of fully sequenced reference genomes enables us to use references that are more closely related to our target sequences, and this is another important factor in enabling us to deconvolute mixtures of relatively closely related organisms.

Bibliography

R. A. Atherton, B. J. McComish, L. D. Shepherd, L. A. Berry, N. W. Albert, and P. J. Lockhart. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. *Plant Methods*, 6:22, 2010.

Biomatters. Geneious version 5.5, 2011. URL `http://www.geneious.com`.

C. A. Corser, P. A. McLenachan, M. J. Pierson, G. L. A. Harrison, and D. Penny. The Q2 mitochondrial haplogroup in Oceania. *PLoS ONE*, 7(12):e52022, 2012.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

B. J. McComish, S. F. K. Hills, P. J. Biggs, and D. Penny. Index-free de novo assembly and deconvolution of mixed mitochondrial genomes. *Genome Biol Evol*, 2:410–424, 2010.

I. Milne, M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright, and D. Marshall. Tablet—next generation sequence assembly visualization. *Bioinformatics*, 26(3):401–402, 2010.

J. H. Robins, P. A. McLenachan, M. J. Phillips, B. J. McComish, E. Matisoo-Smith, and H. A. Ross. Evolutionary relationships and divergence times among the native rats of Australia. *BMC Evol Biol*, 10(1):375, 2010.

G. S. C. Slater and E. Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31, 2005.

M. J. T. N. Timmermans, S. Dodsworth, C. L. Culverwell, L. Bocak, D. Ahrens, D. T. J. Littlewood, J. Pons, and A. P. Vogler. Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Res*, 38(21):e197, 2010.

D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 18(5):821–9, 2008.

Chapter 3

Multiple optima of likelihood

3.1 Preamble

Once we have assembled and annotated sequences, one of the possible uses for them, after they have been aligned, is in the construction of phylogenetic trees. Many tree-building programs use maximum likelihood (ML) as an optimality criterion, and we would like to be certain that it performs as expected.

One potentially problematic property of ML, which could prevent it from performing as we would expect, is that the likelihood function can have several local optima on a given tree topology, that is, several different sets of edge lengths could give locally maximal likelihood values. Because most current ML methods optimise edge lengths by hill-climbing (incrementally changing each edge length, checking whether the new likelihood is higher, and repeating until no higher likelihood is found), the presence of local optima could prevent us from reaching the global optimum, since the algorithm will stop climbing when it reaches the top of a 'hill', regardless of whether there are higher peaks elsewhere in the landscape.

Studies using simulated sequences have found that the best tree (that on which the sequences were simulated) is not usually affected by multiple optima. However, unlike the simulated data, real biological sequences are not always generated on a single tree, and can be affected by conflicting signals arising from a number of sources. It is therefore important to investigate whether (and to what degree) trees constructed from biological sequences are affected by multiple optima.

Here we examine a large number of subsets of taxa from several alignments. For each of these subsets, we select 1,000 random tree topologies, and for each topology generate 1,000 sets of randomly distributed edge lengths. These million trees for each subset of taxa are then used as starting points for hill-climbing optimisation of likelihood, and we compare the resulting trees in order to detect any cases of multiple optima.

We find that while the incidence of multiple optima varies widely between the datasets examined, multiple optima are rarely found on the topologies with the highest likelihoods. These results support those of the simulation studies, and provide reassurance that the performance of ML tree-building methods is not likely to be compromised by the presence of local optima.

This chapter presents the manuscript "Multiple local maxima for likelihoods of phylogenetic trees constructed from biological data", which was submitted to *Systematic Biology* on 11th December, 2012, and has now been accepted with revisions. The manuscript presented here is the revised version.

The R scripts and sequence data used in this manuscript are now freely available for download from `http://sourceforge.net/p/multipleoptima`.

Multiple local maxima for likelihoods of phylogenetic trees constructed from biological data

Bennet J McComish[1], Klaus P Schliep[2], and David Penny[1]

[1] *Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand;*

[2] *Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo 36310, Spain*

Abstract

It is known that the maximum likelihood function can have multiple local optima on a given tree topology. Simulation studies suggest that this is not likely to affect tree-building, but these studies used data generated on a single topology. In contrast, it has been shown that simple mixture models can generate data where multiple optima can occur even on the tree with the highest likelihood. Here we present results using four alignments of biological sequence data. We find that the prevalence of multiple optima varies widely with biological data, and that trees with multiple optima are less likely to occur among the most likely topologies. Overall, our results tend to support those of the simulation studies, providing further reassurance that the value of maximum likelihood as a tree selection criterion is not often compromised by the presence of multiple local optima on a single tree topology.

(Keywords: maximum likelihood, multiple optima, phylogenetic trees, biological sequence data)

3.2 Introduction

Maximum likelihood (ML) is the most widely used tree selection criterion in molecular phylogenetics. Most ML-based phylogenetic methods use a two-step optimisation procedure: first, parameters such as edge weights are optimised on a given topology using an iterative hill-climbing approach; and then secondly the space of trees is searched heuristically for the topology that maximises the overall likelihood (Huelsenbeck and Bollback, 2007). Here we focus on the first optimisation step: optimising the edge weights on a given topology.

There has been considerable uncertainty over the possibility of multiple optima on a specified topology. Initially Fukami and Tateno (1989) suggested a 'proof' that multiple optima would not exist on a single topology under the one-parameter model of nucleotide substitution. However, it was later demonstrated that the likelihood function for a four-taxon phylogenetic tree can have multiple local maxima (Steel, 1994). Because some current ML-based phylogenetic methods rely on an iterative hill-climbing approach to maximise the likelihood for each topology, the existence of multiple optima means that we cannot guarantee that we will find the global maximum for a given topology. However, Steel's example lies at the boundary of parameter space—the two ML points require a substitution probability of $p = 0.5$ on two edges of the four-taxon tree, and $p = 0$ on the other three. Furthermore, the topology that displays multiple optima does not maximise the likelihood function across all trees.

Further analytic studies have identified additional cases of multiple optima in the interior of the parameter space (Chor et al., 2000) for four taxa. In these cases, conflicting signals in the data can lead to multiple optima on all possible tree topologies for four taxa. Schliep (2009) showed that the results of Chor et al. (2000) could arise from a biological process by demonstrating that data generated by simple mixture models can lead to multiple optima for a given topology, including sometimes on the topology with the highest likelihood. Schliep used a mixture of two four-taxon trees with different topologies to either simulate sequence data or to generate expected site frequencies. Depending on the choice of edge lengths, this was found to lead to multiple optima. This can perhaps be explained by analogy to a very simple mixture of two Gaussian distributions, where if the means are near to each other then the mixture is uni-modal, otherwise bimodal. Similarly, if the trees in the mixture have a short internal edge and longer external edges, then we recover a single tree, whereas if the internal edge is longer relative to the external edges, this can lead to multiple optima.

The existence of multiple optima on the topology with the highest overall likelihood would be particularly problematic, as it could lead to a tree with lower likelihood being selected as the maximum likelihood tree. If a topology has several optima, each with a relatively poor likelihood value, the second optimisation step will find another topology with a higher likelihood, so that the tree selection process is unaffected. If, however, a topology has a likelihood optimum that maximises the overall likelihood, but the hill-climbing algorithm finds a different, lower, optimum, then the second optimisation step may select a different topology with a maximum likelihood intermediate between the two.

In contrast to these analytical results, simulation studies have found that, for data generated on a specified tree, multiple optima are not usually found on that tree topology, and the tree that was used to generate the data usually has the highest average likelihood (Rogers and Swofford, 1999). If the same holds true for biological (as opposed to simulated) sequence data, this would imply that multiple optima are, in practice, unlikely to compromise the efficacy of ML as a tree-selection criterion.

However, real biological data are not necessarily generated on a single tree—various evolutionary processes can lead to conflicting signals in the data, so that model misspecification can be a significant problem. For example, lateral transfer of genes between species, introgression, hybridisation, incomplete lineage sorting, gene conversion between paralogous genes, and recombination between genes of different viral strains can all produce genuine historical signals that may differ from that produced by the phylogeny of the taxa.

The second step of optimisation of ML, that is, the search for the tree with highest overall likelihood, can also be affected by multiple optima. Salter (2001) and Morrison (2007) have shown that several islands of topologies can exist in the tree space under both the nearest-neighbour-interchange and tree-bisection-and-reconstruction optimisation strategies. More recently, Money and Whelan (2012) have characterised the problem more thoroughly, and consequently we do not investigate this problem in the present study. A similar result, that of multiple optima on different topologies, has also been shown for parsimony (Hendy et al., 1988).

In the current study, we have focussed on the first optimisation step, with the aim of identifying cases where, with real biological sequence data, multiple optima occur for a given topology, and of attempting to characterise these cases. We hypothesised that multiple optima are more likely to occur in the following cases: for topologies that fit the data poorly; with shorter sequences; when internal edges are long relative to external edges (as shown in Schliep, 2009); when there are conflicting signals in the data; for relatively small numbers of taxa; with simpler substitution models.

In order to test these hypotheses, we took subsets from several large alignments. These included a prokaryote data set (where lateral transfer is likely to be an important factor), a viral dataset (including sequences inferred to be recombinant), and datasets with deep divergences. We looked at subsets of each alignment, with different numbers of taxa, and sequence partitions of different length. For each subset, 1,000 random tree topologies were selected, and 1,000 sets of edge lengths were generated randomly as starting points for each of the 1,000 topologies. From each of these million starting points for each subset, edge lengths were optimised, and the resulting edge lengths and likelihood recorded and compared in order to detect multiple optima.

We found that the incidence of multiple optima varies widely between datasets, being found on anything from zero to 78% of trees on a given alignment. However, the topologies on which we detected multiple optima almost always have relatively low likelihood compared to other topologies on the same data.

3.3 Methods

3.3.1 Datasets

Four datasets were used. These were: the alignments of chloroplast protein-coding genes used by White et al. (2007); an alignment of first and second codon positions of mammalian mitochondrial protein-coding genes from Lin et al. (2002); an alignment of 228 hepatitis B virus strains based on that of Harrison et al. (2011); and a 100-taxon alignment of prokaryote protein sequences extracted from Puigbò et al. (2009). In total, 1,000 replicates were performed on each of over a million trees on subsets of these alignments, giving more than a billion optimisations overall.

For the mammalian dataset, 100 subsets of ten taxa were chosen at random. All the mammalian subsets were analysed using the Jukes-Cantor substitution model (JC—equal base frequencies and all substitutions equally likely—Jukes and Cantor, 1969).

For the chloroplasts, the same four overlapping subsets of 12 taxa (in order of increasing divergence: flowering plants, land plants, green plants and plastids) were used as in White et al. (2007). For each subset, six genes (*psb*B, *atp*B, *rbc*L, *psb*A, *pet*B and *psa*J) were analysed separately, along with the concatenated dataset. All subsets were analysed using the JC substitution model.

For the hepatitis B data, 100 random subsets of ten taxa were analysed using JC as well as two additional substitution models: Hasegawa-Kishino-Yano (HKY—variable base frequencies, one rate for transitions and another for transversions—Hasegawa et al., 1985), and the general time-reversible model (GTR—optimising both base frequencies and substitution rates—Tavaré, 1986). In order to investigate the effects of different numbers of taxa, a further 100 random subsets each of eight, nine, 12, 15 and 20 taxa were examined under the JC model. These random subsets had varying proportions of strains inferred by Harrison et al. (2011) to be recombinant, allowing us to investigate the effects of recombination on the incidence of multiple optima.

In order to keep runtimes to a feasible level, the prokaryote dataset consisted of only four clusters of orthologous genes (COG0098, COG0197, COG0250 and COG0008) selected from those analysed by Puigbò et al. (2009). Puigbò et al. (2009) inferred horizontal gene transfer between bacteria and archaea for one of these, COG0008. These were concatenated into a single alignment for the 100 prokaryote species selected by Puigbò et al. (2009). From this alignment, 100 subsets of nine taxa were selected at random and analysed under the LG substitution model (Le and Gascuel, 2008).

3.3.2 Computational experiments

For each subset of our datasets, an R script was run to:

1. generate 1,000 random tree topologies using the `rtree` function of ape (Paradis et al., 2004);

2. select 1,000 random sets of edge lengths as starting values for each of the 1,000 topologies;

3. from each of these million starting points, optimise the edge lengths (and other parameters, depending on the substitution model used, but not the topology) using the `optim.pml` function of phangorn version 1.6-1 (Schliep, 2011); and finally

4. count the number of distinct optima found on each topology. Two optima were considered distinct if the correlation between their edge length vectors was less than 0.95.

In addition to the random topologies, we also calculated the neighbour-joining tree and optimised its topology to infer a ML tree for each subset, then optimised the edge lengths of this ML topology from 1,000 random starting points. The random edge lengths for starting values for optimisation were drawn from an exponential distribution with rate $\lambda = 4$.

In addition, for the chloroplast dataset, we tested topologies in the neighbourhood of the ML tree inferred for each subset. Rather than selecting random topologies as described above, we began with the inferred ML tree and calculated all topologies one nearest-neighbour interchange (NNI) step away, then proceeded as described in steps 2 and 3 above. This test was carried out using a different version of phangorn (v. 1.7-1), because of a bug in v. 1.6-1 which caused it to crash when calculating topologies one NNI from the ML tree.

The R package phangorn was used due to its speed, and for the convenience of being able to carry out all steps of the analysis within R. However, in order test whether our results are software-dependent, we also carried out smaller-scale analyses using PhyML (Guindon et al., 2010) on selected subsets of each of the four main datasets. These analyses were performed as described above, but with 100 random topologies and 100 random sets of edge lengths for each topology. Optimisation of edge lengths and substitution rate parameters from each starting point was carried out using both phangorn and PhyML, and the resulting likelihood values were compared.

3.4 Results

For any given dataset, the two main results are: (1) the number (or the percentage) of randomly chosen topologies on which we observed multiple optima; and (2) the mean number of optima observed per topology, over all of the topologies examined. The second of these results must be interpreted with caution because, as we shall see, there are trees on which we detect relatively large numbers of optima, and these may not always represent genuinely distinct peaks in the likelihood landscape, so that this number is dependent on the threshold used to distinguish between optima.

In total, we found multiple optima on 6.2% of topologies examined for the chloroplast dataset, 4.2% of topologies for the mammalian mitochondrial data, 9.0% of ten-taxon topologies for the hepatitis B data, and 9.2% of topologies for the prokaryote data. Because the hepatitis B dataset showed more cases of multiple optima than either the chloroplast or the mitochondrial dataset, it was used to explore the effects of the substitution model and the number of taxa on the incidence of multiple optima.

We found that phangorn and PhyML gave the same optima in most cases, but with slightly different domains of attraction. That is, a given starting point does not always lead to the same optimum with both programs. This is likely due to the programs optimising the edge lengths in a different order. In a few cases, an optimum was found for a small number of starting points with one program, but not found with the other, perhaps due to the sparse sampling of starting points used.

3.4.1 Chloroplast dataset

Results for the chloroplast alignments are summarised in Table 3.1. As stated above, 6.2% of the chloroplast tree topologies tested were found to have multiple optima, although, as can be seen in Table 3.1, the number of topologies with multiple optima varied widely between the alignments examined. In the case of the concatenated chloroplast alignment, as well as five of the six genes we examined separately, multiple optima were found most often on the most anciently diverged subset (plastids), and least often on the most recently diverged subset of taxa (flowering plants). This pattern was particularly pronounced for the *rbc*L gene, with the plastid taxon-set showing a much higher incidence of multiple optima than any of the three more recently diverged subsets of taxa (multiple optima were found on 42% of all topologies examined for the plastid taxon-set). The proportion of topologies with multiple optima was higher for *psa*J than for any other gene, but the pattern of increasing incidence of multiple optima with increasing divergence time was absent. However, the *psa*J alignment is very short (138 nucleotides including gaps), so the level of stochastic error is likely to be high in comparison to any signal in the data. Importantly, perhaps, even in those cases where multiple optima were found on relatively large numbers of topologies, they were not found on the topologies with the highest likelihood.

We also tested topologies in the neighbourhood of the inferred ML tree for each subset of the chloroplast data. For each of three alignments (the flowering plant concatenated alignment, the plastid *rbc*L alignment, and the land plant *psa*J alignment), multiple optima were found on two of the topologies one nearest neighbour interchange from the inferred ML tree. In each of these cases, the optima found were effectively identical for the two topologies, because one edge of each tree had zero length. This meant that both topologies in effect represented the same tree with one unresolved trichotomy, and the edge length distributions on the remaining edges were identical for both topologies. In all three cases, the topologies with multiple op-

Table 3.1: Mean numbers of optima per topology, and the percentage of topologies on which more than one optimum was detected, for alignments from the chloroplast dataset.

		Mean no. of optima per topology					% of topologies with multiple optima				
Alignment	Length[a]	FP[b]	LP[b]	GP[b]	PL[b]	mean	FP[b]	LP[b]	GP[b]	PL[b]	mean
concat	22660-23175	1.025	1.028	1.185	1.253	1.1228	2.5	2.4	7.8	9.5	5.6
*psb*B	1527-1539	1.020	1.026	1.139	1.283	1.1170	2.0	1.8	7.1	10.1	5.3
*atp*B	1503-1515	1.029	1.041	1.166	1.249	1.1213	2.9	2.9	7.8	9.9	5.9
*rbc*L	1458-1506	1.014	1.020	1.082	1.677	1.1983	1.3	1.6	4.9	42.0	12.5
*psb*A	1062-1083	1.027	1.001	1.022	1.067	1.0293	2.7	0.1	1.2	4.4	2.1
*pet*B	648	1.001	1.006	1.090	1.157	1.0635	0.1	0.6	5.7	7.5	3.5
*psa*J	135-138	1.026	1.242	1.129	1.193	1.1475	2.6	18.4	6.5	8.5	9.0
mean		1.020	1.052	1.116	1.268	1.1142	2.0	4.0	5.9	13.1	6.2

[a] Alignment lengths (in nucleotides) vary between taxon-sets, so ranges are given.

[b] The taxon-sets are as described in White et al. (2007), in increasing order of divergence: flowering plants (FP), land plants (LP), green plants (GP), and plastids (PL).

tima had lower likelihood than any of the other topologies tested. For the remaining 25 alignments, no cases of multiple optima were detected on any of the topologies one nearest neighbour interchange from the inferred ML tree.

3.4.2 Mammalian mitochondrial dataset

Of the datasets examined, the mammalian mitochondrial dataset was the least prone to multiple optima. Multiple optima were detected in at least one of the 1,000 replicates for each of the 100 random ten-taxon subsets, but in relatively low numbers, being found on a total of 4,191 of the 100,000 tree topologies examined. For individual subsets, the number of topologies with multiple optima ranged from seven to 85 of the 1,000 topologies sampled. The overall mean number of optima found on the 100,000 topologies examined was 1.0616. Multiple optima were not found on the topology with the highest likelihood of those examined for any subset of taxa. Figure 3.1a shows the number of optima found for each topology sampled from the 100 ten-taxon subsets, plotted against the normalised ML value for that topology, that is, the difference between the maximum log-likelihood value of the inferred ML tree on the same dataset, and that of the topology being examined. As with the chloroplast data, more anciently diverged subsets of taxa (i.e. those for which the inferred ML tree had the greatest overall length) showed a higher incidence of multiple optima. This is shown in Figure 3.2a.

3.4.3 Hepatitis B dataset

The hepatitis B data generally showed more cases of multiple optima than either the mammalian mitochondrial or the chloroplast datasets, with multiple optima found on 9,015 of the 100,000 ten-taxon topologies examined under the JC model. For this reason this dataset was used to explore the effects of the substitution model and the number of taxa on the incidence of multiple optima. Figure 3.1b shows the number of optima found for each topology sampled from the 100 ten-taxon subsets, plotted against the normalised ML value for that topology. Overall, multiple optima were detected on 9.0% of topologies sampled on the ten-taxon subsets. For individual ten-taxon alignments, the number of topologies with multiple optima ranged from zero to 492 of the 1,000 topologies sampled. However, as can be seen, those trees with multiple optima on this dataset always had relatively low likelihood. Unlike the chloroplast and mitochondrial datasets, there was no apparent correlation between tree length and the prevalence of multiple optima, as can be seen in Figure 3.2b. As with the mammalian mitochondrial data, multiple optima were not found on the topology with the highest likelihood of those examined for any subset of taxa. Figure

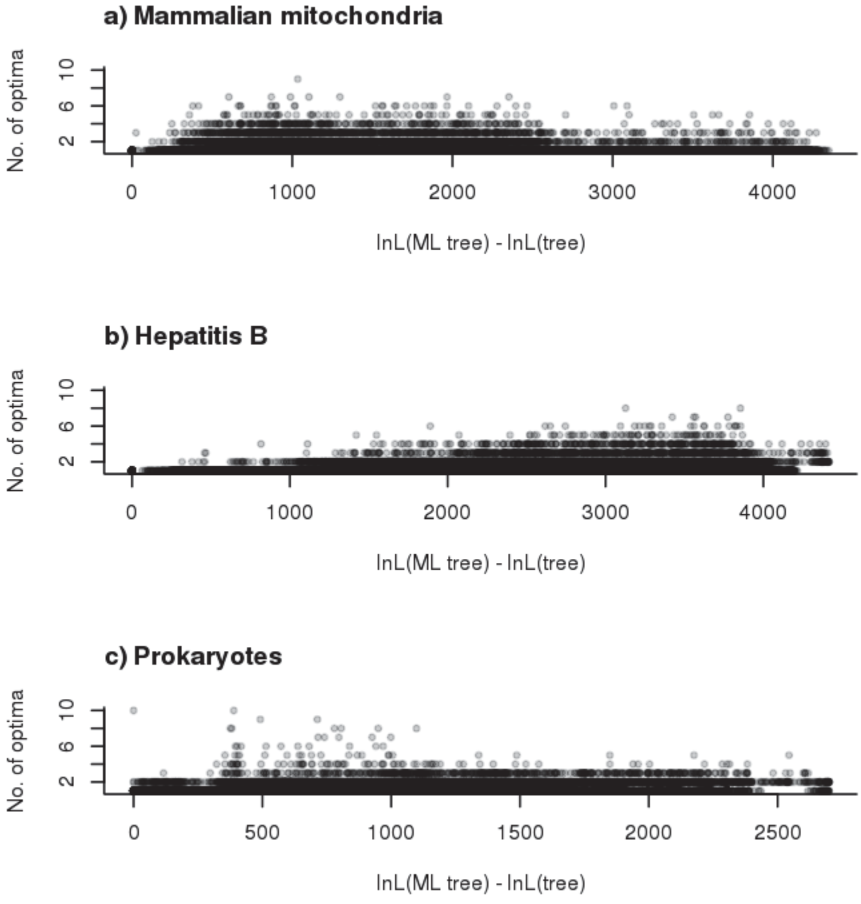


Figure 3.1: **Relative likelihood versus number of optima.**
For 100 subsets from each dataset, the number of optima found for each topology sampled, plotted against the maximum log-likelihood value for that topology relative to the maximum log-likelihood value of the ML tree inferred from the same alignment subset. Lower x-axis values represent better trees (i.e. higher likelihoods), with zero being the log-likelihood value of the inferred ML tree. Points are the numbers of optima for each of 1,000 random topologies on 100 randomly sampled subsets of (a) ten taxa from the mammalian mitochondrial dataset, (b) ten taxa from the hepatitis B alignment, and (c) nine taxa from the prokaryote alignment.

3.3 shows the effect of the number of taxa in a subset on the prevalence of multiple optima. There was little change in numbers of optima observed between eight and 12 taxa, but increasing numbers of optima with 15 and 20 taxa. The substitution model had a very small but statistically significant effect on the prevalence of multiple optima. Wilcoxon signed-rank tests (Wilcoxon, 1945) indicated that the overall mean number of optima per topology was greater under JC than HKY by 0.003 ($p < 0.01$), greater under JC than GTR by 0.002 ($p < 0.01$), and greater under GTR than HKY by 0.001 ($p < 0.01$).

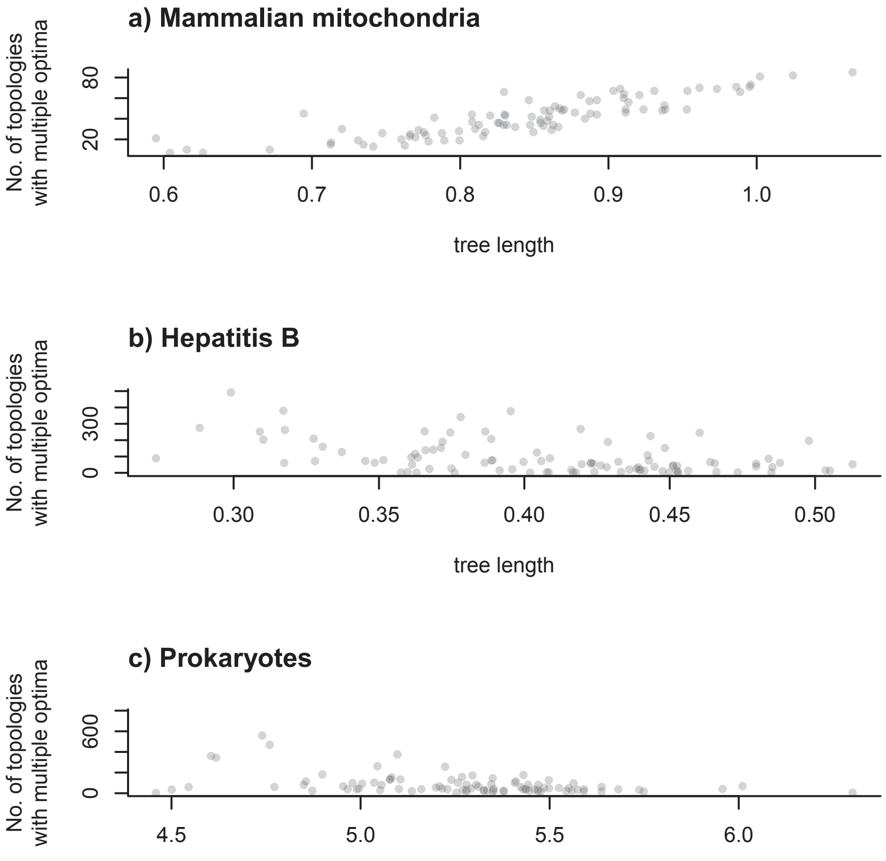


Figure 3.2: **Tree length versus number of topologies with multiple optima.** For 100 subsets from each dataset, the number of topologies (out of 1,000 random topologies) on which multiple optima were detected, plotted against the total length of the ML tree inferred from that alignment subset. Higher x-axis values represent longer trees, that is, deeper divergences. Plots are shown for 100 randomly sampled subsets of (a) ten taxa from the mammalian mitochondrial dataset (b) ten taxa from the hepatitis B alignment, and (c) nine taxa from the prokaryote alignment.

Surprisingly, as shown in Figure 3.4, the prevalence of multiple optima *decreased* as sequences from strains identified by Harrison et al. (2011) as being of recombinant origin were added. This is contrary to what we expected, as the addition of recombinant strains could lead to data similar to that generated by a mixture model, and we discuss this further below.

3.4.4 Prokaryote dataset

The prevalence of multiple optima in the prokaryote data was slightly higher than for the hepatitis B data, as can be seen in Figure 3.1c. Multiple optima were detected on 9,199 of the 100,000 nine-taxon tree topologies we examined. For individual nine-taxon alignments, the number of topologies with multiple optima ranged from three to 782 of the 1,000 topologies sampled. As with the hepatitis B data, there was no apparent correlation between the length of the inferred ML tree and the prevalence

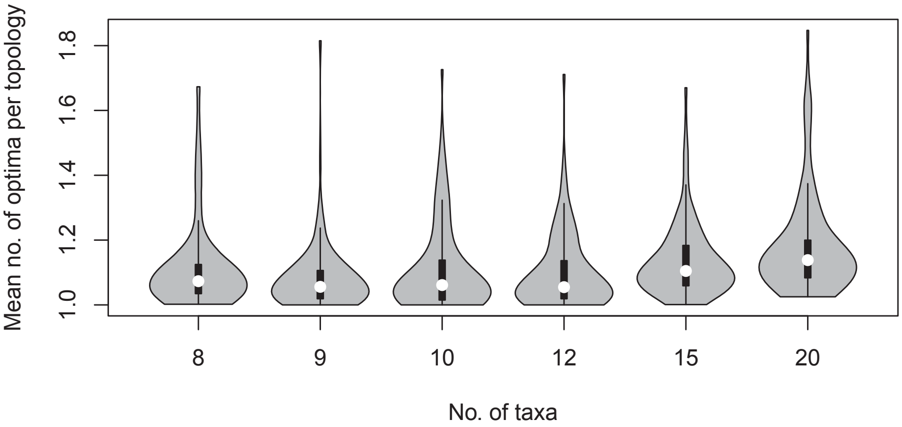


Figure 3.3: **Effect of number of taxa.**
Violin plots (Hintze and Nelson, 1998) of mean numbers of optima per topology sampled, for 100 random subsets each of eight, nine, ten, 12, 15 and 20 hepatitis B strains. The width of each 'violin' represents an estimate of the probability density of the mean number of optima per topology, and the embedded boxplot shows the mean and quartiles of the distribution. In each case, the distribution of mean numbers of optima is tightly clustered about the overall mean, but with a long tail representing a few taxon subsets for which relatively many topologies had multiple optima. The number of taxa appears to have little effect for up to 12 taxa, but there appears to be an increase in the numbers of optima observed for 15 and 20 taxa.

of multiple optima (see Figure 3.2c). Again, however, the topologies with multiple optima usually had relatively low likelihood. In two cases, however, multiple optima were found on the best topology examined (i.e., the topology for which the highest likelihood was observed).

In one of these two cases (taxon subset number 43), two optima were detected. Figure 3.5a shows the log-likelihood for the final trees reached from each of 1,000 starting points, plotted against the length of one edge of the corresponding tree. The point in the top centre of the plot has the highest likelihood, and accounts for 265 of the 1,000 replicates. However, a number of replicates reached optima with slightly lower likelihoods and different edge lengths. It is worth noting that the range of log-likelihood values is small, with maximum -11,361.96 and minimum -11,364.55. For comparison, the highest log-likelihood on any of the random topologies examined on the same data was -11,435.85. It seems likely that the optima detected in this case are not genuinely distinct, but that the slope of the likelihood landscape near the maximum is sufficiently shallow that the hill-climbing algorithm stops climbing prematurely. In such a case, the number of optima detected is somewhat arbitrary, because it depends on the threshold we set for distinguishing between two trees.

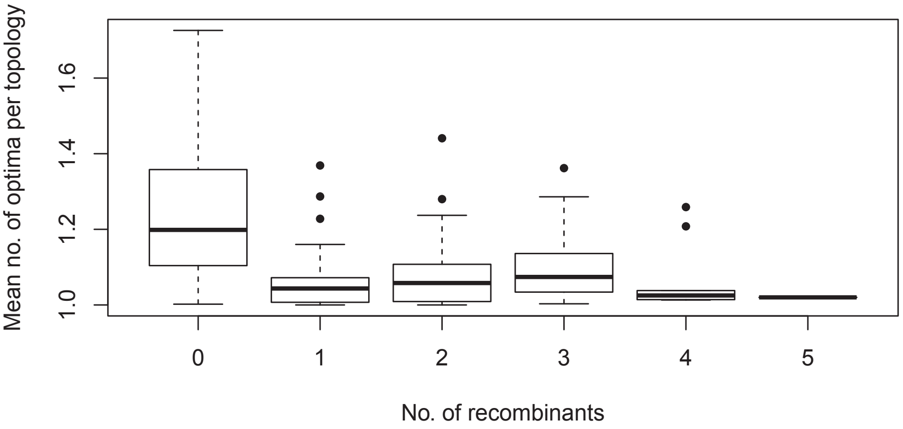


Figure 3.4: **Effect of recombinant sequences.**
Boxplot of mean numbers of optima per topology sampled, for 100 random subsets of ten hepatitis B strains, plotted against the number of recombinant strains in the subset. Surprisingly, there was a reduction in the number of optima per topology as the number of recombinant strains increased.

In the second case (taxon subset number 80), 991 out of 1,000 runs converged to the ML edge lengths for the tree, with the nine remaining runs each finding a separate optimum with considerably lower likelihood. Figure 3.5b shows the log-likelihood for the optima detected on this topology, plotted against the length of one edge of the corresponding tree. In this case, the local optima appear to be distinct, but very localised, as only one out of a thousand random starting points fell in the domain of attraction of each.

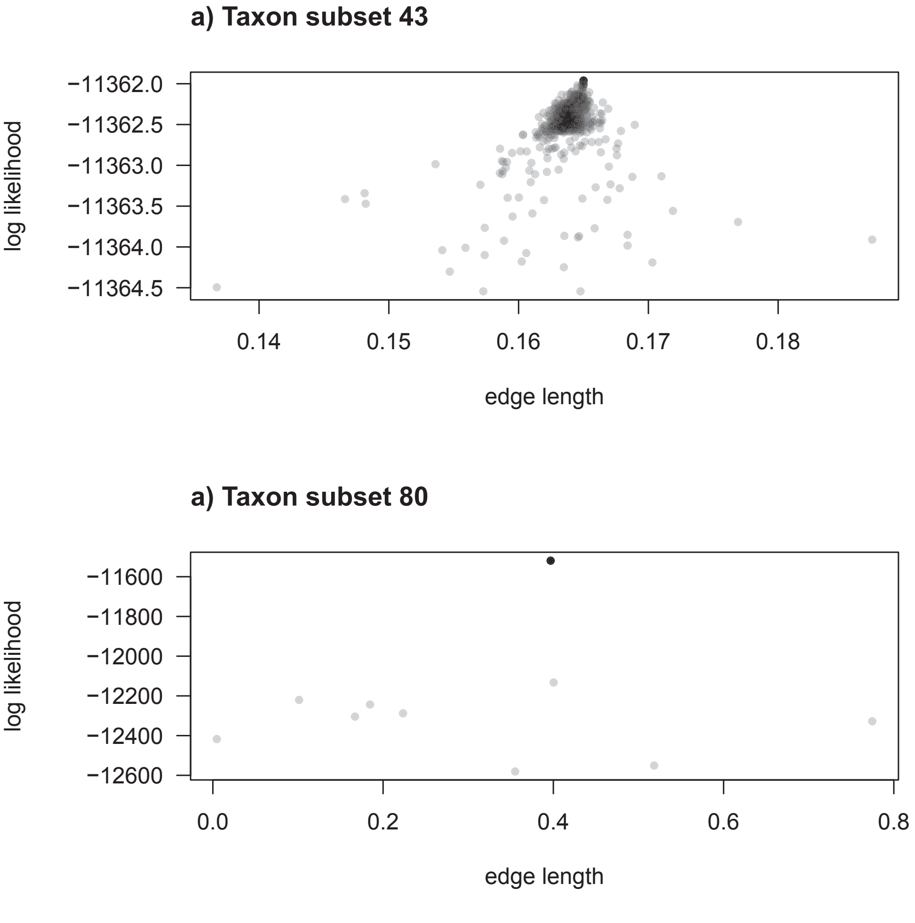


Figure 3.5: **ML trees with multiple optima.**
The log-likelihood reached by the 1000 replicates on the ML tree inferred for taxon subsets number 43 (a) and 80 (b) of the prokaryote data, plotted against the length of one edge of the corresponding tree. In both cases, all fifteen edges show a similar pattern, so the edge used for the figure was chosen arbitrarily. For taxon subset 43, a single main optimum can be seen, with a scatter of minor optima with a range of edge lengths and slightly lower likelihood values. For taxon subset 80, a single main optimum accounts for 991 of 1000 starting points, with each of the remaining nine giving rise to a distinct local optimum with much lower likelihood.

3.4.5 General results

An important point is that, in general, where we have found multiple optima, they have been on relatively poor topologies for the data (see Fig. 3.1). That is, for the topology with highest likelihood on any taxon-set, in almost all cases only a single optimum was found, and topologies on which we detected multiple optima usually all had lower likelihood. This could, in theory, have been a sampling effect, because the distribution of likelihoods on topologies is such that, for any given alignment, there are few topologies with high likelihood values and many more with low to intermediate likelihoods. However, we applied both the Mann-Whitney-Wilcoxon test (MWW, Mann and Whitney, 1947) and the Kolmogorov-Smirnov test (KS, Massey, 1951) to each of our sets of 100 n-taxon datasets for the mammalian, hepatitis B and prokaryote data. The likelihoods on topologies with multiple optima and those for topologies with a single optimum were found to have been sampled from different distributions (KS), with the topologies with a single optimum having higher mean log-likelihood (MWW). The differences between distributions are statistically significant to level $\alpha = 0.001$.

Occasionally, we found topologies with large numbers of optima, such as the case described above for taxon subset number 43 of the prokaryote dataset. These may be cases where, rather than having discrete optima, the likelihood function has a ridge of solutions with near-equal likelihood (or perhaps a terrace analogous to those described in Sanderson et al. (2011), although the Sanderson results apply to optimisation of topology rather than edge lengths). Alternatively, the surface of the likelihood landscape could be either sufficiently rough to prevent the hill-climbing algorithm from finding slightly better solutions in the neighbourhood, or so smooth that the likelihoods of neighbouring solutions differ by less than the cutoff used by the algorithm to decide when to stop searching.

3.5 Discussion

Overall, our results with real biological data are similar to those found by Rogers and Swofford (1999) using sequences simulated on a single tree. That is, topologies with high likelihood rarely had multiple optima, even for alignments with short sequences (as seen with some of the chloroplast alignments), and in cases where we would expect conflicting signals (such as with recombinant strains of hepatitis B). These results provide some reassurance that ML tree reconstruction methods are not likely to be compromised by the presence of multiple local optima of likelihood on the ML topology, except perhaps in very rare instances.

As expected, topologies with multiple optima were found more frequently when sequences were short, and with simpler substitution models. The number of taxa, however, had no discernible effect.

The simulation study of Rogers and Swofford (1999) found multiple optima most often when the trees used to generate the data had very long branches. We found a similar effect, increasing prevalence of multiple optima as the overall length of the inferred ML tree increased, on our mitochondrial and chloroplast datasets but not on the hepatitis B or prokaryote data. It is interesting to note that the prokaryote trees are considerably longer than those for any other dataset, and the numbers of multiple optima observed are higher than for any other dataset.

Contrary to our expectations, when recombinant sequences were included, multiple optima were found less frequently. This was surprising, as we might have expected the presence of recombinant sequences to give data similar to that generated by a mixture model, with different parts of the sequence having different evolutionary histories and supporting different tree topologies. However, Matsen and Steel (2007) have shown (theoretically) that mixtures of two trees can mimic a different tree topology. In this sense, recombination (or, by the same token, concatenation of sequences with differing evolutionary histories) may reduce the incidence of multiple optima, but will not necessarily lead to the right tree.

It is always possible that some of our results may be specific to the order in which the phangorn package optimises edge lengths on a tree, although our tests using

PhyML suggest that this does not affect the number of optima found. The search strategies used by some other implementations of ML, which optimise multiple edge lengths concurrently, may be able to escape local optima in some cases, so our results may represent a worst-case scenario, in that some other ML packages could be less likely to be affected by multiple optima than phangorn.

In summary, our results suggest that multiple optima of likelihood on a given tree topology are not very likely to pose a problem in phylogenetics. Even if multiple optima are encountered on some topologies during optimisation, we can expect that other topologies with higher likelihood will still be found.

Bibliography

B. Chor, M. D. Hendy, B. R. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Mol Biol Evol*, 17(10):1529–1541, 2000.

K. Fukami and Y. Tateno. On the maximum likelihood method for estimating molecular trees: uniqueness of the likelihood point. *J Mol Evol*, 28(5):460–464, 1989.

S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*, 59(3):307–321, 2010.

A. Harrison, P. Lemey, M. Hurles, C. Moyes, S. Horn, J. Pryor, J. Malani, M. Supuri, A. Masta, B. Teriboriki, T. Toatu, D. Penny, A. Rambaut, and B. Shapiro. Genomic analysis of hepatitis B virus reveals antigen state and genotype as sources of evolutionary rate variation. *Viruses*, 3(2):83–101, 2011.

M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174, 1985.

M. D. Hendy, M. A. Steel, D. Penny, and I. M. Henderson. Families of trees and consensus. In H. H. Bock, editor, *Classification and Related Methods of Data Analysis*, pages 355–362. Elsevier Science Publishers B.V., North Holland, 1988.

J. L. Hintze and R. D. Nelson. Violin plots: a box plot-density trace synergism. *Am Stat*, 52(2):181–184, 1998.

J. P. Huelsenbeck and J. P. Bollback. Application of the likelihood function in phylogenetic analysis. In D. J. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, chapter 15, pages 460–488. John Wiley and Sons, Inc., New York, 3rd edition, 2007.

T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian protein metabolism*, pages 21–132. Academic Press, New York, 1969.

S. Q. Le and O. Gascuel. An improved general amino acid replacement matrix. *Mol Biol Evol*, 25(7):1307–1320, 2008.

Y.-H. Lin, P. A. McLenachan, A. R. Gore, M. J. Phillips, R. Ota, M. D. Hendy, and D. Penny. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. *Mol Biol Evol*, 19(12): 2060–2070, 2002.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat*, 18(1):50–60, 1947.

F. J. J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc*, 46 (253):68–78, 1951.

F. A. Matsen and M. Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Syst Biol*, 56(5):767–775, 2007.

D. Money and S. Whelan. Characterizing the phylogenetic tree-search problem. *Syst Biol*, 61(2):228–239, 2012.

D. A. Morrison. Increasing the efficiency of searches for the maximum likelihood tree in a phylogenetic analysis of up to 150 nucleotide sequences. *Syst Biol*, 56(6): 988–1010, 2007.

E. Paradis, J. Claude, and K. Strimmer. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

P. Puigbò, Y. I. Wolf, and E. V. Koonin. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J Biol*, 8(6):59, 2009.

J. S. Rogers and D. L. Swofford. Multiple local maxima for likelihoods of phylogenetic trees: a simulation study. *Mol Biol Evol*, 16(8):1079–1085, 1999.

L. A. Salter. Complexity of the likelihood surface for a large DNA dataset. *Syst Biol*, 50(6):970–978, 2001.

M. J. Sanderson, M. M. McMahon, and M. Steel. Terraces in phylogenetic tree space. *Science*, 333(6041):448–450, 2011.

K. P. Schliep. *Some Applications of Statistical Phylogenetics*. PhD thesis, Massey University, 2009.

K. P. Schliep. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011.

M. Steel. The maximum likelihood point for a phylogenetic tree is not unique. *Syst Biol*, 43(4):560–564, 1994.

S. Tavaré. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. In R. M. Miura, editor, *Some mathematical questions in biologyâĂŤDNA sequence analysis.*, volume 17, pages 57–86. American Mathematical Society, Providence (RI), 1986.

W. T. White, S. F. Hills, R. Gaddam, B. R. Holland, and D. Penny. Treeness triangles: visualizing the loss of phylogenetic signal. *Mol Biol Evol*, 24(9):2029–2039, 2007.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bull*, 1(6):80–83, 1945.

Chapter 4

Investigating mutational mechanisms

4.1 Preamble

Having assembled sequences and constructed a phylogenetic tree, we are now able to ask a whole range of questions about the evolutionary process that has taken place on that tree to produce the sequences we observe. Perhaps the most fundamental of these questions is simply: what process (or processes) has given rise to the pattern of substitutions we observe, and has that process been constant, or has it varied between different parts of the tree. The changes we observe in the sequence data result from mutations, that is, either damage to the DNA or errors in replication, that have not been repaired and that have become fixed in the population. We assume that mutations occur randomly, but this does not imply that they occur with equal probability at all sites of the sequence, or that the different types of mutation all occur at the same rate.

In order to examine these underlying mutational processes, we can estimate the parameters of a Markov model of sequence evolution. Several methods have been proposed for estimating the instantaneous rate matrix of the model, **Q**, and these have been shown to accurately reconstruct **Q** from simulated sequences. Here we use two of these methods to estimate **Q** from several large alignments on phylogenetic trees, in the hope of detecting changes in the underlying mutational processes between different lineages of the trees.

Although we find some consistency between classes of nucleotide change, it is not possible with this data (except perhaps for the mitochondrial alignment) to group taxa into similar classes. These results tend to suggest that underlying mutational mechanisms cannot yet be untangled from other causes of nucleotide substitution using the alignments considered here, and further work will be needed using genome-scale alignments.

This chapter presents a draft manuscript, currently solely authored by myself, although other authors may be added before it is submitted for publication.

The Python scripts used in this manuscript are now freely available for download from `http://sourceforge.net/p/qmatrices`.

Investigating mutational mechanisms by estimating nucleotide substitution rate matrices from alignments

Bennet J McComish

Abstract

The nucleotide substitution rate matrix, **Q**, is a key parameter of molecular evolution. **Q** is often assumed both to be symmetric, and to be constant over a phylogenetic tree, but in reality there are a number of mutational processes that are known to affect nucleotide composition by differentially affecting individual entries of **Q**. Variations in **Q** are therefore important in understanding fundamental processes of molecular evolution, and are also expected to have important effects on phylogeny reconstruction. Several methods for inferring **Q** have been developed, and a number of these have been tested by simulation, but none appear to have been applied to biological data on a large scale before now. Applying these methods to large alignments allows us to empirically estimate **Q** matrices for different lineages, in principle enabling us to detect any changes in mutational processes in particular groups of taxa. Our results suggest that with the alignments used here, except perhaps for the mitochondrial sequences, the underlying mutational processes can not be untangled from other phenomena such as selection, strand asymmetry or site-specific differences in mutation rate. It may be possible to examine these processes more effectively using genome-scale alignments, and this is a promising avenue for further research.

4.2 Introduction

The nucleotide substitution rate matrix, conventionally denoted **Q**, is the instantaneous rate matrix of the continuous-time Markov process used to describe molecular evolution, and contains the rates of change from each type of nucleotide (A, C, G or T) to each other nucleotide, so that

$$\mathbf{Q} = \begin{bmatrix} - & q_{AT} & q_{AC} & q_{AG} \\ q_{TA} & - & q_{TC} & q_{TG} \\ q_{CA} & q_{CT} & - & q_{CG} \\ q_{GA} & q_{GT} & q_{GC} & - \end{bmatrix}$$

where q_{ij} is the rate of replacement of nucleotide i by nucleotide j, and the negative entries on the diagonal are defined by the mathematical requirement that each row must sum to zero (see Liò and Goldman, 1998, for a review of models of molecular evolution). For simplicity, we do not consider here insertions or deletions of nucleotides.

Q is related to the transition matrix **P** by the equation

$$\mathbf{P} = e^{\mathbf{Q}t}$$

where **P** contains the probabilities of transition from one base to another over time t. In theory, since **Q** is an instantaneous rate matrix, it can vary continuously along any edge of a tree. However, we can (usually) only observe sequences at the leaves, and infer ancestral sequences at internal nodes of the tree, so that we estimate a single **Q** for each edge. A change of rate within an edge can affect estimates of divergence times, and this is discussed in Penny et al. (1998).

Despite its central importance in modelling molecular evolution, **Q** is generally treated as a nuisance parameter in phylogenetic reconstruction, and the process is often assumed to be homogeneous, stationary and reversible (see definitions and discussion in Jermiin et al., 2008). This would imply both that **Q** was constant over all edges of the tree, and that the nucleotide frequencies are approximately the same for all taxa. In reality, however, nucleotide composition can vary significantly between genomes (see e.g. Muto and Osawa, 1987), implying that **Q** does not always satisfy these conditions. An example of genomes with different nucleotide composition affecting phylogeny reconstruction is the mitochondrial genome of the hedgehog and its relative the gymnure—these have higher A + T content than other mammals, making it difficult to establish their position in the mammalian tree with any confidence (see Lin et al., 2002; Ota and Penny, 2003). Furthermore, it is likely that the changes we observe in **Q** are caused, at least in part, by changes in the underlying mutational processes, such as changes in the relative efficacy of different DNA replication and repair mechanisms (Herr et al., 2011). If we were able to identify such changes in the mutational processes, these should provide useful insights into broader processes of molecular evolution.

Our basic assumption is that the mutations we observe are a reflection of the error rate in DNA replication and repair. At least 19 distinct DNA polymerases are known in eukaryotes (Hübscher et al., 2002), each with a different role in DNA replication, proofreading and repair (as well as recombination, cell-cycle regulation, and telomere maintenance), and each with its own level of fidelity. In addition to the error-checking mechanisms active during replication, there is on-going repair of double-stranded breaks, cytosine methylation-deamination, and other forms of DNA damage. The net result is estimated to be an average error rate of 10^{-9} to 10^{-10} per base-pair of DNA synthesised (Herr et al., 2011).

Observed substitution patterns do not necessarily reflect the true mutation patterns because of several processes that may confound efforts to identify changes in fundamental mutational processes. These include selection, codon usage bias (Plotkin and Kudla, 2011; Andersson and Kurland, 1990), and strand asymmetries coupled to both transcription and replication (see e.g. Green et al., 2003; Touchon et al., 2005). Chen et al. (2004), however, reported that codon usage bias is determined primarily by genome-wide mutational processes, with selective pressures exerting a much smaller effect. There are also differences in mutation rates in particular regions of the genome (for example, higher rates in the DNA located between nucleosomes), and the mutation rate varies with replication timing (Agier and Fischer, 2012). Similarly, the locations of mutations may not be strictly random, in that repair of double-stranded breaks may lead to a higher rate of mutation in the surrounding region. These phenomena may also affect the relative rates of different substitution types, as they have been shown to in the case of replication timing (Agier and Fischer, 2012). We assume that the substitutions we observe are those cases where an incorrect nucleotide is incorporated, during either DNA replication or repair, and where the resulting mutation is not eliminated by selection or genetic drift. Thus each of the observed substitution rates that are the elements of $\mathbf{Q}$ represents the sum of a number of different mechanisms. Note that neither the organism nor its enzymes can measure their own error rates, so that any control of these error rates must come in the form of indirect selection on the enzymes involved, as increased error rates lead to mutations elsewhere in the genome that are then selected against.

We aim to detect some of these potentially confounding effects, and to isolate the underlying mutational processes, by independently estimating $\mathbf{Q}$ from protein-coding sequences, pseudogenes and other non-coding sequences. In principle, if we consistently find differences in relative rates of different substitution types between lineages for several alignments, it seems reasonable to assume that these reflect differences in the underlying mutational processes. Pseudogenes have been used to infer the frequencies of the different substitution types (along with insertions and deletions) on a large scale in the human genome (Zhang and Gerstein, 2003). Here, however, we are interested in observing changes in those frequencies between lineages over evolutionary time.

A number of methods have been developed for calculating empirical estimates of $\mathbf{Q}$, which we denote $\widehat{\mathbf{Q}}$, on edges of a given phylogenetic tree, and some of these methods have been tested by simulation by Oscamou et al. (2008). Most methods first estimate the transition matrix $\widehat{\mathbf{P}}$ for a given edge, whose entries represent the estimated probabilities of changing from one base to another over the time t spanned by the edge, and $\widehat{\mathbf{Q}}$ is then calculated using the equation $\widehat{\mathbf{P}} = e^{\widehat{\mathbf{Q}}t}$. In practice, we do not usually know the value of t, so that only the elements of $\widehat{\mathbf{Q}}t$ are identifiable, that is, we cannot distinguish the rates along an edge from the length of time elapsed (unless we have independent estimates of divergence times). While Oscamou et al. (2008) found only relatively small differences in accuracy between the methods tested, computational speed differed by orders of magnitude. Overall, they recommended (1) the method of Gojobori et al. (1982) for long sequences and (2) the method of Goldman et al. (1996) for shorter sequences (due to the speed of these simple methods), and (3) the maximum-likelihood-based method of Barry and Hartigan (1987), which provides greater accuracy for very long sequences at the expense of substantially longer computation time.

In the current study, as recommended by Oscamou et al. (2008), we used the method of Gojobori et al. (1982), which simply counts the frequencies of different substitution types, and the maximum likelihood method of Barry and Hartigan (1987). Although Oscamou et al. (2008) found the method of Barry and Hartigan (1987) to be much slower than that of Gojobori et al. (1982), runtimes on our data ranged from a few seconds up to nine minutes for our largest alignment. As such, we did not find computation time to be a major problem for this method. We applied these methods to a variety of datasets, each comprising one or more alignments and a tree, as outlined in Section 4.3 below. Because we can only estimate **Q** on a given tree, our results are dependent on the phylogeny that we assume. Our analysis is therefore restricted to datasets for which there is a relatively good consensus as to the phylogeny.

4.3 Methods

As mentioned above, we have used two methods to estimate **Q**. Following Oscamou et al. (2008), we refer to these as 'Gojobori' (Gojobori et al., 1982) and 'BH' (Barry and Hartigan, 1987). For both methods we assume that sites are independent and identically distributed, so that there is a single **Q** for each edge of a tree.

Initial analyses were carried out using the Gojobori method, as implemented in the PyCogent toolkit (Knight et al., 2007). This method uses rooted triples of sequences ((X,Y),Z). If, for a given position, one of the two sister taxa, X, has the same base as the outgroup Z, and the other sister, Y, has a different base, we can infer a change between Y and its common ancestor with X. This leads to a directional (and not necessarily time-reversible) $\widehat{\mathbf{P}}$, which can then be used to obtain $\widehat{\mathbf{Q}}t$. This method, however, suffers from the drawback that $\widehat{\mathbf{Q}}t$ cannot be measured directly for internal edges of a tree. In theory, $\widehat{\mathbf{Q}}t$ for adjacent edges should be additive, and we could use different rooted triplets to obtain $\widehat{\mathbf{Q}}t$ for paths within the tree, and hence for internal edges. However, this method does not correct for multiple changes, so saturation becomes a problem, especially when we consider older divergences, and particularly for those classes of substitution with higher rates (such as transitions).

The second method used to estimate $\mathbf{Q}$, BH, gives maximum likelihood estimates of joint probability matrices, $\widehat{\mathbf{J}}$, along each edge of a tree of any size, including the internal edges. Each element of $\widehat{\mathbf{J}}_{AB}$, the joint probability matrix for the edge connecting two nodes A and B, is an estimate of the probability that a given site has one nucleotide at node A and another at node B. All elements in $\widehat{\mathbf{J}}_{AB}$ must sum to one, as it is a joint probability matrix. $\widehat{\mathbf{J}}$ can easily be converted to $\widehat{\mathbf{P}}$, and hence $\widehat{\mathbf{Q}}t$. We used the BH algorithm of Jayaswal et al. (2005), as provided at the following website: `http://sydney.edu.au/science/biology/about_us/honorary_staff/jermiin_lars/BH.shtml`. (Unfortunately, different authors in this field use different notation, which can lead to some confusion—a case in point being that the joint probability matrix that we refer to here as $\mathbf{J}$ is called $\mathbf{Q}$ by Jayaswal et al.) A custom Python script was used to run BH using an alignment and a tree in Newick format as input, to extract $\widehat{\mathbf{J}}$ for each edge of the tree from the output, and then to calculate $\widehat{\mathbf{Q}}t$ from each $\widehat{\mathbf{J}}$. The same script also produced 100 bootstrap replicates from the alignment and returned the mean, variance, and maximum and minimum values for each element of $\widehat{\mathbf{Q}}t$ on each edge of the tree.

A potentially serious drawback of the BH method is that its parameters are statistically non-identifiable, that is, different parameter values (of the joint probability matrices, in this case) can give rise to identical distributions of observed data. Zou et al. (2011) have shown that different permutations of the rows of some of the joint probability matrices of the BH model can give exactly the same likelihood. However, the same authors (Zou et al., 2012) have provided a method for fitting a non-stationary general time-reversible (NSGTR) model (which is identifiable for all parameters) to the data. The parameters of the NSGTR model can then be used to identify the permutation of the BH model that provides the closest fit to the NSGTR model. We applied this method to our BH results, and used the resulting joint probability matrix estimates as initial estimates for the BH program—we expect this to ensure that BH converges to the global optimum rather than to a local optimum.

Another potential weakness of the BH model is that it can over-parameterise the data when applied to sequences that have evolved under globally stationary, reversible and homogeneous conditions. In order to check whether this might be a problem with our data, we applied the matched-pairs test of homogeneity developed by Ababneh et al. (2006) to each of the alignments used.

For ease of calculation, and since we are only interested in the 12 off-diagonal elements of $\widehat{\mathbf{Q}}$, we express these as a vector. Because we cannot distinguish between complementary substitutions on the two strands (so, for example, an A → C substitution on the non-sequenced strand will be observed as a T → G substitution on the opposite, sequenced, strand), substitutions between complementary nucleotide pairs (A:T and C:G) are added together, e.g. $q_{A:T \to C:G} = q_{AC} + q_{TG}$. This leaves six substitution types, and so we have the vector

$$\mathbf{q} = (q_{A:T \to T:A}, q_{A:T \to C:G}, q_{A:T \to G:C}, q_{C:G \to A:T}, q_{C:G \to T:A}, q_{C:G \to G:C}).$$

For both the Gojobori and BH methods, each $\mathbf{q}$ vector was normalised by dividing by $|tr(\widehat{\mathbf{Q}})|$ (that is, the absolute value of the sum of the diagonal entries of $\widehat{\mathbf{Q}}$), so that the elements of each resulting vector $\widehat{\mathbf{q}}$ sum to one. This made it easier to compare $\widehat{\mathbf{q}}$ for edges of different lengths in the absence of information about divergence times.

4.3.1 Datasets

We selected alignments comprising several different types of sequence, including coding and non-coding, and nuclear and mitochondrial sequences. Alignments with overlapping sets of taxa were chosen, so that identical (or in some cases nearly identical) subsets of taxa could be used for several types of sequence.

We took three sets of taxa from six alignments where the same taxa or close relatives were available, giving a total of 14 datasets, as listed in Table 4.1. The six

Table 4.1: **Lengths of the alignments used in this study.** Numbers of variable sites are shown in parentheses.

Alignment	apes		primates		mammals	
$\psi\eta$-globin[a]	6,166	(957)	-	-	-	-
Primate mitochondrial[b]	10,053	(3,807)	10,053	(5,138)	-	-
Primate nuclear[b]	5,403	(727)	-	-	-	-
Mammal coding[c]	17,763	(869)	15,838	(4,062)	4,652	(2,233)
fourfold degenerate sites	2,292	(271)	1,831	(1,044)	405	(334)
Mammal non-coding[c]	72,275	(5,464)	46,852	(15,787)	4,689	(3,214)
Laurasiatherian[d]	-	-	-	-	7,306	(3,301)

[a] From Miyamoto et al. (1987).

[b] From Fabre et al. (2009).

[c] From Prasad et al. (2008).

[d] From Lin et al. (2002).

alignments were as follows: (1) The primate $\psi\eta$-globin pseudogene alignment of Miyamoto et al. (1987), as provided at `http://abacus.gene.ucl.ac.uk/ziheng/data.html`. (2) Mitochondrial and (3) nuclear gene alignments from the primate gene supermatrix of Fabre et al. (2009) were concatenated separately. (Because mitochondrial genome replication uses polymerase γ while nuclear DNA replication uses polymerases α, δ and ϵ, we do not expect mitochondrial and nuclear genomes to share the same mutational process.) The nuclear genes were used only for our smallest taxon-set, because most genes are missing for many taxa in the alignment, and neither the Gojobori nor the BH method can accommodate missing sites in the implementations used in this study. (4) The alignment of first and second codon positions of mammalian mitochondrial protein-coding genes of Lin et al. (2002) was used, as available at `http://www.massey.ac.nz/~imbs/download.htm`. Finally, (5) the mammalian nuclear coding and (6) non-coding alignments of Prasad et al. (2008) were used. For the nuclear protein-coding alignment of Prasad et al. (2008) we also extracted all sites that were fourfold degenerate third codon positions in all taxa, and analysed these separately.

From these six alignments, we selected three subsets of taxa. In some cases where one taxon was not present in one of the alignments, we substituted a closely related species. The first taxon-set corresponded to that of Miyamoto et al. (1987), i.e. human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), and rhesus macaque (*Macaca mulatta*), with a New World monkey as outgroup. We will refer to this first taxon-set as "apes". This taxon-set was extracted from all alignments except that of Lin et al. (2002). In the alignments of Miyamoto et al. (1987) and Fabre et al. (2009), the New World monkey was a spider monkey (*Ateles* sp.), while in the alignments of Prasad et al. (2008) an owl monkey (*Aotus nancymai*) was substituted. The second taxon-set consisted of nine primates with rat (*Rattus norvegicus*), mouse (*Mus musculus*) and rabbit (*Oryctolagus cuniculus*) as the outgroup. We will refer to this second taxon-set as "primates". This was extracted from the mitochondrial alignment of Fabre et al. (2009) and both nuclear alignments of Prasad et al. (2008), with galago (*Otolemur garnettii*) being substituted for loris (*Nycticebus coucang*) in the latter. Finally, a set of 20 mammals was extracted from the mitochondrial alignment of Lin et al. (2002) and from both nuclear alignments of Prasad et al. (2008), with the capuchin monkey (*Cebus albifrons*) of Lin et al. (2002) being replaced by a squirrel monkey (*Saimiri boliviensis*) in Prasad et al. (2008). We will refer to this final taxon-set as “mammals”.

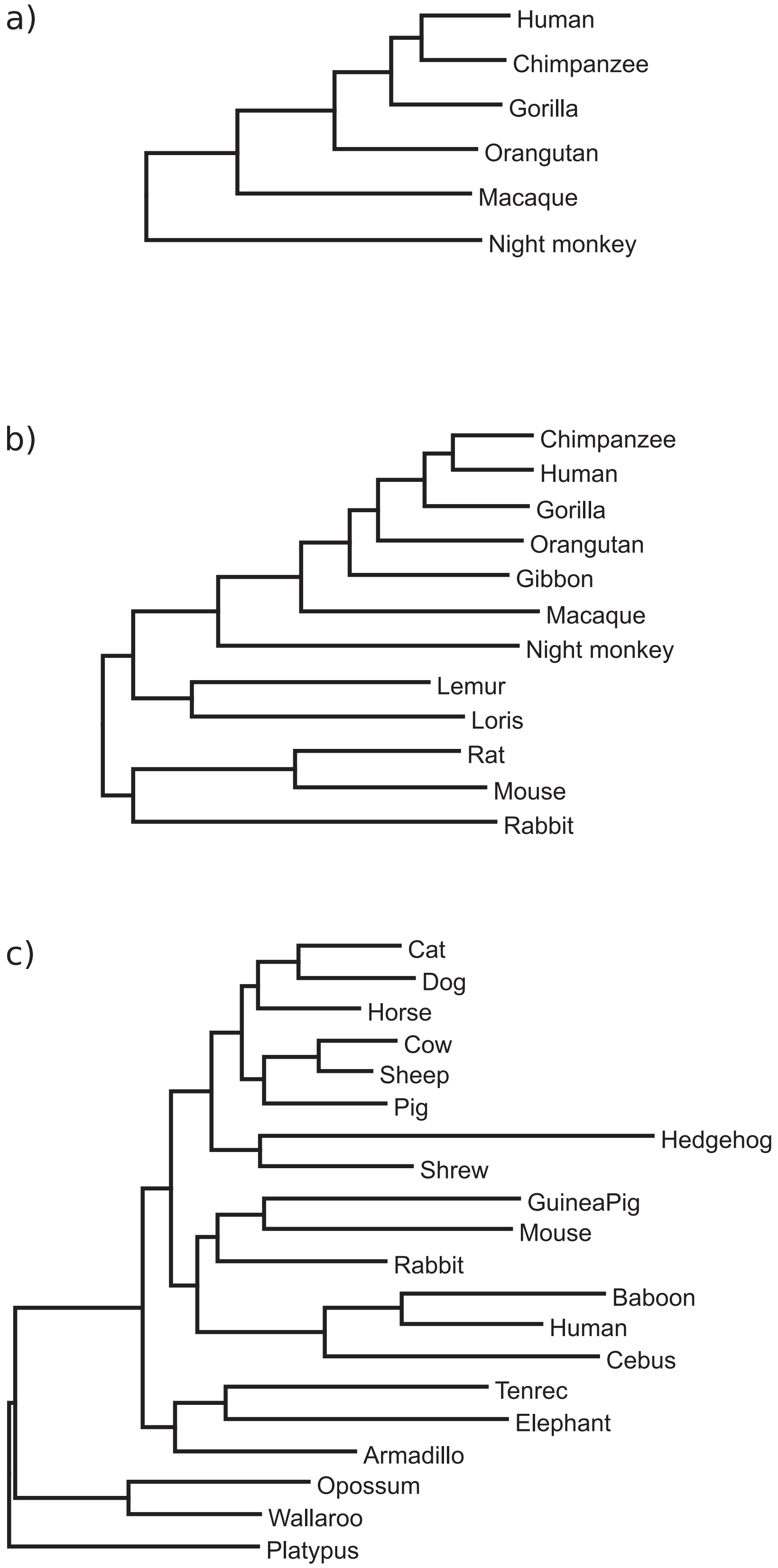
The trees corresponding to our three subsets of taxa are shown in Figure 4.1. The tree for the ape taxon-set is that of Miyamoto et al. (1987). The primate and mammal trees are derived from those of Fabre et al. (2009) and Lin et al. (2002), respectively. These trees were selected for the relatively high degree of consensus in the literature as to their topologies. In all cases, columns with gaps or ambiguous nucleotides were removed from the final alignments.

4.3.2 Visualisation

In order to compare six parameters for each edge of a tree, between multiple datasets, we experimented with several visualisation techniques. The first was simply to use bar charts for each of the six substitution types.

A second approach was to produce heatmaps in the statistical package R (R Development Core Team, 2011). The `heatmap` function of R represents the values of a matrix as colours. The function also computes distances between the rows (and columns) of a matrix, then performs hierarchical cluster analysis on these distances to give dendrograms for the rows and columns, and reorders the matrix to fit the dendrograms. In this case, Euclidean distances were used and clustering was performed

Figure 4.1 *(facing page)*: **Trees for the taxon-sets used in this study.**
(a) The ape taxon-set of Miyamoto et al. (1987). As well as the $\psi\eta$-globin pseudogene alignment of Miyamoto et al. (1987), the same taxon-set was extracted from the nuclear and mitochondrial alignments of Fabre et al. (2009) and, with an owl monkey in place of the spider monkey, from both the coding and non-coding nuclear alignments of Prasad et al. (2008).
(b) A set of nine primates with rat, mouse and rabbit as the outgroup, extracted from the mitochondrial alignment of Fabre et al. (2009). The same taxon-set was extracted from both alignments of Prasad et al. (2008), with the loris replaced by a galago.
(c) A set of 20 mammals taken from the mitochondrial protein-coding alignment of Lin et al. (2002) and both nuclear alignments of Prasad et al. (2008), with a squirrel monkey substituted for the capuchin in the latter.


a)
Human
Chimpanzee
Gorilla
Orangutan
Macaque
Night monkey
b)
Chimpanzee
Human
Gorilla
Orangutan
Gibbon
Macaque
Night monkey
Lemur
Loris
Rat
Mouse
Rabbit
c)
Cat
Dog
Horse
Cow
Sheep
Pig
Hedgehog
Shrew
GuineaPig
Mouse
Rabbit
Baboon
Human
Cebus
Tenrec
Elephant
Armadillo
Opossum
Wallaroo
Platypus

using the complete linkage method (the default settings for `heatmap`).

A third approach, aimed at highlighting similarities in process between different lineages, was to use principal component analysis (PCA), using the `prcomp` function of R, and to plot the first two principal components against each other. PCA converts a set of observations into a set of values of linearly uncorrelated variables called principal components, in such a way that the first component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it be uncorrelated with the preceding components. This allows us to project the data into two dimensions while retaining as much of the variance in the dataset as possible.

4.4 Results

Overall, the BH and Gojobori methods gave very similar results, especially for pendant edges. Correlations between $\widehat{\mathbf{Q}}$ from Gojobori and from BH were greater than 0.9 for all pendant edges tested.

When we used the method of Zou et al. (2012) to set initial parameters for BH, we obtained $\widehat{\mathbf{Q}}$ estimates identical, or almost identical, to those obtained with the default parameters. This suggests that BH is not converging to incorrect permutations of $\widehat{\mathbf{Q}}$ with our data.

The matched-pairs test of homogeneity found that our alignments deviated significantly from homogeneity (with $p < 0.05$) for all of the primate and mammal alignments, and for the ape mitochondrial protein-coding and nuclear non-coding alignments. The remaining ape alignments did not deviate significantly from homogeneity, with p-values of 0.07 for the nuclear coding alignment taken from the data of Prasad et al. (2008), 0.18 for the $\psi\eta$-globin alignment, and 0.75 for the nuclear coding alignment taken from the data of Fabre et al. (2009). It is therefore possible that, for these three alignments, the BH method may have over-parameterised the data.

When the bootstrap samples were examined, the variances of our estimates were found to be very small for all elements of $\widehat{\mathbf{Q}}$. In each case, the variance was smaller than the mean by at least two orders of magnitude. This suggests that our estimates of **Q** are very precise. Our estimates are likely to also be accurate, given that the results of Oscamou et al. (2008) showed that all of the methods tested gave accurate results with simulated data under a wide variety of conditions.

As can be seen in Figure 4.2, results using the "primates" taxon-set in combination with the primate mitochondrial protein-coding alignment showed a clear pattern. When the pendant edges of the tree were examined, the bar chart showed a readily observable gradation of differences between taxa, particularly for T:A $\rightarrow$ C:G, T:A $\rightarrow$ G:C, and C:G $\rightarrow$ A:T (see Figure 4.2a). A similar pattern can be seen in the heatmap (Figure 4.2b), with the African apes and the remaining catarrhines forming two clusters which group together. PCA analysis (Figure 4.2c) showed three relatively tight clusters corresponding to the rodents, the Asian apes and the African apes, separate from the remaining taxa. While there was some agreement between the heatmap and PCA, some of the clustering in the heatmap did not agree well with PCA (for example, night monkey clusters with mouse in the heatmap but the two are quite distant in the PCA plot).

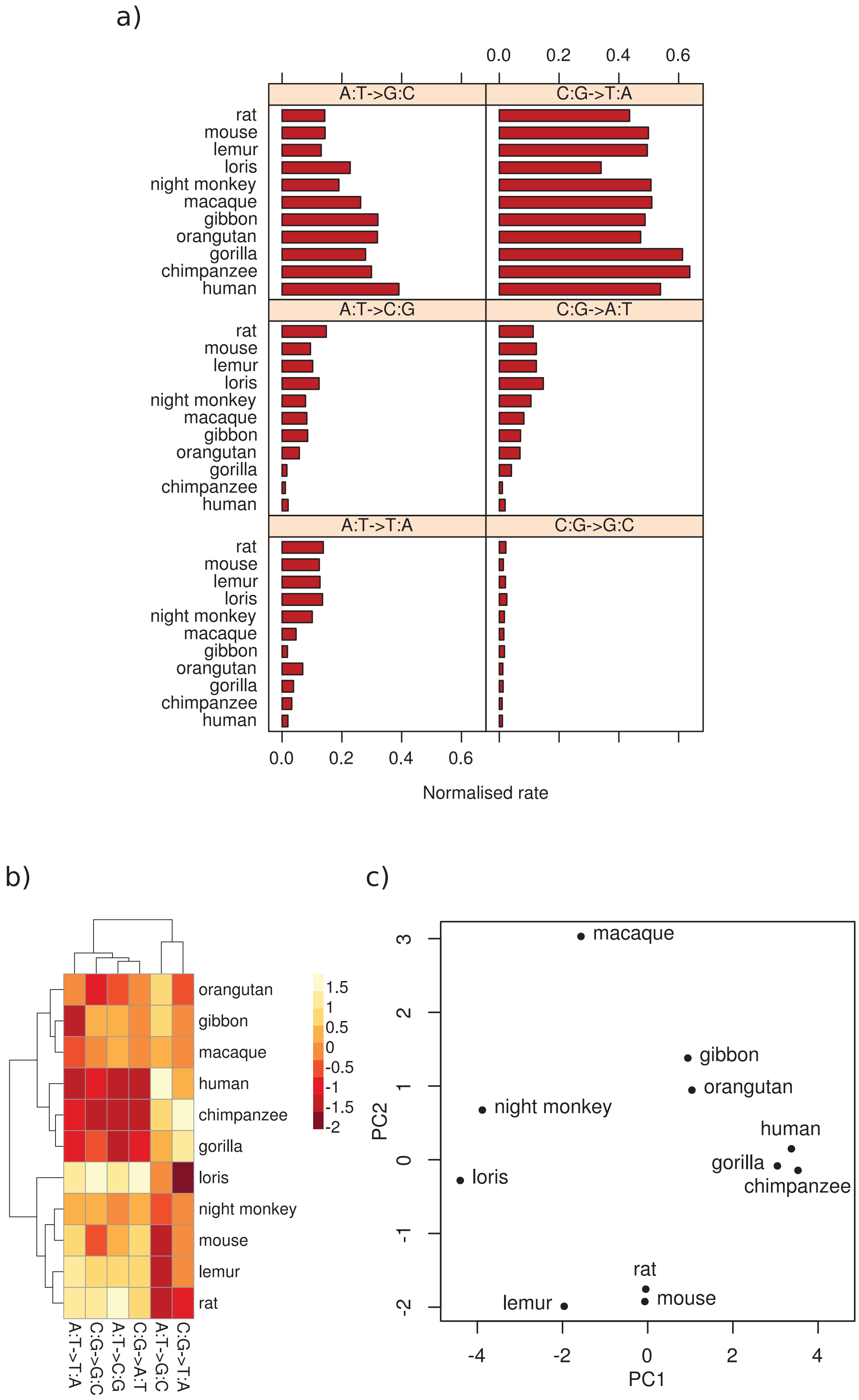
When we examined other alignments for the same taxon-set or for our other taxon-sets, however, we did not detect these groupings in either the PCA plots, heatmaps, or bar charts. Figure 4.3 shows the results for our smallest taxon-set ("apes"), for five different alignments plus the fourfold degenerate third codon positions of the nuclear protein-coding alignment of Prasad et al. (2008). In contrast to what we would expect if we were measuring the same mutational process across all alignments, there is no consistent pattern of differences in relative rates of substitution types between the five taxa. The difference in results between the mitochondrial and nuclear alignments was to be expected, given that different polymerases are involved in mitochondrial and nuclear genome replication, as we mention above. The absence of a common pattern between the nuclear protein-coding alignment of Prasad et al. (2008) and that of Fabre et al. (2009), or between the $\psi\eta$-globin pseudogene alignment of Miyamoto et al. (1987) and the nuclear non-coding alignment of Prasad et al. (2008), however, was more surprising.

We therefore normalised the $\widehat{\mathbf{Q}}$ matrices in other ways in an attempt to identify any changes in mutational process that might be common to all alignments. To control for effects that might be specific to each alignment, we summed all $\widehat{\mathbf{Q}}$s across each tree (excluding the edge leading to the outgroup, as the BH algorithm treats the outgroup as the root node), calculated an overall $\widehat{\mathbf{q}}$, and subtracted this from the $\widehat{\mathbf{q}}$ for each edge. This still did not produce any discernible pattern, or similarity between the alignments, and the results are shown in Figure 4.4.

If there is a detectable effect of the underlying mutational process that is consistent across different alignments, then we would expect that for a given pair of edges a and b, the difference $\widehat{\mathbf{q}}_a - \widehat{\mathbf{q}}_b$ should give a vector with approximately the same direction for each alignment. We checked this for all pairs of edges on the three trees used, and for all pairs of alignments on each taxon-set, by plotting the elements of $\widehat{\mathbf{q}}_a - \widehat{\mathbf{q}}_b$ for one alignment against those for each other alignment. If the vectors have the same direction, the plot should always be linear, but this was not the case for most of the pairs examined.

Overall substitution rates were fastest for the mitochondrial data, as we expected, and slowest for the nuclear protein-coding sequences, with pseudogenes and other nuclear non-coding sequence showing intermediate rates.

Figure 4.2 *(facing page)*: **Substitution rates on the primate mitochondrial tree.** Normalised rates for each of the six pairs of complementary nucleotide substitutions on the pendant edges of the "primates" tree in Figure 1b, calculated using the BH method on the mitochondrial alignment of Fabre et al. (2009). The edge leading to the rabbit is not included because the the BH algorithm roots the tree at the rabbit node, so that $\mathbf{Q}$ for this edge is estimated in the wrong direction. For ease of comparison, rates are normalised such that the rates for each edge sum to one.
(a) Bar chart. The African apes show higher relative rates of transitions (T:A ↔ C:G) than the other taxa, with the rodents and lemuriform primates showing lower transition rates, and the remaining taxa intermediate between the two.
(b) Heatmap of the same data. The columns are scaled so that each has mean zero and standard deviation one. Two of the splits in the dendrogram on the rows correspond to splits in the phylogeny: the split between the African apes and the remaining taxa, and that between the catarrhines and the remaining taxa.
(c) Principal component analysis of the rates shown in the graph. PC1 and PC2 respectively account for 63% and 19% of the total variance. All substitution types have similar weightings on PC1, but PC2 assigns a high weighting to C:G → G:C substitutions and a low weighting to A:T → C:G, with intermediate weightings on the other substitution types. The rodents, the Asian apes and the African apes each form a distinct cluster, with the remaining taxa scattered.

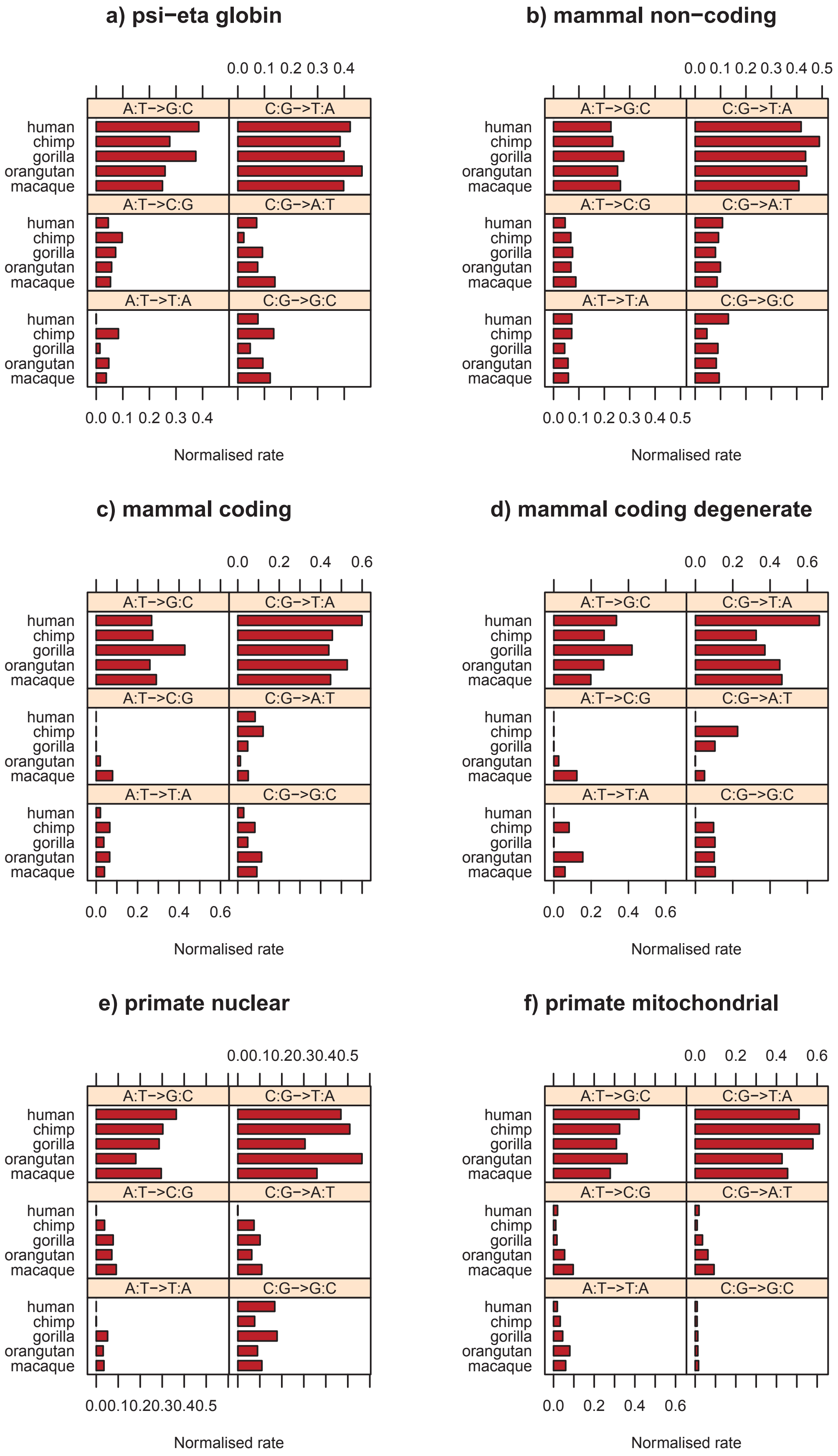

a)
A:T->G:C
C:G->T:A
A:T->C:G
C:G->A:T
A:T->T:A
C:G->G:C
rat
mouse
lemur
loris
night monkey
macaque
gibbon
orangutan
gorilla
chimpanzee
human
0.0
0.2
0.4
0.6
Normalised rate
b)
orangutan
gibbon
macaque
human
chimpanzee
gorilla
loris
night monkey
mouse
lemur
rat
A:T->T:A
C:G->G:C
A:T->C:G
C:G->A:T
A:T->G:C
C:G->T:A
1.5
1
0.5
0
-0.5
-1
-1.5
-2
c)
macaque
gibbon
orangutan
night monkey
human
gorilla
chimpanzee
loris
rat
mouse
lemur
PC1
PC2
-4
-2
0
2
4
-2
-1
0
1
2
3

4.5 Discussion

Our results show some consistency between classes of nucleotide substitution—for example, C:G → T:A substitutions generally have the highest rate on pendant edges of the tree. The results using the mitochondrial sequences on the "primates" dataset also seem to show some conservation between related lineages, of the relative rates of different substitution types. It is possible that the relative clarity of the results using mitochondrial sequences reflects simpler and more homogeneous mutational processes in the mitochondrial genome. This is in line with our expectations given that the mitochondrial genome is replicated by a single DNA polymerase, in contrast to the three main replicative polymerases involved in nuclear DNA replication, along with the many others that carry out proofreading and repair. However, even for this alignment there were some discrepancies between the different visualisations of the results. These discrepancies may arise from a lack of resolution due to the relatively short length of the alignment used, or our visualisation methods may not be the most appropriate for these data.

For the nuclear alignments, we were not able to clearly discern any changes in underlying mutational processes between lineages using our estimates of **Q**. It seems likely that any such effects are masked by phenomena such as selection, strand asymmetries, and differences in mutation rates associated with chromatin structure, all of which can vary across a sequence on a fine scale. The effects of selection in particular will be both site- and lineage-specific, and any sequences that remain alignable

Figure 4.3 *(facing page)*: **Substitution rates for different alignments on the "apes" tree.**
Normalised rates for each of the six pairs of complementary nucleotide substitutions on the pendant edges of the "apes" tree in Figure 1a (not including the outgroup), calculated using:
(a) the $\psi\eta$-globin pseudogene alignment of Miyamoto et al. (1987);
(b) the nuclear non-coding alignment of Prasad et al. (2008);
(c) the nuclear protein-coding alignment of Prasad et al. (2008);
(d) fourfold degenerate sites of the nuclear protein-coding alignment of Prasad et al. (2008);
(e) the nuclear protein-coding alignment of Fabre et al. (2009); and
(f) the mitochondrial alignment of Fabre et al. (2009).



a) psi−eta globin

b) mammal non−coding

c) mammal coding

d) mammal coding degenerate

e) primate nuclear

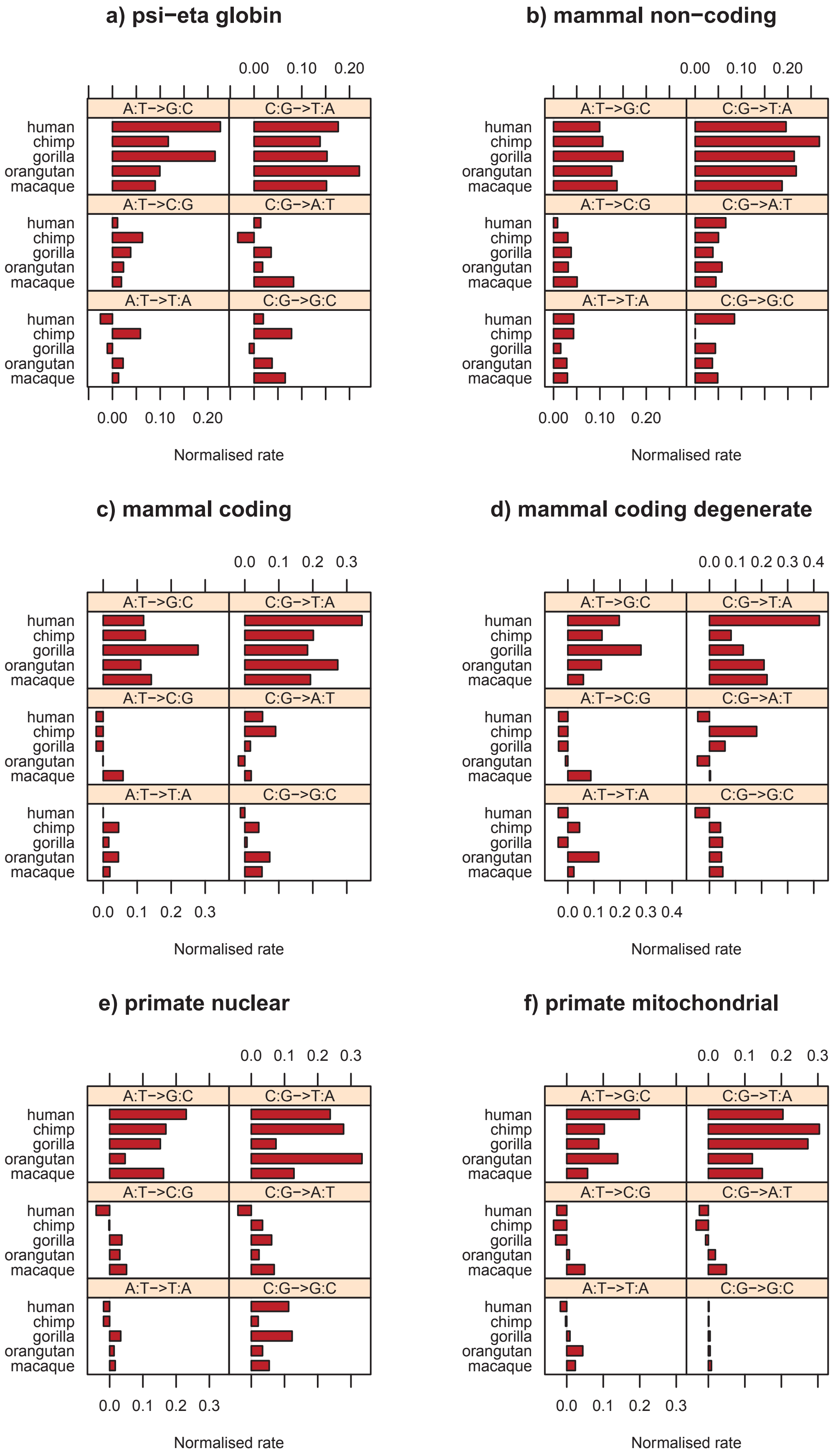
f) primate mitochondrial

over long evolutionary timescales are likely to be under some form of selection, be it direct or indirect.

It is possible that we are simply looking at the wrong scale, either in terms of sequence length or evolutionary divergence. It seems likely that with a more appropriate choice of datasets, it will be possible to obtain estimates of **Q** that reliably reflect an underlying genomic mutational process. On the other hand, it may be the case that the underlying mutational process varies across the genome, even if we could correct for other processes. Replication timing appears to affect the mutational process (Agier and Fischer, 2012), and it is possible that there are other factors that have local effects on DNA replication and error-checking.

It may be possible to avoid some of the confounding effects of selection and other phenomena by using only fourfold degenerate sites, assuming that codon usage bias is simply an effect of the mutational process and not a result of selection, or that we can correct for any such selection. This would require alignments of sites that are fourfold degenerate across a range of taxa, and ideally several large alignments of such sites in order to confirm any consistent pattern; ideally a genome-scale alignment. The 28-way vertebrate alignment produced by Miller et al. (2007) could be a useful resource here, and we intend to carry out further analyses using this alignment.

Figure 4.4 *(facing page)*: **Relative substitution rates for different alignments on the "apes" tree.**
Normalised rates for each of the six pairs of complementary nucleotide substitutions on the pendant edges of the "apes" tree in Figure 1a (not including the outgroup), with the normalised average rate across the tree subtracted, calculated using:
(a) the $\psi\eta$-globin pseudogene alignment of Miyamoto et al. (1987);
(b) the nuclear non-coding alignment of Prasad et al. (2008);
(c) the nuclear protein-coding alignment of Prasad et al. (2008);
(d) fourfold degenerate sites of the nuclear protein-coding alignment of Prasad et al. (2008);
(e) the nuclear protein-coding alignment of Fabre et al. (2009); and
(f) the mitochondrial alignment of Fabre et al. (2009).


a) psi−eta globin
b) mammal non−coding
c) mammal coding
d) mammal coding degenerate
e) primate nuclear
f) primate mitochondrial
A:T−>G:C
C:G−>T:A
A:T−>C:G
C:G−>A:T
A:T−>T:A
C:G−>G:C
human
chimp
gorilla
orangutan
macaque
0.00 0.10 0.20
0.0 0.1 0.2 0.3
0.0 0.1 0.2 0.3 0.4
Normalised rate

Bibliography

F. Ababneh, L. S. Jermiin, C. Ma, and J. Robinson. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics*, 22(10): 1225–1231, 2006.

N. Agier and G. Fischer. The mutational profile of the yeast genome is shaped by replication. *Mol Biol Evol*, 29(3):905–913, 2012.

S. G. E. Andersson and C. G. Kurland. Codon preferences in free-living microorganisms. *Microbiol Rev*, 54(2):198–210, 1990.

D. Barry and J. A. Hartigan. Statistical analysis of hominoid molecular evolution. *Stat Sci*, 2(2):191–210, 1987.

S. L. Chen, W. Lee, A. K. Hottes, L. Shapiro, and H. H. McAdams. Codon usage between genomes is constrained by genome-wide mutational processes. *P Natl Acad Sci USA*, 101(10):3480–3485, 2004.

P.-H. Fabre, A. Rodrigues, and E. J. P. Douzery. Patterns of macroevolution among Primates inferred from a supermatrix of mitochondrial and nuclear DNA. *Mol Phylogenet Evol*, 53(3):808–825, 2009.

T. Gojobori, W.-H. Li, and D. Graur. Patterns of nucleotide substitution in pseudogenes and functional genes. *J Mol Evol*, 18(5):360–369, 1982.

N. Goldman, J. L. Thorne, and D. T. Jones. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J Mol Biol*, 263(2):196–208, 1996.

P. Green, B. Ewing, W. Miller, P. J. Thomas, and E. D. Green. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet*, 33(4):514–517, 2003.

A. J. Herr, L. N. Williams, and B. D. Preston. Antimutator variants of DNA polymerases. *Crit Rev Biochem Mol*, 46(6):548–570, 2011.

U. Hübscher, G. Maga, and S. Spadari. Eukaryotic DNA polymerases. *Annu Rev Biochem*, 71:133–163, 2002.

V. Jayaswal, L. S. Jermiin, and J. Robinson. Estimation of phylogeny using a general Markov model. *Evol Bioinform Online*, 1:62–80, 2005.

L. S. Jermiin, V. Jayaswal, F. Ababneh, and J. Robinson. *Phylogenetic model evaluation.*, volume 452 of *Methods in Molecular Biology*, chapter 16, pages 331–364. Humana Press, Totowa, NJ, 2008.

R. Knight, P. Maxwell, A. Birmingham, J. Carnes, J. G. Caporaso, B. C. Easton, M. Eaton, M. Hamady, H. Lindsay, Z. Liu, C. Lozupone, D. McDonald, M. Robeson, R. Sammut, S. Smit, M. J. Wakefield, J. Widmann, S. Wikman, S. Wilson, H. Ying, and G. A. Huttley. PyCogent: a toolkit for making sense from sequence. *Genome Biol*, 8(8):R171, 2007.

Y.-H. Lin, P. A. McLenachan, A. R. Gore, M. J. Phillips, R. Ota, M. D. Hendy, and D. Penny. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. *Mol Biol Evol*, 19(12): 2060–2070, 2002.

P. Liò and N. Goldman. Models of molecular evolution and phylogeny. *Genome Res*, 8:1233–1244, 1998.

W. Miller, K. Rosenbloom, R. C. Hardison, M. Hou, J. Taylor, B. Raney, R. Burhans, D. C. King, R. Baertsch, D. Blankenberg, S. L. K. Pond, A. Nekrutenko, B. Giardine, R. S. Harris, S. Tyekucheva, M. Diekhans, T. H. Pringle, W. J. Murphy, A. Lesk, G. M. Weinstock, K. Lindblad-toh, R. A. Gibbs, E. S. Lander, A. Siepel, D. Haussler, and W. J. Kent. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*, 17:1797–1808, 2007.

M. M. Miyamoto, J. L. Slightom, and M. Goodman. Phylogenetic relations of humans and African apes from DNA sequences in the $\psi\eta$-globin region. *Science*, 238(4825): 369–373, 1987.

A. Muto and S. Osawa. The guanine and cytosine content of genomic DNA and bacterial evolution. *P Natl Acad Sci USA*, 84(1):166–169, 1987.

M. Oscamou, D. McDonald, V. B. Yap, G. A. Huttley, M. E. Lladser, and R. Knight. Comparison of methods for estimating the nucleotide substitution matrix. *BMC Bioinformatics*, 9:511, 2008.

R. Ota and D. Penny. Estimating changes in mutational mechanisms of evolution. *J Mol Evol*, 57(Suppl 1):S233–240, 2003.

D. Penny, R. P. Murray-McIntosh, and M. D. Hendy. Estimating times of divergence with a change of rate: the orangutan/African ape divergence. *Mol Biol Evol*, 15(5): 608–610, 1998.

J. B. Plotkin and G. Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*, 12(1):32–42, 2011.

A. B. Prasad, M. W. Allard, and E. D. Green. Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol Biol Evol*, 25(9):1795–1808, 2008.

R Development Core Team. R: A Language and Environment for Statistical Computing, 2011. URL `http://www.r-project.org`.

M. Touchon, S. Nicolay, B. Audit, E.-B. Brodie of Brodie, Y. d'Aubenton Carafa, A. Arneodo, and C. Thermes. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *P Natl Acad Sci USA*, 102(28): 9836–9841, 2005.

Z. Zhang and M. Gerstein. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res*, 31(18):5338–5348, 2003.

L. Zou, E. Susko, C. Field, and A. J. Roger. The parameters of the Barry and Hartigan general Markov model are statistically nonidentifiable. *Syst Biol*, 60(6):872–875, 2011.

L. Zou, E. Susko, C. Field, and A. J. Roger. Fitting nonstationary general-time-reversible models to obtain edge-lengths and frequencies for the Barry-Hartigan model. *Syst Biol*, 61(6):927–940, 2012.

Chapter 5

Future Directions

5.1 Assembly of mixed samples

The methods developed in Chapter 2 for assembling and deconvoluting mixed organelle genomes have already proved useful, and I find it encouraging that the paper has been cited several times, including in a review (Pareek et al., 2011), which suggests that the work is of interest to others. As the length of sequence reads and the volume generated in a sequencing run continue to increase, the method becomes more powerful. Increasing read length allows longer contigs to be assembled because overlaps between reads are less likely to be ambiguous. Increasing sequence volume means that even a single indexed sample produces more than enough reads to assemble several organelle genomes, even when genomic DNA is used.

The increasing number of sequenced genomes available means that we will be able to use more closely related reference sequences, enabling us to deconvolute more closely related samples within a mixture. Currently, mitochondrial genomes can easily be separated, and the method should also work well for mixtures of chloroplast genomes, although the presence of larger non-coding regions may mean that chloroplasts require more closely related reference sequences than are needed for mitochondrial genomes.

It is gratifying that this work has been of value to a wide range of other researchers, both colleagues here and collaborators at other institutions. I am continuing to participate in ongoing analysis of the genomes that I and others have assembled, in order to better understand the limitations of the method, as well as to apply further refinements.

One of the most intriguing aspects of the work in Chapter 2 is the non-complementary DNA that we detected. This raises two questions: what is the function of this structure; and how is it maintained *in vivo* through DNA replication? In-depth laboratory-based study will be required in order to fully elucidate these questions.

5.2 Multiple optima of likelihood

We have shown in Chapter 3 that multiple optima will, in practice, seldom be a problem for maximum likelihood phylogenetic methods. However, although we have investigated some properties of alignments that may increase the incidence of multiple optima, there is much information yet to be extracted from our results.

It would be interesting to develop a tree-independent metric that could predict the occurrence of multiple optima given an alignment. Measures of conflict between sites, such as that implemented in TIGER (Cummins and McInerney, 2011), may be helpful, although our results with recombinant strains of hepatitis B suggest that conflict between sites does not necessarily lead to higher incidence of multiple optima. The dataset of over a million trees produced here could be used to test any such metric.

From a mathematical point of view, further characterisation of the likelihood landscape would be interesting, particularly looking at cases where there appears to be either a ridge of optimal values or a very shallow slope around the optimum, as seen in our prokaryote data (see Figure 3.5). Again, the data produced here would easily lend itself to such a mathematical analysis.

5.3 Investigating mutational mechanisms

The results in Chapter 4 tend to suggest that underlying mutational mechanisms cannot yet be untangled from other causes of nucleotide substitution using the alignments considered here. Although we find some consistency between classes of nucleotide change, it is not yet possible (except perhaps for the mitochondrial alignment) to group taxa into similar classes. It may be possible to isolate the mechanisms of mutation by using much larger genomic datasets such as the 28-way vertebrate genome alignment of Miller et al. (2007), and I have begun carrying out further analyses using this alignment. A dataset of this size allows us to compare values of the substitution rate matrix across different parts of the genome, for a wide taxonomic range of species.

It may turn out to be the case that the mutational process varies between parts of the genome on a small enough scale to make it statistically impossible to isolate mechanisms by examining the observed substitution rates. This would be an interesting result in itself, and may mean that these processes could only be characterised using biochemical techniques. However, preliminary results using genome-scale alignments suggest that consistent differences in the elements of $\mathbf{Q}$ can be detected between lineages on a tree.

5.4 Summary

As the volume and complexity of available sequence data continues to grow, so does the need for new sequence analysis techniques and, perhaps more critically, the need to test and to extend the existing techniques. I expect that the work presented in this thesis will be a useful contribution to fulfilling these needs.

Bibliography

C. A. Cummins and J. O. McInerney. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Syst Biol*, 60(6):833–844, 2011.

W. Miller, K. Rosenbloom, R. C. Hardison, M. Hou, J. Taylor, B. Raney, R. Burhans, D. C. King, R. Baertsch, D. Blankenberg, S. L. K. Pond, A. Nekrutenko, B. Giardine, R. S. Harris, S. Tyekucheva, M. Diekhans, T. H. Pringle, W. J. Murphy, A. Lesk, G. M. Weinstock, K. Lindblad-toh, R. A. Gibbs, E. S. Lander, A. Siepel, D. Haussler, and W. J. Kent. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*, 17:1797–1808, 2007.

C. S. Pareek, R. Smoczynski, and A. Tretyn. Sequencing technologies and genome sequencing. *J Appl Genetics*, 52(4):413–435, 2011.

Appendix A

Contribution to publications

This thesis includes two published papers and a submitted manuscript prepared in collaboration with others. Here I outline the extent of each author's contribution to each of these papers.

A.1 Assembly of mixed mitochondrial genomes

The published paper on assembling mixed mitochondrial genomes (McComish et al., 2010) is included in Chapter 2.

A.1.1 Author contributions

- BJM helped with the design of the study, carried out the simulations, wrote the scripts for and carried out *de novo* assembly and deconvolution of the mitochondrial genomes, and identified the non-complementary DNA found in the *Amalda* mitochondrial genome, as well as drafting, contributing to and editing the manuscript;

- SFKH annotated the *Amalda* mitochondrial genome and characterised the secondary structure of its control region, submitted the data to GenBank, performed PCR assays, and contributed to and edited the manuscript;

- PJB helped with the design of the study, wrote the scripts used in the simulations, contributed to and edited the manuscript;

- DP conceived of and designed the study, contributed to, edited and revised several versions of the manuscript, and provided academic guidance to BJM during his PhD study.

- All authors read and approved the final manuscript.

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Bennet McComish

Name/Title of Principal Supervisor: David Penny

Name of Published Research Output and full reference:

"Index-free de novo assembly and deconvolution of mixed mitochondrial genomes."
McComish BJ, Hills SFK, Biggs PJ, Penny D. Genome Biology and Evolution 2010, 2:410-424

In which Chapter is the Published Work: Chapter 2

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  See attached statement.

Candidate's Signature

December 20th, 2012
Date

David Penny
Principal Supervisor's signature

December 20th, 2012
Date

A.2 Multiple optima of likelihood

The manuscript on multiple optima of likelihood is presented in Chapter 3, and has been submitted to *Systematic Biology*.

A.2.1 Author contributions

- BJM helped with the design of the study, wrote the scripts for and carried out all of the analyses, as well as drafting, contributing to and editing the manuscript;
- KPS contributed code to the scripts used in the study, provided a custom version of the phangorn package, and contributed to and edited the manuscript;
- DP conceived of and designed the study, contributed to, edited and revised several versions of the manuscript, and provided academic guidance to BJM during his PhD study.
- All authors read and approved the final manuscript.

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Bennet McComish

Name/Title of Principal Supervisor: David Penny

Name of Published Research Output and full reference:

"Multiple local maxima for likelihoods of phylogenetic trees constructed from biological data."
McComish BJ, Schliep KP, Penny D. Submitted to Systematic Biology.

In which Chapter is the Published Work: Chapter 3

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

See attached statement.

Candidate's Signature

December 20th, 2012
Date

David Penny
Principal Supervisor's signature

December 20th, 2012
Date

A.3 Karaka chloroplast genome

The published paper on the sequencing and assembly of the chloroplast genome of karaka (Atherton et al., 2010) is presented as Appendix B.

A.3.1 Author contributions

- RAA helped with the design of the study, isolated chloroplasts, extracted the DNA, performed RCA amplification of cpDNA, carried out PCR experiments and submitted data to GenBank, as well as drafting, contributing to and editing the manuscript;

- BJM carried out the full *de novo* assembly and annotation of the genome, as well as contributing to and editing the manuscript;

- LDS contributed to, edited and revised several versions of the manuscript and provided academic guidance to RAA during her PhD study.

- LB carried out sample preparation and sequencing on the Illumina GAII;

- NWA designed and performed qPCR assays and contributed to and edited the manuscript;

- PJL conceived of and designed the study, contributed to, edited and revised several versions of the manuscript and provided academic guidance to RAA during her PhD study.

- All authors read and approved the final manuscript.

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Bennet McComish

Name/Title of Principal Supervisor: David Penny

Name of Published Research Output and full reference:

"Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform."
Atherton RA, McComish BJ, Shepherd LD, Berry LA, Albert NW, Lockhart PJ. Plant Methods 2010, 6:22

In which Chapter is the Published Work: Appendix B

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

See attached statement.

Candidate's Signature

December 20th, 2012
Date

David Penny
Principal Supervisor's signature

December 20th, 2012
Date

GRS Version 3– 16 September 2011

Bibliography

R. A. Atherton, B. J. McComish, L. D. Shepherd, L. A. Berry, N. W. Albert, and P. J. Lockhart. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. *Plant Methods*, 6:22, 2010.

B. J. McComish, S. F. K. Hills, P. J. Biggs, and D. Penny. Index-free de novo assembly and deconvolution of mixed mitochondrial genomes. *Genome Biol Evol*, 2:410–424, 2010.

Appendix B

Karaka chloroplast genome

This appendix presents the paper “Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform”, which was published in the journal *Plant Methods* in 2010.

Atherton *et al. Plant Methods* 2010, **6**:22
http://www.plantmethods.com/content/6/1/22

PLANT METHODS

METHODOLOGY Open Access

Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform

Robin A Atherton[1,2*†], Bennet J McComish[1,2†], Lara D Shepherd[1], Lorraine A Berry[3], Nick W Albert[1], Peter J Lockhart[4]

Abstract

Background: Complete chloroplast genome sequences provide a valuable source of molecular markers for studies in molecular ecology and evolution of plants. To obtain complete genome sequences, recent studies have made use of the polymerase chain reaction to amplify overlapping fragments from conserved gene loci. However, this approach is time consuming and can be more difficult to implement where gene organisation differs among plants. An alternative approach is to first isolate chloroplasts and then use the capacity of high-throughput sequencing to obtain complete genome sequences. We report our findings from studies of the latter approach, which used a simple chloroplast isolation procedure, multiply-primed rolling circle amplification of chloroplast DNA, Illumina Genome Analyzer II sequencing, and de novo assembly of paired-end sequence reads.

Results: A modified rapid chloroplast isolation protocol was used to obtain plant DNA that was enriched for chloroplast DNA, but nevertheless contained nuclear and mitochondrial DNA. Multiply-primed rolling circle amplification of this mixed template produced sufficient quantities of chloroplast DNA, even when the amount of starting material was small, and improved the template quality for Illumina Genome Analyzer II (hereafter Illumina GAII) sequencing. We demonstrate, using independent samples of karaka (*Corynocarpus laevigatus*), that there is high fidelity in the sequence obtained from this template. Although less than 20% of our sequenced reads could be mapped to chloroplast genome, it was relatively easy to assemble complete chloroplast genome sequences from the mixture of nuclear, mitochondrial and chloroplast reads.

Conclusions: We report successful whole genome sequencing of chloroplast DNA from karaka, obtained efficiently and with high fidelity.

Background

Chloroplast genomes provide a wealth of information for studies in molecular ecology and evolution. Their conservative gene content and organisation have enabled researchers to isolate homologous loci for comparative studies over different evolutionary time-scales [1-7].

Obtaining the DNA sequence for chloroplast genomes can be achieved by using the polymerase chain reaction (PCR) to amplify chloroplast DNA fragments from genomic DNA (gDNA) extracts. However, this can involve up to 35 amplifications of overlapping chloroplast DNA PCR products [2,8]. While this approach is time consuming [8], it has been preferred over protocols that attempt to first separate chloroplasts from other cellular material. Reasons for this appear to be that chloroplast isolation can be troublesome in some species [9] and because rapid chloroplast isolation protocols often produce template which is still contaminated by large quantities of nuclear DNA [10]. Nevertheless, given the depth of sequencing coverage with the Illumina GAII sequencing platform, we were interested to investigate whether this alternative approach could be used for sequencing whole chloroplast genomes without the need for whole genome PCR amplification. Here we report findings which demonstrate that, even with small amounts of chloroplast DNA, and in the presence of large amounts of nuclear DNA, Illumina short read sequencing provides a practical approach for obtaining complete chloroplast genome sequences.

* Correspondence: r.a.atherton@massey.ac.nz
† Contributed equally
[1]Institute of Molecular BioSciences, Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand
Full list of author information is available at the end of the article

Methods

Chloroplast isolation

Fresh leaf material was obtained from two cultivated karaka trees originating in Rekohu/Chatham Islands and the Kermadec Islands, New Zealand. Leaf material was collected and processed immediately for the sample from the Chatham Islands and within 3 h for the sample from the Kermadec Islands. Leaf samples weighing 2.5-5 g were excised from living trees and processed as follows. Chloroplasts were isolated using a protocol originally designed for isolating chloroplasts from *Arabidopsis thaliana* [11] with minor modifications: (i) the leaf material was homogenised using an Ultra-Turrax homogeniser with an N18 rotor (Janke & Kunkel IKA, Hamburg, Germany); (ii) the homogenate was passed through a double layer of washed and autoclaved nappy (diaper) liner (Johnson & Johnson Ltd.) rather than through Miracloth (Calbiochem); and (iii) the final centrifugation step was carried out using a Sorvall SS32 angled rotor, rather than a swinging bucket rotor. After the final centrifugation step, DNA was extracted from pooled chloroplasts for each tree sample using a DNEasy Plant Mini Kit (Qiagen) following the manufacturer's instructions. Genomic DNA (gDNA) was extracted from silica-dried karaka leaf material from the same accessions using a DNEasy Plant Mini Kit (Qiagen).

Multiply-primed rolling circle amplification

Multiply-primed rolling circle amplification (RCA) was used to produce an abundance of purified chloroplast DNA template in preparation for sequencing [12]. This technique involves isothermal, strand-displacing amplification using multiple primers and is capable of yielding a large amount of product from very little starting DNA template [13]. Phi29, the DNA polymerase used in multiply-primed RCA, is reported to have a very low level of amplification bias making the template suitable for whole genome sequencing [14]. Chloroplast-enriched DNA (cpDNA) from both karaka samples was amplified in this way using a REPLI-g™ Mini Kit (Qiagen) following the manufacturer's instructions, with the exception that samples were incubated at room temperature for 9 min rather than the recommended 3 min. This extension time consistently produced better results with different plant samples. The kit produced ~5 μg of product for each sample.

Confirmation of chloroplast DNA enrichment

Genomic DNA (gDNA), chloroplast-enriched DNA (cpDNA) and RCA amplified chloroplast-enriched DNA (RCAcpDNA) from the Chatham Island sample were quantified fluorometrically using the Quant-iT™ dsDNA HS assay kit on a Qubit™ Quantitation Platform (Invitrogen). The concentration of the gDNA, cpDNA and RCAcpDNA was 110 ngμL^{-1}, 20 ngμL^{-1} and 104 ngμL^{-1}, respectively, in a total volume of 50 μL of AE buffer (Qiagen). The purity of gDNA, cpDNA and RCAcpDNA samples was determined by A_{260}/A_{280} and A_{260}/A_{230} ratios on a NanoDrop (NanoDrop Technologies) spectrophotometer. Enrichment for chloroplast DNA was determined by quantitative real-time PCR (qPCR) with gDNA, cpDNA and RCAcpDNA templates; the quantity of the plastid gene *psbB* was determined relative to nuclear encoded 18S *rRNA* by comparative quantification [15]. Gene-specific primers were designed for *psbB* (*psbB* F 5'GGGGGTTGGAGTATCACAGG3'; *psbB* R 5'CCAAGAAGCACAAGCCAGAA3', 103 bp amplicon) using Primer3 [16] and primers for 18S are described by Zhu and Altmann [17]. qPCR was performed using Lightcycler480 SYBR Green1 Master (Roche Diagnostics) reagents in a Rotor Gene 3000 instrument (Corbett Research) with four technical replicates per sample. Template DNA was diluted 20-fold for cpDNA, and 100-fold for gDNA and RCAcpDNA samples for qPCR. The qPCR cycling conditions were: 95°C 10 min, (95°C 10 s, 60°C 15 s, 72°C 20 s) × 40 cycles with fluorescent detection at 72°C and during the final melt. Melt curve analysis confirmed the amplification of a single product.

Illumina GAII sequencing

The RCAcpDNA samples from both accessions were sequenced by Massey Genome Service (Massey University, Palmerston North, New Zealand). A 75 bp paired-end run was performed on the Illumina GAII with the two samples described here in a single lane along with four other samples from a separate experiment. Samples were prepared for sequencing as follows: genomic DNA libraries were prepared by fragmenting purified genomic DNA using a nebulisation kit (Invitrogen), paired-end index adaptor ligation (Illumina) and 18 cycles of PCR enrichment using the Illumina Paired-End Genomic DNA library preparation kit, Illumina Multiplex Oligonucleotide library preparation kit and Illumina Multiplex Paired-End Genomic DNA library preparation protocol. The enriched libraries were quantified using an ND-1000 NanoDrop spectrophotometer (NanoDrop Technologies) and quality checked by Agilent 2100 Bioanalyzer, DNA 1000 Labchip kit assay. The libraries were then diluted to a 10 nM concentration using EB buffer (Qiagen) and quantified for optimal cluster density using the LightCycler® 480 system Absolute Quantification protocol (Roche Diagnostics) and the LightCycler® 480 SYBR Green I Master kit (Roche Diagnostics). The libraries were pooled at equal molarity and amplified in one flow cell lane on the Illumina Cluster Station instrument at a density of 140,000 clusters per tile and a molarity of 13 pM using the Illumina Paired-

End Cluster Generation kit v2. The amplified libraries were sequenced on the Illumina GAII instrument using 4 Illumina 36 cycle SBS sequencing kits (v3), Illumina Multiplex Sequencing Primers and PhiX control kit v2, on a 75 bp paired-end indexed run. After sequencing, the resulting images were analysed with the proprietary Illumina pipeline v1.3. Reads for each of the indexed samples were then separated using a custom Perl script.

Assembly

Reads from each indexed sample were trimmed to remove poor quality sequence at the 3' end. To determine the optimum trim length, initial de novo assemblies were made for read sets of different length (untrimmed reads, and reads trimmed to 70, 65, 55, 50 bp). These assemblies were carried out using Velvet 0.7 [18] with a range of hash lengths from 33 to 63 and a minimum k-mer coverage of 5×. For these initial assemblies, the data were treated as single reads, that is, the paired-end information was not used. Maximum contig lengths and N50 values were tabulated and the hash lengths that gave the highest N50 for each trimmed set of reads were selected for further optimisation. A second round of assembly was carried out on each trimmed set of reads using the hash length determined above and varying the coverage cut-off parameter from 1 to 100. Finally, paired-end assembly was carried out for each of these read-length/hash-length combinations using the coverage cut-off value that gave the highest N50 value. For these paired-end assemblies, expected coverage was set to the length-weighted median of the coverage values obtained in the initial single read assemblies, and the insert length was estimated as 240 bp. Assembled contigs were aligned to the *Cucumis sativus* chloroplast genome [GenBank: NC_007144; GenBank: DQ119058] using Geneious 4.7 [19].

Four short regions of ambiguous sequence were checked by PCR amplification using the following primers, custom designed using Primer3 [16] unless referenced: Corlaerps2-rpoc2F (TATAGGGTGCCATTCG AGGA), Corlaerps2-rpoc2R GTATCAACAACGGC-CAATCC; CorlaendhAF (GGAATAGGATGGAGA-TAAGAAAGAC), CorlaendhAR (CACGATTCCG ATCCAGAGTA); psbJ ATAGGTACTGTARCYGG-TAT [20], petA AACARTTYGARAAGGTTCAATT [20]; psbAR (CGCGTCTCTCTAAAATTGCAGTCAT) [21], CorlaepsbA-R (ATCCGACTAGTTCCGGGTTC). Figure 1 shows the relative position of the priming sites on the karaka chloroplast genome. The PCR cycling conditions were modified slightly from an existing published protocol [20] as follows: template denaturation at 80°C for 5 min followed by 32 cycles of denaturation at 95°C for 1 min, primer annealing at 50°C for 1 min, followed by a ramp of 0.3°C/s to 65°C, and primer extension at 65°C for 4 min; followed by a final extension step of 5 min at 65°C. Amplified PCR products were sequenced using the BigDye Terminator Cycle Sequencing Kit (Applied Biosystems) and an ABI 3730 automated capillary sequencer at Massey Genome Service (Massey University, Palmerston North, New Zealand). The resulting sequences were visualised and edited using Sequencher 4.9 software for Mac (Gene Codes Corporation, Ann Arbor, MI). Using Geneious [19], the four ambiguous regions of the assembled genome were edited, where necessary, to match the Sanger sequences.

Mapping and annotation

In order to check the de novo assembly, reads were aligned against the assembled genome using BWA [22] with default parameters. Only 19.6% of reads were successfully aligned, but this was sufficient to give a mean coverage of 400×. This mapping enabled us to resolve some short regions of ambiguous sequence in the assembly. The final complete chloroplast genome sequence was annotated using DOGMA [23] and through comparison to published complete chloroplast genome sequences available through GenBank [24].

Results

DNA sequencing template for karaka

The relative quantity of chloroplast DNA in samples of total gDNA, cpDNA and RCAcpDNA preparations was determined by quantitative PCR (Figure 2). The enriched cpDNA sample had 2.6-fold higher levels of chloroplast DNA compared to a standard gDNA preparation prior to RCA and 2.2-fold after RCA. The purity of the DNA preparations was assessed by spectrophotometric A_{260}/A_{280} and A_{260}/A_{230} ratios. The gDNA and cpDNA samples had low ratios, indicating the presence of protein (A_{260}/A_{280} = 1.66 and 1.69, respectively) and other contaminants such as carbohydrates and phenolics (A_{260}/A_{230} = 1.49 and 1.28, respectively). RCA of the cpDNA-enriched sample substantially increased the quantity and quality of template DNA (A_{260}/A_{280} = 1.75, A_{260}/A_{230} = 2.20).

Sequencing and assembly of the karaka chloroplast genomes

Paired-end sequencing of the RCAcpDNA template in a single lane on an Illumina GAII flow cell produced 1.84 and 1.76 million reads for the Chatham Islands sample and Kermadec Islands sample respectively. The Chatham Islands sample was assembled de novo as described in the methods section. The Kermadec Island sample was then mapped to this assembly.

The most useful assembly was achieved for the Chatham Islands sample, with reads trimmed to 50 bp, coverage cut-off of 9 and expected coverage of 40. While

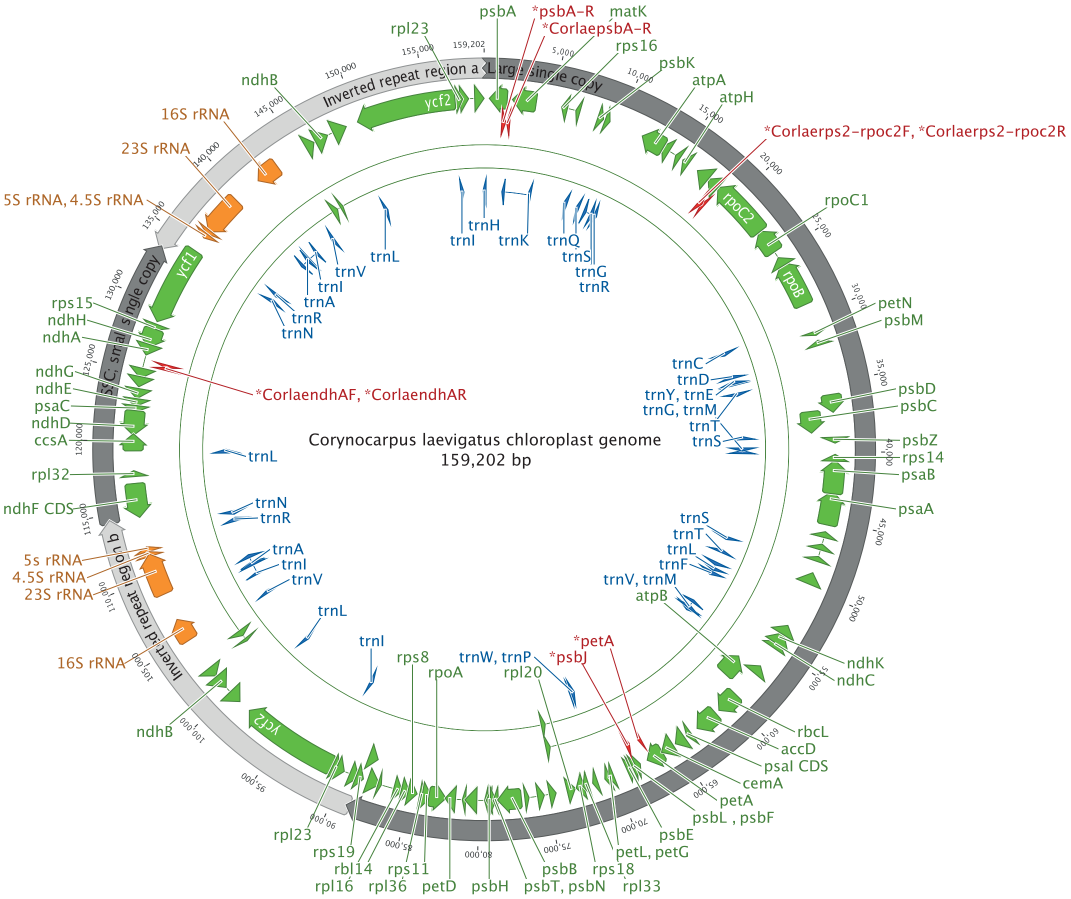


Figure 1 Complete gene map of the entire *Corynocarpus laevigatus* chloroplast genome. Arrowheads depict the direction of transcription. Binding sites for the primers used to generate the PCR products for resolving homopolymers are indicated in red. This figure was produced using Geneious [19].

some of our other assemblies had higher overall N50 values or longer maximum contig lengths, these improved statistics did not reflect any real improvement in the assembly, as the large single-copy region was merged with part of the inverted repeat to form a single long contig at the expense of a more fragmented assembly of the remainder of the inverted repeat and small single-copy region.

The optimal assembly parameters produced a total of 13 contigs, four of which could be mapped to the *Cucumis sativus* chloroplast genome. These four contigs ranged from 7,857 bp to 88,955 bp in length and covered the entire chloroplast genome. The nine remaining contigs were much shorter, ranging from 81 bp to 669 bp. These were checked against the GenBank nucleotide database using the web-based BlastN algorithm [25], and the only significant alignments found were to nuclear ribosomal DNA sequences.

Of the four contigs mapping to the *C. sativus* chloroplast genome, one mapped to the inverted repeat region, one to the large single-copy region, and two to the small single-copy region. The overlaps between contigs at all four junctions between inverted repeat and single-copy regions were 40 bp long, indicating that contig extension was interrupted by the ambiguity of the overlap rather than by insufficient coverage. The overlap between the two contigs that formed the small single-copy region consisted of a polyA-polyT homopolymer. The contig corresponding to the large single-copy region contained six short (1-49 bp) stretches of

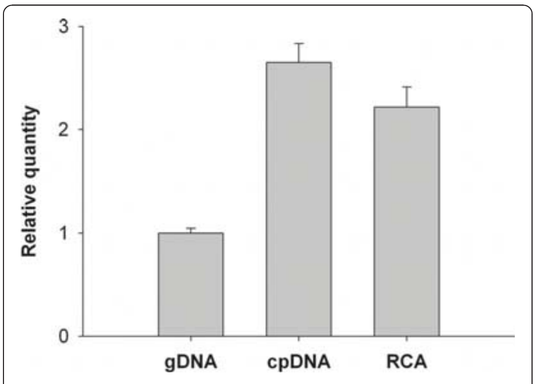


Figure 2 Enrichment of chloroplast DNA from karaka. The relative quantity of chloroplast DNA was determined in preparations of genomic DNA (gDNA), chloroplast-enriched DNA (cpDNA) and RCA amplified chloroplast-enriched cpDNA (RCA). Quantitative PCR amplification of the plastid encoded *psbB* gene was compared to amplification of nuclear 18S *rRNA*. The mean of four technical replicates ± SE has been reported as fold-differences relative to gDNA.

ambiguous bases where Velvet was unable to resolve the sequence due to mono- or dinucleotide repeats. Three of these stretches were resolved by mapping the reads to the assembled sequence as described in the methods. The other three stretches, along with the overlap between the two contigs that made up the small single-copy region, were checked by PCR amplification and Sanger sequencing.

The final assembled chloroplast genome sequence (shown in Figure 1) was checked by mapping the original Illumina reads against the assembled sequence. A total of 344,475 of 1.76 million reads (19.6%) were successfully aligned, suggesting that approximately 80% of the DNA sequenced was of nuclear or mitochondrial origin.

Discussion

We have shown that the modified chloroplast isolation protocol produced DNA template sufficiently enriched for chloroplast sequence to allow de novo assembly of the chloroplast genome. Comparison of the two genomes indicated high fidelity with less than 0.002% error. Whilst RCA of the cpDNA marginally reduced the final ratio of cpDNA/gDNA in the enriched sample, the purity of the DNA was of a higher quality for Illumina sequencing.

The coverage cut-off parameter of the Velvet assembler was crucial for successful assembly, as it allowed the chloroplast sequence reads to be assembled without interference from nuclear sequence. Although over 80% of reads failed to align to our assembled chloroplast genome, and are likely to be of nuclear origin, the much greater size of the nuclear genome means that these reads were present at much lower coverage than the chloroplast reads. A notable exception is nuclear ribosomal DNA, which is present in many copies in the nuclear genome, thus its coverage was comparable to that of the chloroplast genome in our enriched sample.

The lower copy number of the nuclear genome compared to the chloroplast genomes means that nuclear copies of chloroplast DNA sequences are very unlikely to affect our assemblies. In contrast, nuclear-encoded chloroplast DNA may be more difficult to distinguish from chloroplast-encoded sequences if amplified by chloroplast DNA primers. Thus, this is potentially another advantage of the approach we have used for determining complete chloroplast genome sequences.

Finally, although de novo assembly was a feature of our protocol, the availability of a related reference genome did help with our final assembly, allowing us to separate contigs derived from chloroplast DNA from the few short contigs of nuclear origin. This was helpful for determining the arrangement of chloroplast contigs.

Conclusions

We have successfully applied a whole genome sequencing approach to determine the complete chloroplast genome sequence of karaka. We have also applied this approach more recently to a range of New Zealand seed plants (gymnosperms and angiosperms: herbaceous and woody plants), sequencing up to three chloroplast genomes per GAII flow cell lane. Thus we are confident that the approach that we describe here for karaka provides a fast and efficient protocol for obtaining whole chloroplast genome sequences for seed plants.

The fully annotated chloroplast genome sequence of karaka (*Corynocarpus laevigatus*) from the Chatham Islands sample has been deposited in the GenBank database under accession number HQ207704.

Acknowledgements
We would like to acknowledge the help and support of the Hokotehi Moriori Trust Board of Rekohu/Chatham Islands and Ngāti Kuri and Te Aupouri, tangata whenua (indigenous inhabitants) of the Kermadec Islands. We would also like to acknowledge Te Ati Awa (Taranaki Whānui) in the Wellington and Hutt Valley regions of New Zealand, with whom we have ongoing consultations. We thank Peter de Lange (Department of Conservation), Rewi Elliot and Eleanor Burton (Otari Wilton's Bush, Wellington) for assistance with sample collection. Samples were collected under Department of Conservation permit WA-23814-FLO and Otari Wilton's Bush permit 145. Thanks also to Maurice Collins (MGS) for pipeline analysis, Trish McLenachan (IMBS, Massey University) and Jan Binnie (IMBS, Massey University) for laboratory assistance. This study was funded by the New Zealand Marsden Fund (contract numbers MAU050; MAU0709) and the Allan Wilson Centre for Molecular Ecology and Evolution. Robin Atherton acknowledges financial support from a Massey University Doctoral Scholarship, JP Skipworth Scholarship for Plant Biology, Hokotehi Moriori Trust Board, Intellectual Property in Cultural Heritage (IPinCH) and the Allan Wilson Centre for Molecular Ecology and Evolution. Bennet McComish

acknowledges support from an Allan Wilson Centre for Molecular Ecology and Evolution Doctoral Scholarship. Peter Lockhart is supported by a James Cook Research Fellowship from the Royal Society of New Zealand.

Author details

[1]Institute of Molecular BioSciences, Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand. [2]Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand. [3]Massey Genome Service, Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand. [4]Institute of Fundamental Sciences, Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand.

Authors' contributions

RAA helped with the design of the study, isolated chloroplasts, extracted the DNA, performed RCA amplification of cpDNA, carried out PCR experiments and submitted data to GenBank, as well as drafting, contributing to and editing the manuscript; BJM carried out de novo assembly and annotation of the genome, contributed to and edited the manuscript; LDS contributed to, edited and revised several versions of the manuscript and provided academic guidance to RAA during her PhD study. LB carried out sample preparation and sequencing on the Illumina GAII; NWA designed and performed qPCR assays and contributed to and edited the manuscript; PJL conceived of and designed the study, contributed to, edited and revised several versions of the manuscript and provided academic guidance to RAA during her PhD study. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 25 June 2010 Accepted: 28 September 2010
Published: 28 September 2010

References

1. Gruenheit N, Lockhart PJ, Steel M, Martin W: **Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites.** *Molecular Biology and Evolution* 2008, **25(7)**:1512-1520.
2. Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH: **Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm.** *Molecular Biology and Evolution* 2003, **20(9)**:1499-1505.
3. Knapp M, Stockler K, Havell D, Delsuc F, Sebastiani F, Lockhart PJ: **Relaxed molecular clock provides evidence for long-distance dispersal of *Nothofagus* (southern beech).** *Plos Biology* 2005, **3(1)**:38-43.
4. Stehlik I, Blattner FR, Holderegger R, Bachmann K: **Nunatak survival of the high Alpine plant *Eritrichium nanum* (L.) Gaudin in the central Alps during the ice ages.** *Molecular Ecology* 2002, **11(10)**:2027-2036.
5. Ingvarsson PK, Ribstein S, Taylor DR: **Molecular evolution of insertions and deletion in the chloroplast genome of *Silene*.** *Molecular Biology and Evolution* 2003, **20(11)**:1737-1740.
6. Golenberg EM, Clegg MT, Durbin ML, Ma DP: **Evolution of a noncoding region of the chloroplast genome.** *Molecular Phylogenetics and Evolution* 1993, **2(1)**:13.
7. Powell W, Morgante M, Andre C, Mcnicol JW, Machray GC, Doyle JJ, Tingey SV, Rafalski JA: **Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome.** *Current Biology* 1995, **5(9)**:1023-1029.
8. Cronn R, Liston A, Parks M, Gernandt D, Shen R, Mockler T: **Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology.** *Nucleic Acids Research* 2008, **36(19)**:e122-e122.
9. Dhingra A, Folta KM: **ASAP: Amplification, sequencing & annotation of plastomes.** *BMC Genomics* 2005, **6**:176.
10. Jansen RK, Raubeson LA, Boore JL, DePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, *et al*: **Methods for obtaining and analyzing whole chloroplast genome sequences.** *Molecular Evolution: Producing the Biochemical Data, Part B* 2005, **395**:348-384.
11. Aronsson H, Jarvis P: **A simple method for isolating import-competent *Arabidopsis* chloroplasts.** *FEBS Letters* 2002, **529(2-3)**:215-220.
12. Jansen R, Raubeson L, Boore J, dePamphilis C, Chumley T, Haberle R, Wyman S, Alverson A, Peery R, Herman S, *et al*: **Methods for obtaining and analyzing whole chloroplast genome sequences.** In *Methods in Enzymology: Molecular Evolution: Producing the Biochemical Data, Part B.* Edited by: Zimmer E, Roalson E. San Diego: Elsevier Academic Press Inc; 2005:**348-384**:896.
13. Dean FB, Nelson JR, Giesler TL, Lasken RS: **Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification.** *Genome Research* 2001, **11(6)**:1095-1099.
14. Dean FB, Hosono S, Fang LH, Wu XH, Faruqi AF, Bray-Ward P, Sun ZY, Zong QL, Du YF, Du J, *et al*: **Comprehensive human genome amplification using multiple displacement amplification.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99(8)**:5261-5266.
15. Pfaffl MW: **A new mathematical model for relative quantification in real-time RT-PCR.** *Nucleic Acids Research* 2001, **29(9)**:e45.
16. Rozen S, Skaletsky HJ: **Primer3 on the www for general users and for biologist programmers.** In *Bioinformatics Methods and Protocols: Methods in Molecular Biology.* Edited by: Krawetz S, Misener S. Totowa, NJ: Humana Press; 2000:365-386.
17. Zhu L, Altmann SW: **mRNA and 18S-RNA coapplication-reverse transcription for quantitative gene expression analysis.** *Anal Biochem* 2005, **345**:102-109.
18. Zerbino D, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Research* 2008, **18(5)**:821-829.
19. Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A: **Geneious v4.7.** 2009 [http://www.geneious.com].
20. Shaw J, Lickey E, Schilling E, Small R: **Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III.** *American Journal of Botany* 2007, **94(3)**:275.
21. Winkworth RC, Grau J, Robertson A, Lockhart PJ: **The origins and evolution of the genus *Myosotis* L. (Boraginaceae).** *Molecular Phylogenetics and Evolution* 2002, **24(2)**:180-193.
22. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25(14)**:1754-1760.
23. Wyman SK, Jansen RK, Boore JL: **Automatic annotation of organellar genomes with DOGMA.** *Bioinformatics* 2004, **20(17)**:3252-3255.
24. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Research* 2009, **37**:D26-D31.
25. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1997, **25(17)**:3389.

doi:10.1186/1746-4811-6-22
Cite this article as: Atherton *et al.*: Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. *Plant Methods* 2010 **6**:22.