# Novel approaches for large-scale phylogenetics and applications in the context of the amphibian tree of life

A thesis submitted to the National University of Ireland for the Degree of Doctor of Philosophy

NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

Presented by:

Karen Y. Siu Ting Salvatierra M.Sc.

Department of Biology
NUI Maynooth
Maynooth
Co. Kildare, Ireland.

January 2014

**Supervisor:** Dr. Davide Pisani B.Sc., Ph.D. (Bristol)

**Co-supervisor:** Dr. Mark Wilkinson B.Sc., Ph.D. (Bristol)

**Head of Department:** Prof. Paul Moynagh Dip. Biology BA (mod.), PhD (Dublin)

# Table of Contents

# Index of Figures

# Index of Tables

# Declaration

This thesis has not been submitted in whole, or in part, to this, or any other University for any other degree and is, except where otherwise stated, the original work of the author.

Signed: ______________________________

Karen Yvette Siu Ting Salvatierra

# Acknowledgments

It has been a long road to this point, which has brought me through the most challenging and enriching learning experience, both in the academic and personal aspects of my life. I am incredibly thankful to everybody who contributed to make this a great journey.

I would like to express my deep gratitude to my supervisor, Dr. Davide Pisani, for giving me this amazing opportunity to do cool research, being so supporting during the rough times, and guiding me to the completion of all my research tasks, no matter how hard they were. You were the gravity that pulled the loose ideas “down to Earth” to become concrete and real.

I am equally and deeply grateful to my co-supervisor, Dr. Mark Wilkinson (NHM London), for the support, constant encouragement to follow my ideas and instincts, and the constructive and critical insights that pushed me to go beyond my boundaries. You taught me to think outside the box and accept that creativity is an important aspect in science.

I would also like to thank my examiners Dr. Diego San Mauro and Prof. James McInerney for their thorough and thoughtful comments and discussions of the work presented in this thesis.

Next, to Dr. Christopher Creevey, for being the light through this journey and the best companion I could ever have asked for. Your insightful opinions, enthusiasm and inspiration gave me the strength to carry on.

To Dr. Simon Loader (Univ. of Basel) and Dr. David Gower (NHM London) for the opportunity of applying my techniques and skills in such an interesting subject of study (*E. baleensis*) and learning about African Herpetofauna. To Dr. James Tarver (Univ. of Bristol) for your patient guidance during wetlab work and sharing your experience on the interesting and promising miRNAs. To Prof. James McInerney and Dr. David Fitzpatrick (NUIM), for your wise advice and friendship during my entire PhD.

To the former and present members of the Bioinformatics lab at NUIM:
Dr. David Alvarez, Dr. Omar Rota, Therese, Carla, Leanne, Aoife, Sinead and Rob, for providing great advice and thought-provoking discussions. You certainly made the lab a vibrant, welcoming and friendly place to work every day. In particular, I thank Roberto Feuda, Wasiu Akanni and Lahcen Campbell for support and advice in my analyses, and Dr. Misha Paturyan for the help in handling the clusters.

Also in NUI Maynooth, to Prof. Paul Moynagh, to Ms. Michelle Finnegan and Mrs. Terry Roche for the invaluable support in the administrative areas of my PhD. To the Irish Research Council (former IRCSET), for the scholarship to carry out this PhD.

To my friends in Maynooth, especially to Maria, Karl and Jess, for their constant support and providing me a place to crash in the moments of more need. To my friends in Bristol: ‘the gang’ (Rach, Sam, Bri, AJ, Lyndon and Graham), and Martín Ch. and the latin gang for making time in Bristol so enjoyable.

Finally to my Peruvian family and my new Irish family, thank you for your constant support and caring. In particular my parents, to whom I dedicate this work.

*The importance, for classification, of trifling characters, mainly depends on their being correlated with many other characters of more or less importance… Hence also, it has been found that a classification founded on any single character, however, important that may be, has always failed.*

*The Origin of the Species*, Charles Darwin.

# Abbreviations

(12S) Mitochondrial gene 12S rRNA

(16S) Mitochondrial gene 16S rRNA

(28S) Nuclear gene 28S rRNA

(AU) Approximately Unbiased test

(BLAST) Basic Local Alignment Search Tool

(BR) Bootstrap replicates

(BS) Bootstrap support

(CAT) Category-based Bayesian site-heterogeneous model

(CXCR4) C-X-C chemokine receptor type 4 gene

(DNA) Deoxyribonucleic Acid

(EST) Expressed Sequence Tag

(G) Gamma parameter

(GTR) General Time reversible

(H3A) Histone 3a gene

(HKY) Hasegawa-Kishino-Yano 85 model

(JC) Jukes Cantor 69 model

(K2P) Kimura-2-parameter model

(MCMC) Markov Chain Monte Carlo

(ML) Maximum Likelihood

(MR) Matrix Representation

(MS222) Tricaine methanesulfonate

(NGS) Next-generation sequencing

(OTUs) Operational taxonomic units

(PCR) Polymerase chain reaction

(POMC) Pro-opiomelanocortin gene

(RAG1) Recombinase activating protein 1 gene

(RHOD) Rhodopsin gene

(RNA) Ribonucleic Acid

(RT-PCR) Reverse transcription – Polymerase Chain Reaction

(SIA) Seventh-in-absentia gene

(SLC8A1) Solute-carrier family 8 member 1 gene

(SLC8A3) Solute-carrier family 8 member 3 gene

(STR) Safe Taxonomic Reduction

(TYR) Tyrosinase

(UTR) Untranslated Region

(UV) Ultraviolet

(ZNHM-AAU) Natural History Museum of Addis Ababa University, Ethiopia

(bp) Base Pairs

(cDNA) Complementary DNA

(cytb) Cytochrome b gene

(mRNA) Messenger RNA

(miRNA) MicroRNA

(nt) Nucleotide

(pre-miRNA) Long precursor miRNA

(pri-miRNA) Long primary transcript

# Abstract

During this thesis, I addressed some problems associated with large-scale phylogenetic analyses by tackling issues related to missing data and careful handling and addition of novel data in large-scale reconstructions, presenting an application of this approach in the context of amphibian phylogenetics.

I developed a method (called “Concatabominations”) building on the original Safe Taxonomic Reduction method (Wilkinson 1995) as an alternative approach to the issue of identifying rogue taxa. The safe removal of rogue taxa due to missing data can potentially reduce the terraces in tree space search and improve resolution in the final consensus tree. In a pragmatic point of view, the new method can help in targeting taxa that require further sampling during a research design.

Novel sequence data for the rediscovered *Ericabatrachus baleensis* allowed to explore its placement in the Amphibian tree of life. I tested the inclusion of novel data using a backbone alignment from a previous work (*de novo* analysis) and a backbone phylogenetic tree (constrained analysis), after careful curation of gene partitions to include in an analysis. I found that the use of a constrained phylogenetic inference using a previous accepted tree seems to be a practical solution to the rapid phylogenetic placement of a taxon in cases of well-supported relationships. However, a *de novo* analysis might ensure an optimal alignment and avoid risks introduced when adding new data.

Finally, I investigated the evolutionary relationships of the three lineages of the extant amphibians (Anura, Caudata and Gymnophiona) using an independent source of evidence: miRNAs, recently used to help resolve difficult phylogenetic problems. The analyses yielded a high number of shared miRNAs using the *Xenopus tropicalis* genome, contrasting with a lower number of miRNAs discovered using the Axolotl transcriptome. This suggests that not using genomic data is not ideal to validate miRNAs. Nevertheless, in spite of the limitations, I was able to find two potential novel miRNAs: one supporting the monophyly of Lissamphibia, and another supporting the Batrachia hypothesis.

Overall, I hope the work developed in this thesis contributes with new insights into large-scale phylogenetics and in particular to amphibian phylogenetics.

# Chapter 1. Introduction

## 1.1 Large-scale phylogenetics

With the advance of sequencing technologies the availability of sequence data for many organisms is growing at unprecedented rates. As larger numbers of genes for more organisms have become available, researchers included them in their phylogenetic analyses hoping that the inclusion of more data would help resolving the position of problematic taxa that could not be resolved using few genes only. This resulted in increasingly large concatenated data sets being generated (e.g. Hejnol *et al.* 2009; Philippe *et al.* 2009; Meusemann *et al.* 2010; Kocot *et al.* 2011; Rota-Stabelli *et al.* 2011 just to mention a few). However, as pointed out by Philippe *et al.* (2011b), including more data has actually brought up new problems and challenges, such as exacerbation of systematic biases (like long branch attraction), incorrect identification of orthologous genes, an increase in the amount of missing data in the final matrices, inclusion of saturated genes and lack of computational resources to process it in a reasonable time frame. In spite of these problems, there is a general consensus that phylogenetic studies will continue to incorporate more data, as in modern phylogenetics the relationships considered well-supported are mostly those backed by genomic scale data sets (Holton and Pisani 2010).

In the present work I have addressed issues involved in large-scale phylogenetics and phylogenomics, namely identifying unstable taxa due to missing data, adding new taxa in backbone phylogenies, and using novel genomic-scale data sets to resolve phylogenetic relationships.

## 1.2 Large-scale phylogenies of Amphibia

In order to investigate the use of backbone phylogenies to frame new taxa in the postgenomic era (chapter 3), I focused on the species phylogeny of the Amphibia. Also, in chapter 4 I investigated the relationships of the three major extant lineages of the Amphibia. Hence in this section I will briefly summarise the current status of amphibian phylogenetics.

Traditionally, most phylogenetic studies of amphibians have been based on morphological data, hypotheses drawn from nomenclatural and literature based phenotypic classifications until a major shift took place with the inclusion of molecular data in phylogenetics (Frost *et al.* 2006). In the beginning most of the studies focused on generic and infrageneric levels (Frost *et al.* 2006 and references therein). It was only in the 2000's that large-scale studies appeared addressing family level relationships and amphibian diversification (e.g. Biju and Bossuyt 2003; Darst and Cannatella 2004; Roelants and Bossuyt 2005; San Mauro *et al.* 2005; van der Meijden *et al.* 2005; Frost *et al.* 2006; San Mauro 2010; Pyron and Wiens 2011, just to mention some). From these, the work of Frost *et al.* (2006) was the first amphibian tree of life that carried out a broad sampling of the amphibians with an emphasis on including a large number of taxa. In that study, Frost *et al.* (2006) generated a large amount of molecular data (eight loci of mitochondrial and nuclear origin) for a total of 525 species, which was also combined with larval morphological data. But in spite of their enormous efforts to sequence as many taxa as possible for the loci they included, the coverage across the entire data matrix was quite low, with a large proportion (~77%) of missing gene entries. Added to this, the Frost *et al.* (2006) study was also criticized for the methods they used in the simultaneous generation of the alignment and consequent parsimony phylogenetic inference (based on Wheeler *et al.* 2006). A latter study, Pyron and Wiens (2011), using the same principle of including as many taxa as they could, amassed a large data matrix of 12 loci for 2871 taxa, with sequences that mostly originated from the work of Frost *et al.* (2006) and some sequences from Wiens *et al.* (2005). Again, the resulting concatenated matrix contained even higher amounts of missing entries (~80 %) and their phylogenetic tree was inferred using a more accepted model based method (Maximum likelihood).

Including missing data in phylogenetic analyses is an ongoing controversial issue that some regard should be treated with caution (Lemmon *et al.* 2009; Roure *et al.* 2013), and some defend as not having a negative impact (Wiens 2003; Philippe *et al.* 2004; Wiens 2006; Wiens and Morrill 2011). Indeed, in the Pyron and Wiens' (2011) data matrix the overlap across the largest majority of considered species comes from only two out of the 12 genes considered (16S rRNA and 12S rRNA). The arguments that Pyron and Wiens (2011) use to justify this approach can be summarized in three points: (i) that there is an overlap of two loci that allows for a backbone of the placement of taxa (Wiens 2003), (ii) that highly incomplete taxa can be accurately placed in model-based phylogenetic analyses if a large number of characters have been sampled (Wiens 2006; Wiens and Morrill 2011), and (iii) the high-support resulting trees reported in some studies that have also used a supermatrix approach (Driskell *et al.* 2004; Wiens *et al.* 2005; McMahon and Sanderson 2006; Thomson and Shaffer 2010; Pyron *et al.* 2011). Notwithstanding the large amounts of missing data and the controversies regarding the impact of including it (recently discussed in Roure *et al.* 2013), the Pyron and Wiens' (2011) tree represents the most comprehensive phylogeny of Amphibia to date.

In terms of the evolutionary hypotheses at the basal level of the amphibian tree of life, molecular approaches have provided the opportunity for researchers to investigate them with more data (e.g. Hedges and Maxson 1993; Feller and Hedges 1998; Zardoya and Meyer 2001; San Mauro *et al.* 2005; Roelants *et al.* 2007; San Mauro 2010; Pyron 2011). Traditionally studies addressing the evolutionary hypotheses of the main extant lineages of Amphibia (the Anura, Caudata and Gymnophiona) have been based on morphological characters, primarily from fossils (e.g. Milner 1988; Trueb and Cloutier 1991; Vallin and Laurin 2004; Ruta and Coates 2007; Marjanović and Laurin 2008, 2009; Sigurdsen and Green 2011; Maddin and Anderson 2012, just to mention some), so the inclusion of molecular characters promised new insights. However this does not seem to have been the case and there is still controversy regarding the monophyly of the crown group of extant Amphibians (known as Lissamphibia) (Carroll 2007; Anderson *et al.* 2008; Fong *et al.* 2012), and between the two main hypotheses within the Lissamphibia (Batrachia and Procera) (Milner 1988; Trueb and Cloutier 1991; Vallin and Laurin 2004; Ruta and Coates 2007; Marjanović and Laurin 2008, 2009; Sigurdsen and Green 2011; Maddin and Anderson 2012). A review of these issues is presented and addressed in chapter 4.

## 1.3 Inferring phylogenies

Approaches to phylogenetic reconstruction is a major topic in this thesis and in the following sections I will introduce the topics fundamental to this area of research.

### 1.3.1 Homology

Homology is the underpinning concept in comparative biology. Richard Owen defined the term for the first time in 1846 as ‘the same organ under every variety of form and function’. Back then, in pre-Darwinian times, the definition of homology was ruled by structure and location instead of ancestry (Fitch 2000). Nowadays, the definition of homology has been adapted to the Darwinian ideas of descent with modification: the changes observed in the structures (or characters) are interpreted in the light of inheritance from a shared ancestor. So homology can be better defined now as ‘the relationship of two characters that have descended, usually with divergence, from a common ancestral character’ (Fitch 2000).

Homology is usually determined on the basis of the similarity of characters. However this has caveats, because in many cases (i.e morphological, behavioral or functional characters) they could be a product of parallel evolution (that is, the characters did not originate from the same ancestor). Nonetheless, similarity is a useful proxy to infer homology in molecular characters, and the more similar sequences are, the most likely it is that they have evolved from the same ancestor (but this must not be confused by numerical quantification of similarity, homology is categorical, either two characters are homologous or not). Different methods such as BLAST (Altschul *et al.* 1990) are used in molecular biology to identify homologous sequences.

The three major subtypes of homology are orthology, paralogy and xenology. Orthology is where sequence divergence is the product of a speciation event; paralogy is where the sequence divergence is a product of gene duplication; and xenology is where sequence divergence is a product of lateral gene transfer (Fitch 2000). From these three, only orthologous sequences can be used to infer the phylogeny of species.

### 1.3.2 Positional Homology and Multiple Sequence Alignment

In addition to requiring the use of homologous sequences, phylogenetic analyses of sequence data require the identification of positional homologies (Swofford *et al.* 1996). A sequence alignment is a collection of hypotheses of homology for the residues in protein or nucleotide sequences, meaning that the aligned residues are assumed to have diverged from a common ancestral state (Higgins and Lemey 2009). These sites identify positional homologies. The first approach developed to align sequences carried it out in a pairwise manner. Pairwise alignments seek to align two entire homologous regions using a balance of matches and gaps (Hillis *et al.* 1996). Gaps are introduced to account for insertion/deletion events, however because any two sequences can be aligned perfectly together just by adding enough gaps, these are penalized. Additionally, all substitutions are assigned a penalty in an alignment, or a matrix of change costs may be specified (Sankoff and Rousseau 1975). However, given that in phylogenetics multiple species are compared at the same time, pairwise sequence alignments are insufficient. To recover phylogenetic trees, given a set of gene sequences, multiple sequence alignments must be retrieved. A first approach to generate multiple sequence alignments was to use the pairwise alignments of all the sequences and add the sequences together in a stepwise manner by inserting gaps as needed (Hillis *et al.* 1996). However this method depends on the order in which sequences are added, meaning that if sequences are added in different order, different alignments are achieved. As an alternative the progressive alignment was introduced which uses a guide tree to choose the order of addition of sequences (Feng and Doolittle 1987, 1990). This method is one of the most commonly used but it does not question the quality of the guide tree. Another method suggested by Sankoff *et al.* (1973) is to calculate the alignment at the same time as the tree. Even though this last method sounds conceptually good, there have been very few programs implementing this approach and none that did so efficiently due to the intensive computation necessary for such an approach (Wheeler *et al.* 2006). Among the most popular novel variants of the progressive method are the iterative method and the phylogeny aware methods. The iterative method repeatedly realigns the initial sequences and adds new sequences simultaneously to the growing multiple sequence alignment (e.g MUSCLE –Edgar 2004). In phylogeny aware methods, insertions are distinguished from deletions and repeated penalization of insertions are avoided, producing gaps that are phylogenetically consistent but spatially less concentrated in the alignments (e.g. PRANK –Löytynoja and Goldman 2005). When the alignment is complete, it is often necessary to verify the alignment for misaligned regions by eye. In this thesis, I used the iterative method, one of the most popularly used.

### 1.3.3 Phylogenetic methods

Given a multiple sequence alignment there are a variety of methods that can be used to derive a phylogenetic tree.

#### 1.3.3.1 Parsimony

One of the oldest and most widely used numerical methods for inferring phylogenies is Maximum Parsimony (Edwards and Cavalli-Sforza 1963). This method is based on the maximum parsimony principle, which states that simpler hypotheses are preferable over more complicated ones. Translating this concept into phylogenetics: "The most plausible estimate of the evolutionary tree is that which invokes the minimum net amount of evolution" (Edwards and Cavalli-Sforza 1963). Parsimony works by determining the fit of each of a group of characters to a set of trees, with the aim of finding the tree over which the evolution of the considered character can be described using the minimal number of evolutionary changes. Hennig's (1966) work was the first to formalise the concepts of derived (apomorphic) and of ancestral (plesiomorphic) character states (Figure 1.1 a) and to point out that phylogenies should be inferred based only on the information in shared derived characters. In Hennig's work an apomorphic character state is one that can be considered derived (i.e. different from the ancestral state observed in the outgroup). Hence, characters are synapomorphic (the prefix *syn* means "shared") if the derived condition is shared between a group of operational taxonomic units (OTUs) and their common ancestor (Figure 1.1 b). For example: mammary glands are a synapomorphic condition that define mammals but that exclude it from all other vertebrates. In order to be phylogenetically informative, apomorphic characters must be shared (or be synapomorphic) with more than one OTU (Figure 1.1 b), thus autapomorphies (which are derived characters only present in one OTU) are not informative (Figure 1.1 c). On the other hand, plesiomorphic characters are those that retain the ancestral state, and are generally not informative to group OTUs together (Figure 1.1 a), even when shared among a group of OTUs (symplesiomorphic). Using the previous example, even though mammary glands are a synapomorphy for mammals, it is not a character that can help us determine the phylogenetic relationships between the different groups within mammals.

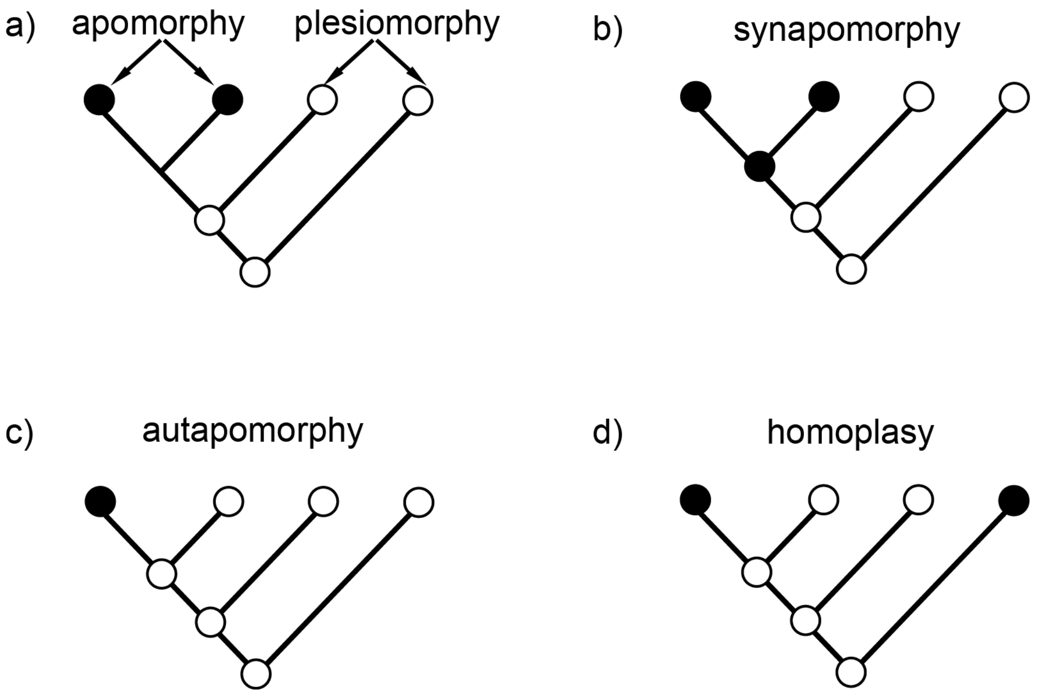


**Figure 1.1** Trees showing the different types of ancestral and derived character states.
White circles represent ancestral states and black circles represent derived states. Modified from Page and Holmes (1998).

In an ideal world, knowing which character state is ancestral and which is derived, and assuming that characters can change only once, it should be straight forward to infer monophyletic groups from synapomorphies. However, in phylogenetics it is very common for different characters to support incompatible phylogenetic relations. In such cases, an underlying "true phylogenetic tree" is derived based on the set of all available characters, and characters that cannot be parsimoniously described on this tree are considered to display homoplasy (to represent independent acquisition of a given state, Figure 1.1 d). Once characters (with their states) are defined (i.e. when a set of sequences have been aligned – in the case of molecular data), OTUs are grouped together based on shared apomorphies using parsimony optimization. The most common algorithm for parsimony optimization being the Fitch parsimony algorithm (Fitch 1971).

Under parsimony, the number of substitutions necessary to explain the distribution of states at the tip of a phylogeny (for each character in a data matrix) is estimated, and the sum of such scores across all characters in a matrix is used to define the length of a tree. The tree of shortest length (assuming the overall minimum amount of evolution) is the most parsimonious tree selected by the parsimony optimization. The maximum parsimony method has been shown to be sensitive to the effect of systematic biases (Felsenstein 1978) and is rarely used now in molecular phylogenetics.  However, it is still the method of choice in morphological data analysis (see chapter 2).

#### 1.3.3.2 Character compatibility

E. O. Wilson described the ideal phylogenetic character as one that “both uniquely defines a set of species and has not been reversed in evolution, so that all existing species which possess this state can be said to have descended from one species in the past that evolved the state” (Wilson 1965). When carrying out a phylogenetic analysis, ideally one would like to know *a priori* which characters in a matrix are noisy in order to exclude them from the analysis. Those characters include those that are homoplastic. A character is defined as compatible with a tree if its evolutionary history can be described on that tree, without the need to assume homoplasy. Building on this one can say that two (or more) characters are compatible if they can be mapped on the same tree without adding homoplasy (Le Quesne 1969). Le Quesne (1969) suggested a simple approach to select the characters to be used for phylogenetic reconstruction, as the largest set of pairwise–compatible characters in a given matrix.

Pairwise compatibility of characters can be determined very simply with a test introduced by Wilson (1965) that assesses the combination of character coding they have. In the example provided (Table 1.1), character 1 and 2 are compared per taxon: Alpha has (1,0), Beta has (0,0), Gamma has (1,1), and so on. In summary, three combinations are found: (1,0), (0,0) and (1,1), which can be annotated in a separate matrix (Table 1.2). According to Wilson (1965), if all possible pairwise combinations are present, then the two characters cannot be compatible because they cannot occur in the same phylogeny without one of the characters changing at least twice. In this sense, from the example in Table 1.1, character 1 and 2 are compatible, however character 1 and 4 are not compatible, given that all four possible combinations of states emerge for these characters (Table 1.3).

Compatibility analysis was used as a tool to derive phylogenies in the early seventies (when it was generally referred to as "Clique analysis" – Estabrook *et al.* 1977; Wilkinson 1994b), however, this method is not used anymore to derive phylogenies because, similarly to parsimony, it is easily affected by systematic artifacts. However, character compatibility is still used to determine specific aspects of the data (e.g. their evolutionary rate – Pisani 2004; Wagner 2012). In the context of this thesis, compatibility analysis is used (in chapter 2) to derive a method to deal with missing data in phylogenetic data sets.

**Table 1.1** Hypothetical data set with four binary coded characters (modified from Felsenstein 2004).

| | Characters | | | |
|---|---|---|---|---|
| Taxon | 1 | 2 | 3 | 4 |
| Alpha | 1 | 0 | 0 | 1 |
| Beta | 0 | 0 | 1 | 0 |
| Gamma | 1 | 1 | 0 | 0 |
| Delta | 1 | 1 | 0 | 1 |
| Epsilon | 0 | 0 | 1 | 1 |
| Omega | 0 | 0 | 0 | 0 |

**Table 1.2** Pairwise combinations found for characters 1 and 2 of Table 1.1

| | Char 2 | |
|---|---|---|
| Char 1 | 0 | 1 |
| 0 | ✓ | |
| 1 | ✓ | ✓ |

**Table 1.3** Pairwise combinations found for characters 1 and 4 of Table 1.1

| | Char 4 | |
|---|---|---|
| Char 1 | 0 | 1 |
| 0 | ✓ | ✓ |
| 1 | ✓ | ✓ |

### 1.3.3.3 Maximum Likelihood

The Maximum Likelihood (ML) method estimates the tree that maximizes the probability of observing the given data (i.e. the alignment) given a model of substitution. Likelihood methods for phylogenetics were first introduced by Edwards and Cavalli-Sforza (1964). Usually, ML implementation in phylogeny reconstruction is mostly focused in molecular sequence data (DNA or amino acids), but there are some models for morphological characters (see Lewis 2001). The likelihood of a molecular phylogeny is calculated by choosing a fixed model of substitution that accounts for the conversion of one sequence into another (Swofford *et al.* 1996). Phylogenetic inference using ML has two steps: the first one determines the tree topology, branch lengths, and parameters of the evolutionary model that maximize the probability of observing the data (typically, the alignment or data set); and second, a tree search for the tree with the highest likelihood is performed using traditional tree search methods that use hill-climbing algorithms (see Figure 1.2 a). Because likelihood values can be very small, these are usually reported as the logarithm of the likelihood (log-likelihood). Hence, the best log-likelihood score is the one with the largest value. An aspect to bear in mind when assessing the trees obtained under this method is that likelihood will not discern if the substitution model used fits the data, it will only assess if the model plus the tree can have generated the data. The 'correctness' of a model can be assessed using other methods, such as Goldman's test (Goldman 1993) and there are also a variety of statistical strategies to help on the selection of the best evolutionary model: for instance the Akaike (Akaike 1974) and Bayesian information criteria (Schwarz 1978), and the likelihood ratio test (Neyman 1971), among the most popular.

Advantages of this method are that it allows the use of an array of models of nucleotide or amino acid substitution to assess the likelihood. The likelihood method is statistically consistent. However, it has the disadvantage of being time-consuming and of requiring a lot of computational effort. Fortunately, in recent years, some programs have been developed to infer ML trees parallelizing the computational load across multiple processors (e.g. RAxML – Stamatakis 2006, Garli – Zwickl 2006) allowing for these complex calculations to be done much more quickly (however still depending on the availability of a considerable amount of processors and memory).

#### 1.3.3.3.1 Assessing best trees

Finding the best trees is the aim in phylogenetics. All of the methods described previously use an optimality criterion to calculate a value for each of the candidate trees and then select the tree with the best score (i.e. the most parsimonious, the highest likelihood). However, it is common that one single tree is not the only solution. One way to assess tree topologies is by using support values, which can be obtained by bootstrapping, jackknife, or even Bayesian MCMC sampling (Schmidt 2009).

In ML the most common way to assess confidence of the results is by carrying out a bootstrap analysis (Figure 1.2 a). Essentially, bootstrapping (Felsenstein 1985) is a general resampling technique. Characters from the original matrix are resampled with replacement to generate a predefined number of bootstrapped data sets, each with the same dimensions as the original. These are generally called pseudoreplicate data sets. For each pseudoreplicate data set a ML analysis is performed and the set of trees generated are summarized using a consensus tree method (explained in section 1.3.3.7). Finally, all the splits from the optimal ML tree (obtained from the original data set) are retrieved across the bootstrapped trees and counted, and their frequencies constitute the bootstrap support (BS) for each cluster in the ML tree. The bootstrap method is very flexible and can be applied to all phylogenetic methods (including Bayesian analysis – Pisani *et al.* 2013).

Several statistical methods are available to carry out hypothesis testing between topologies for parsimony and ML trees, however in this work I only used the ones applied in ML. Those used in ML essentially use the likelihood values for the trees (which is obtained from the product of all the likelihoods per site in an alignment) to carry out a hypothesis testing. The Kishino-Hasegawa (KH) test (Kishino and Hasegawa 1989) compares the log-likelihoods of two *a priori* selected trees (Ta and Tb) and produces a probability value (p-value) for each tree, which is essentially a number ranging from zero to one and that represents the possibility that the tree is the true tree. A higher p-value indicates that the probability that the tree is the true tree is also higher (Shimodaira 2002). Hence it is expected that the true tree will be among the non-rejected hypotheses. However the KH test was devised for *a priori* selected trees and it can be biased when testing if sets of suboptimal trees are equally supported or significantly worse than the best tree (Schmidt 2009), and cannot control for type-1 errors (Shimodaira 2002). The Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa 1999) assesses the likelihoods of a set of *a posteriori* selected trees when the maximum likelihood tree is among the tested trees (Schmidt 2009). The SH test requires that the tree with the maximum likelihood has to be included in the group of tested trees, but a bias that this test has is that it is strongly correlated with the number of trees tested (that is, the more trees are included, the fewer trees will be rejected), rendering it more conservative. In comparison to the KH test, it controls the type-1 error, but it is heavily biased (Shimodaira 2002). The approximately unbiased (AU) test (Shimodaira 2002) is based on a multiscale bootstrap and it allows for a correction of the selection bias. This test essentially resamples the input per-site log-likelihood alignment changing its length (it can be shorter or longer), and then the new bootstrapped probabilities are scaled to the original input alignment length (Schmidt 2009). The number of times the hypothesis is supported by the replicates is counted for each bootstrapped set. The AU test then calculates the p-value (which is approximately unbiased in this case) from the change in the bootstrapped probabilities. This type of comparison makes this method "approximately" less biased to the number of trees (hypotheses) included and it also controls the type-1 error (Shimodaira 2002). Even though the AU test was devised as an alternative to the two previous methods, the results from the AU and KH tests seem to be correlated (Shimodaira 2002; San Mauro *et al.* 2004). The AU test is currently the most used method in ML to assess for best topologies.

#### 1.3.3.4 Bayesian Inference in phylogenetics

Bayesian methods are very old and they can be referred back to a paper published posthumously by Thomas Bayes in 1763. Unfortunately, after its publication, the ideas of Bayes were not particularly popular and they only started to gain new popularity in the last half of the 20th century, in particular in phylogenetics with the works of Rannala and Yang (1996), Yang and Rannala (1997), Mau and Newton (1997), Li *et al.* (2000).

Bayesian methods are very closely related to likelihood methods in the sense that they start with a model, and Bayesian methods also involve the computation of the likelihood of the trees given the data. However, in Bayesian methods the ML function is embedded within the Bayes formula where it is compounded with a prior probability in order to derive posterior probabilities (see formula below and Figure 1.2 b).

$$\Pr[H|D] = \frac{\Pr[H] \times \Pr[D|H]}{\Pr[D]}$$

In this formula the probability (Pr) of the hypothesis "H" given the data "D" is equal to the product of the probability of the hypothesis "H" and the conditional probability of the data "D" given the hypothesis "H", divided by the probability of the data "D".

Even though there are still controversies surrounding the use of Bayesian methods (i.e. the use of priors for unknown parameters, conditional probabilities and complex computational calculations for multidimensional integrals) (Felsenstein 2004), these approaches gained in popularity with the development of the Markov chain Monte Carlo (MCMC) algorithms based on the Metropolis-Hastings algorithm (Metropolis *et al.* 1953; Hastings 1970). The MCMC approach speeds up Bayesian analyses and the estimation of problems that would be intractable otherwise. In the case of phylogenetics, the development of Bayesian MCMC approaches finally allowed the use of complex models to analyse large-scale data sets. This opened the road for the development of more realistic models like the CAT-based models that even today can only be properly implemented within a Bayesian framework.

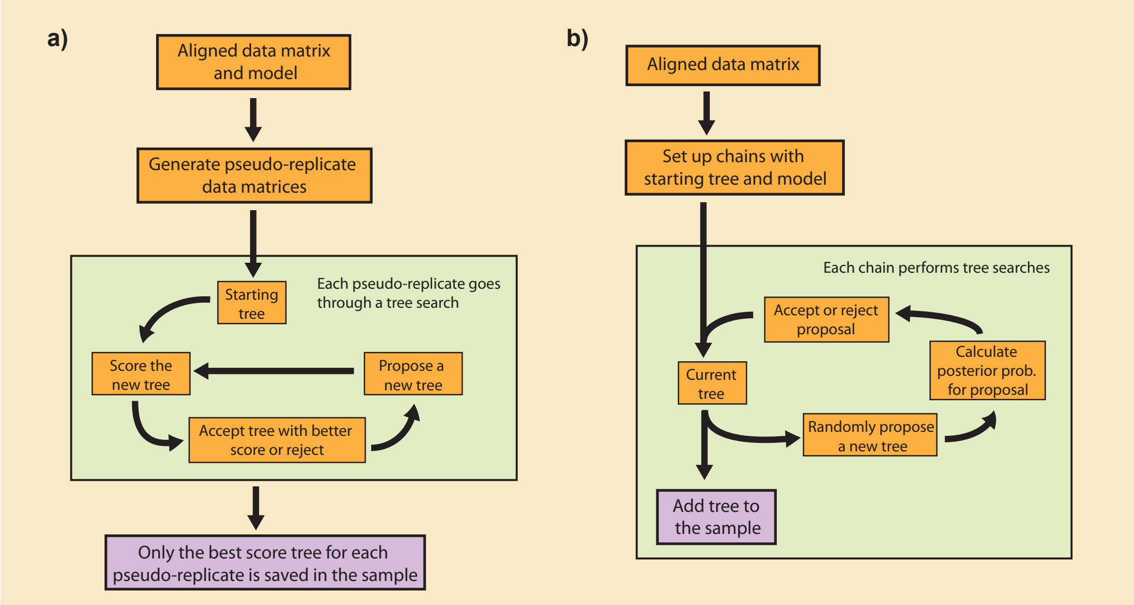

a)
Aligned data matrix and model
Generate pseudo-replicate data matrices
Starting tree
Each pseudo-replicate goes through a tree search
Score the new tree
Propose a new tree
Accept tree with better score or reject
Only the best score tree for each pseudo-replicate is saved in the sample
b)
Aligned data matrix
Set up chains with starting tree and model
Each chain performs tree searches
Accept or reject proposal
Calculate posterior prob. for proposal
Current tree
Randomly propose a new tree
Add tree to the sample

**Figure 1.2** Difference between ML and Bayesian methods in phylogenetic inference.

a) Summarises traditional tree search and bootstrapping to assess support (as carried out in ML methods) and b) summarises Markov-chain Monte Carlo method for tree search as done in Bayesian methods. Modified from Holder and Lewis (2003).

### 1.3.3.5 Models of DNA evolution

Comparing homologous sites from a pair of DNA sequences can be tricky because at a superficial glance, the same site (in two different sequences) is either occupied by the same nucleotide or (amino acid) (Figure 1.3 a) or by two different nucleotides (Figure 1.3 b). Hence, between any two sequences (at a given site) we can only see 1 or 0 differences. From this, the p-distance is calculated as the sum of the differences between two sequences divided by the total number of sites compared. Yet, it is clear that in the evolutionary history of a site, multiple substitutions might have occurred, but these are completely overwritten and hence cannot be directly observed by simply looking at the sequence data. Therefore, this cannot be accounted for using the p-distance (Figure 1.3 b). Given a set of aligned sequences, substitution models can be used to estimate, the real number of substitutions that is likely to have occurred at each site. The substitution models can then be used to inform distance methods (phylogenetic methods based on the use of genetic distances between sequences), maximum likelihood and bayesian methods. Currently, there are many models of nucleotide substitution for DNA sequences, but I will only focus in those that were used in the analyses carried out in this thesis. Standard nucleotide substitution models are composed of two elements. The first is a 4X4 matrix of substitution rates identifying the probability of each nucleotide changing into any other nucleotide. The second is a nucleotide frequency vector indicating the frequency with which we expect each nucleotide to be present in the data set. The simplest model of nucleotide substitution is the Jukes and Cantor model (Jukes and Cantor 1969, JC). In this model all possible nucleotide substitutions are assumed to have the same probability of occurring, and all nucleotides are expected to have the same frequency (Table 1.4).

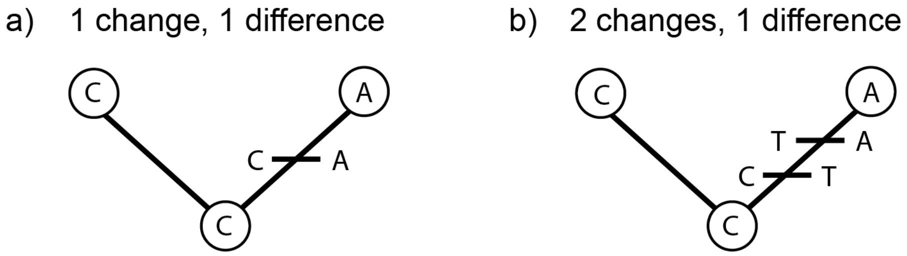


**Figure 1.3** Accounting for nucleotide substitutions when comparing two homologous sequences.

In the first case (a), only one substitution took place, while in the second case (b) two substitutions took place from the ancestral nucleotide C. For b), the real number of substitutions is miscalculated if a simple p-distance had been used. Modified from Page and Holmes (1998).

The Kimura 2-parameter (K2P) model (Kimura 1980) is an extension of the JC model. Basically, the K2P extends the JC model by assuming two rates of nucleotide substitution in the rate substitution matrix. These different rates were introduced to take into account that not all nucleotide substitutions have the same probability of happening, with transitions being much more common than transversions (Table 1.4). Similarly to the K2P model, the HKY85 model (HKY henceforth), proposed by Hasegawa *et al.* (1985), assumes two substitution rates (one for transitions and one for transversions). However, it also assumes that nucleotide frequencies are not all equal. Instead the frequency at which each nucleotide appears in the data set is estimated directly from the data (see Table 1.4). Because of this, the HKY model will be more accurate when modeling sequences with, for instance, higher GC content (as is common in several taxonomic groups), and calculates the rate accordingly. However, one needs to point out that also the HKY model is still essentially unrealistic and that the real substitution process is much more complex than that. This led to the development of more complex models, although, simple models like the K2P and HKY are still popular and easier to understand because of the reduced number of parameters they have.

The most general model that can be considered is the General Time Reversible (GTR) model (Lanave *et al.* 1984). In a GTR model each rate of substitution in the 4X4 rate matrix is unique (generally inferred directly from the data) and all frequencies with which nucleotides are expected to appear in the data are different (and inferred from the data) (Table 1.4). Clearly the JC model can be seen as a special case of a GTR matrix (one in which all substitution rates are the same and all nucleotide frequencies are the same). Similarly, all the other previously described models are special cases representing specific simplifications of the standard GTR model.

The GTR and the HKY are reversible, that is, the rate to convert from a nucleotide A to T is the same as from T to A (as can also be evidenced by the symmetrical matrices these have, shown in Table 1.4). These qualities generally make it a much more realistic model of evolution than the simpler models of evolution. The GTR model can even be made more complex through the use of a gamma correction (Yang 1996), which allows different nucleotide positions to evolve at different rates. Due to its flexibility it has a high parametisation: 6 rates (a,b,c,d,e) + 4 frequencies (all noted as $\pi$) + 1 Gamma correction = 11 parameters.

**Table 1.4** Substitution rate matrices for the JC, K2P, HKY and GTR model.

The frequency distribution used is constant for JC and K2P; and $\pi = (\pi T, \pi C, \pi A, \pi G)$, where $\pi T \neq \pi C \neq \pi A \neq \pi G$ for HKY and GTR. Transitions ($\alpha$) and transversions ($\beta$) are taken into account in the K2P and HKY. The complex GTR model takes into account different rates for the conversion between nucleotides (noted by the different values from “a” to “f”). Table modified from Yang (2006).

| | | To | | | |
|---|---|---|---|---|---|
| Model | From | T | C | A | G |
| JC | T | · | $\lambda$ | $\lambda$ | $\lambda$ |
| | C | $\lambda$ | · | $\lambda$ | $\lambda$ |
| | A | $\lambda$ | $\lambda$ | · | $\lambda$ |
| | G | $\lambda$ | $\lambda$ | $\lambda$ | · |
| K2P | T | · | $\alpha$ | $\beta$ | $\beta$ |
| | C | $\alpha$ | · | $\beta$ | $\beta$ |
| | A | $\beta$ | $\beta$ | · | $\alpha$ |
| | G | $\beta$ | $\beta$ | $\alpha$ | · |
| HKY | T | · | $\alpha\pi C$ | $\beta\pi A$ | $\beta\pi G$ |
| | C | $\alpha\pi T$ | · | $\beta\pi A$ | $\beta\pi G$ |
| | A | $\beta\pi T$ | $\beta\pi C$ | · | $\alpha\pi G$ |
| | G | $\beta\pi T$ | $\beta\pi C$ | $\alpha\pi A$ | · |
| GTR | T | · | $a\pi C$ | $b\pi A$ | $c\pi G$ |
| | C | $a\pi T$ | · | $d\pi A$ | $e\pi G$ |
| | A | $b\pi T$ | $d\pi C$ | · | $f\pi G$ |
| | G | $c\pi T$ | $e\pi C$ | $f\pi A$ | · |

Even though complex models are expected to produce more realistic estimates of substitution rates, these become unreliable when there is strong rate variation among sites (Yang 2006). Trying to model realistic across site rate heterogeneity is difficult and this is generally done by assigning overall substitution rates to sites from a gamma distribution (Yang 1994).

More recently, the "CAT"-based models were introduced by Lartillot and Philippe (2004). This is a class of site heterogeneous mixture models where sites are classified into categories (hence the name "CAT") based on their biochemical composition (as inferred from the alignment). The CAT model clusters columns of the alignment into biochemically specific classes (*K*), each described by its own profile and equilibrium frequency (depending on whether it is amino acid or nucleotide data). Then, each site in the alignment is assigned a category which is then combined with a globally defined set of exchange rates, thus resulting in site-specific substitution processes (Lartillot *et al.* 2009). Once the site-partitioning is obtained, substitution is inferred within each category allowing for different levels of complexity. For example, in the simpler CAT model, substitution within each compositionally defined category is modeled using a Poisson process. In the more complex CAT-GTR model, a GTR model is specified and applied to the compositional partitions. In the even more complex QMM model, a category-specific GTR model is assigned to each category. For each one of these complex models among site rate variation can be further modeled (as in standard GTR models) using the Gamma distribution. Generally, the CAT models define an average of about 130 categories in which to partition the sites in the alignment. In addition to these categories, the parameters of one or more GTR matrices need to be estimated. Hence, the CAT-based models are complex and very parameter rich and therefore can only be effectively used to analyse large alignments (i.e. not shorter than 1000 positions long) (Lartillot and Philippe 2004). The CAT-based models were originally designed for amino acid data but they work well also for nucleotidic data sets (e.g. Rota-Stabelli *et al.* 2013).

An alternative way to achieving variation in evolutionary processes among different portions of an alignment is partitioning data in blocks. Partitioning is very useful when carrying out phylogenetic inference with ML, and is particularly useful when dealing with large sequence data sets. In partitioning, sites that are assumed to have evolved under similar processes, independent substitution models can be applied for each group (Lanfear *et al.* 2012). Choosing an appropriate partitioning scheme is very important, when not using mixture models of evolution (like the CAT model described before – which automatically partition the data in a biochemically meaningful way). When using standard GTR models, partitions have to be defined *a priori* and their division is often based on genes and codon positions. However, this can lead to overparameterization, as codon positions of similar genes can similar rates and patterns of substitution. Overparameterization happens when one ends up with more parameters than can be reliably estimated from the available data (Lanfear *et al.* 2012). To improve phylogenetic reconstruction similar partitions might be better clustered together into a single one, rather than independently.

#### 1.3.3.6 Sequence Saturation

Different genes diverge at different rates according to the their gene-specific rate of substitution. As the evolutionary distance between two sequences increases, the chances of multiple mutations having occurred in the same position in either sequence since they last shared a common ancestor increases. This means that at low evolutionary distances the observed number of mutations is likely to represent all the mutations that have occurred, but as the evolutionary distance increases, so does the chance that "hidden" mutations (those which cannot be observed) have occurred. This can be illustrated as a line graph, where (in a simple model of evolution) the actual number of mutations that have occurred between two sequences since they last shared a common ancestor increases linearly over time (Figure 1.4). When the number of observed mutations is plotted onto the same graph, this number reaches a plateau when it reaches "saturation", in other words, when we cannot observe any more mutations as they are occurring where others have already occurred (Figure 1.4). Evolutionary models account for this in various ways, and given a substitution model and an alignment, the number of mutations that occurred at each site can be estimated from the observed set of mutations (Page and Holmes 1998). Saturation is a problem for phylogenetic reconstruction because in extreme cases when sequences have undergone full saturation, the similarity between the sequences depend entirely on the similarity in nucleotide frequencies (Xia and Lemey 2009), which generally does not reflect phylogenetic relationships but mutational pressures at the genome level (e.g. codon usage biases – Rota-Stabelli *et al.* 2013).

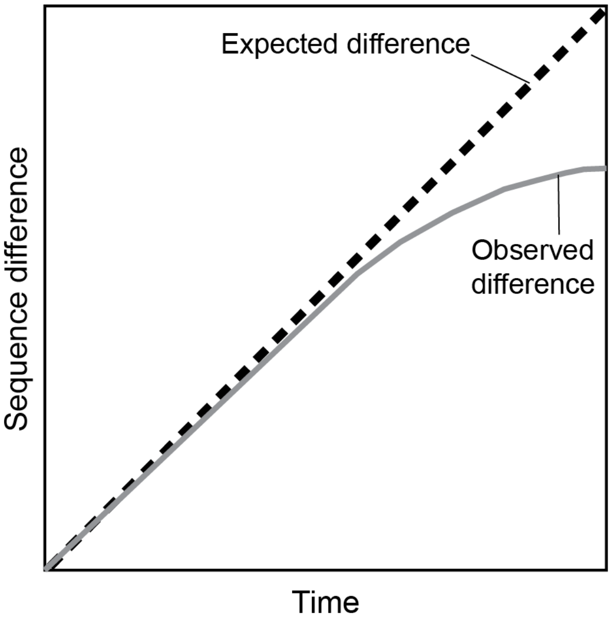


**Figure 1.4** Expected versus observed nucleotide substitutions over time. Nucleotide substitutions are expected to increase linearly over time (dotted line) when using the right model. When the observed differences reach a plateau (grey line), no more differences can be accounted for with the model, hence they have reached "saturation". Modified from Page and Holmes (1998).

When testing for saturation however, the sequences from any two organisms are a single point on this evolutionary continuum, and it would be necessary to have examples of the same sequences from the past and the future in each organism to plot the same saturation graph and determine if they are saturated for change. As a proxy for having this temporal insight, however it is possible to use an entire set of orthologs. In this approach, a phylogenetic tree of the gene family is reconstructed first using observed (p-distances) or distances inferred using a simple model (e.g. JC), and after that using a more complex model that can account (at the least to some extent) for hidden mutations (e.g. HKY or even GTR). The patristic distances (the tip-to-tip distances, Fourment and Gibbs 2006) between all sequences on the tree are calculated and plotted against each other. For those sequences from closely related organisms the number of mutations that have occurred since they last shared a common ancestor is likely to be few, and both the simple and complex models should estimate the same evolutionary distance. However for the comparison between the sequences from two distantly related species, the complex model is likely to predict a larger evolutionary distance than the simple model, as the complex model will better estimate hidden mutations. When all the pairwise calculations from both models are plotted against each other, a plot similar to those shown in Figure 1.5 is produced. If the plot is linear, then the complex model did not predict more mutations for sequences separated by a long evolutionary distance than the simple model and no saturation is predicted to have occurred (Figure 1.5 a). However if the plot reaches a plateau, then at the larger evolutionary distances the complex model calculated a large number of hidden mutations indicating that there is saturation in the data (Figure 1.5 b). This approach can be used to filter gene families that are saturated and that might cause phylogenetic inaccuracies (saturated sequences exacerbate systematic artifacts). I carried out analyses of saturation to analyse the data for inclusion prior to the phylogenetic analysis in chapter 3.

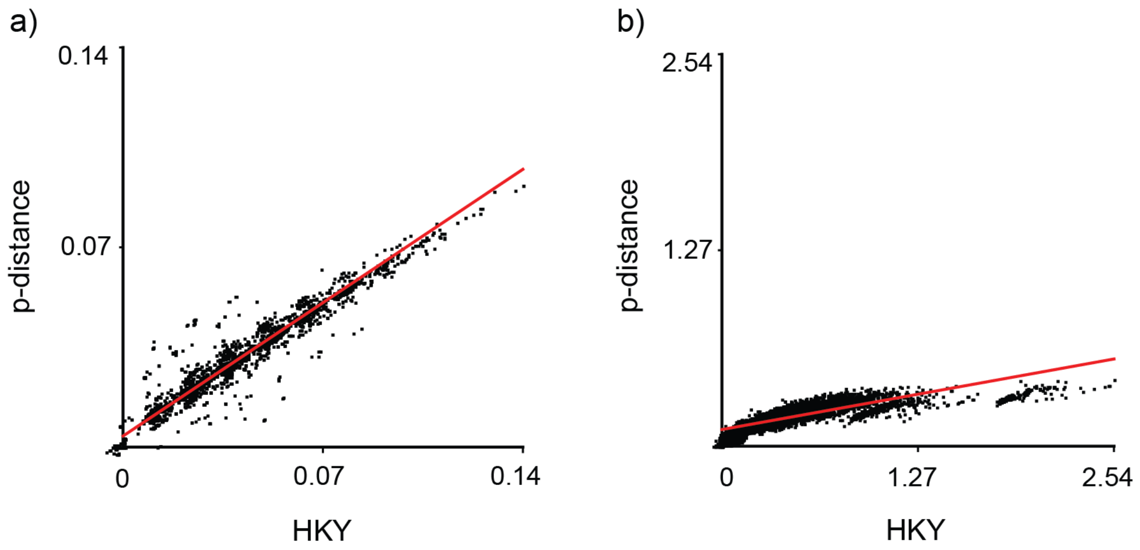


**Figure 1.5** Patristic distance plots of sequences of the gene H3A.

(a) Showing no saturation in the 1st codon position and (b) showing saturation in the 3rd codon position. See chapter 3.

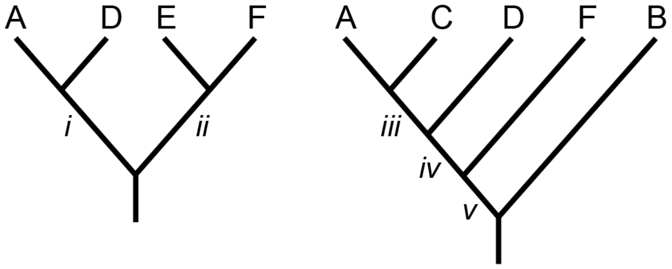
1.3.3.7 Consensus and Supertree methods

Consensus methods summarise a collection of trees to output a single "representative" tree (Bryant 2003). These methods basically consist of a function or algorithm that takes the topological information of all the input trees (which share the same set of taxa) and gives back a resulting single topology for the same set of taxa. Most consensus methods work by identifying common areas in the input trees and then representing these in the final resulting tree. Because of this feature, consensus methods are also very useful for identifying areas of conflict in the input trees. Bryant (2003) summarised these methods as (1) those based on frequency of splits, (2) those based on cluster intersections, (3) based on subtrees and (4) based on recoding. During this thesis I only used methods based on frequency of splits (Strict and Majority Rule extended consensus), and based on recoding (MRP matrices in supertrees).

The strict consensus is a frequency-based method where only monophyletic groups found in all source trees are produced in the resultant tree. Since it only takes those splits that appear in all of the input trees, the resulting consensus tree is often poorly resolved. However, because of this, it is often considered a conservative method, and it is particularly useful for identifying conflicting topologies. I used this type of consensus in chapters 2 and 3 to compare if trees had conflicting topologies and determine splits that were present in all input trees.

The majority rule consensus is also a frequency-based method and it takes clusters or splits that appear in at least half of the input trees. However those splits appearing in less than half of the input trees will appear unresolved and some of the resolution will not be supported in all the input trees. Similarly, the majority rule extended (MRE, but also known as "greedy consensus") will also follow the procedure of the majority rule to form a "backbone" of the resulting tree, and then append the other sets of taxa in order of the frequency with which they have appeared (Felsenstein 1995). This method is usually applied to summarise trees from bootstrap tree searches where the target is to maximize resolution.

Supertree methods are a generalization of the consensus methods in the sense that they accept trees as input to build a single tree that reflects the splits of the input trees, but are more flexible because these can accept trees with different number of taxa (as long as there is a minimum overlap of three) and that have been produced from different data sources. One of the most popular and widely used method of constructing supertrees is the matrix representation using parsimony (MRP) method. In this method, a binary matrix is created where the characters reflect the presence of taxa or OTUs in the splits across all input trees. The matrix coding was devised by Baum 1992 and Ragan 1992. Each internal branch of each rooted input tree is examined and a '1' is assigned to any taxa contained within the clade defined by that internal branch. A '0' is assigned to any taxa that is contained within the input tree, but not in the clade, and a '?' is assigned to any taxa not present in the input tree (Figure 1.6). After the matrix has been inferred from the input trees, it is analysed using parsimony, ending up with a supertree phylogeny (Creevey 2002). Supertree methods are much less popular than supermatrix methods (where the phylogeny is built from a concatenated data matrix), hence for years only parsimony was used for the inference of the tree. However, a resurgence of interest in this approach has been taking place over the last years, as limitations of the supermatrix method have become more evident (i.e. poor availability of adequate models for different types of concatenated data, slow performance when dealing with genome-size data sets and missing data among the most known). This has slowly lead to the development of novel and statistically consistent supertree methods, such as the Maximum Likelihood (Steel and Rodrigo 2008) supertrees and Bayesian supertrees (Ronquist *et al.* 2004).



| | Clades | | | | | |
|---|---|---|---|---|---|---|
| Taxa | *i* | *ii* | .... | *iii* | *iv* | *v* |
| A | 1 | 0 | .... | 1 | 1 | 1 |
| B | ? | ? | .... | 0 | 0 | 0 |
| C | ? | ? | .... | 1 | 1 | 1 |
| D | 1 | 0 | .... | 0 | 1 | 1 |
| E | 0 | 1 | .... | ? | ? | ? |
| F | 0 | 1 | .... | 0 | 0 | 1 |

**Figure 1.6** Generating supertrees by matrix representation.
Columns in the matrix represent splits across all input trees. Taxa are coded by presence (1) or absence (0) in each of the clades. If a taxon is not present in one of the trees (e.g. B and C are not present in the tree on the left), it is denoted by "?".

In this thesis, I only used the MRP matrix coding (that is, only the generation of the matrix, without carrying out the final phylogenetic inference step, from now on called the "Matrix Representation" - MR). The MR matrix was used to combine different gene family trees and different tree sources, which were then assessed for the presence of unstable taxa in chapter 2.

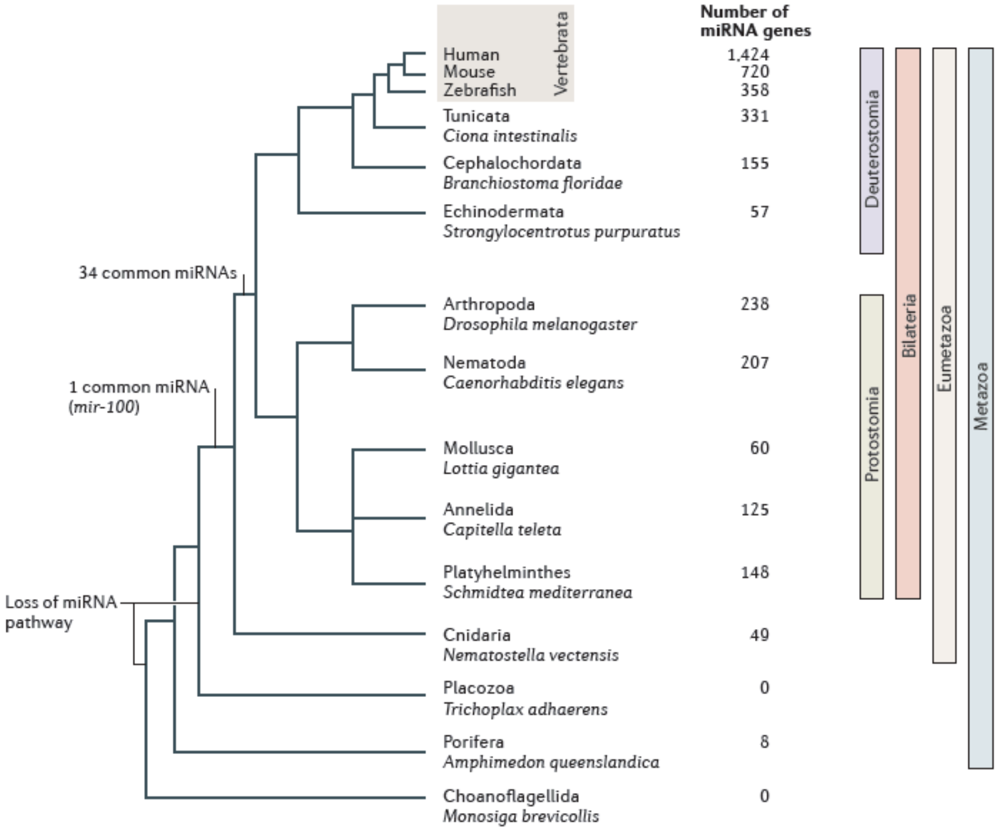
## 1.4 The importance of multiple lines of evidence: microRNAs

### 1.4.1 Discovery and function

Since their discovery, microRNAs (miRNAs) have been shown to be crucial in the regulation of different physiological processes such as developmental timing, neuronal patterning, cell proliferation, apoptosis, tissue differentiation and cell signaling (Bartel 2004). miRNAs are a type of small RNA along with small interfering RNAs (siRNAs) and piwi-interacting RNAs (piRNAs), which have been found in animals, plants, and fungi (Bartel 2009) and are all involved as a guide that controls RNA silencing. However, miRNAs differ from these other types of small RNAs in their biogenesis (i.e. miRNAs originate from endogenous hairpin-looped transcripts, Bartel 2009). miRNAs are single-stranded RNAs of approximately 19-25 nucleotides (nt) in length (Bartel 2004) and were originally identified for their role in developmental timing in *Caenorhabditis elegans* (Lee *et al.* 1993; Reinhart *et al.* 2000). In these works miRNA family lin-4 was identified as the key regulator of the gene product lin-14, and then another major family of miRNA, let-7, was identified in the regulation of developmental timing in *C. elegans*. Then homologs of let-7 that were identified in other bilateral animals including mammals, exhibited temporal expression like the one observed in *C. elegans*. This fact led to infer that let-7 and other small temporal RNAs could be present and have orthologous functions across the animals (Pasquinelli *et al.* 2000). In animals, silencing of messenger RNA (mRNAs) mediated by miRNA is normally carried out by imperfect base paring to the 3' UTR, hence causing to block mRNA from being translated (Lee *et al.* 1993) and also occasionally directing catalytic cleavage (depending on base complementarity degree, Bartel 2009). Individual miRNAs may regulate up to hundreds of different loci that may only become actively expressed at different stages of development in an organism. Therefore, identifying miRNAs (with current sequencing techniques and bioinformatics tools) in a thorough way usually involves sampling at different developmental stages and using species-reference genomes.

### 1.4.2 Importance as deep-phylogeny markers

The development of next-generation sequencing techniques and appropriate computational methods has been a crucial boost in miRNA discovery. This has allowed us to gain a better understanding of which miRNAs occur in certain groups across the animals (Metazoa), and of certain features about its distribution in the animal kingdom. For example, the miRNA family miR-100, thought to be involved in regulating cell differentiation and survival (Zheng *et al.* 2011), is shared across all Eumetazoa (Grimson *et al.* 2008; Berezikov 2011, see Figure 1.7). Then, 34 miRNAs appeared shared between Protostomia and Deuterostomia but not in the other ancestral lineages, suggesting a burst of innovations at the Bilateria which seems to coincide with a "bilaterian expansion" (Hertel *et al.* 2006; Peterson *et al.* 2009; Christodoulou *et al.* 2010; Berezikov 2011). Further on down the tree, other miRNAs begin to appear at the base of the Vertebrata and also at the lineage of placental mammals (Hertel *et al.* 2006; Heimberg *et al.* 2008; Berezikov 2011). One of the main things that has been found about the distribution of conserved miRNAs is that after a miRNA gene emerges in a lineage, it is rarely lost in the descendant lineages (Heimberg *et al.* 2008; Peterson *et al.* 2009; Wheeler *et al.* 2009; Berezikov 2011). The rate of loss is outweighed by the continuous acquisition of novel miRNAs which render these small molecules highly conserved and potentially useful phylogenetic characters. The increasing repertoire of miRNAs seems to be directly correlated to the morphological complexity observed in the animal kingdom, suggesting that miRNA innovation could have an important role in the evolution of complexity in organisms (Berezikov 2011). In this thesis I used miRNAs as an independent line of evidence to investigate early amphibian evolution (chapter 4).



Number of miRNA genes
Vertebrata
Human 1,424
Mouse 720
Zebrafish 358
Tunicata *Ciona intestinalis* 331
Cephalochordata *Branchiostoma floridae* 155
Echinodermata *Strongylocentrotus purpuratus* 57
Arthropoda *Drosophila melanogaster* 238
Nematoda *Caenorhabditis elegans* 207
Mollusca *Lottia gigantea* 60
Annelida *Capitella teleta* 125
Platyhelminthes *Schmidtea mediterranea* 148
Cnidaria *Nematostella vectensis* 49
Placozoa *Trichoplax adhaerens* 0
Porifera *Amphimedon queenslandica* 8
Choanoflagellida *Monosiga brevicollis* 0
34 common miRNAs
1 common miRNA (*mir-100*)
Loss of miRNA pathway
Deuterostomia
Protostomia
Bilateria
Eumetazoa
Metazoa

**Figure 1.7** miRNA innovations across the Metazoa.

When miRNA genes emerge in a lineage, these are rarely lost in the descendant lineages, making them useful phylogenetic characters. Reproduced from Berezikov (2011).

# Chapter 2. Concatabominations – an extension to the *a priori* method of Safe Taxonomic Reduction.

## 2.1 Introduction

One of the main causes of poor phylogenetic resolution is the inclusion of missing data in phylogenetic datasets. This can occur when including taxa with incomplete characters (for instance when increasing the number of taxa for which not all the characters have been sampled) or, on the opposite way, including characters that have not been sampled across all taxa (for instance when combining genes that have not been sampled in all considered taxa). In parsimony, likelihood and Bayesian frameworks, the effects of missing data are still a long-term debate with some authors arguing for or against the inclusion of missing data as being detrimental in tree resolution (Driskell *et al.* 2004; Lemmon *et al.* 2009; Wiens and Morrill 2011; Roure *et al.* 2013). In a parsimony context, having missing entries in a dataset often results in uncertainty in the phylogenetic placement of taxa with missing entries and an increase in alternative most parsimonious trees or MPTs (Nixon and Wheeler 1992; Wilkinson 1995). Taxa that have uncertain placement have been named "rogue" or "unstable" (Swofford 1991; Wilkinson 1994a; Sanderson and Shaffer 2002). Their impact on the resolution of a tree does not only limit to parsimony but also to other methods of analyses, like Bayesian analysis where they will depress posterior probabilities. Similarly, when bootstrapping, rogue taxa will artificially depress support values (Aberer *et al.* 2013). Here I will only consider the parsimony framework, but all my conclusions naturally extend to ML and other methods of analysis

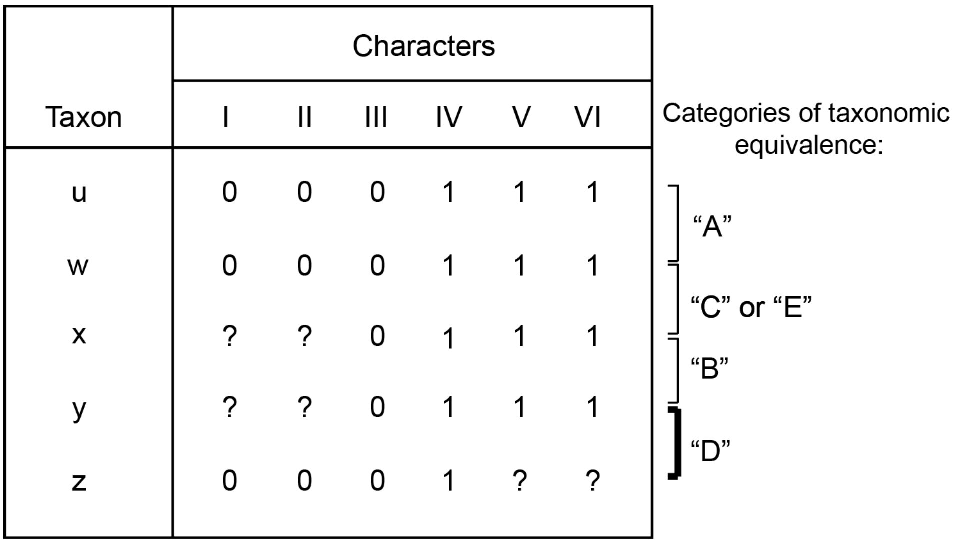
The most common approach used to counter poor resolution due to missing data has been eliminating rogue taxa. Some previous studies, especially in paleontological datasets (Gauthier 1986; Rowe 1988) attempted eliminating rogue taxa without using a method or criterion to justify the choice of taxa for deletion. Following from this, a variety of methods have been developed to identify unstable taxa, which can be classified into either *a priori* or *a posteriori*. *A priori* methods identify the stability of taxa, prior to phylogenetic analyses. The most used method for identifying rogue taxa under this approach is Safe Taxonomic Reduction (STR) by Wilkinson (1995). On the other hand, *a posteriori* methods carry out a phylogenetic reconstruction using the complete dataset, and then identify the unstable taxa to prune using consensus methods. The first attempt by Wilkinson (1996) pointed out the utility of reduced consensus methods to identify unstable taxa. Other researchers have later focused in the identification of unstable taxa using a modification of the previous approach, among some examples of these are the taxon instability index based on the distance of taxa across multiple bootstrap trees (Maddison and Maddison 2010), selective deletion based on instability index scores (Thomson and Shaffer 2010), an algorithm for finding the set of taxa to keep for a reduced consensus based on tree resolutions (Aberer and Stamatakis 2011; Pattengale *et al.* 2011); finding a maximum agreement subtree under the Bayesian framework (Cranston and Rannala 2007); or even using supertree methods (Ranwez *et al.* 2007 and Scornavacca *et al.* 2008). As seen from above, most authors have explored different *a posteriori* alternatives for identifying unstable taxa, but very few have ventured to develop *a priori* methods for this purpose simply because of the difficulty that it seems to pose (Aberer and Stamatakis 2011).

STR has been widely used mainly by palaeontologists who have been confronted with relatively incomplete fossil taxa (ie. Anquetin 2012; Graf 2012; McDonald 2012 for recent examples), but also in the context of the matrix representation with parsimony (Baum 1992; Ragan 1992) approach to supertree construction (eg. Cardillo *et al.* 2004). Additionally, supertree methods are becoming more popular in phylogenomic reconstructions because they seem to be less prone to missing data artifacts (Bininda-Emonds *et al.* 2002; Baum and Ragan 2004). In spite of this, STR is not always as effective as one might hope (eg. Mannion *et al.* 2013). In this work I aim to extend the original method of Wilkinson (1995), suggest a novel way to visualize instability of taxa, and propose a systematic way of choosing taxa to delete.

## 2.2 Methods

### 2.2.1 Safe Taxonomic Reduction

Under parsimony, certain characters are expected to be informative in creating groupings of leaves or taxa based on the similarity and compatibility of their character states. Basically if two taxa share the same informative characters, in a simplistic way they are considered ‘taxonomic equivalents’ of each other. Wilkinson’s (1995) STR method differentiates five different types of taxonomic equivalences based on informative characters shared and the distribution of missing data among a pair of taxa (see Figure 2.1). When two taxa have no missing entries and share the same character information, the two taxa are real equivalents. This type of equivalence is termed “type A”. When missing data is present, two taxa cannot be “real equivalents” and are categorised as a “potential equivalents”. There are four types of potential taxonomic equivalences: where taxa have the same characters coded as missing and the same character states for all phylogenetically informative characters (symmetric potential equivalence, type B); where the missing data are concentrated in one of the pair of taxa (asymmetric potential equivalence one-way, types C and E where both are reciprocal of each other); and where there are missing entries in both pairs of taxa but there is a known and scored entry for their corresponding pair (asymmetric potential equivalence two-ways, type D).



| Taxon | Characters | | | | | |
|---|---|---|---|---|---|---|
| | I | II | III | IV | V | VI |
| u | 0 | 0 | 0 | 1 | 1 | 1 |
| w | 0 | 0 | 0 | 1 | 1 | 1 |
| x | ? | ? | 0 | 1 | 1 | 1 |
| y | ? | ? | 0 | 1 | 1 | 1 |
| z | 0 | 0 | 0 | 1 | ? | ? |

Categories of taxonomic equivalence:
“A”
“C” or “E”
“B”
“D”

**Figure 2.1** Hypothetical character data illustrating relations of taxonomic equivalence (from Wilkinson 1995).

Relationships are shown by pairwise comparison of taxa. A represents actual symmetric equivalents, all the other possible pairs are potential equivalents. B represents potential symmetric equivalents. C and E represent asymmetric potential equivalents one way (and both are reciprocal). D represents asymmetric potential equivalents both ways.

The next step in the STR method is the selection of taxa to delete. This is based on a set of rules proposed by Wilkinson (1995) that have as aim to reduce MPTs without generating an impact on the number of steps in the character reconstruction and hence not affecting the inferred relationships of the other taxa included in the analysis. The rules can be summarized as: 1) Represent all sets of symmetric taxonomic equivalents with a single taxon (which is applicable to symmetric potential equivalents type A and B); and 2) eliminate those potential equivalents that have the greatest number of missing entries when the asymmetry is all one-way (applicable to asymmetric potential equivalents type C and E). In standard STR, cases where the asymmetric potential equivalence is both ways (type D) are considered unsafe to delete. The reason for this is that the criterion solely based on the information presented by the characters (coded and missing) is not enough to determine whether the two taxa are potential equivalent because both taxa present missing data. Therefore, asymmetric potential equivalents both ways cannot be removed according to STR even if they have a negative effect on the resolution of the tree. Here I shall introduce a heuristic approach to extend STR and allow the elimination of taxa falling into the type D category.

### 2.2.2 Recoding strategy

The main problem dealing with the type D taxonomic equivalents lies in not having enough information to determine their equivalence in phylogenetic reconstructions. Hence, a new strategy is required to address this problem. In a type D potential equivalence, the missing data is present in both taxa but in different characters (that is, while one of the taxa has a missing entry for any particular characters, there will be a coded or known corresponding character for the other taxon). In order to test whether each type D pair are potential equivalents the information for the two leaves are forced together. This was achieved by substituting the missing entries with the corresponding information from the other of the pair that has the coded character in all type D pairs (see Figure 2.2). This results in an ‘abominated’ taxon which I termed “concatabomination”. The “concatabominated” taxon is then introduced back into the original matrix, replacing the two taxa that are being tested. This is repeated with all the possible potential equivalents (type D), creating new matrices each time. Seeing it from the opposite perspective, the original taxa can be considered approximations derived from the concatabominated taxon where some of its states have been substituted with missing entries.

|  | Characters |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Taxon | I | II | III | IV | V | VI |
| y | ? | ? | 0 | 1 | 1 | 1 |
|  | ↑ | ↑ |  |  | ↓ | ↓ |
| z | 0 | 0 | 0 | 1 | ? | ? |
| y + z = | 0 | 0 | 0 | 1 | 1 | 1 |

**Figure 2.2** Recoding action performed on a hypothetical character data for an asymmetric potential equivalence two-ways (type D).
The arrows show the direction at which characters replace the missing data yielding a “concatabominated” taxon “y+z”.

### 2.2.3 Use of Compatibility to identify potential equivalents

Compatibility analyses can be used to identify pairs of type D potential equivalents that truly represent potential equivalents from those that do not. After forcing the information of the two leaves together, a pairwise compatibility analysis is performed to the original matrix and to the new matrices containing the concatabominated taxa. The idea is to ask whether the concatabominations introduce homoplasies that did not exist in the original matrix. Pairwise compatibility of characters was introduced by Le Quesne (1969), together with the concept of "uniquely derived character". According to Le Quesne, two characters are compatible if there exists a tree over which they can be mapped without homoplasy. Thus, if there exists no tree upon which both characters can be mapped without assuming homoplasy, then one of the two characters must be homoplastic (i.e. not uniquely derived). If one compares each pair of characters in a data set to test if they are compatible, a lower bound on the amount of homoplasy of a data set will be obtained (which corresponds to the number of pairwise incompatibilities observed). Compatibility is used in the context of the concatabomination approach to evaluate whether two taxa are potential equivalents. If a matrix with a concatabominated taxon has the same amount of homoplasy (pairwise incompatibilities) displayed by the original matrix, then the two concatabominated taxa are potential equivalents. If inclusion of a concatabominated taxon results in an increase in homoplasy (pairwise incompatibilities) then the two taxa fused in the concatabomination are not real equivalents. So, if after concatabomination homoplasy does not increase, one of the two concatabominated taxa can be deleted as the two are potential equivalents.

### 2.2.4 Implementing concatabominations

A pipeline that brings together the different steps of the analysis was assembled in shell scripting (see Figure 2.3) and is freely available at https://code.google.com/p/concatabominations/. The shell scripting language was chosen because it is widely used in command-line programming. The first step of this pipeline is to classify pairs of taxa in taxonomic equivalents of different types (class A to E) *a priori*. This is achieved with the previously developed software PerlEQ (by Jeffery and Wilkinson available at http://www.molekularesystematik.uni-oldenburg.de/33997.html). PerlEQ also provides a list of taxa suggested for deletion as part of its outputs, but this list is generated based on the original STR method of Wilkinson (1995), hence it does not deal with type D equivalences.

The pipeline is designed to assess a dataset for its potential equivalents and further test those potential equivalents that were not safe to delete (type D) using my new recoding strategy that simulates the introduction of "chimeric" taxa, which represent concatabomination of pairs of type D equivalents. The effect of including concatabominations in the data set is assessed based on how incompatible the new matrices are with reference to the original. The recoding strategy was carried out with a program coded in C, "Concatabomb", for the present work (Figure 2.3). To calculate incompatibility, two softwares can be used, depending on the type of matrix. If the matrix is binary coded (informative characters are 0 or 1), then a program called "pairwise_incompatibilities" coded in C (see Figure 2.3) is used in both the original and "concatabominated" matrices. If the matrix had more than two character states, then the software COMPASS (S. Harris, available at http://research.ncl.ac.uk/microbial_eukaryotes/downloads.html) was used to calculate incompatibility for the original and "concatabominated" matrices. The reason why two softwares are used depending on the matrix, is that COMPASS uses too much memory when dealing with large matrices (which would be the case for binary, genomic supertree matrices for instance). The new "pairwise_incompatibilities" software does this job in a more efficient way using less memory, but at the moment it can only deal with binary characters. As a follow up work from this thesis I plan to extend the "pairwise_incompatibilities" software to deal with multistate characters.

The next step in the pipeline is to generate a list of final taxonomic equivalents. This is achieved by comparing the incompatibility scores between the original and "concatabominated" matrices and filtering the cases where the score was increased with respect to the original matrix (following section 2.2.3). Finally, the subset of type D equivalences for which matrix incompatibility is the same for the original and for the concatabominated matrices (referred to as D*) are combined with the A, B, C and E types found in the original STR step. The new list of taxonomic equivalents is then sorted and ranked having the taxa with highest number of taxonomic equivalents first (this represents the most unstable or rogue taxa). The pipeline also prepares a ".sim" file that contains the list of all taxa with their taxonomic equivalences, which can be opened in Cytoscape (Shannon *et al.* 2003). This step allows the user to interactively perform taxa deletions and reduce noise (caused by missing data) in phylogenetic data sets.  This is important to allow reaching an optimal level of resolution whilst at the same time retaining taxa of interest for the specific aims of the study that is being carried out.

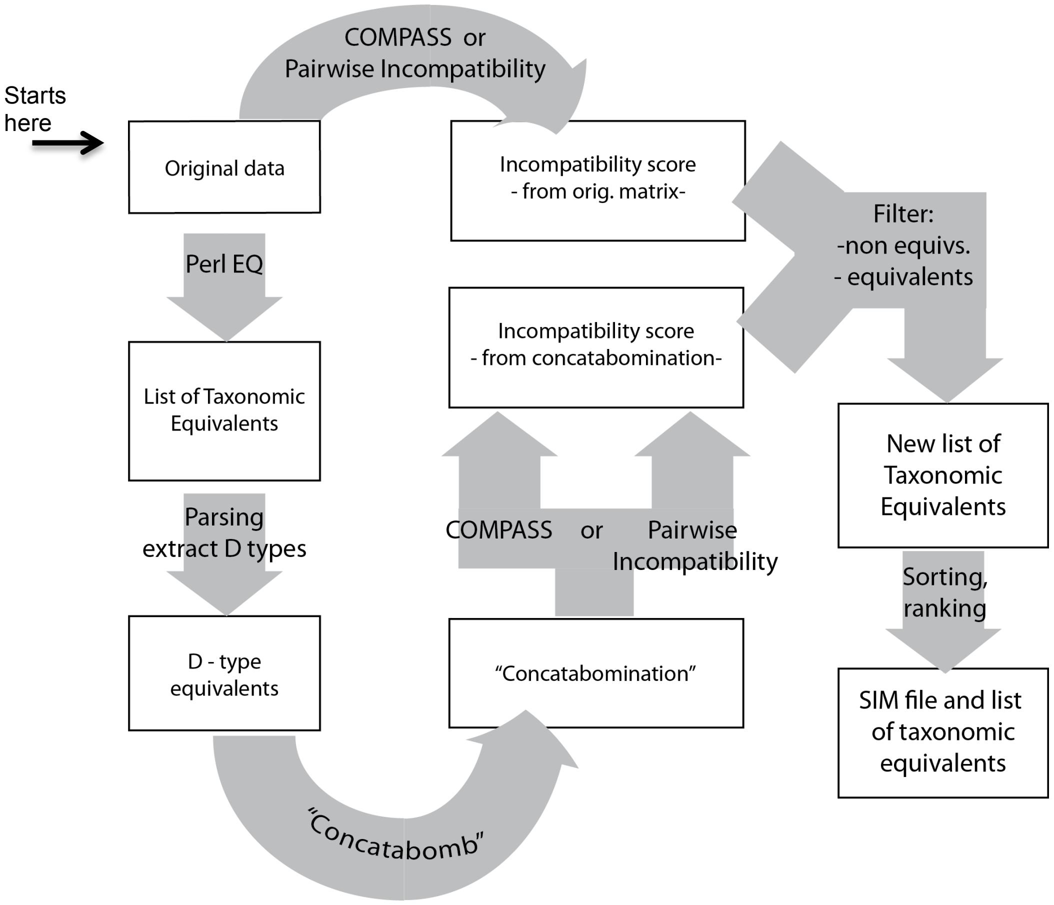

Starts here
COMPASS or Pairwise Incompatibility
Original data
Incompatibility score - from orig. matrix-
Filter: -non equivs. - equivalents
Perl EQ
Incompatibility score - from concatabomination-
List of Taxonomic Equivalents
New list of Taxonomic Equivalents
Parsing extract D types
COMPASS or Pairwise Incompatibility
Sorting, ranking
D - type equivalents
"Concatabomination"
SIM file and list of taxonomic equivalents
"Concatabomb"

**Figure 2.3** Schematic representation of the Concatabomination pipeline, as described in the text.

Boxes represent outputs from the pipeline and arrows represent software or actions performed.

### 2.2.5 Determining which taxa to delete

The taxa to be excluded from the phylogenetic analysis can be chosen using the list of most unstable taxa and the ".sim" file (both represent outputs from the "concatabomination pipeline"). As stated in the previous section, the outputs from both files will help determine which are the most unstable or rogue taxa due to missing data. Therefore, if the aim is to increase the resolution of a phylogeny, following rule 1 proposed by Wilkinson (1995) "*represent all sets of symmetric taxonomic equivalents with a single taxon*", it is logical that the taxa with the highest number of equivalences must be deleted.

By looking at this list of taxonomic equivalents alone, the selection could be possible (however the count of taxonomic equivalents must be updated each time for every taxon after removing one). To facilitate this process, networks present a useful solution to visualize relationships of the taxonomic equivalence among taxa (Figure 2.4). A network is a collection of nodes or vertices that are connected by edges. The nodes represent elements that can share some property or connection (represented by the edges). In this case, the nodes represent taxa and the edges indicate whether two taxa are equivalent. As stated in the previous section, Cytoscape is used for visualization of the connection between taxa and their equivalents. The degree of a node indicates the number of edges incident on it. In this case, for each node its degree indicates the number of taxonomic equivalences in the data set. This measure is used when visualizing and sorting the network of relationships of taxonomic equivalences. Thus, the most unstable taxa are easily identified as the nodes in the network with the highest degrees (Figure 2.4). The advantage of using networks is that clusters appearing in the network after taxa are pruned represent clades and indicate at what level of inclusiveness in the tree instability is emerging. This helps evaluating what taxa to delete and when to stop removing taxa from an analysis (depending on the scope of the considered study).

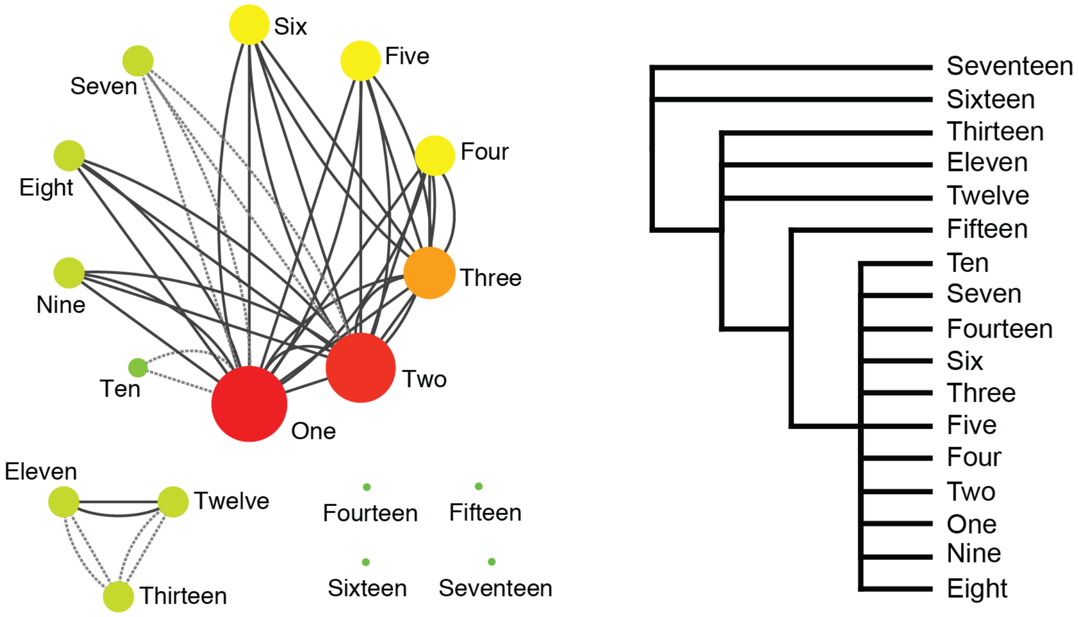


**Figure 2.4** Network diagram of the taxonomic equivalences from a set of taxa and its corresponding strict consensus tree.

The most unstable taxa are represented by a bigger size and colour, with red indicating instability and green stability. Taxa that are disconnected are all unique (they do not have equivalents). Two clusters are present in this network. This indicates two areas of instability one composed of three taxa represent a three-taxon polytomy. The second includes 10 nodes and represent a 10-taxon polytomy. The two polytomies are disconnected (i.e. affect different parts of the tree).

## 2.3 Results: Case studies

To illustrate the utility of the concatabomination approach in the morphological and supertree contexts, we present two different examples: a reanalysis of the Saurischian fossil data matrix by Gauthier (1986), and an analysis of a genomic supertree matrix of Bacteria prepared by Akanni (2014).

### 2.3.1 Application to a morphological data set

Gauthier's (1986) study on Saurischian interrelationships is a classic example of the effect of including taxa with high percentages of missing data. The original study had the primary objective to determine the relationships of Avialae within the Therapoda. Previous studies dealing with the effects on missing data have used it as an example because of its poorly resolved trees (see Wilkinson 1995; Kearney 2002) and use of no systematic criterion to justify which taxa to delete. In particular, it was used as a case study in the original STR paper by Wilkinson (1995), and thus I decided to use it to illustrate the utility of the procedure presented here.

The dataset contains 17 taxa and 84 binary characters (Appendix A) and it was analysed with the new "*Concatabominations pipeline*" described before. Final phylogenetic reconstructions were carried out in PAUP* v.4.0b10 (Swofford 2002). To show the effectiveness of the method, I carried out *a posteriori* phylogenetic analyses with no taxa deletions, and following deletions from the original STR and the *Concatabominations* method.

A parsimony analysis on the complete dataset (making no deletions) gives 832,902 MPTs of 98 steps of which the strict consensus results in three resolved nodes (Figure 2.5 a). When applying the original STR method, the taxa suggested for deletion are *Hulsanpes*, *Liliensternus, Procompsognathus* and *Saurornitholestes*. The parsimony analysis on the remaining dataset clearly decreases the number of MPTs obtained to 197 maintaining the same length, and the strict consensus (Figure 2.5 b) also shows an improvement in the number of nodes retrieved to five. However, it is worth noting that this improvement could have been achieved just by deleting *Hulsanpes* and *Saurornitholestes* alone. Even though deleting *Liliensternus* and

*Procompsognathus* will also reduce the number of MPTs, it does not increase the resolution (denoted by the retrieval of a further split).

Finally, the analysis with the *Concatabominations* method gave a list of taxa with its taxonomic equivalents (Table 2.1). From this table, it can be observed that *Hulsanpes* is the most unstable taxon, followed by *Saurornitholestes*, and others. From this method, it was observed that *Hulsanpes* has actually got nine taxonomic equivalences (of which seven correspond to D*). Likewise, *Saurornitholestes* got eight taxonomic equivalences (of which seven correspond to D*).  A third taxon, *Coelurus*, is also shown with five taxonomic equivalences (all D* tested with the novel method). Obviously, this taxon could not have been identified as unstable using the standard STR approach.

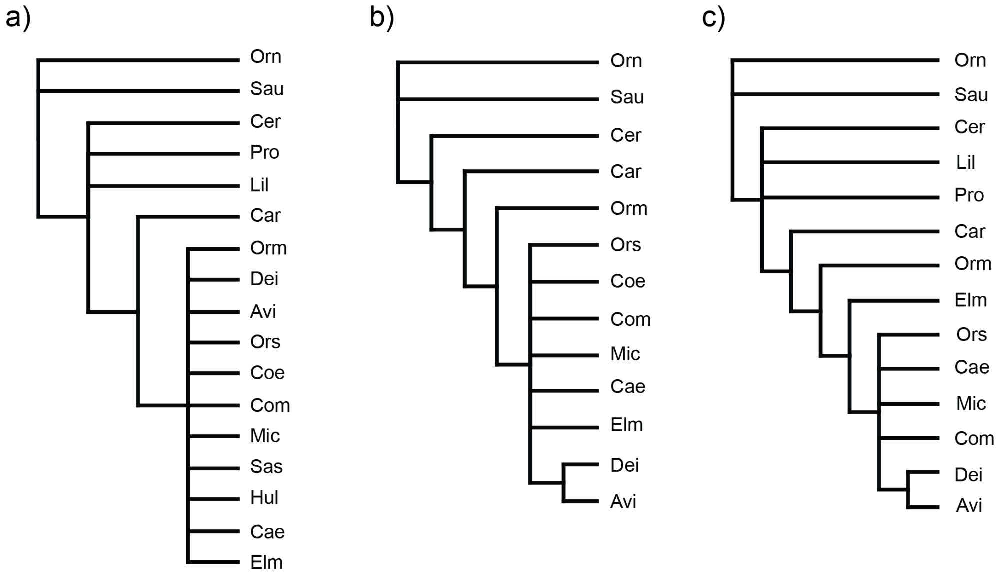


**Figure 2.5** Strict consensus trees from the resulting MP analyses of the Saurischian dataset from Gauthier (1986).

a) Entire dataset (no deletions); b) after deleting taxa following original STR; and c) after deleting taxa from the Concatabomination method. Note: the remaining polytomy (including Ornitholestes -Ors-, Caenagnathidae -Cae-, *Microvenator* -Mic-, *Compsognathus* -Com- and the clade of Deinonychosauria -Dei- and Avialae -Avi-) is not the consequence of the presence of missing data but of homoplasy, and the Concatabomination method cannot resolve it. The polytomy involving *Procompsognathus* (Pro), *Liliensternus* (Lil) and Ceratosauria (Cer) can be resolved removing any of these taxa (see Figure 2.6 d).

**Table 2.1** Results from the concatabomination pipeline analysis of the Gauthier (1986) dataset.

The numbers of D* and ABC scores as well as the percentage of missing entries and abbreviations (Abb.) of taxon names used in the Figures are shown. *Coelurus*[§] the third most unstable taxon could not have been highlighted as unstable using the original STR as it only had D*–type equivalences.

| Taxon | Abb. | % Missing entries | D* | ABC | Total |
|---|---|---|---|---|---|
| *Hulsanpes* | Hul | 81 | 7 | 2 | 9 |
| *Saurornitholestes* | Sas | 72 | 7 | 1 | 8 |
| *Coelurus*[§] | Coe | 72 | 5 | 0 | 5 |
| *Ornitholestes* | Ors | 40 | 3 | 0 | 3 |
| *Compsognathus* | Com | 38 | 3 | 0 | 3 |
| *Microvenator* | Mic | 67 | 3 | 0 | 3 |
| Ceratosauria | Cer | 0 | 0 | 2 | 2 |
| Deinonychosauria | Dei | 6 | 0 | 2 | 2 |
| Caenagnathidae | Cae | 33 | 2 | 0 | 2 |
| Elmisauridae | Elm | 54 | 2 | 0 | 2 |
| *Procompsognathus* | Pro | 64 | 1 | 1 | 2 |
| *Liliensternus* | Lil | 48 | 1 | 1 | 2 |
| Ornithomimidae | Orm | 8 | 0 | 1 | 1 |
| Ornithischia | Orn | 0 | 0 | 0 | 0 |
| Sauropodomorpha | Sau | 0 | 0 | 0 | 0 |
| Carnosauria | Car | 2 | 0 | 0 | 0 |
| Avialae | Avi | 4 | 0 | 0 | 0 |

A network visualization of the taxonomic equivalences allows for a clearer identification of the most unstable taxa (Figure 2.6 a). Clearly, *Hulsanpes* (Hul) and *Saurornitholestes* (Sau) appear to share the most taxonomic equivalences with other taxa. After deletion of *Hulsanpes* (Figure 2.6 b) and *Saurornitholestes* (Figure 2.6 c), this analysis identifies *Coelurus* (Coe) as the new most unstable taxon. Deleting *Coelurus* (Figure 2.6 d) causes a decrease in the number of MPTs to 322 and the strict consensus of these retrieve a further split (bringing it to a total of six nodes – see Figure 2.5 c) that could not be identified using standard STR. Importantly, the tree length is not changed when *Coelurus* is excluded, indicating that character state reconstruction has not changed and that the deletion of *Coelurus* is "safe" (according to standard STR rules), even though standard STR could not identify this taxon as unstable. At this stage, only one more polytomous component or cluster is left which comprises *Procompsognathus* (Pro), *Liliensternus* (Lil) and Ceratosauria (Cer). These in fact form a trichotomy in the final tree (Figure 2.5 c), and the deletion of either is expected to have the same effect, given that all of them are connected to each other (Figure 2.6 d).

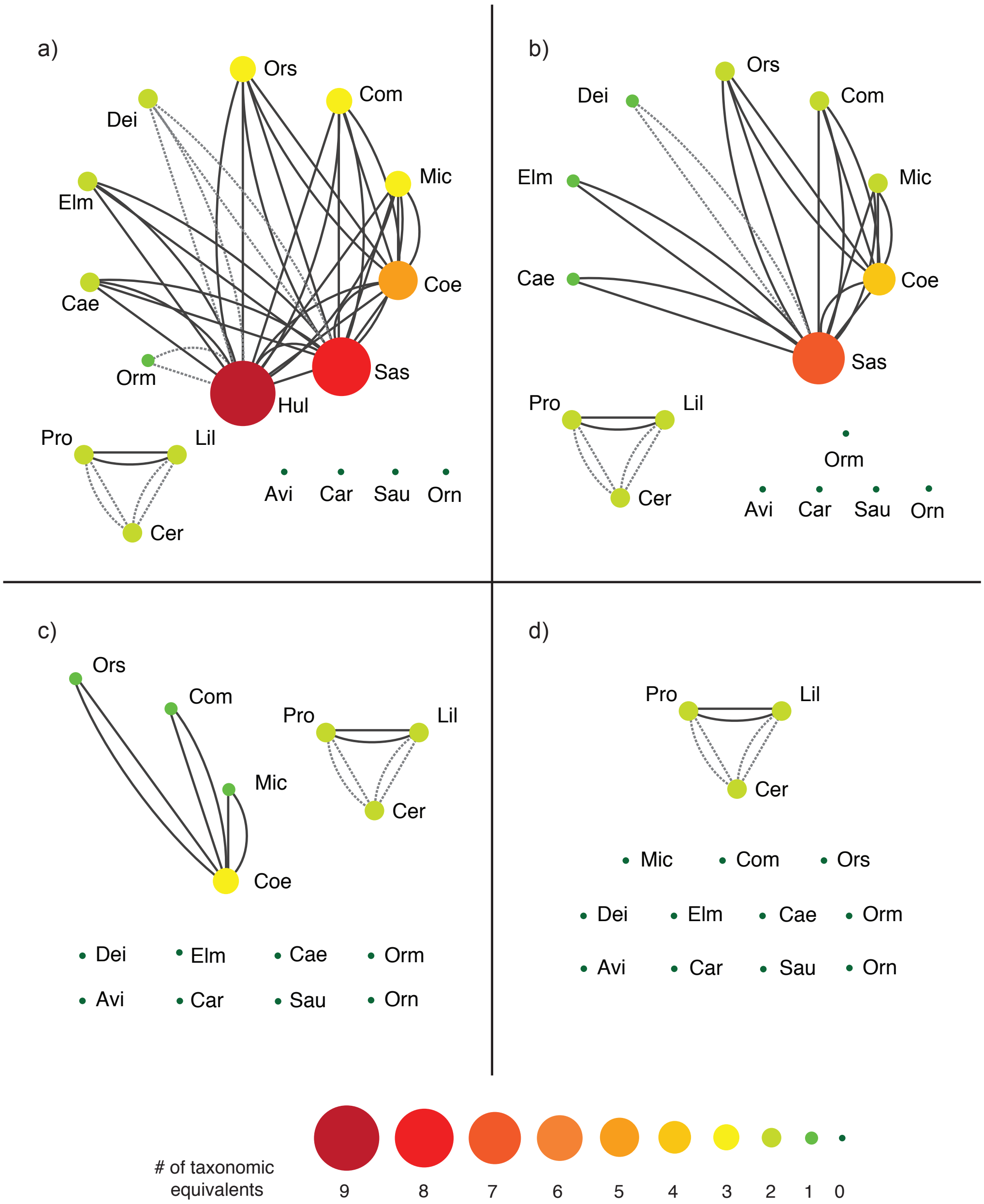


**Figure 2.6** Network representation of the taxonomic equivalences of the Gauthier (1986) data set.

View with no deletions (a), deleting *Hulsanpes* (Hul) (b), deleting *Saurornitholestes* (Sas) (c) and deleting *Coelurus* (Coe) (d). Note that at each step the number of connected nodes decreases and the number of disconnected nodes increases. Disconnected nodes are those without taxonomic equivalences and that are not affected by missing-data driven instability. Note that if further polytomies still exist once all taxa have been disconnected from the network, the underlying instability must be the result of homoplasy not missing data, and cannot be dealt with using the concatabomination approach. Other strategies might need to be used to address such problems but these are not the subject of this thesis.

### 2.3.2 Application to a genomic data set in a supertree context

As an example of the use of this software in a genomic data, I used a data set of Bacteria genomes assembled by Akanni (2014) from gene-family trees for this taxonomic group. In genomic data sets, the Concatabomination method is applied running the concatabomination pipeline on a Matrix Representation (generated using the Baum and Ragan coding scheme) of the collection of gene trees (Baum 1992; Ragan 1992). Note that this does not mean that the supertree (after the matrix has been analysed using the concatabomination pipeline) needs to be built using the Matrix Representation with Parsimony supertree method. Once unstable taxa have been removed, every supertree reconstruction method can be used on the collection of pruned (of unstable taxa) gene trees. Using the Baum and Ragan recoding strategy the presence or absence of a leaf in each split in a tree is recoded using additive binary coding (i.e. 1 equal presence and 0 equal absence). In addition, taxa that are present in other gene trees but not in the one that is currently being recoded is identified as "?". This recoding strategy naturally creates a matrix that is perfectly suited to be analysed using the concatabomination approach to identify taxa that are unstable (equivalent to many other taxa) because they include missing data. The bacterial data set of Akanni (2014) scored 443 taxa and 16463 gene trees. These were recoded with Clann v.3.2.3 (Creevey and McInerney 2005) using Matrix Representation into a rectangular binary matrix of 443 taxa and 69972 characters.

A phylogenetic analysis of this dataset using a Bayesian MCMC supertree method developed by Akanni (2014) was carried out with two chains for 2.7 million iterations sampling every 5000 iterations. After convergence and discarding the burnin phase, I was left with 50 trees per chain, and a majority rule consensus was computed in PAUP* v.4.0b10 for the combined trees. This allowed the generation of a supertree from the original data without concatabominations. The resulting supertree was very poorly resolved (see Figure 2.7 a). The matrix representation of this data set was analysed with the original STR method, but, quite surprisingly, STR could not identify any taxa safe for deletion. Finally, the data set was analysed using the Concatabomination pipeline. A network analysis of the concatabomination results identified a large area of instability, with one genus *Aster*, causing substantial instability because of its high number of taxonomic equivalents (Figure 2.8 a). Interestingly, all equivalents of *Aster* were of type D*, explaining why STR could not identify this taxon as unstable. After deleting *Aster*, the instability was substantially reduced (Figure 2.8 b) and new internal nodes were retrieved (Table 2.2). Following the method described in section 2.2.5, more unstable taxa were deleted reducing the instability in the matrix each time (Figure 2.8 b-d). It is interesting to note, for example, that after deleting eight additional taxa: *Blattabacterium, Orientia, Neorickettsia, Anaplasma, Wolbachia Baumannia, Cyanobacterium* and *Buchnera* (Figure 2.8 c), the large component of taxonomic equivalents divides in two disconnected components. This indicates that whilst the eight deleted taxa were globally unstable (across the entire tree), remaining instability is not global but local to distinct sub-sections of the supertree. Even though these deletions allowed the decomposition of the larger clusters into smaller subclusters of taxonomic equivalences, the resulting strict consensus did not increase the number of retrieved internal nodes (Table 2.2). Deletions were continued following the network visualization, up to the point of reducing the components to only three or two elements connected to see the final effect. In total, 15 taxa (Table 2.2) were deleted, which allowed retrieving 11 new nodes (Figure 2.7 b and Table 2.2.). This indicates that the concatabomination approach can also be useful in phylogenomics. It is important to point out that, exactly as in the case of paleontological data sets, elimination of rogue taxa will not necessarily result in the elimination of all polytomies, it will only eliminate polytomies caused by missing data.  Polytomies caused by homoplasy will still be present in the resulting tree.

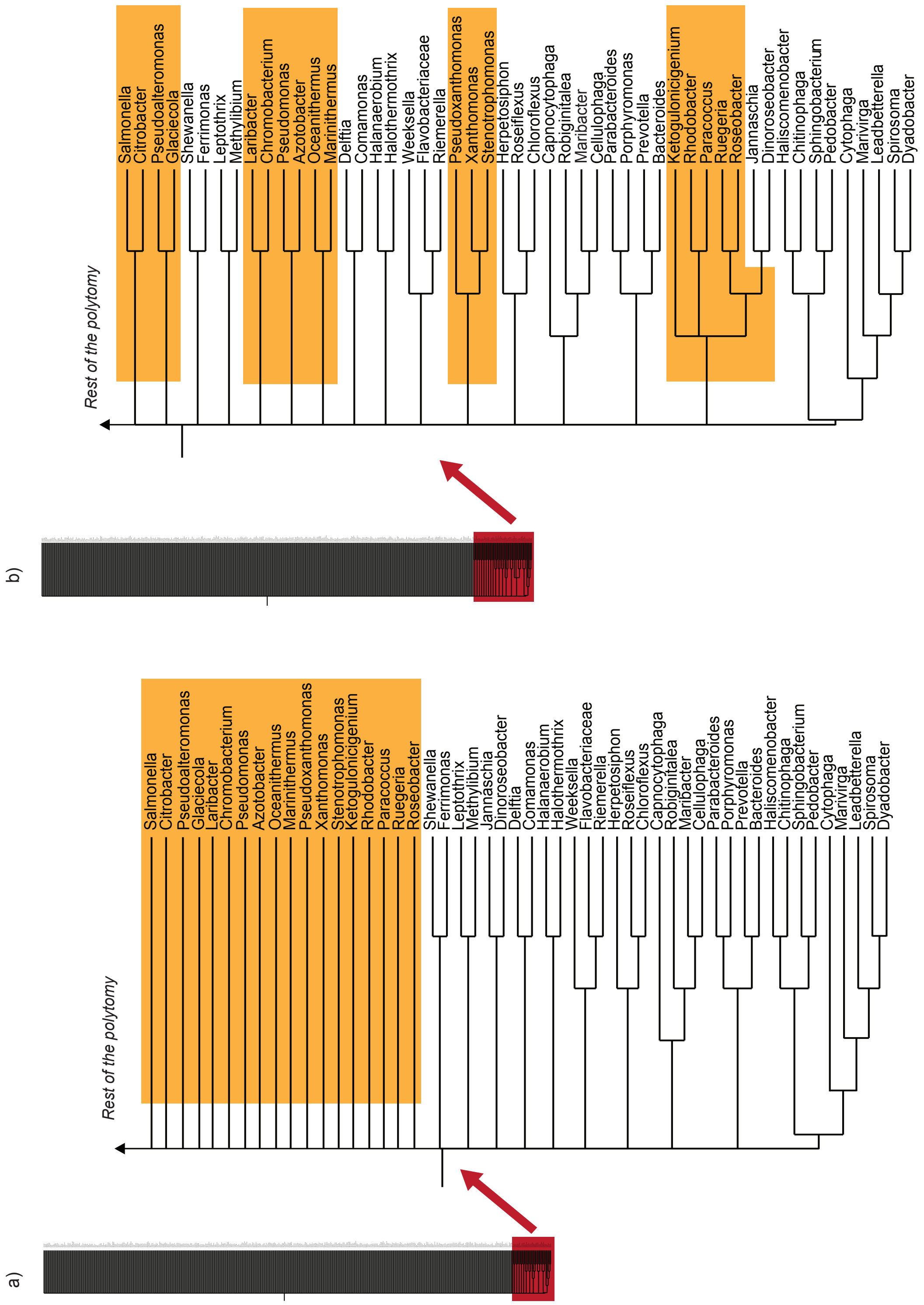


**Figure 2.7** Phylogenomic supertree of Bacteria.

Strict consensus of trees (a) without any deletions and (b) after deleting 15 taxa following the Concatabominations method. Taxa and nodes highlighted in yellow were further resolved using the Concatabominations method.

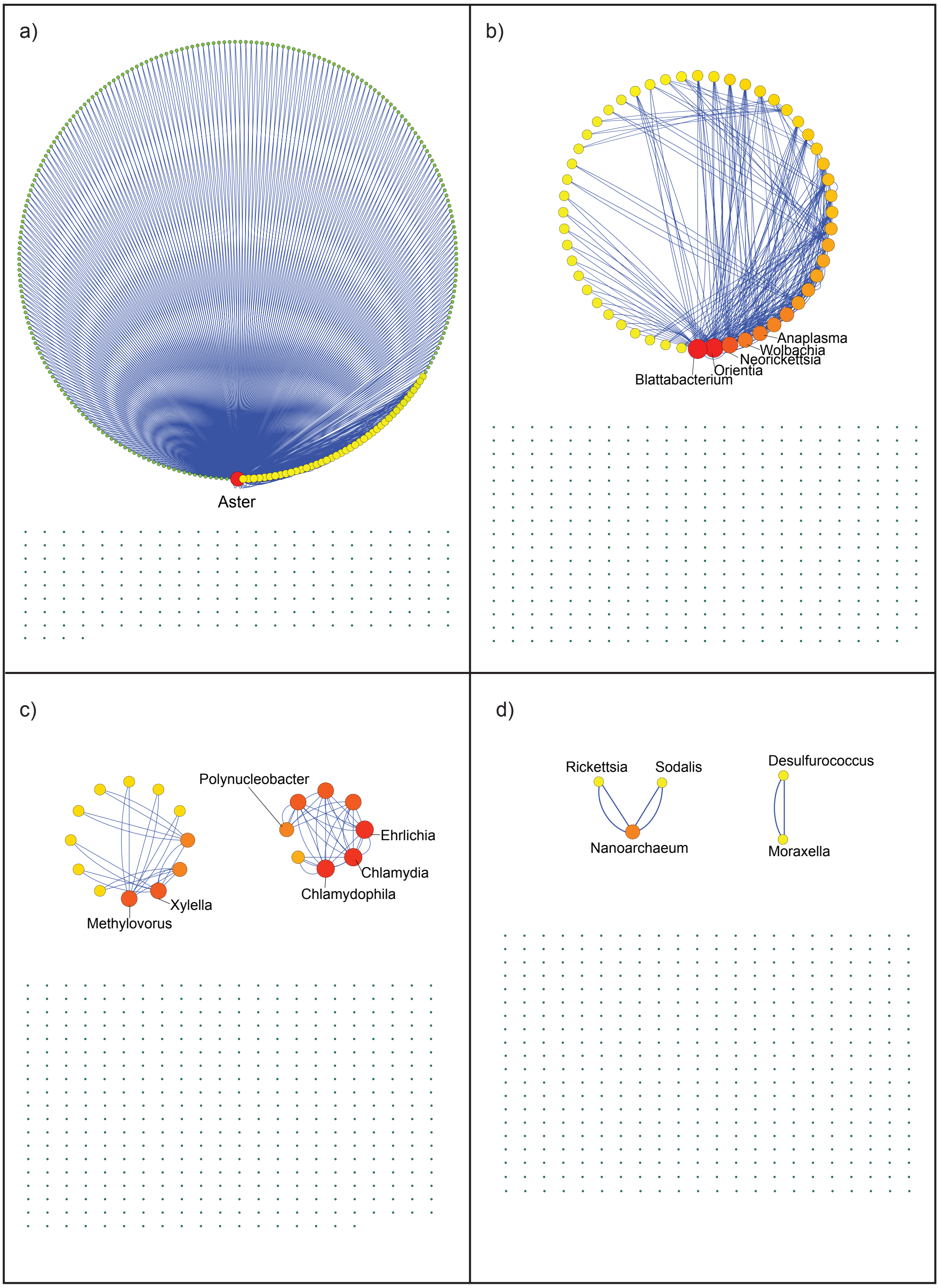


**Figure 2.8** Network representation of the taxonomic equivalences in the bacterial genomic dataset used.

a) Full dataset with no deletions; b) after deleting *Aster*; c) after deleting nine taxa; d) after deleting 15 taxa.

**Table 2.2** List of deleted taxa and number of internal nodes retrieved in the strict consensus (shown by the consensus fork index "CFI") after pruning taxa in a cumulative way following the network visualisation.

| | Cummulative deletion of: | CFI |
|---|---|---|
| 1 | None | 23 |
| 2 | *Aster* | 29 |
| 3 | *Blattabacterium* | 29 |
| 4 | *Orientia* | 29 |
| 5 | *Neorickettsia* | 29 |
| 6 | *Anaplasma* | 29 |
| 7 | *Wolbachia* | 29 |
| 8 | *Baumannia* | 29 |
| 9 | *Cyanobacterium* | 29 |
| 10 | *Buchnera* | 29 |
| 11 | *Chlamydophila* | 29 |
| 12 | *Chlamydia* | 29 |
| 13 | *Polynucleobacter* | 29 |
| 14 | *Ehrlichia* | 30 |
| 15 | *Methylovorus* | 30 |
| 16 | *Xylella* | 34 |

## 2.4 Discussion

STR has been widely used in phylogenetic paleontological studies and also, to a lesser degree, in supertree analyses. One of the advantages it presents is that it can be used *a priori* to detect rogue taxa, as it is certain that STR-identified rogue taxa are always safe to delete (i.e. their exclusion cannot impact the way in which node are resolved and character reconstructed). A straight-forward way to check that this is the case is to compare the length of trees recovered including all taxa against that of trees that do not include STR-identified taxa. The length of these trees is invariably the same as all characters are always correspondingly reconstructed on the pruned trees and on the trees scoring all the taxa. We conjecture that exactly as in the case of the STR, taxa identified for deletion using the concatabomination method are always safe for deletion. However, we have not been able to prove it mathematically and until such a proof can be given we prefer to define our method as a heuristic approach. Accordingly, it is recommended that deletions are confirmed to be safe. This can easily be done *a posteriori* comparing the parsimony length of the tree generated after excluding taxa marked for deletion using the concatabomination method against the parsimony length of the tree generated including all the taxa. If these two trees have the same length then the user can be sure that the deleted taxa are safe to eliminate.

The Concatabominations method has a good potential for diagnosing problems in the long withstanding issue of effective overlap of concatenated gene matrices (Sanderson *et al.* 2011), when treated as matrix representations. In this case, the taxa identified as rogue are highly unstable due to missing data, and obtaining further sequencing for these has the potential to be key in improving the resolution and reducing "terrace sizes" in tree-space. In a rather mundane context, these taxa can be prioritized for sequencing in projects where funding resources or time is limited. Different components in the network can represent clusters in the phylogenetic tree at different levels (when these are not affected by other factors such as conflicting signals, model specifications, etc). In light of this, the pipeline could be used beyond determining the taxa for sequencing, and even to identify loci or genes that could be targeted that would enable to help resolve or reduce the number of possible solutions in particular subtrees. This might require a further step, which will involve analyzing the subtrees at the levels wanted, but this will be left for future developments.

The interactive graphical representation of the results implemented in the present pipeline offers the user an easy way to understand the distribution and connections among taxa (nodes) and their type of taxonomic equivalence (edges) in the datasets analysed. Identifying highly connected taxa (representing the most unstable) and updating the number of remaining connections after deletions are easily carried out using this tool, enabling the user to go through them in a methodological way. The stopping points during experimental deletion can be when formerly connected components completely separate, when connected taxa cannot be safely deleted or when their safe deletion does not retrieve further internal nodes in the consensus. Nevertheless, the choice of taxa to delete relies in the end in the user. If a taxon of particular interest to the user appears to be unstable (and suggested to be deleted), it can be chosen to remain (although, bearing the obvious consequence of keeping it in the analysis). Users can decide whether to continue testing deletions in an *a posteriori* way to fit the purposes of their objectives better.

## 2.5 Conclusions

In the last years, there has been growing interest in the detection of rogue taxa in large-scale phylogenetics mostly using purely *a posteriori* approaches (for example Cranston and Rannala 2007; Aberer and Stamatakis 2011; Aberer *et al.* 2013, just to name a few). The Concatabominations method, which sits somewhat between the pure *a priori* approach of STR and purely *a posteriori* approaches such as leaf stability (Thorley and Wilkinson 1999) or reduced consensus (Wilkinson 1994a) offers another approach to this problem.

The Concatabominations method presents an improvement from its predecessor (original STR). The heuristic approach in this method can identify new taxa that were not evident in the original STR and outperforms it by further testing the ambiguous type D that the original STR failed to determine for deletion. This approach could be particularly useful for paleontological datasets (that present numerous missing data) and phylogenomic supertree constructions from concatenated genes available in electronic repositories. The safe removal of rogue taxa due to missing data can potentially reduce the terraces in tree space search and improve resolution in the final consensus tree. When seen from a pragmatic point of view, the new method can help in targeting taxa that require further sampling or sequencing during a research design.

# Chapter 3. Placing taxa in the context of a large-scale phylogeny

## 3.1 Introduction

Substantial steps have recently been made in resolving amphibian phylogenetic relationships at all taxonomic scales (e.g. Frost *et al.* 2006; Pyron and Wiens 2011). In particular, Pyron and Wiens (2011) assembled a superalignment for 2,872 amphibians, potentially providing a useful starting base for investigating the phylogenetic position of previously unsampled and unincluded taxa (Blackburn and Wake 2011). However, what might constitute best use of prior phylogenetic work and resources is not necessarily obvious. For example, should we simply append or shoehorn data for new taxa into an existing superalignment, thereby accepting previous strategies employed in marker selection, alignments and masking or should we re-evaluate some or all of these aspects? Should we accept previous phylogenetic conclusions and use these as topological constraints in order to expedite efficient placement of the newly included taxa or should we begin time-consuming unconstrained analyses *de novo*?

In the present work, I attempt to find the taxonomic placement of *Ericabatrachus baleensis* Largen 1991, the sole member of its genus. This is an enigmatic and critically endangered frog known only from the Harenna Forest in the Bale Mountains of Ethiopia and, until recently, only from the original collection made in 1986 (Hillman 1988; Largen and Drewes 1989; Largen 2001; see also Gower *et al.* 2013). In his description of the genus and species, Largen (1991) noted that although he intended to study comparative osteology, which might have provided compelling insights into the evolutionary affinities of *E. baleensis,* this was not completed in time. Largen tentatively concluded, on the basis of shared external features such as terminally T-shaped ("bifid") phalanges, that it is a petropedetine (= petropedetid of some classifications) and thus most closely related to the East African *Arthroleptides* Nieden, 1911 (= *Petropedetes* Reichenow, 1874), East and West African *Petropedetes* Reichenow, 1874*,* and Central/West African *Phrynodon,* Parker, 1935. Petropedetinae/dae is a putatively monophyletic group nested within the large clade of "True Frogs" here termed "ranids" (Dubois 1992, 2005).

Uncertainty over the affinities of *Ericabatrachus* is reflected in a period of taxonomic instability from 2005 until present (see Appendix A and Figure 3.1 for summary). Dubois (2005) suggested an affiliation between *Ericabatrachus* and *Phrynobatrachus* but this was not based on explicit data or analyses (Figure 3.1 b). The same year, Scott (2005) published the first broad scale analysis of ranid phylogeny based on both morphology (predominantly osteology derived from x-ray photography and clearing and staining techniques) and DNA sequence data. Only morphological data were available for *E. baleensis* and Scott's (2005) analyses recovered *Ericabatrachus* within the primarily southern African cacosternids, separate from phrynobatrachines and only distantly related to petropedetines (see Figure 3.1 c). Subsequently, substantial changes to amphibian classification have been proposed by Frost *et al.* (2006) and Pyron and Wiens (2011) on the basis of large-scale phylogenetic analyses of mostly DNA sequence data. Neither of these studies included *Ericabatrachus* in their phylogenetic analysis, but Frost *et al.* (2006) included *Ericabatrachus* within Phrynobatrachidae (Figure 3.1 d), considering it likely that it nests within *Phrynobatrachus*, and Pyron and Wiens (2011) included

*Ericabatrachus* within Pyxicephalidae (Figure 3.1 e), the former based only on it being "*Phrynobatrachus* like" (Largen 1991) and the latter based on Scott's (2005) findings. Note that, although Largen (1991) remarked that *Ericabatrachus* was "reminiscent of *Phrynobatrachus*" (p. 147) with "habitus *Phrynobatrachus*-like" (p. 141), he actually considered it to be a petropedetine. In summary, over the past 22 years *Ericabatrachus* has been treated as a member of three different families based entirely on more or less superficial considerations of its morphology.

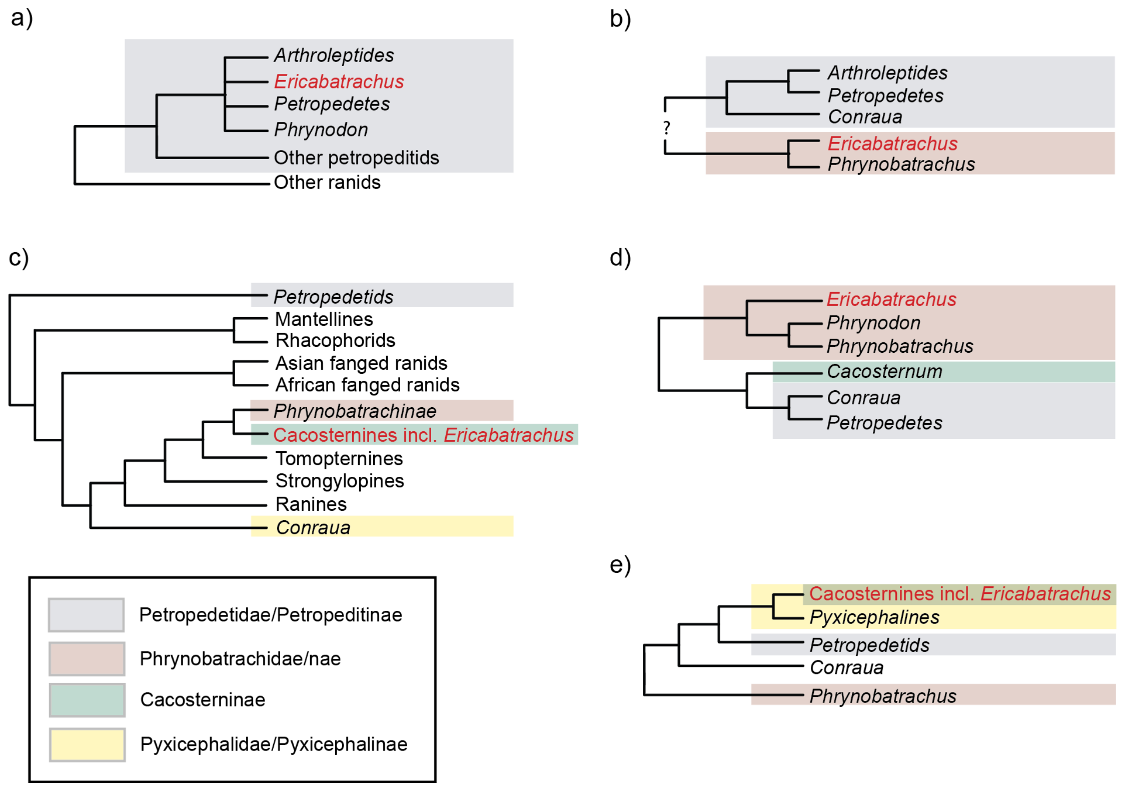


**Figure 3.1** Alternative hypotheses of the relationships of *Ericabatrachus baleensis* and its sister groups (see also Appendix A).

The hypotheses are derived from different sources, which were at the time not necessarily phylogenetic hypotheses but nomenclatural resolutions. Topology a) Largen (1991), b) Dubois (2005) which focused on nomenclatural issues, c) Scott, (2005) based on her figure 4, the consensus of morphological and molecular analyses and the revised classification in appendix 7, d) Frost *et al.* (2006), and, e) Pyron and Wiens (2011).

With a newly collected specimen of *Ericabatrachus baleensis* (see Gower *et al.* 2012; Gower *et al.* 2013), DNA sequence data can, for the first time, be used to investigate the phylogenetic relationships of this challenging taxon. Inferring the phylogenetic relationships of *Ericabatrachus* has important implications for both biogeography and conservation. If the phylogenetic relationship of *Ericabatrachus* indicates an affiliation with Petropedetidae, this would support the continuity of the recognised Afromontane region (White 1978). Alternatively, relationships shared with predominantly southern African taxa (either Pyxicephalidae or Phrynobatrachidae) would provide evidence of a less well-recognized biogeographical association. Phylogeny is an important consideration in conservation prioritization (e.g. Isaac *et al.* 2012), and resolution of the relationships of *E. baleensis* will shed light on the validity of *Ericabatrachus* as a monotypic genus and the degree to which this now Critically Endangered (IUCN 2013) frog contributes to the genetic distinctiveness of conservation targets in the generally threatened (Gower *et al.* 2013) Bale Mountains of Ethiopia.

Here I use the newly generated DNA sequence data to investigate the phylogenetic relationships of *Ericabatrachus* and some of the possible strategies for incorporating previously unsampled taxa into large-scale phylogenetic analyses.

## 3.2 Material and Methods

### 3.2.1 Sampling and DNA extraction

Fieldwork was conducted in July to August in 2008 in southeastern Ethiopia (Figure 3.2 a-b) and June 2009, in Harenna Forest in Bale Mountains National Park. Harenna Forest is the type locality of *Ericabatrachus baleensis* Largen 1991, and comprises patchy, montane, primary rain forest and secondary vegetation (Largen 1991; Miehe and Miehe 1994; Gower *et al.* 2012; Gower *et al.* 2013).

In 2008, collected amphibian specimens, including a single sample of *Ericabatrachus baleensis* (ZNHM-AAU-A2013-003) found under rock, just beside stream on 2nd August at 11.35am in cloudy but not rainy conditions. The specimen (Figure 3.2 c) was collected at a site in Harrena Forest called "Fute" (6.76474 N 39.751661 E, at

3208m). Almost one year later (20th June 2009), a further two specimens (ZNHM-AAU-A2013-001, ZNHM-AAU-A2013-002) phenotypically similar to the first specimens and those of Largen (1991), were secured at the same locality (Gower *et al.* 2013). All specimens were euthanized by submerging them in MS222 dissolved in water. These were then fixed in ca. 5% formalin, rinsed in water and stored in 70% ethanol in the collections of the Natural History Museum of Addis Ababa, Ethiopia. Tissue samples (liver) were taken from specimens prior to fixation and preserved in absolute ethanol.

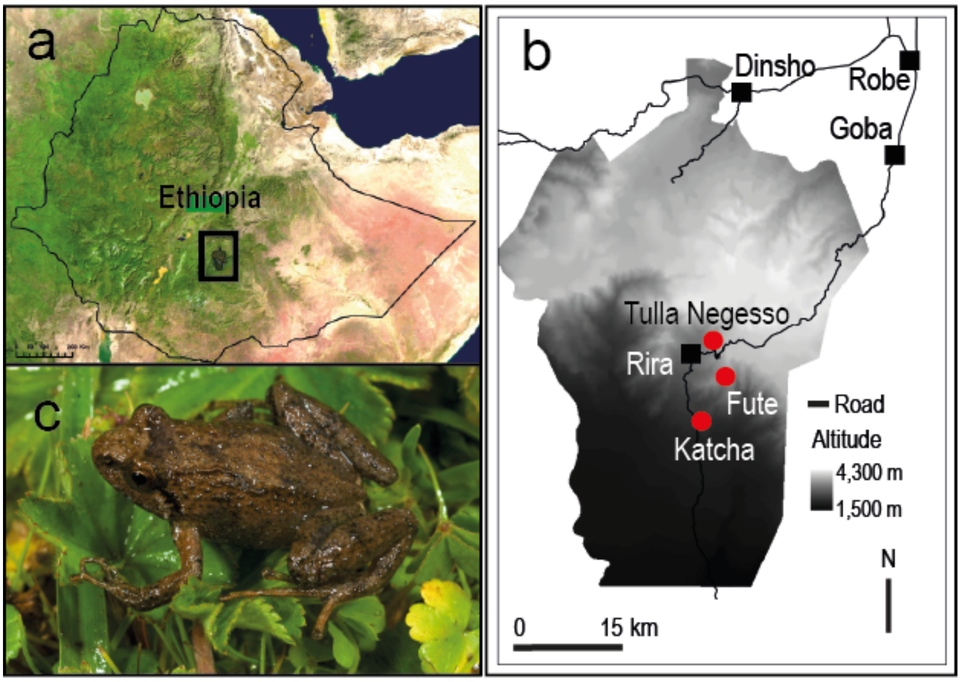


**Figure 3.2** *Ericabatrachus baleensis* and its reported localities.

a) Map showing the Bale Mountains National Park in Ethiopia. b) Close-up of the Bale Mountains National Park showing the geographic position of the type locality and other sites where *E. baleensis* was found. c) A specimen of *E. baleensis* found in the recent surveys (Gower *et al.* 2013).

Genomic DNA was extracted from each of the three *Ericabatrachus baleensis* liver samples using a commercial kit (Qiagen). The 2008 sample was amplified by PCR and two partial mitochondrial (mt) genes, 12SrRNA (*12S*), and 16SrRNA (*16S*) and three nuclear (nu) genes, 28S ribosomal RNA (*28S*), Histone H3a (*H3A*), and recombinase activating protein 1 (*RAG1*) were sequenced. In addition, *12S*, *16S* and *RAG1* were sequenced for the 2009 samples to test for conspecific differences. See Appendix B for all primers used and Genbank accession numbers.

### 3.2.2 Data Matrix

To investigate the phylogenetic relationships of *Ericabatrachus baleensis* I added the genes we sequenced to the superalignment of Pyron and Wiens (2011). Pyron and Wiens' (2011) data set covers the entirety of the Amphibia but it seems reasonable to suppose that including sequence information for non-anuran amphibians or for some groups within Anura to which *Ericabatrachus* clearly does not belong would not be helpful. Inclusion of distantly related sequences (e.g. for salamanders and caecilians) would be at a cost of increased computational complexity and would potentially lead to suboptimal model selection for the phylogenetic problem at hand. Accordingly, the attention was restricted to Ranoidea (sensu Pyron and Wiens 2011 and Frost *et al.* 2006), as there seem to be little doubt that *Ericabatrachus* is a member of this taxon (see Largen 1991; Scott 2005).

The Ranoidea superalignment derived from Pyron and Wiens (2011) was decomposed into its constituent genes. For each gene, taxa with only missing data and empty columns (alignment gaps) were deleted. For all protein coding genes, first, second, and third codon positions were identified, and reading frames verified using Mega v.5 (Tamura *et al.* 2011). In the case of the non-coding *12S* and *16S* partitions, the alignments were only inspected by eye, no obvious, problems were found and, no further testing was performed.

New sequences for *Ericabatrachus* and for some other potentially highly relevant species that were not in Pyron and Wiens (2011) work, namely the *16S* genes for *Petropedetes euskircheni, P. perreti, P. juliawurstnerae, P. vulpiae* and *P. johnstoni* were added (Appendix B) to the corresponding alignments using the profile method in Muscle (Edgar 2004). The data were further extended by the addition of 28S rRNA (*28S*) sequences for all the included species for which this nuclear marker was available using the structure-based alignment of Mallatt *et al.* (2010) as a reference (after having deleted all non-amphibian species and having removed all gap-only columns). A final round of verification was performed during which the alignments were opened in Jalview (Waterhouse *et al.* 2009), inspected by eye and modified as necessary. Ultimately, the initial concatenated, pruned and extended Ranoidea superalignment included the following markers (and numbers of sequences): *12S* (645), *16S* (795), *cytb* (244), *28S* (144), *H3A* (141), *RAG1* (258), *CXCR4* (56), *SLC8A1* (73), *POMC* (45), *RHOD* (340), *SIA* (114), *SLC8A3* (52) and *TYR* (301), that were amassed for a total of 858 species, even though obviously, not all species had data for all the markers.

### 3.2.3 Saturation Analysis

Saturation was investigated in alternative data partitions (genes and codon positions) using saturation plots generated using the program Patristic v.2 (Fourment and Gibbs 2006) from tip-to-tip distances for corresponding pairs of taxa on trees derived using uncorrected distances (p-distance) and the HKY85 + G model. Partitions that did not display substantial deviations from a linear regression pattern between the observed (p) distances and the HKY85 distances are not saturated. In contrast, a plateau (i.e. increasing HKY85 distances correspond to non-increasing observed distance) is indicative of sequence saturation (e.g. Sperling *et al.* 2009; Rota-Stabelli *et al.* 2011). Saturation plots also allow the identification of sequences that are highly dissimilar from their putative homologs in the data set (probably due to poor curation or contamination). Saturated partitions and outlier sequences (with extremely high tip-to-tip distances with respect to all the other sequences in the data set) were excluded in an attempt to minimize the potential emergence of saturation-driven tree reconstruction artifacts (as previous works on amphibian phylogeny have done, e.g. Zhang *et al.* 2013).

### 3.2.4 Phylogenetic Analysis: A two-tiered approach

Given previous disagreement and uncertainty over the phylogenetic placement of *Ericabatrachus*, a "large-scale" approach was initially employed (including all Ranoidea). The large-scale dataset comprised 858 taxa and 9960 basepairs (bp). Maximum likelihood (ML) inferences and non-parametric bootstrapping were carried out using RAxML (Stamatakis 2006). For this analysis, unlinked GTR + G (GTRGAMMA) models were used across the different gene partitions. The family Hemisotidae (represented by *Hemisus marmoratus*) was used as outgroup in this analysis, based on the species position, being recognized as one of the basal taxa in the Ranoidea in previous works (Frost *et al.* 2006; Roelants *et al.* 2007; Pyron and Wiens 2011). Additionally, I investigated the use of a partitioned model, identified using PartitionFinder (Lanfear *et al.* 2012), which suggested that some of the partitions initially defined should be merged. The PartitionFinder model separated the data according to codon position and whether they had mitochondrial or nuclear origin. For comparison, I conducted a parallel large-scale analysis in which the Ranoidea section of the Pyron and Wiens (2011) tree was used as a topological constraint, with only the positions of the newly introduced taxa (*Ericabatrachus* and some *Petropedetes*) unconstrained.

Subsequent, "small-scale" analyses were performed using a subset of taxa, selected on the basis of the large-scale ML analyses and their relative completeness, to better contextualize and further investigate the phylogenetic relationships of *Ericabatrachus*. The small-scale data set (66 taxa and 8216 bp) included all species belonging to Petropedetidae, Pyxicephalidae (comprising Pyxicephalinae + Cacosterninae), Conrauidae and Micrixalidae. Additionally, two representatives (chosen such as to minimise missing data) from each of the Ptychadenidae, Phrynobatrachidae, Ceratobatrachidae, Dicroglossidae, Mantellidae, Ranidae and Rhacophoridae clades were included. Using this small-scale data set allowed missing entries to be reduced (from having 78% missing entries in the large-scale data set to 65% in the small-scale data set) and the use of Bayesian inference under the often better-fitting CAT-based models in PhyloBayes v.3.3 (Lartillot *et al.* 2009). Three separate Bayesian analyses were performed. These used GTR + G, CAT + G, and CAT-GTR + G. For each analysis two runs were performed. The initial 1000 trees (~10%) sampled in each MCMC run were discarded as the burn-in. For comparison, a ML GTR + G analysis of this data set was also performed (using RAxML). In all ML analyses performed, support values were estimated using non-parametric bootstrap (100 replicates). All trees were visualized and handled in iTOL (Letunic and Bork 2011).

Approximately Unbiased (AU) tests of two trees were used to compare the fit to the small-scale data of our new and the previously proposed (Figures 3.1 b-e) hypotheses of the relationships of *Ericabatrachus* not including Largen's (1991) very incompletely resolved hypothesis (Figure 3.1 a). A total of eight trees were tested:

those in Figures 3.1 b-e, plus my Bayesian (GTR + G, CAT + G and CAT-GTR + G) trees and ML (GTR + G) tree. To compare trees in Figures 3.1 b-e with my results, a preliminary series of AU tests was performed (under GTR + G) including only the trees generated from my analyses. Site-wise log-likelihoods were recalculated (for each of these topologies under GTR + G) in RAxML, and these likelihood values were used to estimate significance in CONSEL v.0.2 (Shimodaira and Hasegawa 2001). The tree with the best overall fit was my Bayesian GTR + G tree. This tree was then selected as the backbone to generate (by manually editing the position of *Ericabatrachus* and other taxa), trees representing the hypotheses in Figures 3.1 b-e. By using the tree that provided the best fit to the data (from my preliminary AU analyses) I avoided introducing a potential bias that might have disfavored previous hypotheses not on the grounds of their placement of *Ericabatrachus* but because of the relationships they displayed for other irrelevant taxa. The trees representing the previous hypotheses and the trees from my original analyses were then subjected to another round of AU tests (under GTR + G). Additionally, I pruned the newly added taxa (*Ericabatrachus* and some *Petropedetes* species) from my Bayesian tree, used the strict consensus to compare this topology with that of the Pyron and Wiens' (2011) tree restricted to the common taxa, and used AU tests to compare the fit of these two trees to my and to Pyron and Wiens' (2011) data (under GTR + G) restricted to the subset of taxa.

## 3.3 Results

### 3.3.1 Saturation analysis

Saturation plots (Table 3.1 and for actual plots go to Appendix C) supported the inclusion of the following partitions in the large-scale phylogenetic analysis: *RAG1* codon positions 1, 2 and 3; *H3A* codon positions 1 and 2; *16S*; *12S*; *28S*; *CXCR4* codon positions 1, 2 and 3; *SLC8A1* codon positions 1, 2 and 3; *POMC* codon positions 1, 2 and 3; *RHOD* codon positions 1 and 2; *SIA* codon position 2; *SLC8A3* codon positions 1, 2 and 3; *TYR* codon positions1 and 2; and *cytb* codon positions 1 and 2. An outlier species was detected in the *28S* saturation plot, *Fejervarya limnocharis*, which was excluded from the analysis.

**Table 3.1** Summary of saturated (S) and non-saturated (NS) partitions resulting from the Saturation analysis.

| Protein-coding partitions | | | |
|---|---|---|---|
| | Codon position | | |
| | 1 | 2 | 3 |
| CXCR4 | NS | NS | NS |
| SLC8A1 | NS | NS | NS |
| POMC | NS | NS | NS |
| RHOD | NS | NS | S |
| SIA | S | NS | S |
| SLC8A3 | NS | NS | NS |
| TYR | NS | NS | S |
| cytb | NS | NS | S |
| RAG1 | NS | NS | NS |
| H3A | NS | NS | S |
| Non protein-coding partitions | | | |
| 16S | NS | | |
| 12S | NS | | |
| 28S | NS | | |

### 3.3.2 Phylogenetic analyses

The major clade relationships and topologies obtained from the ML large-scale (858-taxon data set) analysis resemble that of previous works (van der Meijden *et al.* 2005; Roelants *et al.* 2007; Pyron and Wiens 2011; Zhang *et al.* 2013). In particular, the sister group relation between the clades Petropedetidae and Pyxicephalidae is also in agreement with previous hypotheses (Frost *et al.* 2006; Roelants *et al.* 2007; Pyron and Wiens 2011; Zhang *et al.* 2013). The large-scale ML analysis recovered *Ericabatrachus* as the sister taxon of *Petropedetes* with a bootstrap support (BS) of 59% (see Figure 3.3 a). This low BS is primarily a consequence of *Ericabatrachus* being associated with other clades in 35% of the bootstrap replicates (BR) (Table 3.2) but is contributed also by the instability of *Petropedetes newtoni* which was found outside of *Petropedetes* + *Ericabatrachus* in 9% of the BR. Hence, the effective support for an *Ericabatrachus-Petropedetes* (with exclusion of *P. newtoni*) relationship is 65% (Table 3.2). The second most frequent position (25% of the BR) place *Ericabatrachus* as a sister group of or nested inside Pyxicephalinae (the clade composed of *Aubria* + *Pyxicephalus*). Taken together these results circumscribe a relatively well-defined area of the tree in Figure 3.3 a and Table 3.2 within Ranoidae (of van der Meijden *et al.* 2005, or Natatanura of Frost *et al.* 2006), in which *Ericabatrachus* occurs with a cumulative bootstrap proportion of ~99%. This allows narrowing the set of plausible relationships for *Ericabatrachus,* and permits more focused analyses to be performed. Using the Pyron and Wien's (2011) tree as a topological constraint produced very similar results with respect to the position of *Ericabatrachus* including similar BS scores (Figure 3.3b).

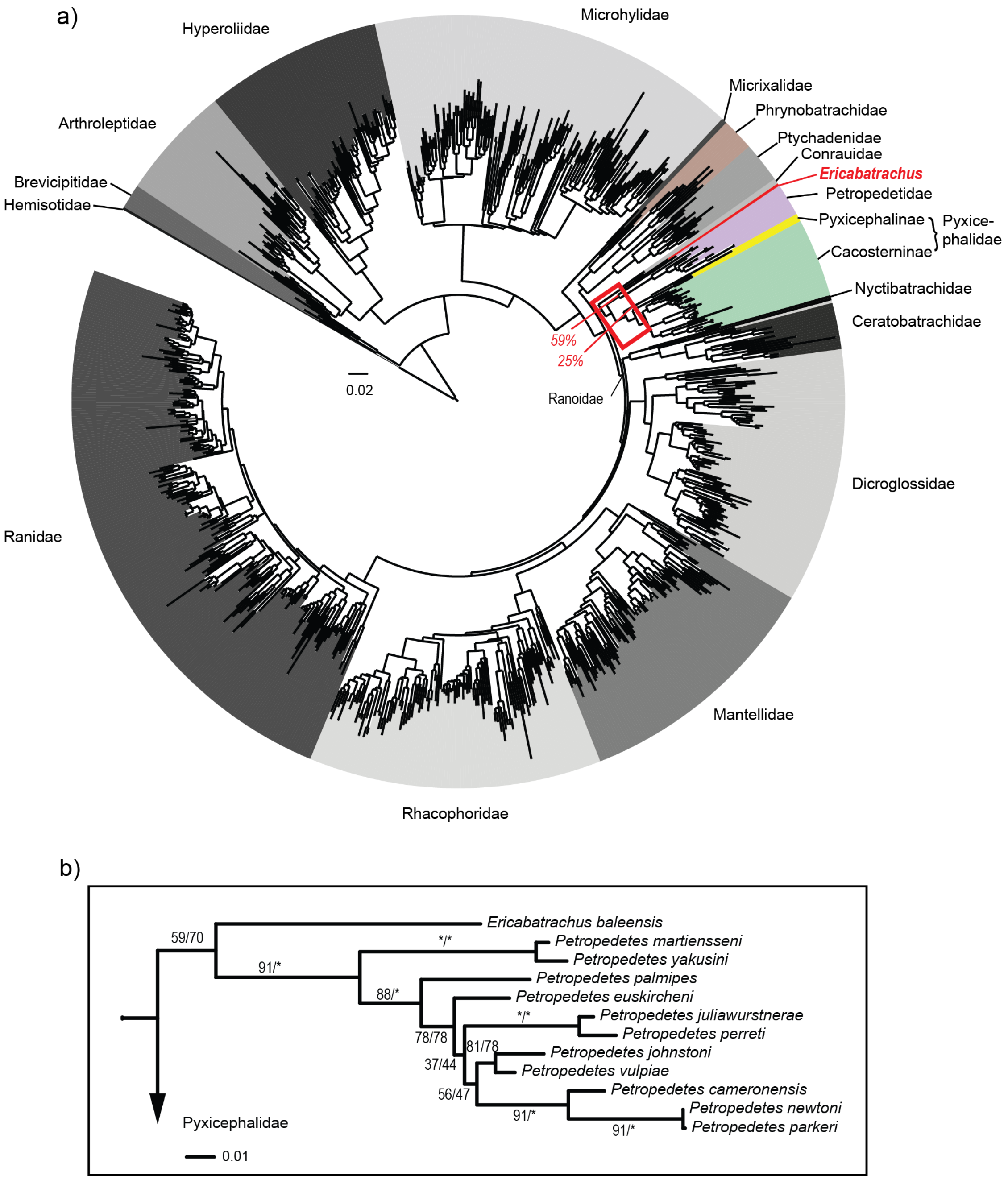


**Figure 3.3** Optimal ML tree from the large-scale analysis of Ranoidea.

a) Most frequent placement for *Ericabatrachus* with the corresponding percentages are shown in red. The red square denotes the narrowed down area where *Ericabatrachus* is most likely to be positioned 99% of the times (see text). b) Close up view of position of *Ericabatrachus* as the sister taxon of *Petropedetes*. BS values of each branch correspond to the *de novo* analysis (left) and to the constrained analysis (right).

**Table 3.2** Table summarizing positions for *Ericabatrachus* in the bootstrap replicates (BR) of the large-scale reconstruction.

| Position | Number of BR supporting position |
|---|---|
| Sister taxon of Petropedetidae | 65 |
| Sister taxon of Conrauidae + Petropedetidae | 3 |
| Sister taxon of Pyxicephalinae | 22 |
| Sister taxon of Conrauidae | 4 |
| Sister taxon of Phrynobatrachidae | 1 |
| Sister taxon of Pyxicephalidae | 1 |
| Nested in Pyxicephalinae | 3 |
| Sister taxon of Petropedetidae + Pyxicephalidae | 1 |

Focused, small-scale Bayesian and ML analyses (66-taxon data set) confirmed *Ericabatrachus* as the most likely sister group to *Petropedetes* (Figure 3.4). The posterior probability for this position under the GTR + G, CAT + G, or CAT-GTR + G models is invariably equal to one. ML bootstrap support is only marginally increased (to ~ 67%). The topologies obtained in different analyses of the 66-taxon data set are almost identical, varying only in the positions of *Occidoziga lima*, *Phrynobatrachus kreffti* and the *Micrixalus* clade. AU tests show that the phylogenetic placement of *Ericabatrachus* obtained in our Bayesian and ML results fits the 66-taxon data significantly better than any previously proposed hypotheses (Table 3.3).

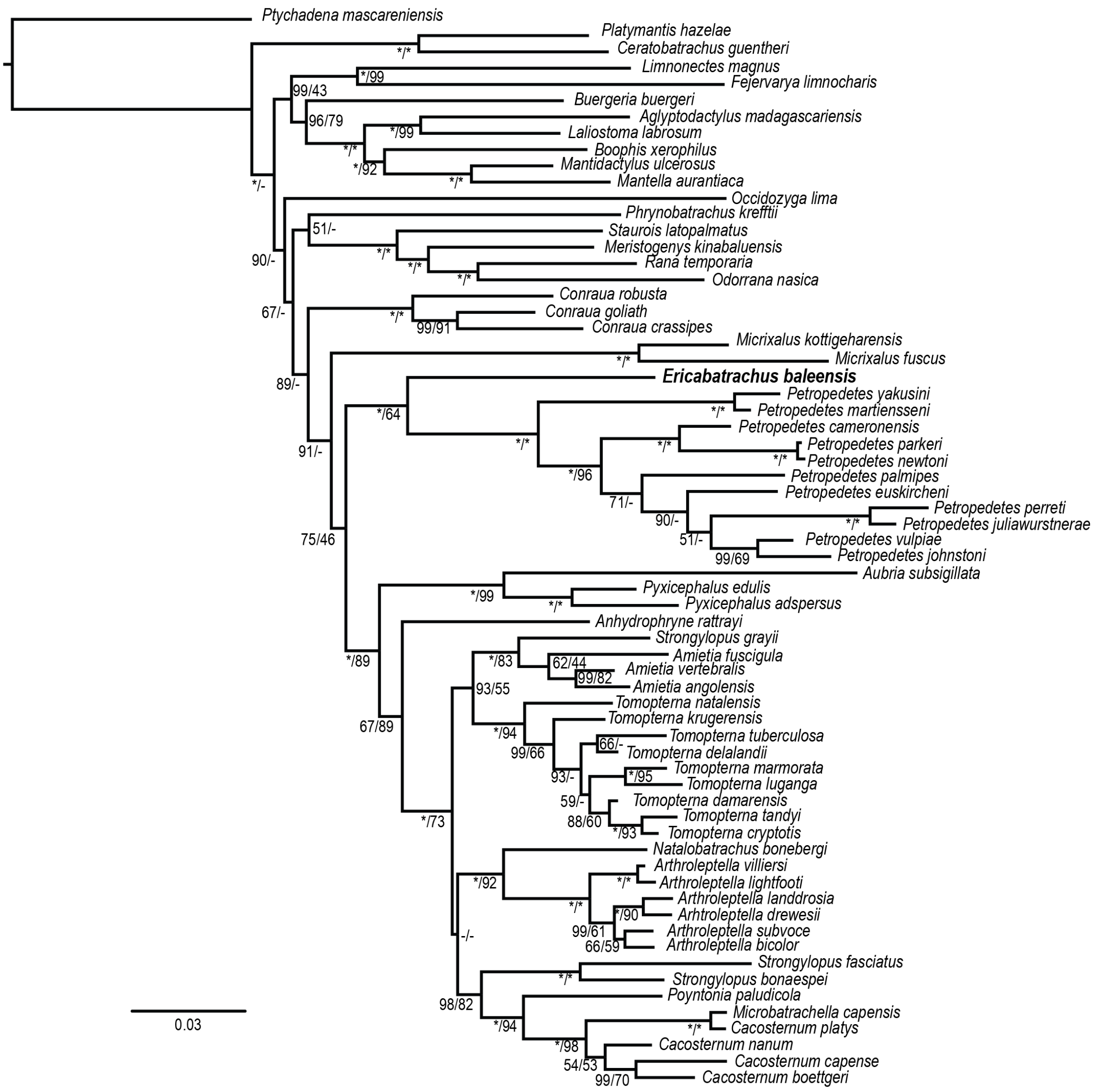


**Figure 3.4** Bayesian tree under GTR model of a subgroup of Ranoidea showing the phylogenetic placement of *Ericabatrachus baleensis* (in bold letters).

Support values for the nodes correspond to posterior probabilities (left) and non-parametric bootstraps (right). Values with "*" represent full support (100%), values lower than 40% are denoted by "-".

**Table 3.3** Hypotheses testing results from CONSEL (Shimodaira and Hasegawa 2001).
Values shown for the Approximately-Unbiased test (AU test). Dotted line separates the non-rejected hypotheses (above line) from the rejected hypotheses (below line).

| Rank | Item | AU test |
|---|---|---|
| 1 | present work, Bayesian GTR Tree | 0.853 |
| 2 | present work, ML Tree | 0.262 |
| 3 | Dubois (2005) hypothesis | 0.008 |
| 4 | Pyron and Wiens (2011) hypothesis | 1.00E-05 |
| 5 | Frost *et al.* (2006) hypothesis | 1.00E-05 |
| 6 | Scott (2005) hypothesis | 4.00E-08 |

The strict consensus of our small-scale Bayesian tree and the Pyron and Wiens' (2011) tree (both restricted to the common taxa) includes a large basal polytomy but is well-resolved in the area where the new taxa (*Ericabatrachus* and some *Petropedetes*) join the tree (Figure 3.5). There is a more substantial difference in log-likelihoods between these two trees with the present work alignment (24.2) than with the Pyron and Wiens' (2011) alignment (8.3), but results of AU tests of these restricted topologies using either this work's alignment or that of Pyron and Wiens' (2011) were not significant (p = 0.089 and p = 0.331 respectively).

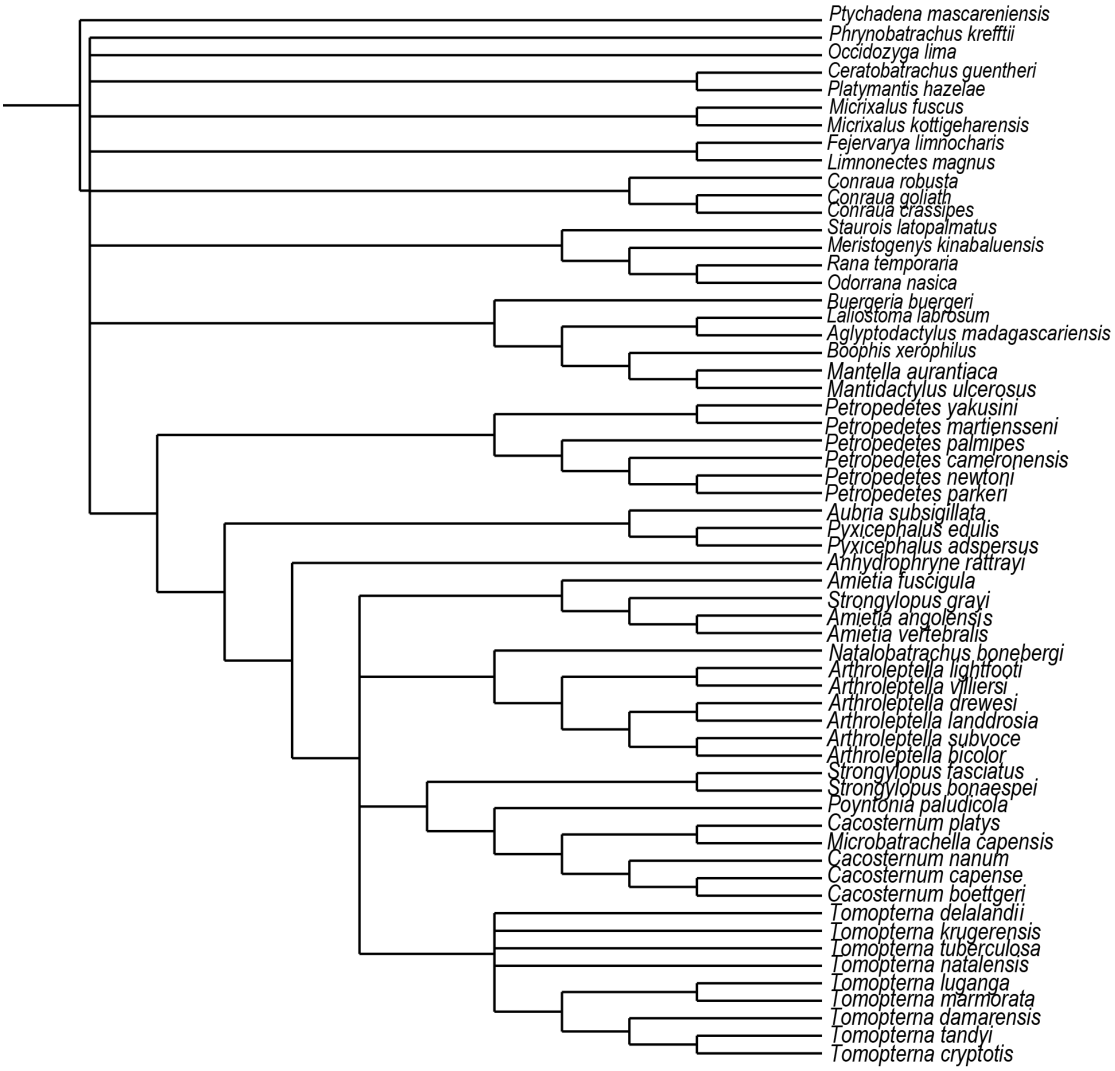


**Figure 3.5** The strict consensus of the present work small-scale Bayesian tree and the Pyron and Wiens' (2011) tree (both restricted to the common taxa).

Polytomies represent relationships that were in disagreement between the two trees.

## 3.4 Discussion

### 3.4.1 Taxonomy, phylogeny and biogeography

Comprehensive phylogenetic analyses of newly acquired molecular data for the rare and Critically Endangered *Ericabatrachus baleensis* provide good support for a sister-group relationship with *Petropedetes* Reichenow, 1874. The results of this work support Largen's (1991) original assignment of *Ericabatrachus* to the family Petropedetidae (although his concept of "Petropedetidae" was somewhat different from current taxonomy). Alternative groupings proposed more recently by other authors (i.e. with Phrynobatrachidae and Cacosterninae - Scott 2005; Frost *et al.* 2006; Pyron and Wiens 2011) are not supported by my concluding analyses. *Ericabatrachus* is placed at the base of Phrynobatrachidae (Frost *et al.* 2006) only once in the bootstrap replicates, and never in Cacosterninae (Scott 2005). In terms of evolutionary relationships within "ranids", in my analysis Petropedetidae forms a sister group to a southern African radiation of ranids (Pyxicephalidae), with Conrauidae lying outside this pairing, which is in agreement with previous hypotheses (i.e. Frost *et al.* 2006; Roelants *et al.* 2007; Pyron and Wiens 2011; Zhang *et al.* 2013). Other possible resolutions are rejected by AU tests (see Table 3.3).

The genus *Petropedetes sensu* Scott (2005) comprises 12 nominal species distributed in both East and Central Africa. Largen (1991) was aware of the high degree of morphological dissimilarity between *E. baleensis* and other petropedetids (*Petropedetes*, *Arthroleptides* (=*Petropedetes*) and *Phrynodon (=Phrynonbatrachus sandersoni)*) and he was not drawn on any particular putative sister-group relationship. It might have been suspected that, given the geographical proximity of the highlands of Kenya and Tanzania, and the relative biogeographical continuity of this area with the Ethiopian highlands, *E. baleensis* was most closely related to *Petropedetes* from East Africa paralleling suspected relationships for other eastern African montane frogs such as *Balebreviceps* and *Nectophrynoides* (see Grandison 1978; Largen and Drewes 1989; Largen 1991). This East African unit (*Ericabatrachus, P. martiennseni, P. yakusini*) is not supported in the analyses of this work. Sampling of *Petropedetes* is almost complete, but data are lacking for *P. dutoiti* and *P. natator* from West Africa and this awaits to be tested (Barej *et al.* 2014), which might alter our understanding of the relationship of *E. baleensis* relative to all known *Petropedetes*. From the results of this work, *Petropedetes* (excluding *P. dutoiti* and *P. natator*) forms a well supported clade with 91% bootstrap value in the large-scale ML tree (which is caused by the instability of *P. newtoni*), and full support in the small-scale ML and Bayesian analyses (Figure 3.4).

*Ericabatrachus* has been one of the most problematic genera of African ranids to classify. Efforts were hampered by the lack of molecular data since its description in 1991 but uncertainty was compounded by the fact that *Ericabatrachus* has a suite of morphological characters that have confused understanding of its evolutionary relationships. Characters that might have supported Largen's suspicion that *Ericabatrachus* was a petropedetid were seemingly not revealed in Scott's (2005) analysis, who placed it in Cacosterninae. Only the presence of dorsal scutes provided a potential unique synapomorphy for such a grouping (*Ericabatrachus, Arthroleptides* and *Petropedetes*). Among the remaining non-unique synapomorphic characters that grouped *Arthroleptides* and *Petropedetes,* they are either coded in *Ericabatrachus* as unknown, inapplicable, or varying (see Scott 2005 characters 31, 32, 48, 62, 64, 74, 98, 140). The reported "aberrant" character states (Scott 2005: p.

532) might require an investigation for those unknown and potentially a re-assessment in light of the findings presented here. What is clear is that the definition of the genus *Petropedetes* and family Petropedetidae based on morphological characters is still unclear and requires further work. Characters that have been reported to differentiate *Ericabatrachus* from the genera *Petropedetes* and *Arthroleptides* are conspicuously reduced first finger, weak subarticular tubercles, hidden or barely visible tympanum and striking ventral colours (Largen 1991). The functional significance of morphological features in African ranids can now be better understood against those derived from common ancestry. As previously noted by Largen (1991; p.151) *Ericabatrachus* would appear to be an interesting taxon to add to studies on correlated patterns of evolution in geographically isolated localities in riverine adapted African ranid species. Further research into the still rather complex, and fluctuating taxonomy of African ranids will be necessary before a full and suitable nomenclatural resolution can be made.

Biogeographically, *Ericabatrachus* has fascinated herpetologists since its original description. It is restricted to the high montane forest of the Bale Mountains, part of the fragmented chain of the Afromontane region (Gower *et al.* 2013). Ethiopia is the most northerly, and therefore isolated part of an extensive chain of mountains in subSaharan Africa (habitats that stretch across Africa north to south). Notable endemics are known from Ethiopia and have given rise to the impression that the region is a refuge for old and divergent taxa – often referred to as palaeoendemics. For example, a remarkable assemblage of endemic monotypic amphibian genera were described from Largen's original 1986 collection in Harenna Forest (see summary in Largen 2001). Based on branch lengths in my inferred phylogenies, I suspect that the divergence of *Ericabatrachus* from its closest extant relatives is very old given previous estimates of divergence times with closely related pairings in Petropedetidae, Pyxicephalidae and Conrauidae (e.g., van der Meijden *et al.* 2005; Roelants *et al.* 2007). The phylogenetic results reported support the idea that this species is a palaeo-endemic species. In light of the other putative palaeo-endemic taxa (e.g. *Balebreviceps,* and *Altiphrynoides*) the Bale Mountains of Ethiopia appears to have an intriguing ancient biogeographic history (Loader *et al.* accepted).

### 3.4.2 Conservation

*Ericabatrachus baleensis* has declined substantially since its description. It has not been recorded at its type locality since 1986 (Tulla Negesso) and the only other known historical collecting site (Katcha) since its original collection (Gower *et al.* 2013) and it is now assessed as Critically Endangered on the IUCN Red List. The declines in these localities are likely to be in association with substantial human-induced habitat degradation in the Rira catchment area (Gower *et al.* 2013), but also possibly the emergent infectious disease amphibian chytridiomycosis (Gower *et al.* 2012). Collaborators of this work were only able to locate *E. baleensis* in Fute, a new locality close to Rira, a more pristine habitat. The phylogenetic results from this work demonstrate that the extinction of this frog would be a considerable loss of evolutionary history, thus adding to the demand (Gower *et al.* 2013) that urgent conservation action is taken. This could include both ex situ or in situ approaches, but given the co-occurrence of other distinctive, potentially palaeo-endemic taxa in this locality, a more integrated in situ conservation action would seem to be preferable.

### 3.4.3 Incorporating previously unsampled taxa into large-scale phylogenetic analyses

With the collection of previously unsampled and enigmatic taxa of uncertain phylogenetic relationships, such as *Ericabatrachus,* then (ignoring the choice of markers) one might try to find related taxa to include in a phylogenetic analysis with a BLAST (Altschul *et al.* 1990) search database query, produce an alignment and analyse it as exhaustively as seems worthwhile. However, in the age of large-scale phylogeny projects, researchers are increasingly likely to have access to relevant mega-alignments and trees from previous phylogenetic studies, such as those available in repositories like Data Dryad (www.datadryad.org) and TreeBASE (Piel *et al.* 2002) just to mention a few. Such resources might greatly simplify and speed up the inference of phylogenetic relationships of previously unsampled taxa. For example, expanding the data through profile alignment and using previous trees as topological constraints, have the potential to produce fast results.

However, relying upon previous alignments and trees carries the risk that they are not optimal, particularly given the inclusion of additional taxa (and genes) and the potential such addition has to change the inferred interrelationships of other taxa. We might consider *de novo* alignment and unconstrained phylogenetic analyses to be the optimal use of the new data because it would avoid such risks. But when resources are limited, seeking to use previous results to speed up analyses can provide a practical solution.

Here, the main strategy was to use the previous study of Pyron and Wiens (2011) as a convenient source of aligned data and as a guide to the taxonomic content of a major clade whose background knowledge or assumptions suggested included

*Ericabatrachus*. The alignment was expanded with taxa and an additional marker and *de novo* large-scale analyses were conducted, which in turn informed taxon selection for further small-scale analyses using additional methods and models. Different from Pyron and Wiens (2011), these *de novo* analyses included removal of seemingly saturated data partitions which is generally considered to be helpful in phylogenetic analyses (Sperling *et al.* 2009; Rota-Stabelli *et al.* 2011; Rota-Stabelli *et al.* 2013) and has even been carried out in recent Amphibian phylogenetics (e.g. Zhang *et al.* 2013). Substantial topological differences between the Pyron and Wiens (2011) tree and my tree (Figure 3.5) result from these differences in the data and its analyses. Although AU tests do not allow rejection of either tree, the topological differences highlight that many relationships within the tree are probably best considered uncertain. In turn this might be taken to suggest that the alternative strategy, of using the Pyron and Wiens (2011) tree as a topological constraint, would be problematic. However, this is not the case in this instance. Both this work's *de novo* analyses and use of a topological constraint recovered the same relationships of *Ericabatrachus*. This can be a fortuitous consequence of the incongruences between the tree in this work and the Pyron and Wiens' (2011) tree being concentrated in areas that are least relevant to the relationships of the previously unsampled *Ericabatrachus*.

## 3.5 Conclusions

Novel sequenced data for the rediscovered *Ericabatrachus baleensis* allowed to explore its placement in the Amphibian tree of life. The recent alignment of Pyron and Wiens (2011) was used as a backbone for phylogenetic inferences with ML and Bayesian methods, after careful curation of gene partitions to be included. A two-tiered approach of phylogenetic analyses using ML and Bayesian methods showed that *Ericabatrachus* is the sister group of *Petropedetes*, which is supported by limited morphological evidence. All previous hypotheses of placement are statistically rejected based on the present work data set. Using a constrained tree yields the same phylogenetic position for *Ericabatrachus* demonstrating how this approach may obviate the need for time-consuming *de novo* analyses. In general, constraints should be relied upon only when they are very well-supported.

In the biogeographic context, the current results do not support the hypothesis of African continuity and suggest that *Ericabatrachus baleensis* is a palaeoendemic species, as has been observed in other sympatric species distributed in the Bale Mountains of Ethiopia. These facts reinforce the need to target and prioritize conservation efforts in this enigmatic species and the Bale Mountains.

# Chapter 4. Investigating the phylogenetic relationships of the extant orders of Amphibia with novel genetic evidence

## 4.1 Introduction

### 4.1.1 Conflicting phylogenetic hypotheses of the extant Amphibia

Amphibia is a group of vertebrates that comprise the living representatives (known as Lissamphibia) and its fossil relatives. Lissamphibia includes the crown groups: Anura (frogs), Caudata (salamanders and newts) and Gymnophiona (caecilians), all of which are monophyletic. Anura (also called Salientia when including its fossil relatives) is the most speciose clade with currently 6347 species (AmphibiaWeb 2014, http://amphibiaweb.org/) distributed across all continents (except Antarctica). Caudata is the second most speciose clade, with 655 species (Amphibiaweb), and the distribution of its members is mostly limited to palearctic Eurasia, north of Africa and all of America (Frost 2013). The Gymnophiona is the least studied of the three lineages, and comprises 199 species (Amphibiaweb) that are mostly distributed through the tropics, except for Oceania and Australia (Frost 2013).

The evolutionary relationships among the three orders that form the Lissamphibia still represents a highly controversial question in vertebrate evolution. Most debates centre around two main hypotheses (summarised in Figure 4.1). The first one, the Procera hypothesis, proposes a close relationship between Gymnophiona and Caudata, with the Anura as a sister group to these two. Most early analyses of mitochondrial DNA (Hedges and Maxson 1993; Feller and Hedges 1998), a combination of nuclear and mitochondrial markers (Hedges *et al.* 1990; Zhang *et al.* 2003), and expression sequence tags, termed “ESTs” (Fong *et al.* 2012) have supported this hypothesis, as well as morphological data analysis (Vallin and Laurin 2004 and possibly Pyron 2011). The Procera hypothesis (Figure 4.1 a) seems to have advantages for interpreting distribution patterns and the fossil record of the three orders, given that salamanders and caecilians have strong Laurasian and Gondwanan distribution patterns respectively (Cannatella *et al.* 2009). However such an inference is strongly dependent on the time of divergence estimated for these nodes (i.e. before or after the break-up of Pangaea). Furthermore, most of the molecular analyses that proposed this hypothesis are now considered to have been misled by uninformative data and poor taxon sampling (Cannatella *et al.* 2009), both of which can lead to the recovery of incorrect trees.

The second one, the Batrachia hypothesis (Figure 4.1 b), proposes a close relationship between Caudata and Anura, with the Gymnophiona as a sister group to these two. This hypothesis is the most accepted one, with most morphological analyses that include paleontological data (Milner 1988; Trueb and Cloutier 1991; Ruta and Coates 2007; Marjanović and Laurin 2008, 2009; Sigurdsen and Green 2011; Maddin *et al.* 2012) and recent molecular analyses (Zardoya and Meyer 2001; San Mauro *et al.* 2005; Zhang *et al.* 2005; Roelants *et al.* 2007; San Mauro 2010 and the combined analysis of Pyron 2011 and Shen *et al.* 2013) supporting it. Based on this hypothesis, understanding the current distributions of the major clades of Amphibians becomes more challenging. From a palaeontological perspective the majority of the studies supporting this hypothesis place the divergence of Batrachia and Gymnophiona between the early Carboniferous and the mid Permian (before the break up of Pangaea).

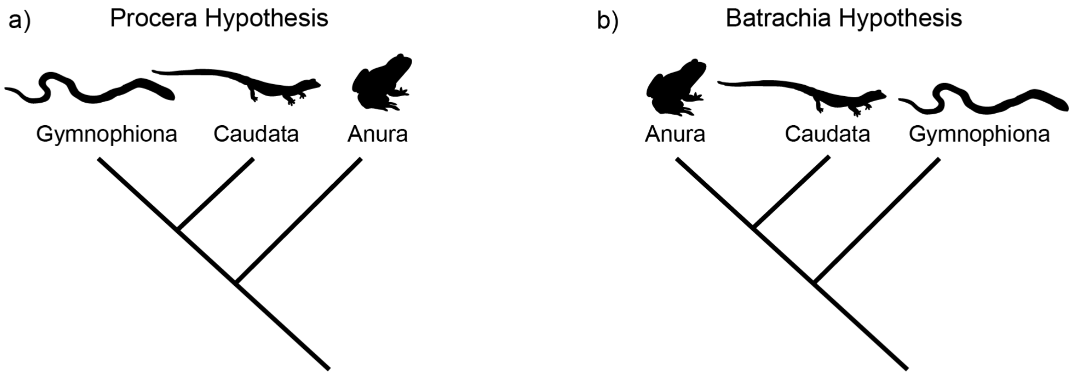


**Figure 4.1** The two hypothesized evolutionary relationships in the Lissamphibia. a) The Procera hypothesis, where the salamanders and caecilians are more closely related to each other. b) The Batrachia hypothesis, where the frogs and salamanders are more closely related to each other. Both hypotheses assume the monophyly of amphibians.

Investigating the evolutionary relationships of the three major groups of Lissamphibia is also strongly tied to understanding their origin and the validation of Amphibia as a monophyletic group. With regards to the monophyly of Amphibia, most morphological studies that have addressed this question using paleontological data of the fossil relatives of Lissamphibia, have reached two possible scenarios: (1) the "temnospondyl hypothesis" (Ruta and Coates 2007; Sigurdsen and Green 2011; Maddin and Anderson 2012; Maddin *et al.* 2012) and the "lepospondyl hypothesis" (Vallin and Laurin 2004; Marjanović and Laurin 2008, 2009; Pyron 2011; Marjanović and Laurin 2013). Both support a monophyletic origin of the lissamphibian lineages but from within different groups of fossil amphibians. (2) A polyphyletic origin of the three orders of modern amphibians from different group of Palaeozoic tetrapods which invalidates a monophyletic Lissamphibia (Carroll 2007; Anderson *et al.* 2008). This second scenario has recently obtained further support from molecular data, as Fong *et al.* (2012), using EST data from different vertebrate species, recovered a tree where Lissamphibia is not monophyletic. Fong *et al.* (2012) presented a large amount of surveyed genes across vertebrates and after carrying out several tests, they concluded that the Lissamphibia are not monophyletic with a Caudata plus Gymnophiona clade representing the sister group of Amniota and the Anura representing the sister group of the latter (Figure 4.2).

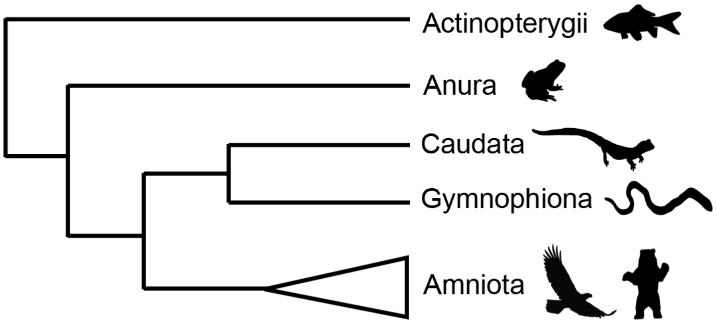


**Figure 4.2** Paraphyletic Lissamphibia hypothesis of Fong *et al.* (2012).
In this work they propose a Caudata-Gymnophiona hypothesis (such as the one in the Procera hypothesis), but with this clade being more closely related to Amniota than to Anura (rendering Lissamphibia paraphyletic).

### 4.1.2 Using miRNAs to investigate the basal relationships of Lissamphibia

MicroRNAs (miRNAs) are small, noncoding regulatory genes implicated in the control of cellular differentiation and homeostasis, and as such might be involved in the evolution of organism complexity (Heimberg *et al.* 2008; Peterson *et al.* 2009;

Christodoulou *et al.* 2010; Heimberg *et al.* 2010). Understanding the biogenesis process of miRNAs is fundamental for enabling their *in silico* predictions. Basically, miRNAs are ~18-22 nucleotides (nt) long, and these are generally transcribed from intergenic regions, (but can also be found in introns and exons) as a long primary transcript (pri-miRNA) which is then folded into a hairpin structure (see Figure 4.3) (Tarver *et al.* 2013). Before been transported out of the nucleus, pri-miRNAs are recognized by a microprocessor–enzyme complex involving the Drosha enzyme (Krol *et al.* 2010), which cleaves the pri-miRNA into a ~70 nt long precursor miRNA (pre-miRNA) (Tarver *et al.* 2013). The pre-miRNA is then transported into the cytoplasm where it is further processed by an enzyme called Dicer (Figure 4.3), cleaving the loop end of the hairpin to form a ~22 nt long RNA duplex with two nucleotide overhangs at each 3'-end (Tarver *et al.* 2013). The duplex molecule is then separated into two strands, the 5'-end and the 3'-end, and one of the strands joins a group of proteins forming a miRNA-protein complex, usually with the Argonaute proteins (Figure 4.3) (Hui *et al.* 2013). There seems to be a preference for one of the strands, which is often called the "mature", while the opposing strand is termed the "star" sequence (Tarver *et al.* 2013). The mature strand is the one that negatively regulates the translation of protein coding genes by binding with imperfect complementarity to sites in the 3'- UTR of the messenger RNAs (mRNAs) (Tarver *et al.* 2013), resulting in the blocking and degradation of the mRNA (Huntzinger and Izaurralde 2011). The seed and the 3'- complimentary motifs are the two most highly conserved regions of mature sequences (Wheeler *et al.* 2009) because they are the most critical for target recognition (Grimson *et al.* 2007; Tarver *et al.* 2013).

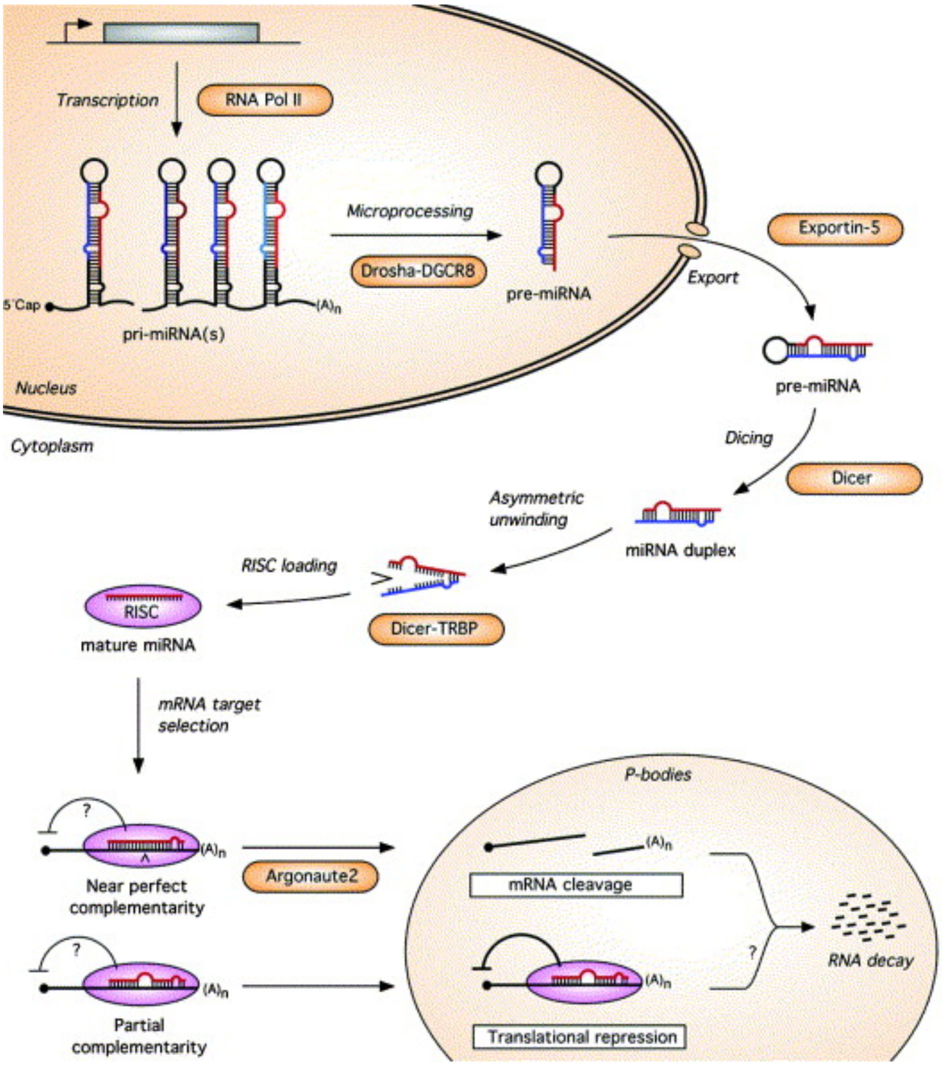


**Figure 4.3** Model of the miRNA biogenesis pathway in animals.

Reproduced from Wienholds and Plasterk (2005). Description of the whole process is given in the text.

A true miRNA with a hairpin structure (Figure 4.4 a), will have consistent Dicer processing sites, which will result in the production of small RNA reads that follow a particular pattern (Friedländer *et al.* 2008; Berezikov 2011). This pattern aids in the *in silico* discrimination of true miRNAs from other RNAs with hairpin structures. The most abundant reads from true miRNAs usually correspond to the 22 nt long mature region, and the less abundant reads will correspond to the star and loop sequences (Friedländer *et al.* 2008; Berezikov 2011). The RNA 5'- ends that correspond to the Dicer cleavage sites should be more conserved, therefore the reads are expected to align uniformly (Figure 4.4 a). The mature region is mismatched from the opposite star region by two nucleotides, which accounts for the overhang end at the 3'- end. On the other hand, when small RNAs are derived from a hairpin by a process different from the precise excision by Dicer, the alignment of the small RNA reads over the hairpin will not give a clear alignment pattern, looking more random, and the 3′- end overhang regions will not be evident (Figure 4.4 b) (Friedländer *et al.* 2008; Berezikov 2011).

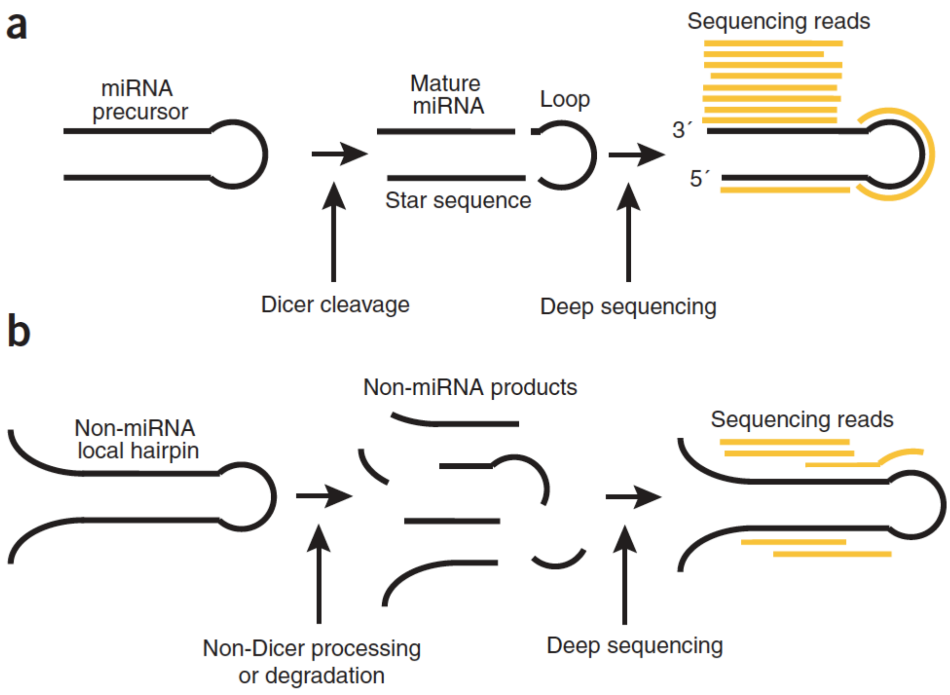


**Figure 4.4** Identification of miRNA based on its biogenesis.

a) Hairpin structures processed by Dicer, as occurs in miRNA biogenesis. b) Hairpin structure not processed by Dicer. Reproduced from Friedländer *et al.* (2008).

The discovery of shared miRNAs across lineages in the animal kingdom has revealed that when these emerge in a particular lineage, they are rarely lost in descendent lineages, hence probably having a key role in phenotypic diversity (Berezikov 2011; Tarver *et al.* 2013). In terms of phylogeny reconstruction, this implies that nearly every animal clade thus far investigated can be characterized by at least one new miRNA family acquisition, making these characters very useful to support phylogenetic relationships. Additional to the continuous increase of miRNA families in animal lineages, miRNAs accumulate mutations very slowly, hence the probability of independent convergent evolution of these molecules in separate lineages is actually very low (yet, they are not homoplasy free, e.g. Philippe *et al.* 2011a). However, because of their size (approx. 18-20 nt long), these are used in phylogenies by annotating their presence or absence in the studied lineages. Using the presence or absence of miRNAs has been criticized because proving that a miRNA is absent is difficult. The inability to identify a miRNA could be due to several reasons, for instance that the miRNA was not active during the developmental stage or in the tissue sampled. Poor quality of sequencing and lack of a reference genome in non-model organisms can also result in the inability to identify a miRNA. In spite of this, and with all their limitations, it is clear that miRNAs can represent an interesting alternative line of genomic evidence to corroborate or reject phylogenetic relationships (Heimberg *et al.* 2010; Campbell *et al.* 2011; Philippe *et al.* 2011a; Rota-Stabelli *et al.* 2011, among some examples).

In this study I investigated the evolutionary hypotheses between the three lineages of Lissamphibia using newly sequenced miRNA data with the current bioinformatics tools available. Identifying miRNAs that are exclusive to the lissamphibian lineages could help in determining whether they are monophyletic or not and also to discriminate between the alternative hypotheses of lissamphibian relationships.

## 4.2 Material and Methods

In this section I will outline the methods I used for generating the RNA libraries that I then sequenced using Illumina technology to obtain the repertoire of miRNA genes for a representative of each of the three extant groups of Amphibia (Anura, Gymnophiona and Caudata). Further on, I will explain the analyses carried out with this data set and the results obtained.

### 4.2.1 Materials

| Reagents / Kit | Supplier |
|---|---|
| Tricaine methanesulfonate (MS222) | Sigma |
| TRIzol® | Life Technologies |
| Liquid nitrogen | Various suppliers |
| Chloroform | Sigma |
| Isopropanol | Sigma |
| DNA ladder & Loading dye | Invitrogen |
| RNase Free water | Sigma |
| Ethidium Bromide | Sigma |
| T4 RNA Ligase 2, Truncated | Sigma |
| 5X First Strand Buffer | Life Technologies |
| SuperScript II Reverse Transcriptase | Invitrogen |
| TruSeq kit | Illumina |
| Gel Breaker Tubes | IST Engineering Inc. |
| 5X Novex TBE Buffer | Life Technologies |
| 6% Novex TBE PAGE Gel | Life Technologies |
| 5 um filter tube | IST Engineering Inc. |
| Ethanol | Sigma |

### 4.2.2 Methods

#### 4.2.2.1 Sample preparation

In order to carry out this study, I obtained living samples of one representative of each of the extant lineages within the Amphibia. The choice of species was based on the availability of samples and data. For the Anura (frogs), I used the African clawed frog, *Xenopus tropicalis* (hereafter referred to as "Xenopus"), because it is the only amphibian species with a sequenced genome. For the Caudata (salamanders) I used the aquatic *Ambystoma mexicanum* (hereafter referred to as "Axolotl"), a common species found in pet shops. Finally, for the Gymnophiona (caecilians) I used the aquatic *Typhlonectes compressicauda* (hereafter referred to as "Caecilian"), which was donated by Mark Wilkinson (The Natural History Museum, London). All specimens were at a juvenile developmental stage.

The specimens were euthanized by letting them swim in a solution of MS222 and water. As soon as the specimens stopped moving, these were deep frozen in liquid nitrogen. Due to the size of the specimens, an initial tissue homogenization step was performed to obtain a full breakdown of the tissue. This was done using a pestle and mortar filled with liquid nitrogen where the sample was placed. By doing this, the tissues snap froze and these were gradually grounded with the pestle to obtain a fine powder, which can be further homogenized or stored at -15 °C.

#### 4.2.2.2 RNA extraction

For the RNA extraction we used the standard RNA extraction methods as outlined in the TRIzol® (Life Technologies, Thermo Fisher Scientific Inc.) reagent manual (http://tools.lifetechnologies.com/content/sfs/manuals/trizol_reagent.pdf). I used a very small amount of the ground specimens to carry out a further homogenization step with TRIzol® reagent measured as 1ml of TRIzol® per 5-100 mg of tissue. TRIzol® works by maintaining the RNA integrity during tissue homogenization, while at the same time disrupting and breaking down cells and cell components. The homogenized tissue was then incubated for five minutes at room temperature to allow complete diassociation of nucleoprotein complexes. Then added 0.2 ml of chloroform per 1 ml of TRIzol® (in this case, I used 5 ml of chloroform as I used 25 ml of TRIzol® per sample). Then the capped tubes were shaken for 15 seconds and left to stand for approximately 6-7 minutes at room temperature. Tubes were then centrifuged for 17 minutes at 8000 x *g* (although the protocol states that ideally should be 12000 x *g* if the equipment allows it). After centrifugation, the mixture separated into three phases: an aqueous phase at the top, an interphase in the middle, and a red phenol-chloroform phase. The top aqueous phase is the one that contained the RNA, so it was transferred into a new tube. Then 0.5 ml of Isopropanol per 1 ml of TRIzol® was added (12.5 ml per sample in my case). The samples were then mixed and left to rest at room temperature for 10 minutes, and then put through another centrifugation cycle at 8000 x *g* for 10 minutes at 4ºC. From this last step I obtained the RNA pellets in each tube. I poured off the Isopropanol and the pellet was put into centrifugation at 7500 x *g* for 5 minutes at 4ºC. I poured out the supernatant and left the pellets to dry for 10 mins in room temperature. The RNA pellets were resuspended in milliq (or RNase free water), and left for 1 or 2 hours in a fridge.

#### 4.2.2.3 Small RNA libraries generation

The sample preparation and RNA libraries generation steps were carried out under the supervision of Dr. James Tarver at the University of Bristol facilities. I followed the TruSeq Small RNA Sample Preparation protocol (available at http://supportres.illumina.com/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqsmallrna/truseq_smallrna_sampleprep_guide_15004197_e.pdf) for Illumina sequencing. In summary the steps involved were: (1) ligate the RNA 3' adapter; (2) ligate the RNA 5' adapter; (3) Reverse transcription – Polymerase Chain Reaction (RT-PCR) amplification; (4) gel purification of small RNA Library, as summarized in Figure 4.5.

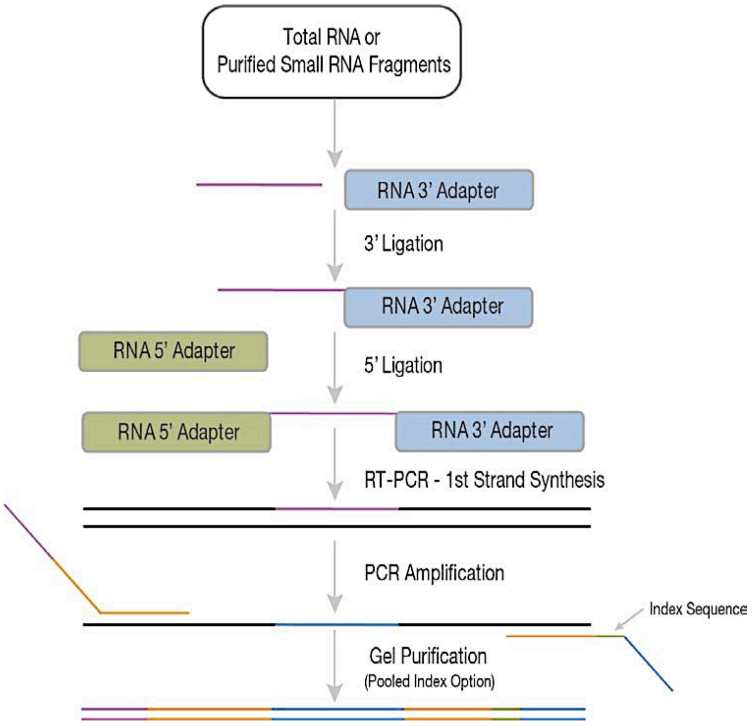


**Figure 4.5** Workflow of TruSeq Small RNA Sample preparation.
From Illumina Sequencing TruSeq Small RNA preparation guide.

The small RNA fractions were isolated with fluorescein-labeled DNA oligonucleotides equivalent to 21 and 27 nt in molecular weight and combined with 200 – 500 mg of total RNA and electrophoresed on a 15% urea-polyacrylamide gel. I carried out the sequential ligations (as explained in the TruSeq preparation guide) of the RNA 3' and 5' RNA adapters, which guide the excision of the 3' and 5' ends of the RNAs.  Then I carried out the Reverse Transcription (RT) and PCR Amplification, which created the complementary DNA (cDNA) constructs based on the small RNA ligated with 3' and 5' adapters. By doing this step, I selectively enriched those fragments that had the adapter molecules at both ends. The PCR step was performed with two primers that anneal to the ends of the adapters. These primers included a unique 4 nt barcode so that the source of the sequence could be identified after sequencing and the TruSeq kit Illumina primers. PCR amplification of the small RNA cDNA was performed with an initial denaturation at 98°C for 30 seconds (sec) followed by 11 cycles of 98°C (for 10 sec), 60°C (30 sec) and 72°C (15 sec). Then, the sample was given a final extension time of 10 min at 72°C, to be finally held indefinitely at 4°C.

The resultant amplified cDNA constructs from the PCR step were then purified using gel electrophoresis and a DNA loading dye to observe band migration (as explained in the TruSeq preparation guide). The gel was visualized in a UV transilluminator and only the fraction that corresponded to the 22 nt and 30 nt small RNA fragments were excised (these corresponded to the area between the 145 bp and 160 bp bands in the custom ladder, see Figure 4.6).

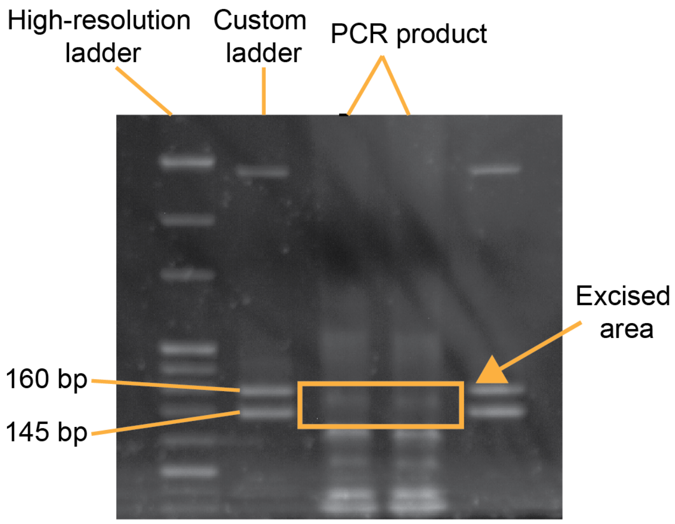


**Figure 4.6** An example of an agarose gel electrophoresis used to target miRNA regions.

This gel corresponded to the *Xenopus tropicalis* sample. Square in orange indicates excised area of the gel containing miRNA bands.

The excised gel fragment was placed in a gel breaker tube and centrifuged at 20,000 x *g* for 2 mins at room temperature. DNA concentrations were measured using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific Inc). Libraries with different barcodes were pooled and submitted to the Bristol Genomics and Proteomics Sequencing Facility for sequencing.

#### 4.2.2.4 Identifying miRNAs

A pre-processing of the raw Illumina sequenced data was carried out using next-generation sequence (NGS) Quality Control and manipulation tools in the online version of Galaxy (Blankenberg *et al.* 2010). The pre-processing involved first clipping the 5' and 3' adapter sequences using an 18 nt cut off on both ends of each sequence read and sorting reads by taxa using the specific barcode identifiers, which are afterwards clipped. The program enforces a cut off of 18 nt minimum length to retain sequence reads (meaning that all the reads that are smaller than 18 nt and that did not match the 5' and 3' adapters were removed). Reads whose average quality was below 20 were discarded. These sequences (originally in fastQ) were finally exported into fastA format, ready for the identification step.

The process of identifying miRNAs was done using the mirDeep2 algorithm (Friedländer *et al.* 2008) implemented in a pipeline in perl (Friedländer *et al.* 2012). The normal procedure usually involves using an indexed reference genome to which the RNA sequencing reads are mapped into. Then the miRDeep algorithm uses a probabilistic model of miRNA genesis to score the compatibility of the position and frequency of sequenced RNA based on the secondary structure of the miRNA precursor (Friedländer *et al.* 2008). Basically, after the sequencing reads are aligned into the genome, the algorithm excises the genomic DNA for aligned regions, and computes their secondary RNA structure. The plausible miRNA precursors are then scored for their likelihood to be real miRNA precursors, yielding a final list of novel and known miRNAs (Friedländer *et al.* 2008). The predicted miRNAs can then be further checked by eye to ensure that the secondary structure and the alignment of the reads and bases match those of a real miRNA, this was done following criteria proposed by Wheeler *et al.* (2009) and Tarver *et al.* (2013).

As stated before, for most of the current bioinformatics tools available, using a reference genome is essential in the process of miRNA prediction and identification. This is because these tools use algorithms that map the short reads into the reference genome. This works well in cases like that of Xenopus, where a reference genome is available. However, in the case of non-model organisms like the Axolotl and the Caecilian, identifying miRNAs is a challenge, because no reference genome is available. Some studies have suggested different alternatives like employing the increasingly available *de novo* transcriptome data (Hornett and Wheat 2012), or the genome of a closely related species (Etebari and Asgari 2014) as proxy for the genomic reference. Even though both studies found that the amount of predicted and identified miRNAs is lower than if using the species reference genome, they also agree that using *de novo* transcriptomes of related species (up to 100 million years divergence – Hornett and Wheat 2012) can still yield results almost as good as using the same species.  This is due to the great level of conservation observed in mature miRNA sequences. However, it is clear that bias and error in miRNA detection increase with the evolutionary distance (Hornett and Wheat 2012; Etebari and Asgari 2014). In my case, as only the *Xenopus tropicalis* genome (Hellsten *et al.* 2010) and the *Ambystoma mexicanum* transcriptome (Monaghan *et al.* 2009) were available, I was forced to use only these two sources of genomic information in Amphibia. Additionally, I considered the recently published genome of the coelacanth *Latimeria chalumnae* (Amemiya *et al.* 2013) in my analyses. The latter was used as an outgroup to investigate miRNA presences and absences.

#### 4.2.2.5 Shared miRNA analysis

In this study, the main objective was to find miRNAs that could help in inferring phylogenetic relationships. These could only be those that are novel (i.e. that arose in the Amphibia) and are not found outside of this lineage. miRNA that are found in all amphibian lineages could be used as characters supporting the monophyly of this group. miRNA that are found only in two out of three amphibian lineages can be used to investigate relationships within Amphibia. Given that the priority was to find phylogenetic informative characters, I used the mirDeep pipeline and algorithm to map the RNA data for all my samples against each of the reference genomes or transcriptome data (i.e. the Xenopus, Axolotl and the Coelacanth outgroup). The way this was done was by providing all three RNA data sets in to the mapper.pl script in mirDeep (using the –d option), which combined them and labeled each read according to the source data set. This way, any miRNAs that were predicted by miRDeep would have the information of all the reads that aligned against the genome assessed, inferring presence of that miRNA in those organisms. From all the predicted miRNAs, I then focused on those that could be shared between two lineages of the ingroup, (especially those that presented a star and a mature region and that had a total score higher or equal to zero). I counted them and then verified their secondary structure by eye. Finally, after narrowing down to those that are shared, a BLAST search (Altschul *et al.* 1990) was performed to identify miRNAs known in the miRBase database (www.mirbase.org).

## 4.3 Results

### 4.3.1 Identified shared miRNAs

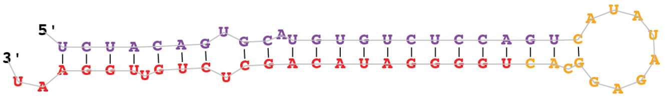
The miRDeep pipeline identified a total of 515 putative miRNAs across all three considered samples using the Xenopus genome as reference; 30 miRNAs across all three considered samples using the Axolotl transcriptome as reference; and 138 miRNAs across all three considered samples using the Coelacanth genome as reference. Of the putative miRNAs found when using the Xenopus genome as a reference, 175 were unique to the Xenopus, only 5 had an identifiable ortholog in Axolotl, 16 had an identifiable ortholog in the Caecilian, and 319 miRNAs were shared between all three lineages (Table 4.1). Using the Axolotl transcriptome as reference, among those identified, I found 29 miRNAs found only in the Axolotl sample and only 1 was shared across all three lineages. Finally, using the Coelacanth genome as reference, only 2 putative miRNAs were found that had a putative hit within Amphibia, but these were only present in the Caecilian. Table 4.1 summarises all the preliminary findings.

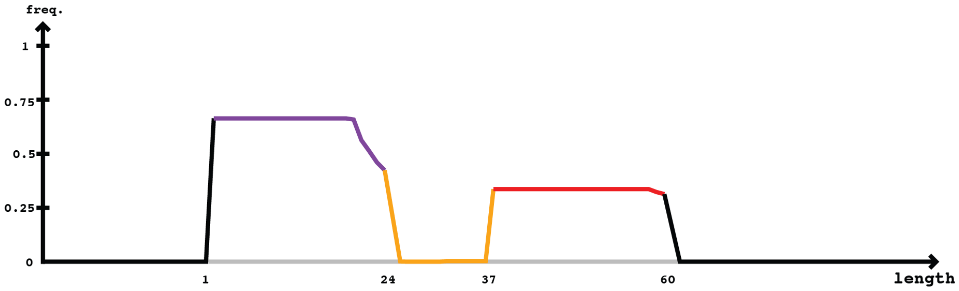
**Table 4.1** Summary of uniquely shared miRNAs from the sequenced RNA between pairs of the lineages of Lissamphibia and those found across all three.
These numbers are prior to comparisons against the miRBase database and other vertebrate genomes.

| | miRNAs predicted from RNA | | | | |
|---|---|---|---|---|---|
| Reference genome/transcriptome | Xenopus | Axolotl | Caecilian | Across all three | Total |
| Xenopus | 175 | 5 | 16 | 319 | 515 |
| Axolotl | 0 | 29 | 0 | 1 | 30 |
| Coelacanth | 0 | 0 | 2 | 136 | 138 |

A BLAST analysis of all the predicted miRNAs from Table 4.1 was carried out against miRBase (the database of known miRNAs) and determined that almost all of these were known (hence not exclusive to the Amphibia). Furthermore, the BLAST analysis showed that all the miRNAs identified as known were present across almost all vertebrates. Only 1 miRNA (Figure 4.7) found across all three lissamphibian lineages is potentially novel and amphibian-specific. In the BLAST search, it hit two similar miRNAs found outside the Amphibia (miR-139 and miR-4660) but the most similar (E-value=0.006) had two mismatches, including one in the first 5 positions where specificity is key to the binding efficiency of the miRNA (see Appendix D). Additionally, only 1 miRNA (Figure 4.8) shared between Xenopus and Axolotl was found to be novel, not hitting any of the known miRNAs. To further verify that these two miRNAs were not present in other vertebrate lineages, I performed a BLAST search of their mature sequence against the coelacanth, human, Anolis and pufferfish genomes, without finding any significant hits.

| | | |
|---|---|---|
| Provisional ID | : | scaffold_995_11919 |
| Score total | : | 377.8 |
| Score for star read(s) | : | 3.9 |
| Score for read counts | : | 370.8 |
| Score for mfe | : | 2.2 |
| Score for randfold | : | 1.6 |
| Score for cons. seed | : | -0.6 |
| Total read count | : | 739 |
| Mature read count | : | 247 |
| Loop read count | : | 0 |
| Star read count | : | 492 |





Star Mature

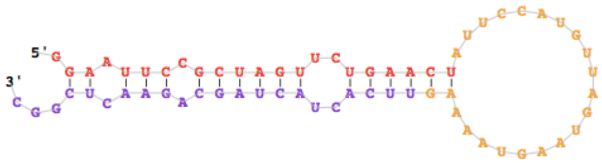
```
5'- ccggugaggccuggguguauucuacagugcaugugucuccagucauauagaggcacuggggauacagcucuguuggaauaacaaucagugccgcguucagcgcuuuauccacc  -3'   obs
    ccggugaggccuggguguauucuacagugcaugugucuccagucauauagaggcacuggggauacagcucuguuggaauaacaaucagugccgcguucagcgcuuuauccacc         exp

    ..((((.((((((..(((((((((((.((.(((((((((((((........).)))))))))))))).))).)))))).)))..))).)))(((.....)))......))))   reads     mm       sample
    ....................ucuacagugcaugugucuc..........................................................................   21        0        Xen
    ....................ucuacagugcaugugucucc.........................................................................   16        0        Xen
    ....................ucuacagugcaugugucucca........................................................................   18        0        Xen
    ....................ucuacagugcaugugucuccU........................................................................   3         1        Xen
    ....................ucuacagugcaugugucuccag.......................................................................   6         0        Xen
    ....................ucuacagugcauguUucuccagu......................................................................   1         1        Xen
    ....................ucuacagugcaugugucucAagu......................................................................   1         1        Xen
    ....................ucuacagugcaugugucucUagu......................................................................   1         1        Xen
    ....................ucuacagugcauguguGuccagu......................................................................   1         1        Xen
    ....................ucuacagugcaugugucuccaUu......................................................................   3         1        Xen
    ....................ucuacagugcaugugucuccagu......................................................................   205       0        Xen
    ....................ucuacagugcauguguUuccagu......................................................................   1         1        Xen
    ....................ucuacagugcaugugucuccagG......................................................................   1         1        Xen
    ....................ucuacagugcaGgugucuccagu......................................................................   1         1        Xen
    ....................ucuacagGgcaugugucuccagu......................................................................   1         1        Xen
    ....................ucuacagugcauguguAuccagu......................................................................   2         1        Xen
    ....................ucuacagugcaugugucuccaguU.....................................................................   2         1        Xen
    ..................................................aggcacuggggauacagcucuguuggaa...................................   2         0        Xen
    ........................................................uggggauacagcucuguugga....................................   10        0        Xen
    ........................................................uggggauacagcucuguuggaa...................................   5         0        Xen
    ........................................................uggggauacagcucugGuggaau..................................   1         1        Xen
    ........................................................uggggauacagcucuguuggaGu..................................   1         1        Xen
    ........................................................uggCgauacagcucuguuggaau..................................   1         1        Xen
    ........................................................Aggggauacagcucuguuggaau..................................   1         1        Xen
    ........................................................uggggauacagcucuguuggaau..................................   223       0        Xen
    ........................................................uggggauacagcucugCuggaau..................................   2         1        Xen
    ........................................................uggggauacagcucuguuggaaA..................................   1         1        Xen
    ........................................................uggggauacagcucuguuggaauU.................................   2         1        Xen

    ....................ucuacagugcaugugucu...........................................................................   1         0        Axo
    ....................ucuacagugcaugugucA...........................................................................   1         1        Axo
    ....................ucuacagugcaugugucuc..........................................................................   18        0        Axo
    ....................ucuacagugcaugugucucA.........................................................................   3         1        Axo
    ....................ucuacagugcaugugucucc.........................................................................   4         0        Axo
    ....................ucuacagugcaugugucuAca........................................................................   1         1        Axo
    ....................ucuacagugcaugugucucca........................................................................   9         0        Axo
    ....................ucuacagugcaugugucuccaA.......................................................................   1         1        Axo
    ....................ucuacagugcaugugucuccag.......................................................................   9         0        Axo
    ....................ucuacagugcaugugucuccagu......................................................................   28        0        Axo

    ....................ucuacagugcaugugucu...........................................................................   2         0        Cae
    ....................ucuacaAugcaugugucuc..........................................................................   1         1        Cae
    ....................ucuacagugcaugugucuc..........................................................................   30        0        Cae
    ....................ucuacagugcaugGgucuc..........................................................................   1         1        Cae
    ....................ucuacagugcaugugucucA.........................................................................   2         1        Cae
    ....................ucuacagugcaugugucucc.........................................................................   12        0        Cae
    ....................ucuacagugcaugugucucca........................................................................   8         0        Cae
    ....................ucuacagugcaugugucucccCg......................................................................   1         1        Cae
    ....................ucuacagugcaugugucuccaA.......................................................................   1         1        Cae
    ....................ucuacagugcaugugucuccag.......................................................................   9         0        Cae
    ....................ucuacagugcaugugucuccaAu......................................................................   1         1        Cae
    ....................ucuacagugcaugugucuccagu......................................................................   57        0        Cae
    ....................ucuacagugcaugugucuccagC......................................................................   1         1        Cae
    ....................ucuGcagugcaugugucuccagu......................................................................   1         1        Cae
    ....................ucuacagugcaugugucuccaUu......................................................................   1         1        Cae
    ....................ucuacagugcaugugucuccagA......................................................................   2         1        Cae
    ....................ucuaAagugcaugugucuccagu......................................................................   1         1        Cae
    ....................ucuacagugcaugugucuccaguU.....................................................................   1         1        Cae
    ....................ucuacagugcaugugucuccaguAa....................................................................   1         1        Cae
```

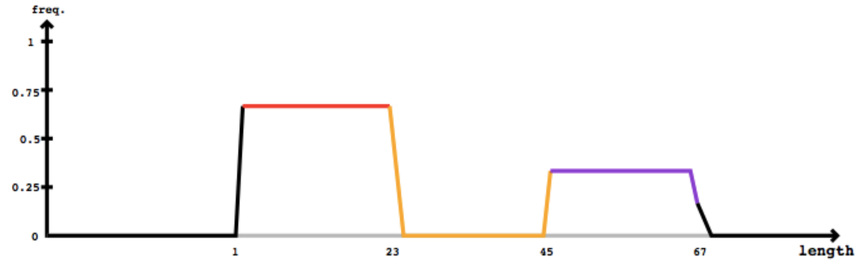
← (Previous page):

**Figure 4.7** Potential novel miRNA shared within the three lineages of Lissamphibia.

On the top left, a summary of scores used in miRDeep, and to the right a schematic representation of the inferred secondary structure of the hairpin with the mature region (in red) and the star region (in purple). The graph below shows the amount of reads that align to each section. At the bottom, the alignment of all the reads from each of the three datasets into the mature and star sections of the predicted and observed miRNA.

| | | |
|---|---|---|
| Provisional ID | : | scaffold_275_6764 |
| Score total | : | 1.7 |
| Score for star read(s) | : | 3.9 |
| Score for read counts | : | 0.1 |
| Score for mfe | : | 0.5 |
| Score for randfold | : | -2.2 |
| Score for cons. seed | : | -0.6 |
| Total read count | : | 12 |
| Mature read count | : | 8 |
| Loop read count | : | 0 |
| Star read count | : | 4 |





```
                          Mature                                                 Star
5'- ucuccuguuccugugggugcuugucaggaauuccgcuaguucugaacuauuccauguuaguaaguaaaaguucacuacuagcagaacucggccacguaccacacuugcugc  -3'   obs
    ucuccuguuccugugggugcuugucaggaauuccgcuaguucugaacuauuccauguuaguaaguaaaaguucacuacuagcagaacucggccacguaccacacuugcugc        exp
    ......((....((((((((.((.(..((.(((.((((((..((((((....................))))))..)))))).))).)).).)).))))).)))..))...    reads      mm        sample
    ..........................ggaGuuccgcuaguucugaacu...............................................................    4          1         Xen
    ......................................................................uucacuacuagcagaacucggA...................    2          1         Xen

    ..........................ggaGuuccgcuaguucugaacu...............................................................    4          1         Axo
    ......................................................................Cucacuacuagcagaacucgg....................    2          1         Axo
```

**Figure 4.8** Potential novel miRNA shared only between the Anura and Caudata (Batrachia).

On the top left, a summary of scores used in miRDeep, and to the right a schematic representation of the inferred secondary structure of the hairpin with the mature region (in red) and the star region (in purple). The graph below shows the amount of reads that align to each section. At the bottom, the alignment of all the reads from each of the three datasets into the mature and star sections of the predicted and observed miRNA.

## 4.4 Discussion

### 4.4.1 Batrachia hypothesis and the monophyly of Amphibia.

Finding a valid novel miRNA that is shared between Xenopus and the Axolotl provide further support to the Batrachia hypothesis (Figure 4.9 a), however, it is clear that only one novel marker does not indicate a significant result. In any case, it is clear that Batrachia is the most parsimonious solution as it suggests that this miRNA was acquired only once in the lineage leading to the Caudata and the Anura (under the Batrachia hypothesis, Figure 4.9 b), instead of being acquired and then lost (under the Procera hypothesis). The fact that this result agrees with several studies that arrived at the same conclusion using independent lines of evidence: paleontological data (Milner 1988; Trueb and Cloutier 1991; Ruta and Coates 2007; Marjanović and Laurin 2008, 2009; Sigurdsen and Green 2011; Maddin *et al.* 2012); recent molecular data (Zardoya and Meyer 2001; San Mauro *et al.* 2005; Zhang *et al.* 2005; Roelants *et al.* 2007; San Mauro 2010 and the combined analysis of Pyron 2011; Shen *et al.* 2013), however suggest that there is a substantial congruence of data that are starting to be accumulated and that all seem to agree.

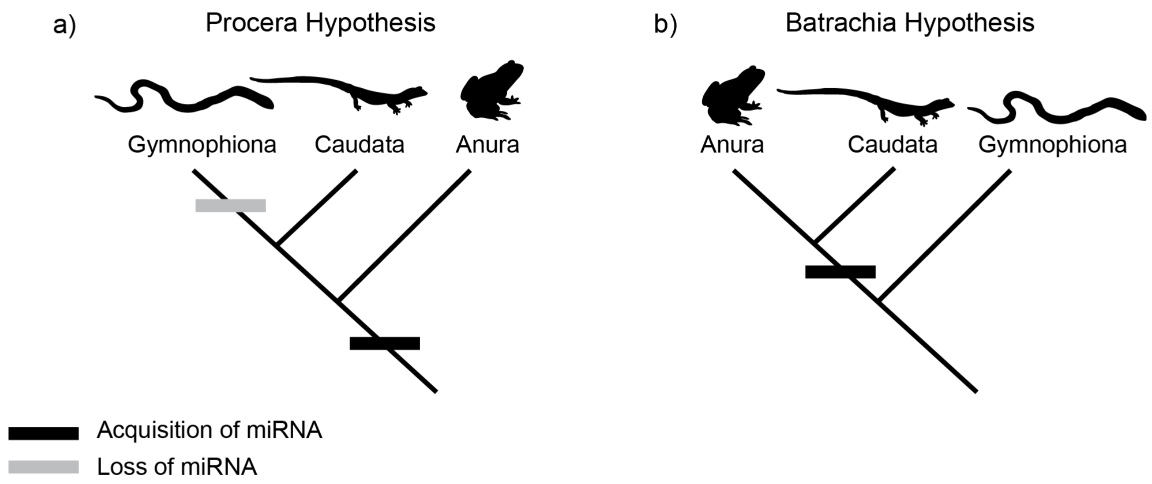


**Figure 4.9** Character reconstruction of miRNA "scaffold_275_6764"
(a) under the Procera hypothesis and (b) the Batrachia hypothesis. The most parsimonious reconstruction supports the Batrachia hypothesis (1 acquisition).

Additionally, my results by supporting the Batrachia hypothesis, are in direct disagreement with those of Fong *et al.* (2012) that on the contrary found Gymnophiona and Caudata to form a clade to the exclusion of Anura. I could not find any evidence for this specific result in my data. The monophyly of the Lissamphibia is further supported by the presence of one novel miRNA found across all three lineages (Figure 4.10 b), which argues against the paraphyletic hypothesis of Fong *et al.* (2012) given that it is the least parsimonious (Figure 4.10 a). However, also in this case, the result can hardly be considered significant. Yet our results suggest that perhaps further reanalyses of the Fong *et al.* (2012) data should be carried out. Unfortunately I did not have time to pursue this avenue of research as part of my PhD.

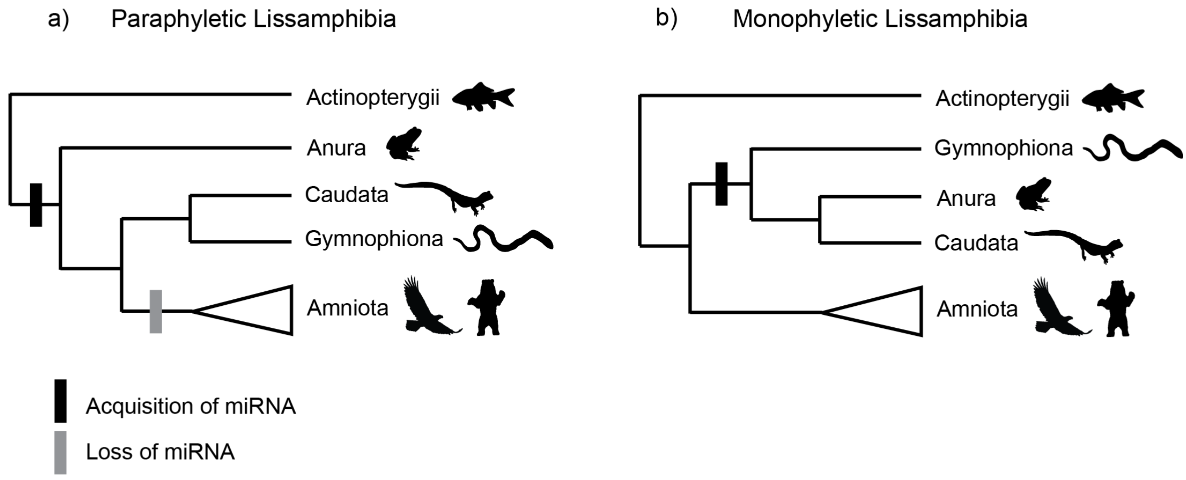


**Figure 4.10** Character reconstruction of the miRNA "scaffold_995_11919"
(a) under the paraphyletic Lissamphibia hypothesis and (b) the monophyletic Lissamphibia hypothesis. The most parsimonious reconstruction supports the monophyletic Lissamphibia hypothesis (1 acquisition).

In terms of biogeography, the Batrachia hypothesis poses a challenge to understand the historical distributions of the lissamphibian lineages. Previous studies (San Mauro *et al.* 2005; Roelants *et al.* 2007; San Mauro 2011; Zhang and Wake 2009) proposed that the divergence between Batrachia and Gymnophiona dates from 368 - 294 mya, placing the divergence of these clades well before the break up of Pangaea (early-middle Jurassic).

### 4.4.2 Caveats of using miRNA data in phylogenies

There are two types of caveats when identifying miRNAs. One is the type and quality of the sequenced sample. Some miRNAs may be specific to developmental stages or to tissue types. We addressed the tissue type problem by isolating the RNA from the entire animal for each of the study organisms, and by obtaining high quality samples. However, since I only sampled the organisms in the juvenile stage, it is likely that some (probably a small number) of miRNAs that are only expressed at different developmental stages were missed. During the analysis of shared miRNAs, some putative miRNAs were found in Caecilian and Axolotl (which were also found outside the Amphibia) that were not found to be expressed in the Xenopus. However, when comparing these mature sequences against the genome of Xenopus, hits could be found indicating that these microRNAs are present in Xenopus but were not expressed at the specific stage at which I sampled this taxon. The same problem could be implied have happened in the cases of Axolotl and Caecilian samples, but unfortunately, the lack of a genome reference for this taxa, implies that certainty cannot be reached with reference to these species.

The other main caveat is lacking the reference genomes of the studied organisms. Even though some studies state that it is probably fine to use alternative reference data (such as the transcriptome or the genome of a related species - Hornett and Wheat 2012; Etebari and Asgari 2014) the present study shows that for those samples the numbers of inferred miRNAs are considerably lower (i.e. 515 miRNAs identified using the Xenopus genome versus 30 miRNAs identified using the Axolotl transcriptome). The underestimation of miRNAs in Axolotl and Caecilian data might add a bias in the analysis that will tend to favour either of the Axolotl or Caecilian to be related to the Xenopus. In this sense, it could be expected that maybe the miRNA that was only present in Axolotl and Xenopus, could be as well in the Caecilian genome (but was not captured in the miRNA library). This could only be tested with a Caecilian genome available. Therefore, until proved otherwise, the evidence of this miRNA being present only in Axolotl and Xenopus holds true.

Even though the results from this work yielded a very small number of miRNAs that could be informative of the relationships among the considered taxa, it is worth pointing out that it is not uncommon to find studies that have used very few miRNAs to support clades. Some examples include: the work of Campbell *et al.* (2011) which explores the phylogenetic position of the Tardigrada and Onychophora with respect to the Arthropods using miRNA and EST data and where they found only one miRNA family (miR-305) supporting the position of Onychophora as the sister group of the arthropods, and one miRNA family (miR-276) to support the position of the Tardigrada as a sister group of the latter clade. Similarly, the work of Rota-Stabelli *et al.* (2011) which also used miRNA and EST data, only found one novel miRNA to support the monophyly of Chelicerata (which they called "Arthropod-Novel-1"), one novel miRNA to support the monophyly of Myriapoda (Arthropod-Novel-2") and one known miRNA family that only occurred within the Pancrustacea (miR-286); and the study of Philippe *et al.* (2011a) where one miRNA family (miR-103) supports the monophyly of the Deuterostomes. However, the mentioned studies combined this evidence with that of EST data and based their inference on the consilience of these lines of evidence rather than solely on miRNA data. Indeed, the only work that used few miRNA in isolation was that of Lyson *et al.* (2012) where one miRNA family was found to support the monophyly of the archosaur clade (miR-1791), and one to support the monophyly of the reptiles, including birds (miR-1677), but it is my opinion that the practice of using few miRNA in isolation (as done by Lyson and co-workers) should be stigmatized as bad practice.

In any case, having found one potentially novel miRNA family unique to the Amphibians and one potentially novel miRNA family unique to the Batrachia (until further evidence of presence in the caecilians is found), still presents a very optimistic result that shows that miRNAs can be useful as independent line of evidence in phylogenetics, and that when more genomic data becomes available, it will help resolving this dilemma. As for my work, I intend to reanalyze the data of Fong *et al.* (2012) in order to investigate whether they had errors that might have induced the topology they presented.

## 4.5 Conclusions

miRNAs have the potential to be very useful in resolving difficult phylogenetic problems. In the present study, I addressed the historically controversial phylogenetic relationships of the three extant lineages of the Amphibia: Anura, Caudata and Gymnophiona by using high-throughput sequencing targeted to identify miRNAs. For the purposes of resolving this question, only the shared miRNA families unique to the Lissamphibia were going to be phylogenetically informative, hence, the aim was to identify potential novel miRNAs unique to the three orders of Lissamphibia. The analyses carried out showed a very high number of shared miRNAs discovered using the *Xenopus tropicalis* genome, which contrasted with a lower number of miRNAs discovered using the Axolotl transcriptome, suggesting that not using genomic data is not ideal to validate miRNAs. Nevertheless, in spite of the limitations encountered (lack of reference genome for Axolotl and Caecilian), I was able to find two potential novel miRNAs, one that supports the monophyly of the Lissamphibia, and another that supports the Batrachia hypothesis. These findings show that even though the number of miRNAs supporting the monophyly of Lissamphibia and Batrachia was indeed small, there is potential to get better results and finally resolve the phylogenetic problems of the three orders of Amphibia when more reference genomes become available.

# Chapter 5. General discussion and conclusions

During this thesis I addressed some problems associated with large-scale phylogenetic analyses by tackling issues related to missing data and careful handling and addition of novel data in large-scale reconstructions, presenting an application of this approach in the context of amphibian phylogenetics. To do so, I explored the phylogenetic placement of a newly sequenced taxon in the Amphibian tree of life using a previously published data set as backbone. Furthermore, I developed a method for identifying rogue taxa. Given that the safe removal of taxa that are rogue because they are missing too many characters can potentially reduce the dimension of terraces in tree space (Sanderson *et al.* 2011), a method to identify and remove such taxa has the advantage of improving tree searches, increasing the likelihood of finding the optimal tree(s), and in the case that multiple optimal trees exist, improving the resolution in the final consensus tree.  Finally, I investigated the evolutionary relationships of the three lineages of the extant amphibians (Anura, Caudata and Gymnophiona), using an independent source of evidence: miRNAs, which have been recently used to help resolve difficult phylogenetic problems. In the following sections I will address the implications of my findings and how this can potentially help in our understanding and current uses in large-scale phylogenetics.

## 5.1 Large-scale phylogenies and the road to phylogenomics

It is now common to see an increasing number of studies generating ever larger data sets (i.e. genomes and transcriptomes) to resolve problematic relationships. Given the rate at which such studies are increasing, it is clear that most researchers are convinced that "more data" will ultimately allow the resolution of all problems in phylogenetics. However, more recent large-scale studies have shown that by including more data, the problems actually become far more complex and difficult to handle (Philippe *et al.* 2011b and works referred therein, also some recent examples include Wiens *et al.* 2012; Pyron *et al.* 2013). More broadly, one wonders: do we really need all these large-scale genome-size data sets to resolve phylogenies?

First, we should probably start by asking why do we concatenate genes into a massive superalignment. In an ideal situation, where we assume an ortholog gene is used to build a phylogeny, we can infer that the phylogenetic information from the sites of that gene will nicely arrive to an agreement of a phylogenetic hypothesis. However, using simply one gene is simply not enough because of a variety of reasons. First, depending on the gene evolutionary rate it is possible that the considered marker can only resolve a phylogenetic tree up to a certain level. Furthermore, it is not known *a priori* whether the signal of one gene will agree or not with that of other genes. Therefore, studies started to sequence and concatenate into a single alignment more genes with different evolutionary rates, in the hope that this will help to resolve all areas of a phylogenetic tree. Additionally, it was hoped that by using many genes a genomic consensus could be reached. Of great relevance was a paper published by Rokas *et al.* (2003) that suggested that the inclusion of more genes would have almost "magically" overruled incongruent signals generated by misleading sites in some or many markers. This ideal scenario (and the fact that sequencing technologies have become more affordable and available) drove researchers to go down the "phylogenomics" alley. However, as pointed out by Philippe *et al.* (2011b) the reality is much more complex than that: "more genes do not necessarily mean more resolution". Adding more genes, has the effect of making the problem even bigger and this can result in non-phylogenetic signal becoming dominant and produce incorrect, yet statistically highly supported phylogenomic trees. As pointed out by Philippe and co-workers, the negative effect of non-phylogenetic signal in large-scale data sets is a consequence of incorrect identification of orthologs, erroneous alignments, or the incorrect reconstruction of multiple site substitutions (Philippe *et al.* 2011b). Additionally, this is further exacerbated by the increment in the amount of missing data. Hence, it is not about including all the data that one can obtain (although such an approach has been defended in some studies, such as Wiens 2003; Wiens 2006; Wiens and Morrill 2011). What is important is to use better methods and greater curation of the data.

Realistically, only a few sites in every locus might be truly informative, that is, if their substitution rates are adequately modeled, they change, but change slow enough to trace back a divergence event. In chapter 3 I have shown that careful analysis and management of the data (which is nowhere near a genome-size data set) taking into account the issues described above can probably be enough to resolve some of these problems. But that prompts us to ask if there might be a "tipping point" of phylogenetic information, and whether it is possible to estimate the amount of data that will be enough to correctly resolve the relationships in a phylogeny given the correct models of evolution. An interesting idea was proposed by Sanderson *et al.* (2010), fractional decisiveness, which is an index tied to the impact of missing data in tree construction, and that allows to make estimations of the amount of loci needed to reconstruct a unique tree on all taxa irrespective of what the tree is.

Maybe this should be the starting point before beginning a phylogenetic analysis. Perhaps for many unresolved phylogenetic problems we have already met this minimum requirement, but not enough effort has been made to identify problems and fix them. While the pace at which genetic information from model organisms will continue to increase, there is always going to be a lack of data for less well-studied species. It is for these organisms that these methods will be most important.

## 5.2 Pragmatic solutions to unequal sampling and the continuous increase of molecular data

### 5.2.1 Targeting problematic areas in a tree

For non-model organisms it is often the case that only relatively small amount of data are available (e.g. no genomes or transcriptomes) than for more studied organisms. Hence, when combining data from these unequally sampled taxa, concatenated matrices will have missing entries (caused by the missing loci). Including missing data has been widely discussed in several studies, where some regard the negative impact of it optimistically (Wiens 2003; Philippe *et al.* 2005; Wiens and Morrill 2011) to cautiously (Philippe *et al.* 2004; Lemmon *et al.* 2009; Roure *et al.* 2013). Regardless of the view, it is perhaps more useful to employ more clever strategies for sampling markers or taxa before carrying out phylogenetic analyses in cases of unequal sampling. One of the aspects I investigated during this thesis was identifying rogue taxa in morphological and phylogenomic data, where I developed a new method as an expansion of the *a priori* Safe Taxonomic Reduction method (Wilkinson 1995). One of the most interesting features of this method is that it allows recognising rogue taxa due to missing data and distinguishes it from unstable taxa originated by conflicting signal. Most *a posteriori* methods (leaf stability, consensus methods) are not able to detect the cause of instability. This method will be very useful in phylogenetic studies dealing with paleontological data (which usually suffers from many missing characters), and also with genomic data not evenly sampled. In the case of phylogenomics, the approach uses matrix representations (the same used in the matrix representation with parsimony supertree method) in order to generate presence/absence matrices that are then analysed to identify taxa that are unstable because of missing data. In this context, removal of missing data will result in the generation of "decisive" matrices from non-decisive ones. Another potential use of the approach I designed is the targeted identification of loci for sequencing that could help resolve the tree efficiently. A couple of studies have recently advocated this type of approach but used *a posteriori* approaches (e.g. a variation of maximum agreement subtrees -Sanderson *et al.* 2011). My approach, being *a priori*, has the potential to detect the most unstable taxa at an early stage in the development of a project and can allow for maximal efficiency by targeting them for sequencing, instead of trying to obtain sequence data for everything that is missing.

### 5.2.2 Building up on previous knowledge

Given that a lot of data (alignments and trees) is available online, it seems reasonable to make use of it. In the future, when even larger datasets will be put together, and hopefully these data sets will be properly curated, it will not be necessary to carry out a full analysis from the start (a *de novo* inference). In chapter 3 I explored ways to add novel sequences using a backbone alignment (with and without a *de novo* inference) to find the phylogenetic placement of a newly sequenced taxon. I found that using either a backbone phylogeny and a backbone alignment was equally as good (at least for the case we assessed), since the hypotheses obtained could not be discerned from each other. However, taking into consideration all the errors that can be added when using fixed topologies, at this stage, it is still probably preferable to carry out a *de novo* inference to ensure control of such biases. Also, the use of a backbone phylogeny would only be advisable in cases where the relationships are reasonably well supported and accepted. Investigating further on this type of approach could be very useful in the long term, when the relationships of taxa become better established, but current experiments (e.g. that of the "open tree of life" initiative – http://blog.opentreeoflife.org/), are still quite likely to generate phylogenies of constantly increasing (snowballing) error level.

## 5.3 Amphibian phylogenetics

There is a need to reassess a large-scale phylogeny of Amphibians. The past efforts (Frost *et al.* 2006; Pyron and Wiens 2011) represent excellent starting points from where a massive and important production and gathering of data has been carried out. However, future studies should improve data curation and also generate data. For instance, assessing that the signal retrieved from the sites included is not random. Also, unjustified *a priori* exclusion of available markers, as in the case of Pyron and Wiens' (2011) *a priori* exclusion of 28S sequences should be avoided. In terms of methodology, using more effective and accurate software to compute better-fitting substitution models (e.g. CAT-based model) will probably prove fundamental.

An important aspect to remember when thinking about groups like the amphibians, is that a very small proportion of taxa have been thoroughly sampled (in terms of both genes and taxon coverage). The two large-scale phylogenies mentioned before (Frost *et al.* 2006; Pyron and Wiens 2011) are characterized by large amounts of missing entries (77% and 80% respectively). Even though large efforts were done in both studies (particularly in the Frost *et al.* 2006 work) to try to fill in spaces by sequencing, the majority of the data in both comes from independent, focused, small-scale phylogenetic studies. These studies only sample a very few amount of loci in few taxa, and not always the sampled genes coincide, which generates the poor overlap for the considered taxa (and huge proportion of missing data). Perhaps the most complete data sets at the moment are those that have carried out phylogenies using mitochondrial genomes (e.g. San Mauro *et al.* 2004; Zhang *et al.* 2013). However these data are only available for a few representative species, and if concatenated into a data set including a large proportion of amphibian species, they will only further exacerbate the missing data problem.

Finally, the phylogenetic relationships among the three main extant lineages of Amphibia (Anura, Caudata and Gymnophiona) were addressed in chapter 4 using independent evidence (miRNAs). Based on the results I obtained, the monophyly of the Amphibia and of Batrachia (Anura+Caudata) were supported. Unfortunately, due to the limitations of the reference data (i.e. the fact that no genome is available for Caudata and Gymnophiona), the number of miRNAs identified was lower than expected. Nonetheless, my results suggest that there is a high potential for the use of miRNAs as an independent source of genomic evidence in amphibian phylogenetics. In terms of the implications it has in our understanding of the relationships among the basal lineages of amphibians, it would be important to repeat an analysis of the controversial EST data set of Fong *et al.* (2012). Unfortunately I did not have time to do this. Perhaps another way to address the problem of the relationships of the three extant orders of Amphibia (among each other and within the context of the phylogenetic relationships of the Tetrapoda) would be by combining the morphological data of the extinct lineages of Amphibia along with the molecular data of the representatives of the extant ones. Pyron (2011) attempted to do this but his results were highly biased towards the results obtained using only the RAG1 gene. A separate work by San Mauro (2010) carried out a multiple locus phylogeny and molecular clock analysis, nevertheless in this last work fossil taxa were not included.

# Chapter 6. Future work

I hope that with the work presented in this thesis, I have contributed to overcome issues in large-scale (both in terms of taxonomic and gene sampling) phylogenetics, in amphibian phylogenetics and theoretical phylogenetics. However, the results I obtained have only opened the door to more questions and ideas. Follow up research could include building up from the method I developed in chapter 2 to identify key loci that could help resolve phylogenetic trees, perhaps using Maximum Agreement Subtrees as proposed recently by Sanderson *et al.* (2011). Another could be building a large-scale phylogeny of amphibians using a combination of methods (mixed models in a superalignment or single-genes combined with supertree approaches) that will minimize impact of missing data, whilst achieving an optimal use of the phylogenetic information. In large-scale phylogenetics, another important aspect to evaluate could be the impact of taxon sampling in model selection when concatenating large matrices. Finally, another interesting aspect to follow would be that of the evolutionary relationships of the extant linages of Lissamphibia.  This could be done including genomic data for many species and fossil data into a single analysis.

# Chapter 7. Bibliography

Aberer A.J., Krompass D., Stamatakis A. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Systematic Biology 62:162-166.

Aberer A.J., Stamatakis A. 2011. A simple and accurate method for rogue taxon identification. IEEE International Conference on Bioinformatics and Biomedicine; Atlanta (GA), IEEE, p. 118-122

Akaike H. 1974. A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19:716-723.

Akanni W.A. 2014. Developing and applying supertree methods in Phylogenomics and Macroevolution. PhD Thesis. Biology Department. Maynooth, National University of Ireland, Maynooth.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403-410.

Amemiya C.T., Alföldi J., Lee A.P., Fan S., Philippe H., MacCallum I., Braasch I., Manousaki T., Schneider I., Rohner N. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature 496:311-316.

AmphibiaWeb. 2014. AmphibiaWeb: Information on amphibian biology and conservation. Available at http://amphibiaweb.org/.

Anderson J.S., Reisz R.R., Scott D., Fröbisch N.B., Sumida S.S. 2008. A stem batrachian from the Early Permian of Texas and the origin of frogs and salamanders. Nature 453:515-518.

Anquetin J. 2012. Reassessment of the phylogenetic interrelationships of basal turtles (Testudinata). Journal of Systematic Palaeontology 10:3-45.

Barej M.F., Rödel M.-O., Loader S.P., Menegon M., Gonwouo N.L., Penner J., Gvoždík V., Günther R., Bell R.C., Nagel P. 2014. Light shines through the spindrift–Phylogeny of African torrent frogs (Amphibia, Anura, Petropedetidae). Molecular Phylogenetics and Evolution 71:261-273.

Bartel D.P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281-297.

Bartel D.P. 2009. MicroRNAs: target recognition and regulatory functions. Cell 136:215-233.

Baum B. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3-10.

Baum B., Ragan M. 2004. The MRP Method. In: Bininda-Emonds O.R. editor. Phylogenetic Supertrees. Netherlands: Springer. p. 17-34.

Berezikov E. 2011. Evolution of microRNA diversity and regulation in animals. Nature Reviews Genetics 12:846-860.

Biju S., Bossuyt F. 2003. New frog family from India reveals an ancient biogeographical link with the Seychelles. Nature 425:711-714.

Bininda-Emonds O.R., Gittleman J.L., Steel M.A. 2002. The (super) tree of life: procedures, problems, and prospects. Annual Review of Ecology and Systematics 33:265-289.

Blackburn D., Wake D. 2011. Class Amphibia Gray, 1825. Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. Zootaxa 3148:39-55.

Blankenberg D., Kuster G.V., Coraor N., Ananda G., Lazarus R., Mangan M., Nekrutenko A., Taylor J. 2010. Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. Current Protocols in Molecular Biology:19.10. 11-19.10. 21.

Bryant D. 2003. A classification of consensus methods for phylogenetics. DIMACS Series in Discrete Mathematics and Theoretical Computer Science 61:163-184.

Campbell L.I., Rota-Stabelli O., Edgecombe G.D., Marchioro T., Longhorn S.J., Telford M.J., Philippe H., Rebecchi L., Peterson K.J., Pisani D. 2011. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. Proceedings of the National Academy of Sciences 108:15920-15924.

Cannatella D.C., Vieites D.R., Zhang P., Wake M.H., Wake D.B. 2009. Amphibians (Lissamphibia). In: Hedges S.B., Kumar S. editors. The timetree of life. Oxford: Oxford University Press. p. 353-356.

Cardillo M., Bininda-Emonds R., Boakes E., Purvis A. 2004. A species-level phylogenetic supertree of marsupials. Journal of Zoology 264:11-31.

Carroll R.L. 2007. The Palaeozoic ancestry of salamanders, frogs and caecilians. Zoological Journal of the Linnean Society 150:1-140.

Christodoulou F., Raible F., Tomer R., Simakov O., Trachana K., Klaus S., Snyman H., Hannon G.J., Bork P., Arendt D. 2010. Ancient animal microRNAs and the evolution of tissue identity. Nature 463:1084-1088.

Cranston K.A., Rannala B. 2007. Summarizing a posterior distribution of trees using agreement subtrees. Systematic Biology 56:578-590.

Creevey C.J. 2002. Algorithms for simulating and detecting adaptive evolution and reconstructing supertrees from genomic data. PhD Thesis. Department of Biology. Maynooth, Ireland, National University of Ireland, Maynooth. 165pp.

Creevey C.J., McInerney J.O. 2005. Clann: investigating phylogenetic information through supertree analyses. Bioinformatics 21:390-392.

Darst C.R., Cannatella D.C. 2004. Novel relationships among hyloid frogs inferred from 12S and 16S mitochondrial DNA sequences. Molecular Phylogenetics and Evolution 31:462-475.

Driskell A.C., Ané C., Burleigh J.G., McMahon M.M., O'Meara B.C., Sanderson M.J. 2004. Prospects for building the tree of life from large sequence databases. Science 306:1172-1174.

Dubois A. 1992. Notes sur la classification des Ranidae (Amphibiens Anoures). Bulletin mensuel de la société linnéenne de Lyon 61:305-352.

Dubois A. 2005. Amphibia Mundi. 1.1. An ergotaxonomy of recent amphibians. Alytes 23:1-24.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

Edwards A.W., Cavalli-Sforza L. 1963. The reconstruction of evolution. Heredity 18:553.

Edwards A.W., Cavalli-Sforza L. 1964. Reconstruction of Evolutionary Trees. In: Heywood V., McNeill J. editors. Phenetic and Phylogenetic Classification. London: Systematics Association.

Estabrook G.F., Strauch J.G., Fiala K.L. 1977. An application of compatibility analysis to the Blackiths' data on orthopteroid insects. Systematic Biology 26:269-276.

Etebari K., Asgari S. 2014. Accuracy of MicroRNA Discovery Pipelines in Non-Model Organisms Using Closely Related Species Genomes. PLoS ONE 9:e84747.

Feller A.E., Hedges S.B. 1998. Molecular evidence for the early history of living amphibians. Molecular Phylogenetics and Evolution 9:509-516.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Biology 27:401-410.

Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the Bootstrap. Evolution 39:783-791.

Felsenstein J. 1995. Phylogenetic Inference Package (PHYLIP). Available at http://evolution.genetics.washington.edu/phylip.html.

Felsenstein J. 2004. Inferring Phylogenies. eds. Sunderland, Massachusetts, Sinauer Associates pp.

Feng D.-F., Doolittle R.F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. Journal of Molecular Evolution 25:351-360.

Feng D.-F., Doolittle R.F. 1990. Progressive alignment and phylogenetic tree construction of protein sequences. Methods in Enzymology 183:375-387.

Fitch W.M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Systematic Biology 20:406-416.

Fitch W.M. 2000. Homology: a personal view on some of the problems. Trends in Genetics 16:227-231.

Fong J.J., Brown J.M., Fujita M.K., Boussau B. 2012. A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic Lissamphibia. PLoS ONE 7:e48990.

Fourment M., Gibbs M.J. 2006. PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. BMC Evolutionary Biology 6:1.

Friedländer M.R., Chen W., Adamidi C., Maaskola J., Einspanier R., Knespel S., Rajewsky N. 2008. Discovering microRNAs from deep sequencing data using miRDeep. Nature Biotechnology 26:407-415.

Friedländer M.R., Mackowiak S.D., Li N., Chen W., Rajewsky N. 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Research 40:37-52.

Frost D.R. 2013. Amphibian Species of the World: an Online Reference. Available at http://research.amnh.org/herpetology/amphibia/index.php.

Frost D.R., Grant T., Faivovich J., Bain R.H., Haas A., Haddad C.F., De Sá R.O., Channing A., Wilkinson M., Donnellan S.C. 2006. The amphibian tree of life. Bulletin of the American Museum of Natural History:1-291.

Gauthier J.A. 1986. Saurischian monophyly and the origin of birds. Memoirs of the California Academy of Sciences 8:1-47.

Goldman N. 1993. Statistical tests of models of DNA substitution. Journal of Molecular Evolution 36:182-198.

Gower D.J., Aberra R.K., Schwaller S., Largen M.J., Collen B., Spawls S., Menegon M., Zimkus B.M., de Sá R., Mengistu A., et al. 2013. Long-term data for endemic frog genera reveal potential conservation crisis in the Bale Mountains, Ethiopia. Oryx 47:56-59.

Gower D.J., Doherty-Bone T.M., Aberra R.K., Mengistu A., Schwaller S., Menegon M., de Sá R., Saber S.A., Cunningham A.A., Loader S.P. 2012. High prevalence of the amphibian chytrid fungus (Batrachochytrium dendrobatidis) across multiple taxa and localities in the highlands of Ethiopia. Herpetological Journal 22:225-233.

Graf J. 2012. A new Early Cretaceous coelacanth from Texas. Historical Biology 24:441-452.

Grandison A.G. 1978. The occurrence of Nectophrynoides (Anura: Bufonidae) in Ethiopia. A new concept of the genus with a description of a new species. Monitore Zoologico Italiano (N.S) Suppl. 11:119-172.

Grimson A., Farh K.K.-H., Johnston W.K., Garrett-Engele P., Lim L.P., Bartel D.P. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Molecular Cell 27:91-105.

Grimson A., Srivastava M., Fahey B., Woodcroft B.J., Chiang H.R., King N., Degnan B.M., Rokhsar D.S., Bartel D.P. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature 455:1193-1197.

Hasegawa M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22:160-174.

Hastings W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109.

Hedges S.B., Maxson L.R. 1993. A molecular perspective on lissamphibian phylogeny. Herpetological Monographs 7:27-42.

Hedges S.B., Moberg K.D., Maxson L.R. 1990. Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences and a review of the evidence for amniote relationships. Molecular Biology and Evolution 7:607-633.

Heimberg A.M., Cowper-Sal R., Sémon M., Donoghue P.C., Peterson K.J. 2010. microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. Proceedings of the National Academy of Sciences 107:19379-19383.

Heimberg A.M., Sempere L.F., Moy V.N., Donoghue P.C., Peterson K.J. 2008. MicroRNAs and the advent of vertebrate morphological complexity. Proceedings of the National Academy of Sciences 105:2946-2950.

Hejnol A., Obst M., Stamatakis A., Ott M., Rouse G.W., Edgecombe G.D., Martinez P., Baguñà J., Bailly X., Jondelius U. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proceedings of the Royal Society B: Biological Sciences 276:4261-4270.

Hellsten U., Harland R.M., Gilchrist M.J., Hendrix D., Jurka J., Kapitonov V., Ovcharenko I., Putnam N.H., Shu S., Taher L. 2010. The genome of the Western clawed frog *Xenopus tropicalis*. Science 328:633-636.

Hennig W. 1966. Phylogenetic systematics. eds. Urbana, University of Illinois Press.

Hertel J., Lindemeyer M., Missal K., Fried C., Tanzer A., Flamm C., Hofacker I.L., Stadler P.F. 2006. The expansion of the metazoan microRNA repertoire. BMC Genomics 7:25.

Higgins D., Lemey P. 2009. Multiple sequence alignment. In: Lemey P., Salemi M., Vandamme A.-M. editors. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge, UK: Cambridge University Press.

Hillis D., Mable B., Larson A., Davis S., Zimmer E. 1996. Nucleic Acids IV: Sequencing and cloning. In: Hillis D.M., Moritz C., Mable B.K. editors. Molecular Systematics. Sunderland, Massachusetts: Sinauer Associates. p. 321-381.

Hillman J.C. 1988. The Bale Mountains National Park area, southeast Ethiopia, and its management. Mountain Research and Development 8:253-258.

Holder M., Lewis P.O. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nature Reviews Genetics 4:275-284.

Holton T.A., Pisani D. 2010. Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single-and multigene families analyzed under a supertree and supermatrix paradigm. Genome Biology and Evolution 2:310.

Hornett E.A., Wheat C.W. 2012. Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. BMC Genomics 13:361.

Hui J.H., Marco A., Hunt S., Melling J., Griffiths-Jones S., Ronshaugen M. 2013. Structure, evolution and function of the bi-directionally transcribed iab-4/iab-8 microRNA locus in arthropods. Nucleic Acids Research 41:3352-3361.

Huntzinger E., Izaurralde E. 2011. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nature Reviews Genetics 12:99-110.

Isaac N.J.B., Redding D.W., Meredith H.M., Safi K. 2012. Phylogenetically-Informed Priorities for Amphibian Conservation. PLoS ONE 7:e43912.

IUCN. 2013. IUCN Red List of threatened species. Version 2010.3. Available at http://www.iucnredlist.org.

Jukes T., Cantor C. 1969. Evolution of protein molecules. In: Munro M. editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.

Kearney M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Systematic Biology 51:369-381.

Kimura M. 1980. A simple method for estfimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16:111-120.

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

Kocot K.M., Cannon J.T., Todt C., Citarella M.R., Kohn A.B., Meyer A., Santos S.R., Schander C., Moroz L.L., Lieb B. 2011. Phylogenomics reveals deep molluscan relationships. Nature 477:452-456.

Krol J., Loedige I., Filipowicz W. 2010. The widespread regulation of microRNA biogenesis, function and decay. Nature Reviews Genetics 11:597-610.

Lanave C., Preparata G., Sacone C., Serio G. 1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution 20:86-93.

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

Largen M.J. 1991. A new genus and species of petropedetine frog (Amphibia Anura Ranidae) from high altitude in the mountains of Ethiopia. Tropical Zoology 4:139-152.

Largen M.J. 2001. Catalogue of the amphibians of Ethiopia, including a key for their identification. Tropical Zoology 14:307-402.

Largen M.J., Drewes R.C. 1989. A new genus and species of brevicipitine frog (Amphibia Anura Microhylidae) from high altitude in the mountains of Ethiopia. Tropical Zoology 2:13-30.

Lartillot N., Lepage T., Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286-2288.

Lartillot N., Philippe H. 2004. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Molecular Biology and Evolution 21:1095-1109.

Le Quesne W.J. 1969. A method of selection of characters in numerical taxonomy. Systematic Biology 18:201-205.

Lee R.C., Feinbaum R.L., Ambros V. 1993. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843-854.

Lemmon A.R., Brown J.M., Stanger-Hall K., Lemmon E.M. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Systematic Biology 58:130-145.

Letunic I., Bork P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Research 39:W475-W478.

Lewis P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50:913-925.

Li S., Pearl D.K., Doss H. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. Journal of the American Statistical Association 95:493-508.

Löytynoja A., Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences of the United States of America 102:10557-10562.

Lyson T.R., Sperling E.A., Heimberg A.M., Gauthier J.A., King B.L., Peterson K.J. 2012. MicroRNAs support a turtle+ lizard clade. Biology letters 8:104-107.

Maddin H.C., Anderson J.S. 2012. Evolution of the amphibian ear with implications for lissamphibian phylogeny: insight gained from the caecilian inner ear. Fieldiana Life and Earth Sciences 5:59-76.

Maddin H.C., Jenkins Jr F.A., Anderson J.S. 2012. The braincase of *Eocaecilia micropodia* (Lissamphibia, Gymnophiona) and the origin of Caecilians. PLoS ONE 7:e50743.

Maddison W., Maddison D. 2010. Mesquite: a modular system for evolutionary analysis. Available at http://mesquiteproject.org/mesquite/download/download.html.

Mallatt J., Craig C.W., Yoder M.J. 2010. Nearly complete rRNA genes assembled from across the metazoan animals: Effects of more taxa, a structure-based alignment, and paired-sites evolutionary models on phylogeny reconstruction. Molecular Phylogenetics and Evolution 55:1-17.

Mannion P.D., Upchurch P., Barnes R.N., Mateus O. 2013. Osteology of the Late Jurassic Portuguese sauropod dinosaur *Lusotitan atalaiensis* (Macronaria) and the evolutionary history of basal titanosauriforms. Zoological Journal of the Linnean Society 168:98-206.

Marjanović D., Laurin M. 2008. A reevaluation of the evidence supporting an unorthodox hypothesis on the origin of extant amphibians. Contributions to Zoology 77:149-199.

Marjanović D., Laurin M. 2009. The origin (s) of modern amphibians: a commentary. Evolutionary Biology 36:336-338.

Marjanović D., Laurin M. 2013. The origin (s) of extant amphibians: a review with emphasis on the "lepospondyl hypothesis". Geodiversitas 35:207-272.

Mau B., Newton M.A. 1997. Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. Journal of Computational and Graphical Statistics 6:122-131.

McDonald A.T. 2012. Phylogeny of basal iguanodonts (Dinosauria: Ornithischia): an update. PLoS ONE 7:e36745.

McMahon M.M., Sanderson M.J. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. Systematic Biology 55:818-836.

Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E. 1953. Equation of state calculations by fast computing machines. The journal of chemical physics 21:1087-1092.

Meusemann K., von Reumont B.M., Simon S., Roeding F., Strauss S., Kück P., Ebersberger I., Walzl M., Pass G., Breuers S. 2010. A phylogenomic approach to resolve the arthropod tree of life. Molecular Biology and Evolution 27:2451-2464.

Miehe S., Miehe G. 1994. Ericaceous forests and heathlands in the Bale Mountains of South Ethiopia. Stiftung walderhaltung in Afrika and Bundesforschungsansalt für Forst- und Holzwirtschaft. Hamburg, Germany.

Milner A.R. 1988. The relationships and origin of living amphibians. In: Benton M.J. editor. The Phylogeny and Classification of Tetrapods. Oxford, UK: Clarendon Press. p. 59-102.

Monaghan J., Epp L., Putta S., Page R., Walker J., Beachy C., Zhu W., Pao G., Verma I., Hunter T. 2009. Microarray and cDNA sequence analysis of transcription during nerve-dependent limb regeneration. BMC Biology 7:1.

Neyman J. 1971. Molecular studies of evolution: a source of novel statistical problems. In: Gupta S., Yackel J. editors. Statistical decision theory and related topics. New York: Academic Press. p. 1-27.

Nixon K.C., Wheeler Q.D. 1992. Extinction and the origin of species. In: Novacek M.J., Wheeler Q.D. editors. Extinction and Phylogeny. New York: Columbia University Press. p. 119-143.

Page R., Holmes E. 1998. Molecular Evolution: A Phylogenetic Approach. eds. Oxford, UK, Blackwell Publishing, 346 pp.

Pasquinelli A.E., Reinhart B.J., Slack F., Martindale M.Q., Kuroda M.I., Maller B., Hayward D.C., Ball E.E., Degnan B., Müller P. 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature 408:86-89.

Pattengale N., Aberer A., Swenson K., Stamatakis A., Moret B. 2011. Uncovering Hidden Phylogenetic Consensus in Large Data Sets. IEEE/ACM Transactions on Computational Biology and Bioinformatics; IEEE, p. 902-911

Peterson K.J., Dietrich M.R., McPeek M.A. 2009. MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. Bioessays 31:736-747.

Philippe H., Brinkmann H., Copley R.R., Moroz L.L., Nakano H., Poustka A.J., Wallberg A., Peterson K.J., Telford M.J. 2011a. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature 470:255-258.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011b. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS biology 9:e1000602.

Philippe H., Derelle R., Lopez P., Pick K., Borchiellini C., Boury-Esnault N., Vacelet J., Renard E., Houliston E., Quéinnec E. 2009. Phylogenomics revives traditional views on deep animal relationships. Current Biology 19:706-712.

Philippe H., Lartillot N., Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Molecular Biology and Evolution 22:1246-1253.

Philippe H., Snell E.A., Bapteste E., Lopez P., Holland P.W., Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Molecular Biology and Evolution 21:1740-1752.

Piel W.H., Donoghue M.J., Sanderson M.J. 2002. TreeBASE: a database of phylogenetic knowledge. In: Shimura J., Wilson K.L., Gordon D.editors. To the interoperable "Catalog of Life" with partners Species 2000 Asia Oceanea.Research

Report from the National Institute for Environmental Studies No. 171, Tsukuba, Japan.

Pisani D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Systematic Biology 53:978-989.

Pisani D., Carton R., Campbell L., Akanni W., Mulville E., Rota-Stabelli O. 2013. An Overview of Arthropod Genomics, Mitogenomics, and the Evolutionary Origins of the Arthropod Proteome. In: Minelli A., Boxshall G., Fusco G. editors. Arthropod Biology and Evolution: Springer Berlin Heidelberg. p. 41-61.

Pyron R.A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Systematic Biology 60:466-481.

Pyron R.A., Burbrink F.T., Colli G.R., De Oca A.N.M., Vitt L.J., Kuczynski C.A., Wiens J.J. 2011. The phylogeny of advanced snakes (Colubroidea), with discovery of a new subfamily and comparison of support methods for likelihood trees. Molecular Phylogenetics and Evolution 58:329-342.

Pyron R.A., Burbrink F.T., Wiens J.J. 2013. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. BMC Evolutionary Biology 13:93.

Pyron R.A., Wiens J.J. 2011. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. Molecular Phylogenetics and Evolution 61:543-583.

Ragan M. 1992. Phylogenetic inference based on matrix representation of trees. Molecular Phylogenetics and Evolution 1:53-58.

Rannala B., Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution 43:304-311.

Ranwez V., Berry V., Criscuolo A., Fabre P.H., Guillemot S., Scornavacca C., Douzery E.J.P. 2007. PhySIC: a veto supertree method with desirable properties. Systematic Biology 56:798-817.

Reinhart B.J., Slack F.J., Basson M., Pasquinelli A.E., Bettinger J.C., Rougvie A.E., Horvitz H.R., Ruvkun G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. Nature 403:901-906.

Roelants K., Bossuyt F. 2005. Archaeobatrachian paraphyly and Pangaean diversification of crown-group frogs. Systematic Biology 54:111-126.

Roelants K., Gower D.J., Wilkinson M., Loader S.P., Biju S., Guillaume K., Moriau L., Bossuyt F. 2007. Global patterns of diversification in the history of modern amphibians. Proceedings of the National Academy of Sciences 104:887-892.

Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804.

Ronquist F., Huelsenbeck J., Britton T. 2004. Bayesian Supertrees. In: Bininda-Emonds O.P. editor. Phylogenetic Supertrees: Springer Netherlands. p. 193-224.

Rota-Stabelli O., Campbell L., Brinkmann H., Edgecombe G.D., Longhorn S.J., Peterson K.J., Pisani D., Philippe H., Telford M.J. 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proceedings of the Royal Society B: Biological Sciences 278:298-306.

Rota-Stabelli O., Lartillot N., Philippe H., Pisani D. 2013. Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study. Systematic Biology 62:121-133.

Roure B., Baurain D., Philippe H. 2013. Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets. Molecular Biology and Evolution 30:197-214.

Rowe T. 1988. Definition, diagnosis, and origin of Mammalia. Journal of Vertebrate Paleontology 8:241-264.

Ruta M., Coates M.I. 2007. Dates, nodes and character conflict: addressing the lissamphibian origin problem. Journal of Systematic Palaeontology 5:69-122.

San Mauro D. 2010. A multilocus timescale for the origin of extant amphibians. Molecular Phylogenetics and Evolution 56:554-561.

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

San Mauro D., Vences M., Alcobendas M., Zardoya R., Meyer A. 2005. Initial diversification of living amphibians predated the breakup of Pangaea. The American Naturalist 165:590-599.

Sanderson M.J., McMahon M.M., Steel M. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evolutionary Biology 10:155.

Sanderson M.J., McMahon M.M., Steel M. 2011. Terraces in phylogenetic tree space. Science 333:448-450.

Sanderson M.J., Shaffer H.B. 2002. Troubleshooting molecular phylogenetic analyses. Annual Review of Ecology and Systematics 33:49-72.

Sankoff D., Morel C., Cedergren R. 1973. Evolution of 5S RNA and the non-randomness of base replacement. Nature 245:232-234.

Sankoff D., Rousseau P. 1975. Locating the vertices of a Steiner tree in an arbitrary metric space. Mathematical Programming 9:240-246.

Schmidt H.A. 2009. Testing tree topologies. In: Lemey P., Salemi M., Vandamme A.-M. editors. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge, UK: Cambridge University Press. p. 381-404.

Schwarz G. 1978. Estimating the dimension of a model. The annals of statistics 6:461-464.

Scornavacca C., Berry V., Lefort V., Douzery E.J., Ranwez V. 2008. PhySIC_IST: cleaning source trees to infer more informative supertrees. BMC Bioinformatics 9:413.

Scott E. 2005. A phylogeny of ranid frogs (Anura : Ranoidea : Ranidae), based on a simultaneous analysis of morphological and molecular data. Cladistics 21:507-574.

Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 13:2498-2504.

Shen X.X., Liang D., Feng Y.J., Chen M.Y., Zhang P. 2013. A Versatile and Highly Efficient Toolkit Including 102 Nuclear Markers for Vertebrate Phylogenomics, Tested by Resolving the Higher Level Relationships of the Caudata. Molecular Biology and Evolution 30:2235-2248.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Systematic Biology 51:492-508.

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

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

Sigurdsen T., Green D.M. 2011. The origin of modern amphibians: a re-evaluation. Zoological Journal of the Linnean Society 162:457-469.

Sperling E.A., Peterson K.J., Pisani D. 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Molecular Biology and Evolution 26:2261-2274.

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

Steel M., Rodrigo A. 2008. Maximum likelihood supertrees. Systematic Biology 57:243-250.

Swofford D. 1991. When are Phylogeny Estimates From Molecular and Morphological Data Incongruent? In: Miyamoto M., Cracraft J. editors. Phylogenetic Analysis of DNA Sequences. New York: Oxford University Press. p. 295-333.

Swofford D. 2002. PAUP 4.0 b10: Phylogenetic analysis using parsimony. eds. Sunderland, Massachusetts, Sinauer Associates.

Swofford D., Olsen G., Waddell P., Hillis D.M. 1996. Phylogenetic inference. In: Hillis D.M., Moritz C., Mable B.K. editors. Molecular Systematics. Sunderland, Massachusetts: Sinauer Associates. p. 407–514.

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28:2731-2739.

Tarver J.E., Sperling E.A., Nailor A., Heimberg A.M., Robinson J.M., King B.L., Pisani D., Donoghue P.C., Peterson K.J. 2013. miRNAs: small genes with big potential in metazoan phylogenetics. Molecular Biology and Evolution 30:2369-2382.

Thomson R.C., Shaffer H.B. 2010. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Systematic Biology 59:42-58.

Thorley J.L., Wilkinson M. 1999. Testing the phylogenetic stability of early tetrapods. Journal of Theoretical Biology 200:343-344.

Trueb L., Cloutier R. 1991. A phylogenetic investigation of the inter-and intrarelationships of the Lissamphibia (Amphibia: Temnospondyli). eds., Cornell University Press, New York.

Vallin G., Laurin M. 2004. Cranial morphology and affinities of *Microbrachis*, and a reappraisal of the phylogeny and lifestyle of the first amphibians. Journal of Vertebrate Paleontology 24:56-72.

van der Meijden A., Vences M., Hoegg S., Meyer A. 2005. A previously unrecognized radiation of ranid frogs in Southern Africa revealed by nuclear and mitochondrial DNA sequences. Molecular Phylogenetics and Evolution 37:674-685.

Wagner P.J. 2012. Modelling rate distributions using character compatibility: implications for morphological evolution among fossil invertebrates. Biology letters 8:143-146.

Waterhouse A.M., Procter J.B., Martin D.M., Clamp M., Barton G.J. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189-1191.

Wheeler B.M., Heimberg A.M., Moy V.N., Sperling E.A., Holstein T.W., Heber S., Peterson K.J. 2009. The deep evolution of metazoan microRNAs. Evolution & development 11:50-68.

Wheeler W., Aagesen L., Arango C.P., Faivovich J., Grant T., D'Haese C., Janies D., Smith W.L., Varon A., Giribet G. 2006. Dynamic Homology and Phylogenetic Systematics: A Unified Approach Using POY. New York, American Museum of Natural History, 365 pp.

White F. 1978. The Afromontane Region. In: Werger M.J.A. editor. Biogeography and Ecology of Southern Africa. The Hague: Springer Netherlands. p. 463-513.

Wienholds E., Plasterk R.H. 2005. MicroRNA function in animal development. FEBS letters 579:5911-5922.

Wiens J.J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Systematic Biology 52:528-538.

Wiens J.J. 2006. Missing data and the design of phylogenetic analyses. Journal of biomedical informatics 39:34-42.

Wiens J.J., Fetzner J.W., Parkinson C.L., Reeder T.W. 2005. Hylid frog phylogeny and sampling strategies for speciose clades. Systematic Biology 54:778-807.

Wiens J.J., Hutter C.R., Mulcahy D.G., Noonan B.P., Townsend T.M., Sites J.W., Reeder T.W. 2012. Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. Biology letters, 10.1098/rsbl.2012.0703.

Wiens J.J., Morrill M. 2011. Missing Data in Phylogenetic Analysis: Reconciling Results from Simulations and Empirical Data. Systematic Biology 60:719-731.

Wilkinson M. 1994a. Common Cladistic Information and its Consensus Representation: Reduced Adams and Reduced Cladistic Consensus Trees and Profiles. Systematic Biology 43:343-368.

Wilkinson M. 1994b. The permutation method and character compatibility. Systematic Biology 43:274-277.

Wilkinson M. 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Systematic Biology 44:501-514.

Wilkinson M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. Molecular Biology and Evolution 13:437-444.

Wilson E.O. 1965. A consistency test for phylogenies based on contemporaneous species. Systematic Zoology 14:214-220.

Xia X., Lemey P. 2009. Assessing substitution saturation with DAMBE. In: Lemey P., Salemi M., Vandamme A.-M. editors. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge, UK: Cambridge University Press. p. 611-626.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39:306-314.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology and Evolution 11:367–372.

Yang Z. 2006. Computational Molecular Evolution. Oxford University Press, Oxford, UK. 357 pp.

Yang Z., Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Molecular Biology and Evolution 14:717-724.

Zardoya R., Meyer A. 2001. On the origin of and phylogenetic relationships among living amphibians. Proceedings of the National Academy of Sciences 98:7380-7383.

Zhang P., Chen Y., Zhou H., Wang X., Qu L. 2003. The complete mitochondrial genome of a relic salamander, *Ranodon sibiricus* (Amphibia: Caudata) and implications for amphibian phylogeny. Molecular Phylogenetics and Evolution 28:620-626.

Zhang P., Liang D., Mao R.-L., Hillis D.M., Wake D.B., Cannatella D.C. 2013. Efficient sequencing of anuran mtDNAs and a mitogenomic exploration of the phylogeny and evolution of frogs. Molecular Biology and Evolution 30:1899-1915.

Zhang P., Wake D. 2009. Higher-level salamander relationships and divergence dates inferred from complete mitochondrial genomes. Molecular Phylogenetics and Evolution 53:492-508.

Zhang P., Zhou H., Chen Y.-Q., Liu Y.-F., Qu L.-H. 2005. Mitogenomic perspectives on the origin and phylogeny of living amphibians. Systematic Biology 54:391-400.

Zheng Y., Zhang H., Zhang X., Feng D., Luo X., Zeng C., Lin K., Zhou H., Qu L., Zhang P. 2011. MiR-100 regulates cell differentiation and survival by targeting RBSP3, a phosphatase-like tumxor suppressor in acute myeloid leukemia. Oncogene 31:80-92.

Zwickl D. 2006. GARLI: genetic algorithm for rapid likelihood inference. Available at http://www.bio.utexas.edu/faculty/antisense/garli/Garli.html.

# Appendices

# Appendix A

Chronological account of the taxonomic arrangement of *Ericabatrachus baleensis,* Largen 1991. **Bold** text indicates family placement of *Ericabatrachus*; **bold underlined** placement of *Ericabatrachus*

| Author, Year | Family | Subfamily | Genera included |
|---|---|---|---|
| | | | |
| Frost, 1985 | **Petropedetidae Noble, 1931** | | *Anhydrophryne* Hewitt, 1919;<br>*Arthroleptella* Hewitt, 1926;<br>*Arthroleptides,* Nieden 1910;<br>*Cacosternum* Boulenger, 1887;<br>*Dimorphognathus* Boulenger, 1906;<br>*Microbatrachella* Hewitt, 1926;<br>*Natalobatrachus* Hewitt & Methuen, 1913;<br>*Nothophryne* Poynton, 1963;<br>*Petropedetes* Reichenow, 1874<br>*Phrynobatrachus* Günther, 1862;<br>*Phrynodon* Parker, 1935; |
| | | | |
| Largen, 1991 | **Petropedetidae Noble, 1931** | | *Anhydrophryne* Hewitt, 1919;<br>*Arthroleptella* Hewitt, 1926;<br>*Arthroleptides,* Nieden 1910;<br>*Cacosternum* Boulenger, 1887;<br>*Dimorphognathus* Boulenger, 1906;<br>***Ericabatrachus* Largen, 1991;**<br>*Microbatrachella* Hewitt, 1926;<br>*Natalobatrachus* Hewitt & Methuen, 1913;<br>*Nothophryne* Poynton, 1963;<br>*Petropedetes* Reichenow, 1874;<br>*Phrynodon*, Parker, 1935;<br>*Poyntonia* Channing & Boycott, 1989; |
| | | | |
| Dubois, 2005 | Petropedetidae Noble, 1931 | | *Arthroleptides* Nieden 1910;<br>*Conraua* Nieden, 1908;<br>*Petropedetes* Reichenow, 1874. |
| | **Phrynobatrachidae Laurent, 1941** | | *Phrynobatrachus* Günther, 1862;<br>***Ericabatrachus* Largen, 1991;** |
| | | | |
| Scott, 2005 | Ranidae Rafinesque-Schmaltz, 1814 | Phrynobatrachinae Laurent, 1941 | *Natalobatrachus* Hewitt & Methuen, 1913;<br>*Phrynobatrachus* Günther, 1862; |
| | | Petropedetinae Noble, 1931 | *Petropedetes* Reichenow, 1874;<br>> (synonmized *Arthroleptides*., Nieden, 1910) |
| | | **Cacosterninae Noble, 1931** | *Anhydrophryne* Hewitt, 1919;<br>*Arthroleptella* Hewitt, 1926;<br>*Cacosternum* Boulenger, 1887;<br>***Ericabatrachus* Largen, 1991**;<br>*Microbatrachella* Hewitt, 1926;<br>*Nothophryne* Poynton, 1963;<br>*Poyntonia* Channing & Boycott, 1989. |
| | | | |

|  |  |  |  |
|---|---|---|---|
| Frost, 2006 | Petropedetidae Noble, 1931 |  | *Conraua,* Nieden, 1908;<br>*Petropedetes* Reichenow, 1874;<br>*Indirana* Laurent, 1986 |
|  | Phrynobatrachidae Laurent, 1941 |  | ***Ericabatrachus* Largen, 1991**;<br>*Phrynobatrachus* Günther, 1862;<br>(*Phrynodon* Parker, 1935) * placed in synonymy of *Phrynobatrachus* |
|  |  |  |  |
| Roelants, et al. 2007 | Petropedetidae, Noble, 1931 |  | *Conraua,* Nieden, 1908;<br>*Petropedetes* Reichenow, 1874;<br>> (removal of *Indirana* Laurent, 1986) |
|  |  |  |  |
| Pyron and Wiens, 2011 | Petropedetidae, Noble, 1931 |  | *Petropedetes* Reichenow, 1874; |
|  | Phrynobatrachidae, **Laurent, 1941** |  | *Phrynobatrachus* Günther, 1862; |
|  | **Pyxicephalidae, Bonaparte, 1850** | **Cacosterninae** | *Amietia* Dubois, 1987<br>*Anhydrophryne* Hewitt, 1919;<br>*Arthroleptella* Hewitt, 1926;<br>*Cacosternum* Boulenger, 1887;<br>***Ericabatrachus* Largen, 1991**;<br>*Microbatrachella* Hewitt, 1926;<br>*Natalobatrachus* Hewitt & Methuen, 1913;<br>*Nothophryne* Poynton, 1963;<br>*Poyntonia* Channing & Boycott, 1989.<br>*Strongylopus* Tschudi, 1838<br>*Tomopterna* Duméril and Bibron, 1841 |
|  |  | Pyxicephalinae | *Aubria* Boulenger, 1917<br>*Pyxicephalus* Tschudi, 1838 |
|  | Conrauidae |  | *Conraua* Nieden, 1908; |

# Appendix B.

## Generated sequences.

GenBank accession numbers for the *Ericabatrachus baleensis* sequences generated in this study.

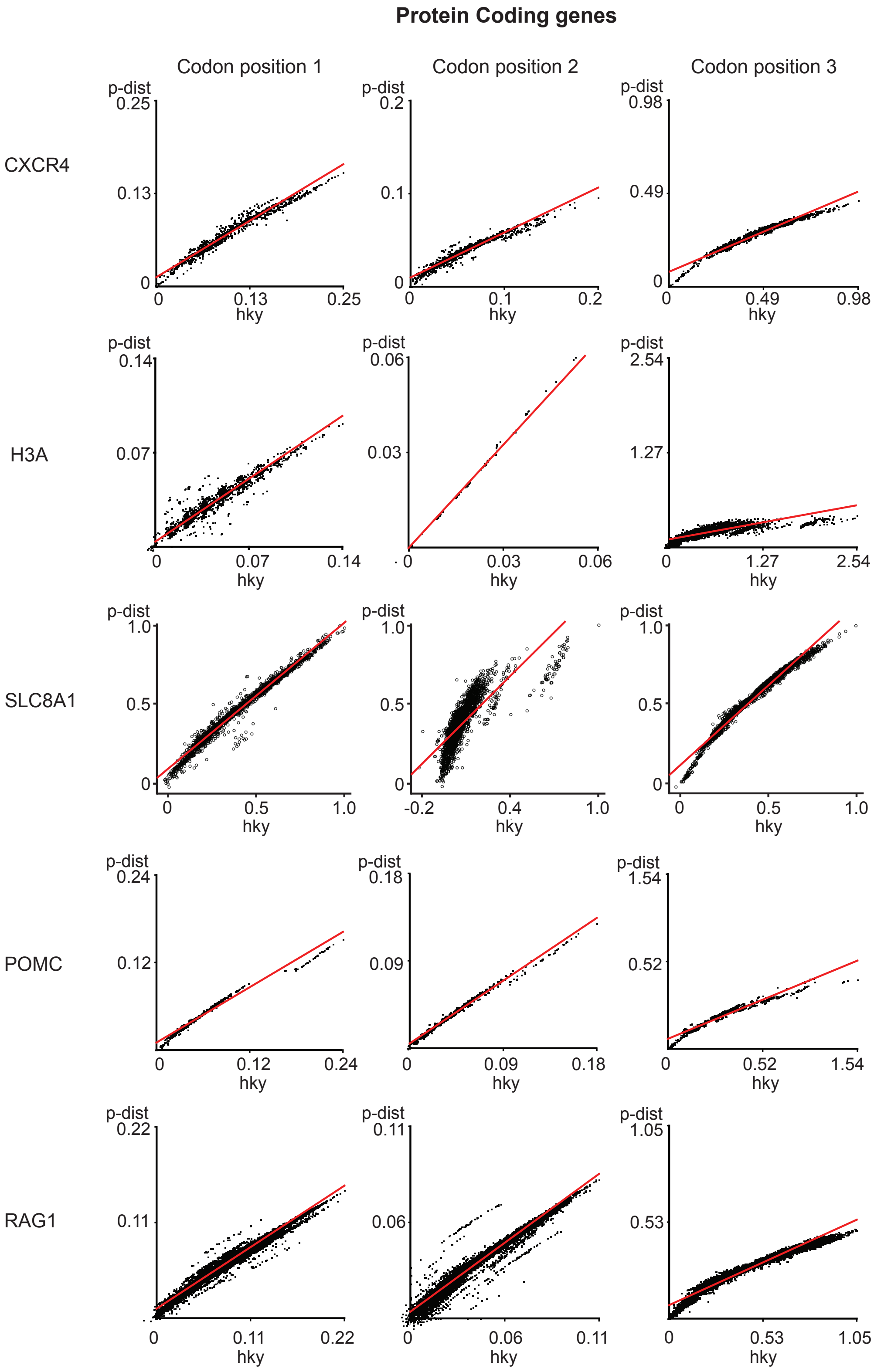
| No./ Molecular Accession | Voucher and Field Number | Locality | 12S | 16S | 28S | H3A | RAG1 |
|---|---|---|---|---|---|---|---|
| T880 | ZNHM-AAU-A2013-003<br>SL 065 | Fute, Harena Forest, Bale Mountains | KF938362 | KF938365 | KF938368 | KF938369 | KF938370 |
| T1083 | ZNHM-AAU-A2013-001<br>AK 2020 | Fute, Harena Forest, Bale Mountains | KF938363 | KF938366 | - | - | KF938371 |
| T1084 | ZNHM-AAU-A2013-002<br>AK 2022 | Fute, Harena Forest, Bale Mountains | KF938364 | KF938367 | - | - | KF938372 |

Primers used in this study.

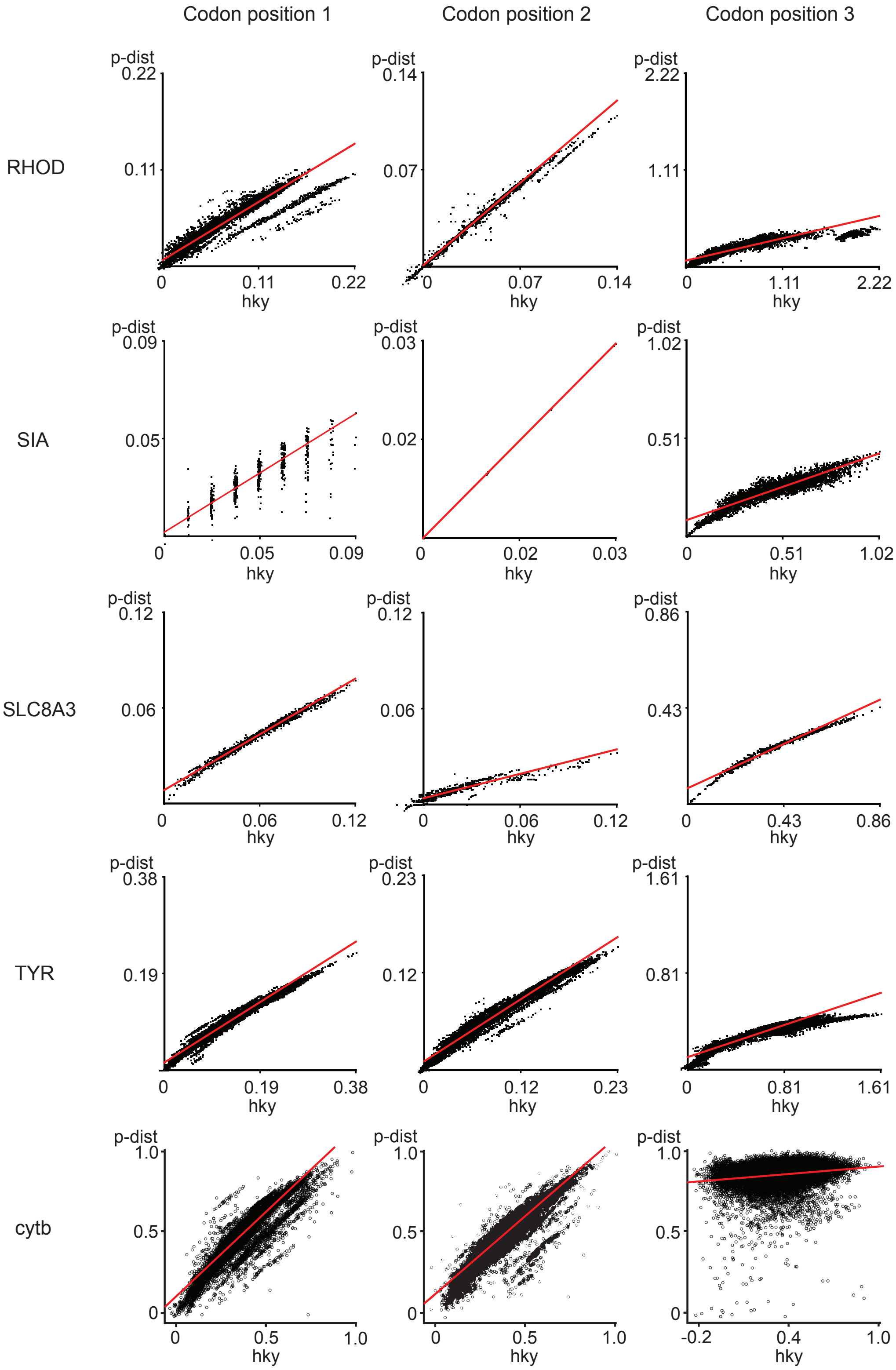
| Gene | Primer |
|---|---|
| *12S* | 12S A-L: AAACTGGGATTAGATACCCCACTAT<br>12S F-H: CTTGGCTCGTAGTTCCCTGGCG |
| *16S* | 16AR: CGCCTGTTTATCAAAAACAT<br>16Br: CCGGTCTGAACTCAGATCACGT |
| *RAG1* | RAG 1c: GGAGATGTTAGTGAGAARCAYGG<br>RAG 1e: TCCGCTGCATTTCCRATGTCRCA |
| *28S* | 28Sv: AAGGTAGCCAAATGCCTCATC<br>28Sjj: AGTAGGGTAAAACTAACCT |
| *H3A* | h3F: ATGGCTCGTACCAAGCAGACVGC<br>h3R: TATCCTTRGGCATRATRGTGAC |

# Appendix C.

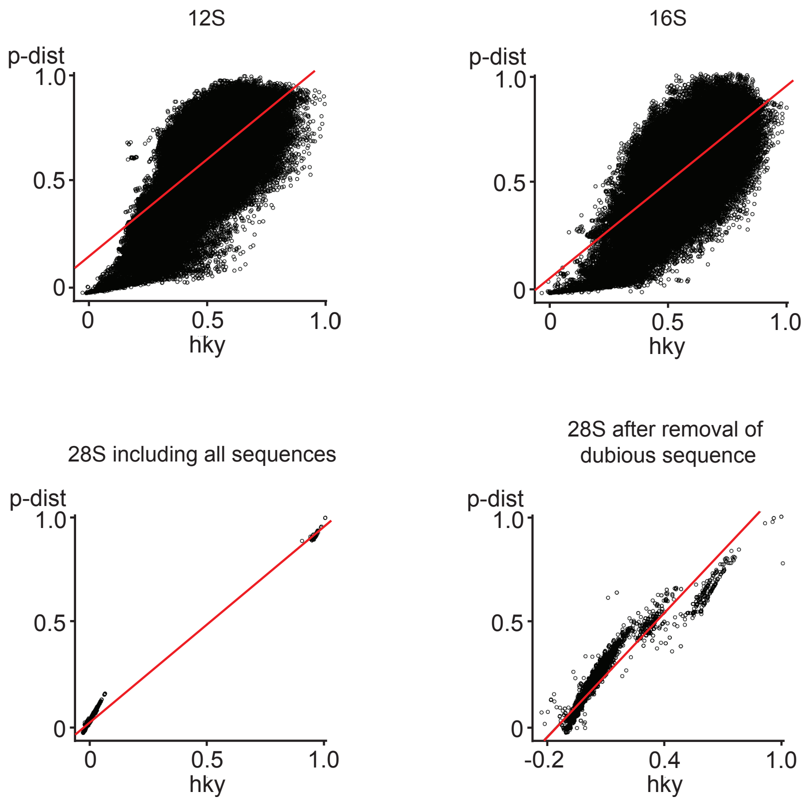
## Saturation plots for all the tested gene partitions.



## Protein Coding genes (continued)

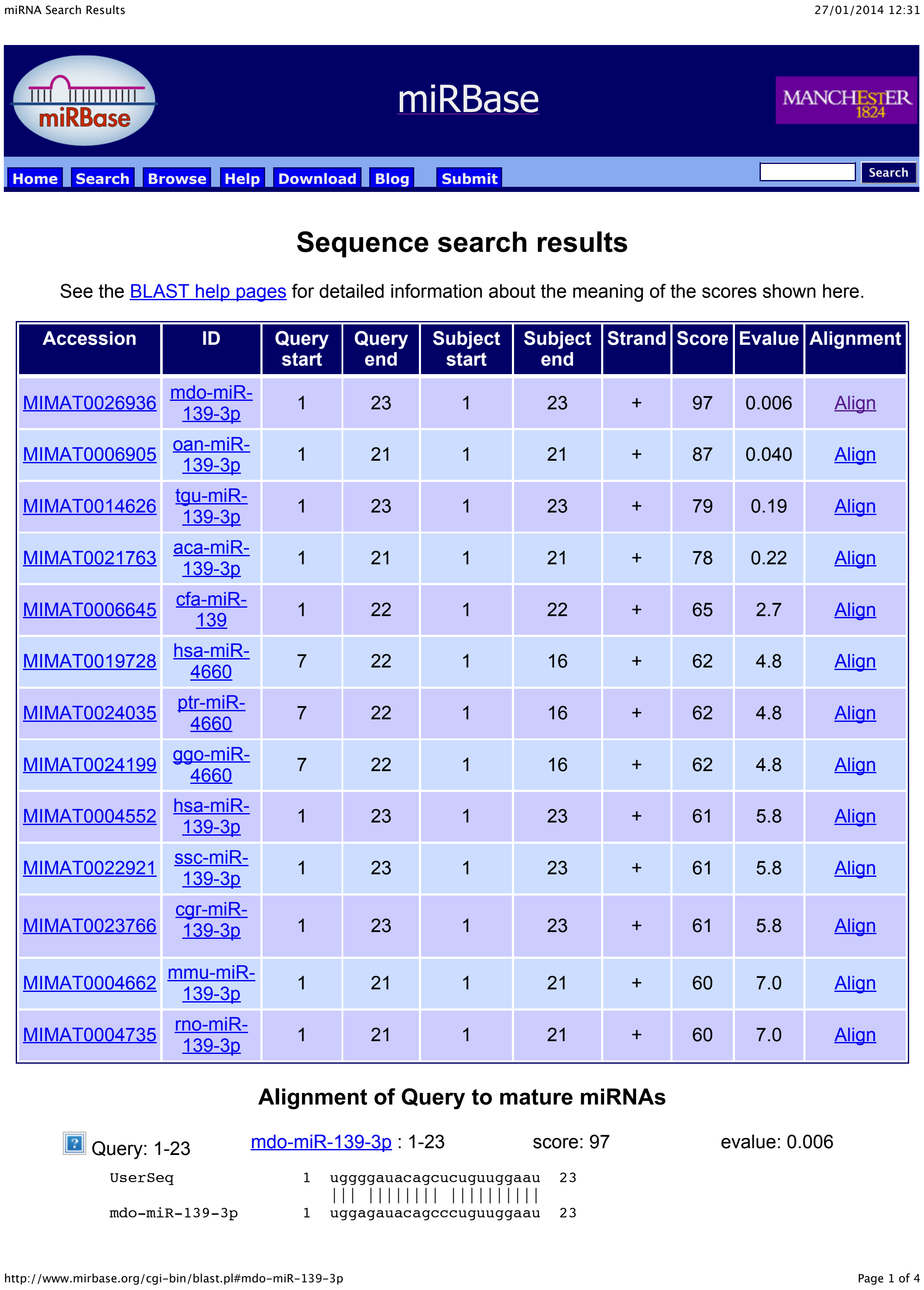


## Non Protein Coding Genes



# Appendix D

BLAST search results for the predicted miRNA “scaffold_995_11919” against the miRBase database.



miRNA Search Results 27/01/2014 12:31

## Sequence search results

See the BLAST help pages for detailed information about the meaning of the scores shown here.

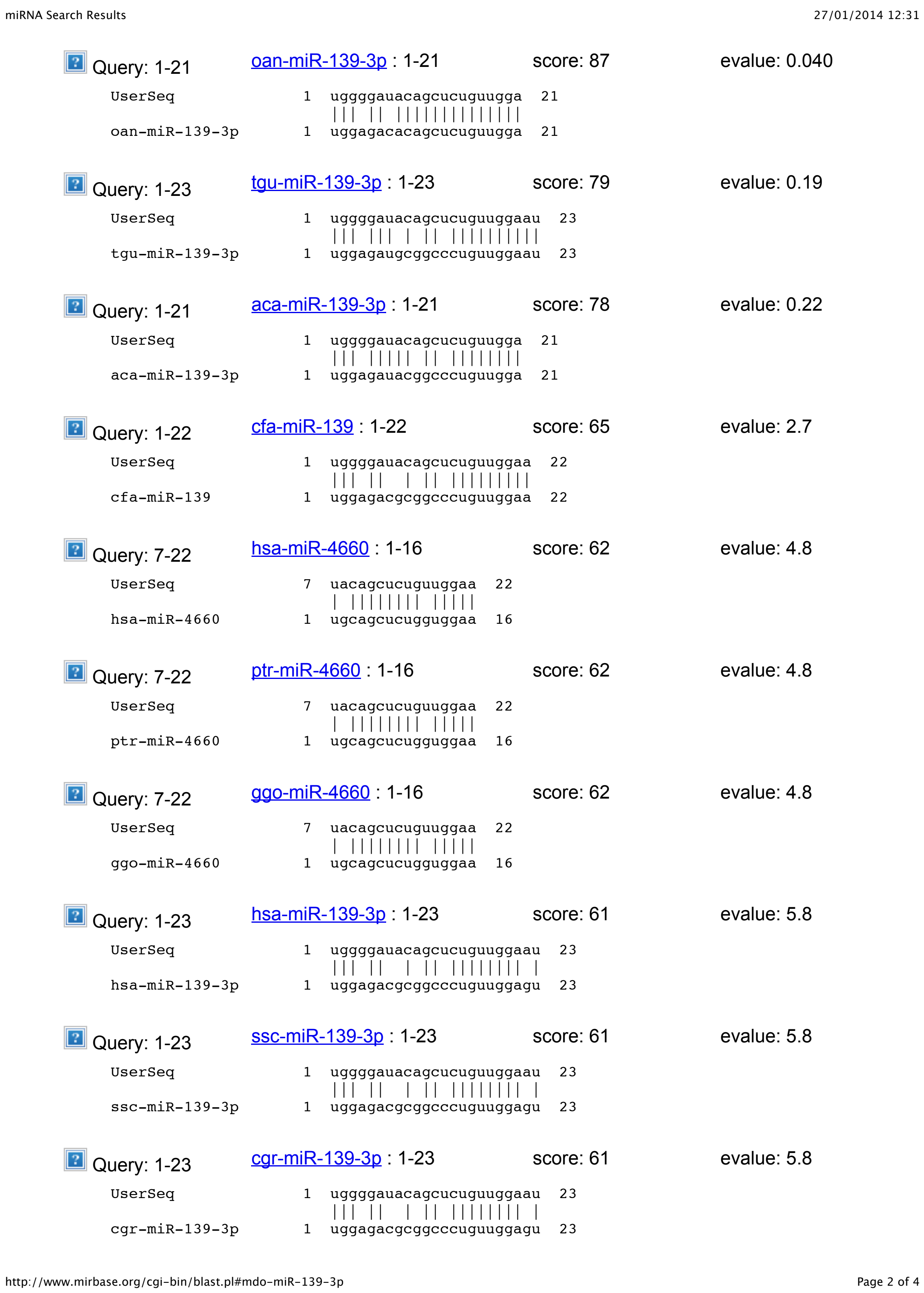
| Accession | ID | Query start | Query end | Subject start | Subject end | Strand | Score | Evalue | Alignment |
|---|---|---|---|---|---|---|---|---|---|
| MIMAT0026936 | mdo-miR-139-3p | 1 | 23 | 1 | 23 | + | 97 | 0.006 | Align |
| MIMAT0006905 | oan-miR-139-3p | 1 | 21 | 1 | 21 | + | 87 | 0.040 | Align |
| MIMAT0014626 | tgu-miR-139-3p | 1 | 23 | 1 | 23 | + | 79 | 0.19 | Align |
| MIMAT0021763 | aca-miR-139-3p | 1 | 21 | 1 | 21 | + | 78 | 0.22 | Align |
| MIMAT0006645 | cfa-miR-139 | 1 | 22 | 1 | 22 | + | 65 | 2.7 | Align |
| MIMAT0019728 | hsa-miR-4660 | 7 | 22 | 1 | 16 | + | 62 | 4.8 | Align |
| MIMAT0024035 | ptr-miR-4660 | 7 | 22 | 1 | 16 | + | 62 | 4.8 | Align |
| MIMAT0024199 | ggo-miR-4660 | 7 | 22 | 1 | 16 | + | 62 | 4.8 | Align |
| MIMAT0004552 | hsa-miR-139-3p | 1 | 23 | 1 | 23 | + | 61 | 5.8 | Align |
| MIMAT0022921 | ssc-miR-139-3p | 1 | 23 | 1 | 23 | + | 61 | 5.8 | Align |
| MIMAT0023766 | cgr-miR-139-3p | 1 | 23 | 1 | 23 | + | 61 | 5.8 | Align |
| MIMAT0004662 | mmu-miR-139-3p | 1 | 21 | 1 | 21 | + | 60 | 7.0 | Align |
| MIMAT0004735 | rno-miR-139-3p | 1 | 21 | 1 | 21 | + | 60 | 7.0 | Align |

## Alignment of Query to mature miRNAs

Query: 1-23   mdo-miR-139-3p : 1-23   score: 97   evalue: 0.006

```
UserSeq           1  uggggauacagcucuguuggaau  23
                     ||| |||||||| ||||||||||
mdo-miR-139-3p    1  uggagauacagcccuguuggaau  23
```

http://www.mirbase.org/cgi-bin/blast.pl#mdo-miR-139-3p Page 1 of 4



miRNA Search Results 27/01/2014 12:31

Query: 1-21 oan-miR-139-3p : 1-21 score: 87 evalue: 0.040

```
UserSeq              1  uggggauacagcucuguugga  21
                        ||| || ||||||||||||||
oan-miR-139-3p       1  uggagacacagcucuguugga  21
```

Query: 1-23 tgu-miR-139-3p : 1-23 score: 79 evalue: 0.19

```
UserSeq              1  uggggauacagcucuguuggaau  23
                        ||| ||| | || ||||||||||
tgu-miR-139-3p       1  uggagaugcggcccuguuggaau  23
```

Query: 1-21 aca-miR-139-3p : 1-21 score: 78 evalue: 0.22

```
UserSeq              1  uggggauacagcucuguugga  21
                        ||| ||||| || ||||||||
aca-miR-139-3p       1  uggagauacggcccuguugga  21
```

Query: 1-22 cfa-miR-139 : 1-22 score: 65 evalue: 2.7

```
UserSeq              1  uggggauacagcucuguuggaa  22
                        ||| ||  | || |||||||||
cfa-miR-139          1  uggagacgcggcccuguuggaa  22
```

Query: 7-22 hsa-miR-4660 : 1-16 score: 62 evalue: 4.8

```
UserSeq              7  uacagcucuguuggaa  22
                        | |||||||| |||||
hsa-miR-4660         1  ugcagcucugguggaa  16
```

Query: 7-22 ptr-miR-4660 : 1-16 score: 62 evalue: 4.8

```
UserSeq              7  uacagcucuguuggaa  22
                        | |||||||| |||||
ptr-miR-4660         1  ugcagcucugguggaa  16
```

Query: 7-22 ggo-miR-4660 : 1-16 score: 62 evalue: 4.8

```
UserSeq              7  uacagcucuguuggaa  22
                        | |||||||| |||||
ggo-miR-4660         1  ugcagcucugguggaa  16
```

Query: 1-23 hsa-miR-139-3p : 1-23 score: 61 evalue: 5.8

```
UserSeq              1  uggggauacagcucuguuggaau  23
                        ||| ||  | || |||||||| |
hsa-miR-139-3p       1  uggagacgcggcccuguuggagu  23
```

Query: 1-23 ssc-miR-139-3p : 1-23 score: 61 evalue: 5.8

```
UserSeq              1  uggggauacagcucuguuggaau  23
                        ||| ||  | || |||||||| |
ssc-miR-139-3p       1  uggagacgcggcccuguuggagu  23
```

Query: 1-23 cgr-miR-139-3p : 1-23 score: 61 evalue: 5.8

```
UserSeq              1  uggggauacagcucuguuggaau  23
                        ||| ||  | || |||||||| |
cgr-miR-139-3p       1  uggagacgcggcccuguuggagu  23
```

http://www.mirbase.org/cgi-bin/blast.pl#mdo-miR-139-3p Page 2 of 4

# Publications

BMC
Evolutionary Biology

# Evolutionary relationships of the Critically Endangered frog *Ericabatrachus baleensis* Largen, 1991 with notes on incorporating previously unsampled taxa into large-scale phylogenetic analyses

Siu-Ting *et al.*

BMC
Evolutionary Biology

RESEARCH ARTICLE Open Access

# Evolutionary relationships of the Critically Endangered frog *Ericabatrachus baleensis* Largen, 1991 with notes on incorporating previously unsampled taxa into large-scale phylogenetic analyses

Karen Siu-Ting[1,2], David J Gower[3], Davide Pisani[1,2], Roman Kassahun[4], Fikirte Gebresenbet[5], Michele Menegon[6], Abebe A Mengistu[7], Samy A Saber[8], Rafael de Sá[9], Mark Wilkinson[3] and Simon P Loader[7*]

## Abstract

**Background:** The phylogenetic relationships of many taxa remain poorly known because of a lack of appropriate data and/or analyses. Despite substantial recent advances, amphibian phylogeny remains poorly resolved in many instances. The phylogenetic relationships of the Ethiopian endemic monotypic genus *Ericabatrachus* has been addressed thus far only with phenotypic data and remains contentious.

**Results:** We obtained fresh samples of the now rare and Critically Endangered *Ericabatrachus baleensis* and generated DNA sequences for two mitochondrial and four nuclear genes. Analyses of these new data using *de novo* and constrained-tree phylogenetic reconstructions strongly support a close relationship between *Ericabatrachus* and *Petropedetes*, and allow us to reject previously proposed alternative hypotheses of a close relationship with cacosternines or *Phrynobatrachus*.

**Conclusions:** We discuss the implications of our results for the taxonomy, biogeography and conservation of *E. baleensis*, and suggest a two-tiered approach to the inclusion and analyses of new data in order to assess the phylogenetic relationships of previously unsampled taxa. Such approaches will be important in the future given the increasing availability of relevant mega-alignments and potential framework phylogenies.

**Keywords:** Africa, Amphibia, Eastern Afromontane, Ethiopia, *Petropedetes*, phylogenetics

## Background

*Ericabatrachus baleensis* Largen, 1991, the sole member of its genus, is poorly known and Critically Endangered frog known only from the Harenna Forest in the Bale Mountains of Ethiopia and, until recently, only from the original collection made in 1986 [1-4]. In his description of the genus and species, Largen [5] noted that he intended but had not yet managed to study comparative osteology, which might have provided compelling insights into the evolutionary affinities of *Ericabatrachus*. Instead, while noting that *Ericabatrachus* was "reminiscent of *Phrynobatrachus*" (p. 147) with "habitus *Phrynobatrachus*-like" (p. 141), Largen tentatively concluded, on the basis of shared external features such as terminally T-shaped ("bifid") phalanges, that *Ericabatrachus* is a petropedetine (= petropedetid of some classifications). Thus, in this view, *Ericabatrachus* is most closely related to the East African *Arthroleptides* Nieden, 1911 (= *Petropedetes* Reichenow, 1874), East and West African *Petropedetes* Reichenow, 1874, and Central/West African *Phrynodon*, Parker, 1935. Petropedetinae/dae is a putatively monophyletic group nested within the large clade of "True Frogs" [6] termed "ranids" [7,8].

* Correspondence: simon.loader@unibas.ch
[7]Department of Environmental Sciences, University of Basel, Biogeography Research Group, Basel 4056, Switzerland
Full list of author information is available at the end of the article

Uncertainty over the affinities of *Ericabatrachus* is reflected in a period of taxonomic instability from 2005 until present (summary in Table 1 and Figure 1). Dubois [8] suggested an affiliation between *Ericabatrachus* and *Phrynobatrachus*, presumably based on the similar habitus and superficial resemblance noted by Largen [5]. The same year, Scott [9] published the first broad-scale analysis of ranid phylogeny based on both morphology (predominantly osteology, including the first data for *Ericabatrachus*) and DNA sequence data (lacking for *Ericabatrachus*). Scott's [9] analyses recovered *Ericabatrachus* within the primarily southern African cacosternids, separate from phrynobatrachines and only distantly related to petropedetines (Figure 1). Subsequently, substantial changes to amphibian classification were proposed [6,10] on the basis of large-scale phylogenetic analyses of mostly or entirely DNA sequence data, respectively. Neither of these studies included *Ericabatrachus* in their phylogenetic analysis, but *Ericabatrachus* was alternatively classified within Phrynobatrachidae, considered as likely nesting within *Phrynobatrachus* based on Largen's [5] comment that the taxon was "*Phrynobatrachus*-like" [6] or classified within Pyxicephalidae based on Scott's [9] findings [10]. In summary, over the past 22 years *Ericabatrachus* has been treated as a member of three different families.

With newly collected specimens of *Ericabatrachus baleensis* (see [4,11]), DNA sequence data can, for the first time, be used to investigate the phylogenetic relationships of this challenging taxon. Inferring the phylogenetic relationships of *Ericabatrachus* has important implications for both its biogeography and conservation. If *Ericabatrachus* is closely related to Petropedetidae, this would support the Afromontane biogeographic region [12]. Alternatively, relationships shared with predominantly southern African taxa (either Pyxicephalidae or Phrynobatrachidae) would provide evidence of an unusual biogeographical association. Phylogeny is an important consideration in conservation prioritization (e.g. [13]) and resolution of the relationships of *Ericabatrachus* will shed light on the validity of *Ericabatrachus* as a monotypic genus and the degree to which this now Critically Endangered frog [14] contributes to the genetic distinctiveness of conservation targets in the generally threatened [4] Bale Mountains of Ethiopia.

Substantial steps have recently been made in resolving amphibian phylogenetic relationships [6,10]. The existence of a large and relatively well–sampled mega-alignment including more than 2,800 amphibians [10], potentially provides a useful basis for investigating the phylogenetic position of previously unsampled taxa [15] such as *Ericabatrachus*. However, what might constitute best use of prior phylogenetic work and resources is not necessarily obvious. For example, should we simply append or shoehorn data for new taxa into an existing mega-alignment, thereby accepting previous strategies employed in marker selection, alignments, and masking or should we re-evaluate some or all of these? Should we accept previous phylogenetic conclusions and use these as topological constraints in order to expedite efficient placement of the newly included taxa or should we begin time-consuming unconstrained analyses *de novo*? Here we use newly generated DNA sequence data to investigate the phylogenetic relationships of *Ericabatrachus* and some of the possible strategies for incorporating previously unsampled taxa into large-scale phylogenetic analyses.

## Results

### Saturation analysis

Saturation plots (reported in Additional file 1) supported the inclusion of the following partitions in the large-scale phylogenetic analysis: *RAG1* codon positions 1, 2 and 3; *H3A* codon positions 1 and 2; *16S*; *12S; 28S; CXCR4* codon positions 1, 2 and 3; *SLC8A1* codon positions 1, 2 and 3; *POMC* codon positions 1, 2 and 3; *RHOD* codon positions 1 and 2; *SIA* codon position 2; *SLC8A3* codon positions 1, 2 and 3; *TYR* codon positions1 and 2; and *cytb* codon positions 1 and 2. An outlier species was detected in the *28S* saturation plot, *Fejervarya limnocharis*, and this marker was excluded for this taxon from the analysis.

### Phylogenetic analyses

The large-scale (858-taxon data set), unconstrained ML analysis recovered *Ericabatrachus* as the sister taxon of *Petropedetes* with a bootstrap support of 59% (Figure 2A). This low bootstrap support is primarily a consequence of *Ericabatrachus* being associated with other clades in 35% of the bootstrap replicates (BR) (Additional file 2) but is contributed to also by the instability of *Petropedetes newtoni* which was found outside of *Petropedetes* + *Ericabatrachus* in 9% of the BR. Hence, the effective support for an *Ericabatrachus-Petropedetes* (with exclusion of *P. newtoni*) relationship is 65% (Additional file 2). The second most frequent position (25% of the BR) places *Ericabatrachus* as the sister to or nested inside Pyxicephalinae (the clade composed of *Aubria* + *Pyxicephalus*). Taken together these results circumscribe a relatively well-defined area of the tree within the Ranoidae (or Natatanura [6]) in Figure 2A, including the following lineages: Pyxicephalidae + Petropedetidae + Conrauidae, in which *Ericabatrachus* occurs with a cumulative bootstrap proportion of ~99% (Additional file 2). This allows narrowing the set of plausible relationships for *Ericabatrachus*, and permits more focused analyses to be performed. Using Pyron and Wiens' [10] tree as a topological constraint produced very similar results with respect to the position of

**Table 1 Chronological account of the taxonomic arrangement of *Ericabatrachus baleensis* Largen, 1991**

| Author, Year | Family | Subfamily | Genera included |
|---|---|---|---|
| Frost, 1985 | **Petropedetidae Noble, 1931** | | *Anhydrophryne* Hewitt, 1919; |
| | | | *Arthroleptella* Hewitt, 1926; |
| | | | *Arthroleptides* Nieden, 1910; |
| | | | *Cacosternum* Boulenger, 1887; |
| | | | *Dimorphognathus* Boulenger, 1906; |
| | | | *Microbatrachella* Hewitt, 1926; |
| | | | *Natalobatrachus* Hewitt & Methuen, 1913; |
| | | | *Nothophryne* Poynton, 1963; |
| | | | *Petropedetes* Reichenow, 1874 |
| | | | *Phrynobatrachus* Günther, 1862; |
| | | | *Phrynodon* Parker, 1935. |
| Largen, 1991 | **Petropedetidae Noble, 1931** | | *Anhydrophryne* Hewitt, 1919; |
| | | | *Arthroleptella* Hewitt, 1926; |
| | | | *Arthroleptides*, Nieden, 1910; |
| | | | *Cacosternum* Boulenger, 1887; |
| | | | *Dimorphognathus* Boulenger, 1906; |
| | | | ***Ericabatrachus* Largen, 1991;** |
| | | | *Microbatrachella* Hewitt, 1926; |
| | | | *Natalobatrachus* Hewitt & Methuen, 1913; |
| | | | *Nothophryne* Poynton, 1963; |
| | | | *Petropedetes* Reichenow, 1874; |
| | | | *Phrynodon* Parker, 1935; |
| | | | *Poyntonia* Channing & Boycott, 1989. |
| Dubois, 2005 | Petropedetidae Noble, 1931 | | *Arthroleptides* Nieden, 1910; |
| | | | *Conraua* Nieden, 1908; |
| | | | *Petropedetes* Reichenow, 1874; |
| | **Phrynobatrachidae Laurent, 1941** | | *Phrynobatrachus* Günther, 1862; |
| | | | ***Ericabatrachus* Largen, 1991.** |
| Scott, 2005 | Ranidae Rafinesque-Schmaltz, 1814 | Phrynobatrachinae Laurent, 1941 | *Natalobatrachus* Hewitt & Methuen, 1913; |
| | | | *Phrynobatrachus* Günther, 1862; |
| | | Petropedetinae Noble, 1931 | *Petropedetes* Reichenow, 1874; > (synonymized *Arthroleptides* Nieden, 1910); |
| | | **Cacosterninae Noble, 1931** | *Anhydrophryne* Hewitt, 1919; |
| | | | *Arthroleptella* Hewitt, 1926; |
| | | | *Cacosternum* Boulenger, 1887; |
| | | | ***Ericabatrachus* Largen, 1991**; |
| | | | *Microbatrachella* Hewitt, 1926; |
| | | | *Nothophryne* Poynton, 1963; |
| | | | *Poyntonia* Channing & Boycott, 1989. |

**Table 1 Chronological account of the taxonomic arrangement of *Ericabatrachus baleensis* Largen, 1991** *(Continued)*

| | | | |
|---|---|---|---|
| Frost, 2006 | Petropedetidae Noble, 1931 | | *Conraua* Nieden, 1908; |
| | | | *Petropedetes* Reichenow, 1874; |
| | | | Indirana Laurent, 1986; |
| | Phrynobatrachidae Laurent, 1941 | | ***Ericabatrachus* Largen, 1991**; |
| | | | *Phrynobatrachus* Günther, 1862; |
| | | | (*Phrynodon* Parker, 1935) * placed in synonymy of *Phrynobatrachus*. |
| Roelants et al., 2007 | Petropedetidae Noble, 1931 | | *Conraua* Nieden, 1908; |
| | | | *Petropedetes* Reichenow, 1874; > (removal of *Indirana* Laurent, 1986). |
| Pyron and Wiens, 2011 | Petropedetidae Noble, 1931 | | *Petropedetes* Reichenow, 1874; |
| | Phrynobatrachidae Laurent, 1941 | | *Phrynobatrachus* Günther, 1862; |
| | **Pyxicephalidae Bonaparte, 1850** | **Cacosterninae** | *Amietia* Dubois, 1987; |
| | | | *Anhydrophryne* Hewitt, 1919; |
| | | | *Arthroleptella* Hewitt, 1926; |
| | | | *Cacosternum* Boulenger, 1887; |
| | | | ***Ericabatrachus* Largen, 1991**; |
| | | | *Microbatrachella* Hewitt, 1926; |
| | | | *Natalobatrachus* Hewitt & Methuen, 1913; |
| | | | *Nothophryne* Poynton, 1963; |
| | | | Poyntonia Channing & Boycott, 1989; |
| | | | *Strongylopus* Tschudi, 1838; |
| | | | *Tomopterna* Duméril and Bibron, 1841; |
| | | Pyxicephalinae | *Aubria* Boulenger, 1917; |
| | | | *Pyxicephalus* Tschudi, 1838; |
| | Conrauidae | | *Conraua* Nieden, 1908. |

Text in bold indicates placement of *E. baleensis*.

*Ericabatrachus* and the added *Petropedetes* species, including similar bootstrap support scores (Figure 2B).

Focused, smaller-scale Bayesian and ML analyses (66-taxon data set) recover *Ericabatrachus* as most likely the sister group to *Petropedetes* (Figure 3). The posterior probability for this position under the GTR + G, CAT + G, or CAT-GTR + G models is invariably equal to one. ML bootstrap support is only marginally increased (to ~ 67%). The topologies obtained in different analyses of the 66-taxon data set are almost identical, varying only in the positions of Occidoziga lima, *Phrynobatrachus kreffti*, and *Micrixalus*. AU tests show that the phylogenetic placement of *Ericabatrachus* obtained in our Bayesian and ML results fits the 66-taxon data significantly better than any previously proposed hypothesis.

The strict consensus of our small-scale Bayesian tree and the Pyron and Wiens' [10] tree (both restricted to the common taxa) includes a large basal polytomy but is well resolved in the area where the new taxa (*Ericabatrachus* and some *Petropedetes* species) join the tree (Figure 4). There is a more substantial difference in log-likelihoods between these two trees with our alignment (24.2) than with the Pyron and Wiens' [10] alignment (8.3), but results of AU tests of these restricted topologies using either our alignment or that of Pyron and Wiens [10] were not significant (p = 0.089 and p = 0.331 respectively).

## Discussion

### Taxonomy, phylogeny and biogeography

Thorough phylogenetic analyses of newly acquired molecular data for the rare and Critically Endangered *Ericabatrachus baleensis* provide good support for a sister-group relationship with *Petropedetes* Reichenow, 1874. Our results support Largen's [5] original assignment of *Ericabatrachus* to the family Petropedetidae (although his concept of "Petropedetidae" was somewhat different from current

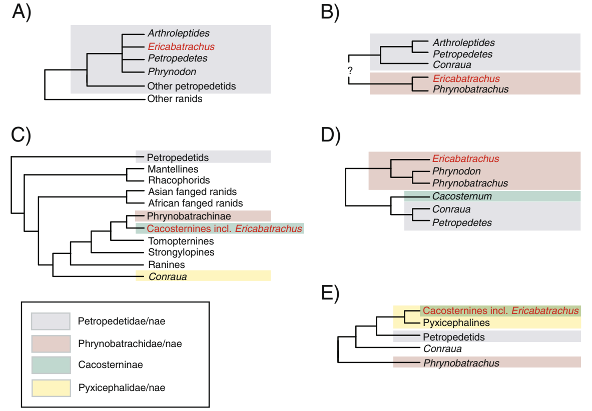


**Figure 1 Alternative hypotheses of the relationships of *Ericabatrachus baleensis* and its sister groups.** The hypotheses are derived from different sources, which were at the time not necessarily presented as trees (see also Table 1). **A)** Largen [5], **B)** Dubois [8], **C)** Scott [9] (based on her Figure 4, the consensus of morphological and molecular analyses and the revised classification in Appendix seven), **D)** Frost et al. [6], and **E)** Pyron and Wiens [10].

taxonomy), rather than Scott's ([9]: p. 532) conclusion that there is "no doubt that… *Ericabatrachus* is a cacosternine, not a petropedetine". Largen suspected this petropedetid relationship on the basis of the presence of terminally T-shaped phalanges ([5,9]). Alternative groupings proposed more recently by other authors are not supported by our analyses. In the bootstrap replicates *Ericabatrachus* joins the tree only once at the base of Phrynobatrachidae (as proposed by Frost et al. [6]) and never in Cacosterninae (as proposed by Scott [9]). In terms of evolutionary relationships within "ranids", in our analysis Petropedetidae forms a strongly supported sister group to a southern African radiation of ranids (Pyxicephalidae), with Conrauidae lying outside this pairing. Other possible resolutions are rejected by the AU tests (Table 2).

The genus *Petropedetes sensu* [9] comprises 12 nominal species distributed in both East and Central Africa. Largen [5] was aware of the high degree of morphological dissimilarity between *Ericabatrachus* and other petropedetids (*Petropedetes*, *Arthroleptides* (=*Petropedetes*) and *Phrynodon* (=*Phrynonbatrachus sandersoni*)) and he was not drawn on any particular putative sister-group relationship. It might have been suspected that, given the geographical proximity of the highlands of Kenya and Tanzania, and the relative but fragmented biogeographic continuity of this area with the Ethiopian highlands, *Ericabatrachus* was most closely related to *Petropedetes* from East Africa (paralleling suspected relationships for other eastern African montane frogs such as in brevicipitids and bufonids [3,5,16,17]). An East African unit (*Ericabatrachus*, *P. martiennseni*, *P. yakusini*) is not supported in our phylogenetic analyses. Sampling of *Petropedetes* is almost complete, but data are lacking for *P. dutoiti* and *P. natator*the sister of Petropedates and the monophyly of *Petropedetes* awaits to be tested fully [18], and might alter our understanding of the relationship of *E. baleensis* relative to all known *Petropedetes*.

*Ericabatrachus* has been one of the most difficult genera of African ranids to classify. Efforts were hampered by the lack of molecular data, and uncertainty was compounded by the fact that *Ericabatrachus* has a suite of morphological characters that have seemingly confused understanding of its evolutionary relationships. Characters that might have supported Largen's conclusion that *Ericabatrachus* was a petropedetid were seemingly not revealed in Scott's [9] analysis. A re-assessment of the morphology of *Ericabatrachus* would clearly be interesting, particularly given the still incomplete knowledge of *Ericabatrachus*. Furthermore, as previously noted by Largen ([5]; p.151), *Ericabatrachus* would appear to be an interesting taxon to include in investigations of correlated patterns of evolution in geographically isolated localities in riverine adapted African ranid species.

On morphological grounds, *Ericabatrachus* seems to be highly divergent from many other close relatives (see

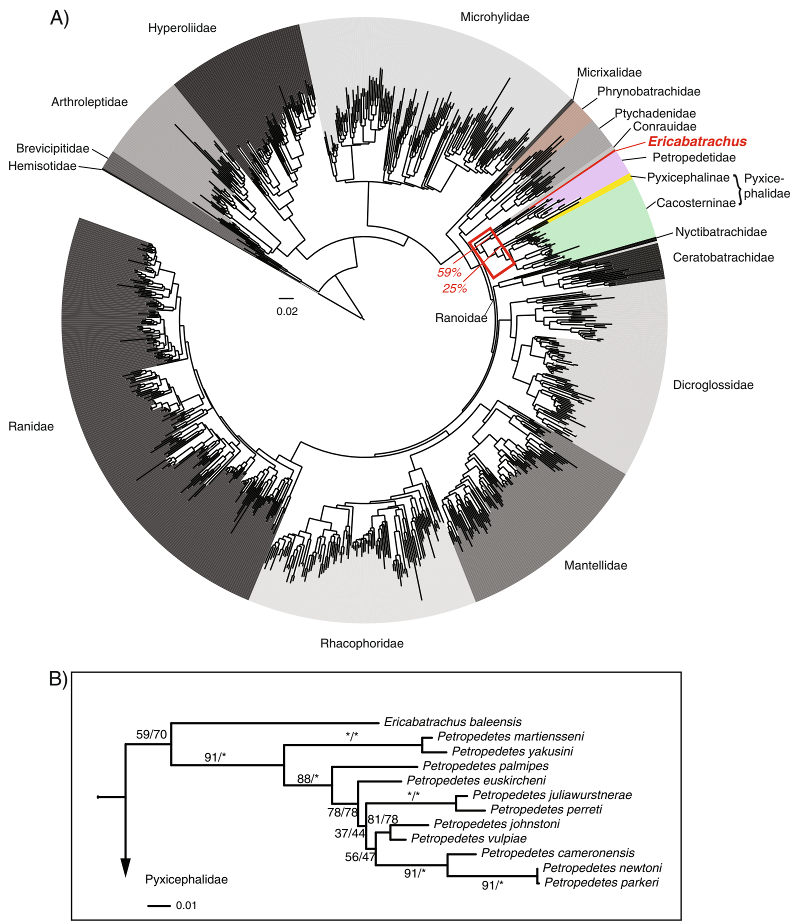


**Figure 2 Phylogenetic relationships of *Ericabatrachus baleensis*. A)** ML tree from the large-scale analysis of Ranoidea. Most frequent placements for *Ericabatrachus* with the corresponding bootstrap percentages are shown (red arrows). The red square denotes the area where *Ericabatrachus* joins the tree 99% of the times (see text). **B)** Close up view of position of *Ericabatrachus* as the sister of *Petropedetes*. Support values for each branch correspond to (left) the *de novo* analysis and (right) the constrained analysis.

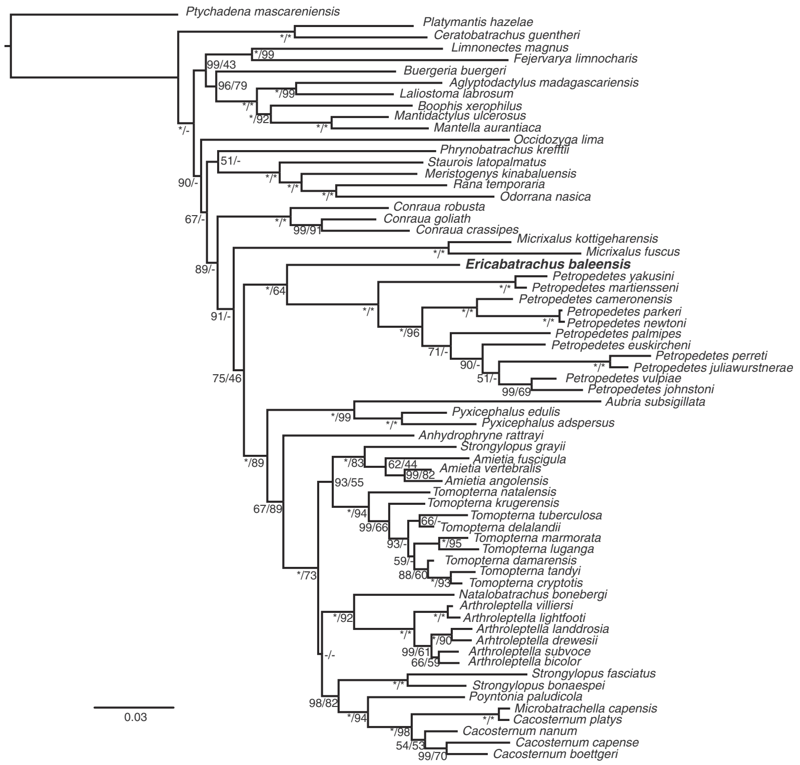


**Figure 3 Small-scale Bayesian tree under GTR model showing the phylogenetic placement of *Ericabatrachus baleensis* (in bold).** Support values for the nodes correspond to posterior probabilities (left) and non-parametric bootstraps (right). Values with "*" represent maximal support (100%), values lower than 40% are denoted by "-".

Appendix four in [9] for list of characters), and this is further supported by molecular differences outlined in this study. A phylogenetic position outside of *Petropedetes*, the morphological distinctiveness of the taxon, and likely long period of divergence from its closest relatives (based on sequence differences for standard genetic markers) agree with Largen's [5] original hypothesis that *Ericabatrachus* should be recognized as a distinct genus. Further research into the still rather complex, and fluctuating taxonomy of African ranids, will be necessary before a full and suitable nomenclatural resolution of Petropedetidae can be made [18].

Biogeographically *Ericabatrachus* has fascinated herpetologists since its original description. It is restricted to the high montane forest of the Bale Mountains, part of the fragmented chain of the Afromontane region [12]. Ethiopia is the most northerly, and therefore isolated part of an extensive chain of mountains in subSaharan Africa. In addition to *Ericabatrachus*, other monotypic amphibian endemics are known from Ethiopia and,

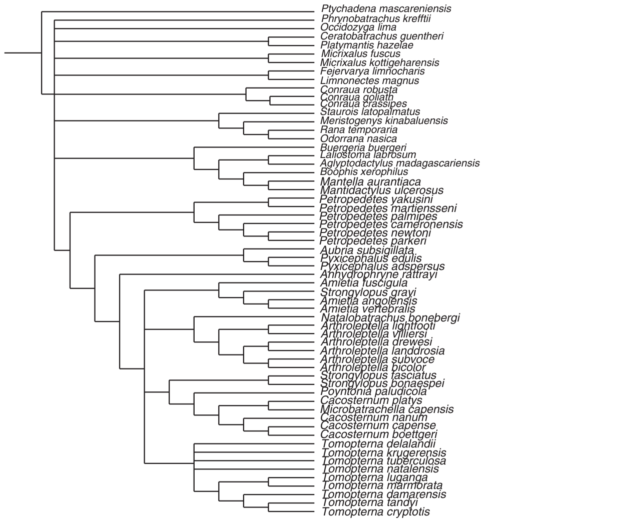


**Figure 4 Strict consensus of our small-scale Bayesian tree and Pyron and Wiens' [10] tree.** Both trees were restricted to the common taxa. Polytomies represent relations that were in disagreement between the two trees.

along with other animal and plant groups, give rise to the impression that the region is a refuge for old and divergent taxa – often referred to as palaeoendemics. Based on branch lengths in our inferred phylogenies, we suspect that the divergence of *Ericabatrachus* from its closest extant relatives is very old given previous estimates of divergence times with closely related pairings in Petropedetidae, Pyxicephalidae and Conrauidae (e.g., [19,20]). The phylogenetic results reported here provide support for the idea that *E. baleensis* is a palaeoendemic species. In light of the other putative palaeoendemic taxa (e.g. *Balebreviceps* and *Altiphrynoides*), the Bale Mountains of Ethiopia appear to have an intriguing, ancient biogeographic history [17].

**Table 2 Hypothesis-testing results**

| Rank | Item | AU test |
|---|---|---|
| 1 | present work, Bayesian GTR Tree | 0.853 |
| 2 | present work, ML Tree | 0.262 |
| **Hypotheses with AU test values lower than 0.05 are rejected** | | |
| 3 | Dubois [8] hypothesis | 0.008 |
| 4 | Pyron and Wiens [10] hypothesis | 1.00E-05 |
| 5 | Frost et al. [10] hypothesis | 1.00E-05 |
| 6 | Scott [10] hypothesis | 4.00E-08 |

Values shown for the Approximately Unbiased test (AU test) from CONSEL [35] tested with the 66-taxon data set (explained in detail in the Methods section). Dotted line separates the non-rejected hypotheses (above) from the rejected hypotheses (below).

### Conservation

*Ericabatrachus baleensis* has declined substantially since its description, it has not been recorded at its type locality (Tulla Negesso) since 1986 or at the only other known historical locality (Katcha) since its original collection [4] and it has recently been re-assessed as Critically Endangered on the IUCN Red List [14]. The declines in these localities are likely to be in association with substantial human-induced habitat degradation in the Rira catchment area [4], but also possibly the emergent infectious disease amphibian chytridiomycosis [11]. We were able to locate *E. baleensis* only in Fute, a new locality in less degraded habitat than nearby Tulla Negesso. Our phylogenetic results demonstrate that the extinction of this frog would be a considerable loss of evolutionary history, thus adding to the demand [4] that urgent conservation action is taken. This could include both *ex situ* or *in situ* approaches, but given the co-occurrence of other distinctive, potentially

palaeoendemic taxa in this locality – a more integrated *in situ* conservation action would be welcomed.

### Incorporating previously unsampled taxa into large-scale phylogenetic analyses

With the collection of previously unsampled taxa of quite uncertain phylogenetic relationships, such as *Ericabatrachus*, then (ignoring the choice of markers) one might try to find closely related taxa to include in a phylogenetic analysis with a BLAST search database query, produce an alignment, and analyse it as exhaustively as seems worthwhile. However, in the age of large-scale phylogeny projects, researchers are increasingly likely to have access to relevant mega-alignments and trees from previous phylogenetic studies. Such resources might greatly simplify and speed up the inference of phylogenetic relationships of previously unsampled taxa. For example, expanding the data through profile alignment and using previous trees as topological constraints can greatly reduce the computational complexity and expense of large-scale phylogenetic inference.

Of course, relying upon previous alignments and trees carries the risk that they are not optimal, particularly given that the inclusion of additional taxa (and markers) has the potential to change the inferred interrelationships of other taxa. In the absence of resource limitations (time, computer power, energy) we might consider *de novo* alignment and unconstrained phylogenetic analyses to be the optimal use of the new data because it would avoid such risks. But resources are always limited. Practical strategies must address the trade off between seeking to use previous results to speed up analysis (and avoid squandering resources) and seeking to avoid sub-optimal inferences.

Here, our main strategy was to use the previous study of Pyron and Wiens [10] as a convenient source of aligned data and as a guide as to the taxonomic content of a major clade that background knowledge suggested included *Ericabatrachus*. We expanded Pyron and Wiens' alignment with taxa and an additional marker and conducted *de novo* large-scale analyses that, in turn, informed taxon selection for subsequent smaller-scale analyses using additional methods and models. Different from [10], our *de novo* analyses included removal of seemingly saturated data partitions, which is generally considered to be helpful in phylogenetic analyses [21-24]. Substantial topological differences between Pyron and Wiens' [10] and our tree (Figure 4) result from these differences in the data and its analyses. Although AU tests do not allow rejection of either tree, the topological differences highlight relationships that are probably best considered uncertain. In turn, this might be taken to suggest that the alternative strategy, of using the Pyron and Wiens [10] tree as a topological constraint, would be problematic. However, this is not the case in this instance. Both our *de novo* analyses and use of Pyron and Wiens' [10] tree as a topological constraint recovered the same relationships of *Ericabatrachus*. We consider the agreement in this particular case to be a fortuitous consequence of the fact that incongruence between our and Pyron and Wiens' [10] trees is concentrated in areas that are least relevant to the relationships of the previously unsampled *Ericabatrachus*.

Eventually it will be neither practical nor sensible to conduct large-scale *de novo* analyses each time a new sequence is added to an alignment. Thus, we anticipate that the use of topological constraints in phylogenetic analyses aimed at placing previously unsampled taxa will increase. We recommend use of topological constraints particularly where relationships have been recovered in multiple unconstrained analyses and appear to be well supported. Conversely we would advise against uncritical acceptance of previous topologies that are not well-corroborated.

When adding novel sequences for genes already present in an existing alignment, we recommend that the inclusion of new data is followed by either an analysis of saturation (as was carried out in this work) or a "quick and dirty" phylogenetic analysis for each gene partition to detect potential sequencing errors or contaminations. If adding entire new gene partitions, then we recommend conducting a BLAST search of the available sequences for that gene, followed by an analysis of saturation and a "quick and dirty" phylogenetic tree of each gene.

## Conclusions

The existence of relatively well-sampled large-scale alignments provided a potentially useful backbone to analyse the taxonomic placement of the poorly known Ethiopian frog *Ericabatrachus baleensis*. A two-tiered approach of phylogenetic analyses using ML and Bayesian methods showed that *Ericabatrachus* is the sister group of *Petropedetes*, which is supported by limited morphological evidence. All previous hypotheses of placement are statistically rejected based on our data set. Using a constrained tree yields the same phylogenetic position for *Ericabatrachus* demonstrating how this approach may obviate the need for time consuming *de novo* analyses. In general, constraints should be relied upon only when they are very well-supported. The sister-group relationship of *Ericabatrachus* and *Petropedetes* and the validity of *Ericabatrachus* as a separate and divergent genus support the contiguity of the Afromontane region and reinforces the importance of continuing conservation efforts in the Bale Mountains of Ethiopia.

## Methods

### Sampling and DNA extraction

Our survey of amphibians in Bale Mountains was given permission by federal and regional authorities in Ethiopia. Permission to collect and export material was facilitated by the Ethiopian Wildlife Conservation Authority. The project was part of a broader project to understanding Ethiopian amphibians in which a memorandum of understanding between University of Basel and Addis Ababa University was signed.

Fieldwork was conducted in July to August in 2008 in southeastern Ethiopia (Figures 5a–b) and June 2009, in Harenna Forest in Bale Mountains National Park. Harenna Forest is the type locality of *Ericabatrachus baleensis* [5], and comprises patchy, montane, primary rain forest, and secondary vegetation [4,5,11,25]. Herpetofaunal surveys carried out consisted primarily of visual encounter including rolling logs/stones and searching through leaf litter. All specimens for this study were collected in accordance with animal ethics guidelines established in the University of Basel.

In 2008, collected amphibian specimens, including a single sample of *Ericabatrachus baleensis* (ZNHM-AAU-A2013-003). The specimen was collected at a site in Harrena Forest called "Fute" (6.76474 N 39.751661 E, at 3208 m). Almost one year later (20th June 2009), a further two specimens (ZNHM-AAU-A2013-001, ZNHM-AAU-A2013-002) phenotypically similar to the first specimen and those of Largen [5], were secured at the same locality [4] (see Figure 5c). Specimens were anaesthetized using MS222, fixed in ca. 5% formalin, rinsed in water and stored in 70% ethanol in the collections of the Natural History Museum of Addis Ababa, Ethiopia. Tissue samples (liver) were taken from specimens prior to fixation and preserved in absolute ethanol.

Genomic DNA was extracted from each of the three *Ericabatrachus baleensis* liver samples with a Qiagen DNeasy kit using the protocol for purification of Total DNA from Animal Tissues. For the 2008 sample, we amplified and sequenced two partial mitochondrial genes, 12SrRNA (12S), and *16S*rRNA (*16S*) and three nuclear genes, 28SrRNA (28S), Histone *H3a* (*H3A*), and recombinase activating protein 1 (*RAG1*). In addition, we sequenced 12S, *16S* and *RAG1* for the 2009 samples to assess intraspecific variation (see Additional file 3 for details), where no major differences were found. Primers used in this study are given in Additional file 3.

### Data Matrix

To investigate the phylogenetic relationships of *Ericabatrachus* we added the newly sequenced data to the mega-alignment of Pyron and Wiens [10]. Pyron and Wiens' [10] data set covers the entirety of the Amphibia but it seems reasonable to suppose that including sequence information for non-anuran amphibians or for some groups within Anura to which *Ericabatrachus* clearly does not belong would not be helpful. Inclusion of distantly related sequences (e.g. salamanders and caecilians) would be at a cost of increased computational complexity and would potentially lead to suboptimal model selection for the phylogenetic problem at hand. Accordingly, we restricted our attention to Ranoidea (*sensu* [6,10]), because there seems to be little doubt that *Ericabatrachus* is a member of this taxon [5,9].

The Ranoidea mega-alignment derived from Pyron and Wiens [10] was decomposed into its constituent genes. Names of samples in the alignment were preserved according to those given by Pyron and Wiens [10]. For each gene, taxa with only missing data and empty columns (alignment gaps) were deleted. For all protein-coding genes, first, second, and third codon positions were identified, and reading frames verified using Mega v.5 [26]. In the case of the non-coding *12S* and *16S* partitions, the alignments were only inspected by eye and no obvious problems were found.

New sequences for *Ericabatrachus* and for some other potentially highly relevant species that were not included in Pyron and Wiens' [10] original work, namely the *16S* data for *Petropedetes euskircheni*, *P. perreti*, *P. juliawurstnerae*, *P. vulpiae*, and *P. johnstoni* (retrieved from GenBank, see Additional file 4) were added to the corresponding alignments using the profile method in Muscle v.3.7 [27]. The data were further extended by the addition of *28S* sequences for all the included species for which this nuclear marker was available in GenBank (see Additional file 4) using the structure-based alignment of Mallat et al. [28] as a reference (after having deleted all non-amphibian species and having removing all gap-only columns). A final round of verification was performed during which the alignments were opened in Jalview v.2.6 [29], inspected by eye and modified as necessary, and single-gene trees were built to test for possible sequencing errors in the newly added data (by looking for unusual resolutions of the new taxon). Only low supported conflicts were observed from this analysis (reported in Additional file 5), and so the new sequences were incorporated. Ultimately, our initial concatenated, pruned and extended Ranoidea mega-alignment included the following markers (and numbers of sequences in parentheses) for a total of 858 species: *12S* (645), *16S* (795), *cytb* (244), *28S* (144), *H3A* (141), *RAG1* (258), *CXCR4* (56), *SLC8A1* (73), *POMC* (45), *RHOD* (340), *SIA* (114), *SLC8A3* (52) and *TYR* (301).

### Saturation analysis

We investigated saturation in alternative data partitions (genes and codon positions) using saturation plots generated using the program Patristic v.2 [30] from tip-to-tip distances for corresponding pairs of taxa on trees derived using uncorrected distances (p-distance) and the HKY85 + Gamma (G) model. Partitions that did not display substantial deviations from a linear regression pattern between the observed p-distances and the HKY85 distances are not saturated. In contrast, a plateau (i.e. increasing HKY85 distances correspond to non-increasing observed distance) is indicative of sequence saturation [21,23]. Saturation plots also allow the identification of sequences that are highly dissimilar from their putative homologs in the data set (probably due to poor curation or contamination). Saturated partitions and outlier sequences (with extremely high tip-to-tip distances with respect to all the other sequences in the data set) were excluded in an attempt to minimize the potential emergence of saturation-driven tree reconstruction artifacts.

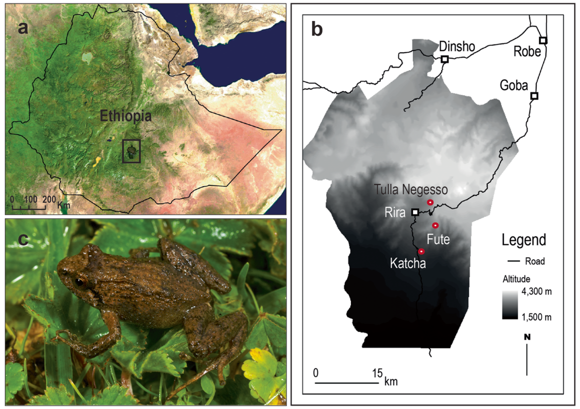


**Figure 5** ***Ericabatrachus baleensis*** **and its reported localities. A)** Map showing the Bale Mountains National Park in Ethiopia. **B)** Close-up of the Bale Mountains National Park showing the geographic position of the type locality (Tulla Negesso) and other sites (circles; squares indicate main human settlements). **C)** A specimen of *E. baleensis* found in the recent surveys [4], photograph by MM.

### Phylogenetic analysis: a two-tiered approach

Given previous disagreement and uncertainty over the phylogenetic placement of *Ericabatrachus*, a "large-scale" approach was initially employed (including all 858 species in our Ranoidea alignment) rooted at Hemisotidae (arising from one of the basalmost splits within Ranoidea following [6,10,19]). Maximum likelihood (ML) inferences and non-parametric bootstrapping were carried out using RAxML v.7.2.6 [31]. For this analysis, unlinked GTR + GAMMA (GTR + G) models were used across the different gene and codon partitions. Additionally, we investigated the use of a partitioned model, identified using PartitionFinder v.1.0.0 [32], which suggested that some of the partitions we initially defined should be merged. The PartitionFinder model separated the data according to codon position and whether they had mitochondrial or nuclear origin. For comparison, we conducted a parallel large-scale analysis in which we used the Ranoidea section of the Pyron and Wiens [10] tree as a topological constraint, with only the positions of the newly introduced taxa (*Ericabatrachus* and some *Petropedetes*) unconstrained.

Subsequent, "small-scale" analyses were performed using a subset of taxa, selected on the basis of the large-scale ML analyses and their relative completeness, to better contextualize and further investigate the phylogenetic relationships of *Ericabatrachus*. The small-scale data set (66 taxa and 8216 bp) included *E. baleensis* and all species belonging to Petropedetidae, Pyxicephalidae (comprising Pyxicephalinae + Cacosterninae), Conrauidae and Micrixalidae. Additionally, two representatives (chosen such as to minimise missing data) from each of the Ptychadenidae, Phrynobatrachidae, Ceratobatrachidae, Dicroglossidae, Mantellidae, Ranidae and Rhacophoridae clades were included as outgroups, based on results of the large-scale analysis. Using this small-scale data set allowed missing data to be reduced (from 78% in the large scale dataset to 65% in the smaller dataset) and the use of Bayesian inference under the often better-fitting CAT-based models in PhyloBayes v.3.3 [33]. Three separate Bayesian analyses were performed with GTR + G, CAT + G, and CAT-GTR + G models. Two chains of 11230, 10900 and 22900 cycles were performed, respectively. Convergence was assessed for each analysis, with a sampling frequency of 100 and the initial 1000 trees (~10%) in each Monte Carlo Markov Chain run being discarded as burn-in. For comparison, a ML GTR + G analysis of this data set was also performed (using RAxML). In all ML analyses support values were estimated using non-parametric bootstrap (100 replicates) and all phylogenetic trees were visualized in iTOL v.2.1 [34].

Approximately Unbiased (AU) tests of two trees were used to compare the fit to the small-scale data of our new and the previously proposed (Figures 1B-E) hypotheses of the relationships of *Ericabatrachus* not including Largen's [5] very incompletely resolved hypothesis (Figure 1A). A total of eight trees was tested: those in Figures 1B to 1E, plus our Bayesian (GTR + G, CAT + G and CAT-GTR + G) trees and ML (GTR + G) tree. To compare trees in Figures 1B to 1E with our results, a preliminary series of AU tests was performed (under GTR + G) including only the trees generated from our analyses. Site-wise log-likelihoods were recalculated (for each of these topologies under GTR + G) in RAxML, and these likelihood values were used to estimate significance in CONSEL v.0.2 [35]. The tree with the best overall fit was our Bayesian GTR + G tree. This tree was then selected as the backbone to generate (by manually editing the position of *Ericabatrachus* and other taxa), trees representing the hypotheses in Figures 1B to 1E. By using the tree that provided the best fit to the data (from our preliminary AU analyses) we avoided introducing a potential bias that might have disfavored previous hypotheses not on the grounds of their placement of *Ericabatrachus* but because of the relationships they displayed for other irrelevant taxa. The trees representing the previous hypotheses and the trees from our original analyses were then subjected to another round of AU tests (under GTR + G). Additionally, we pruned the newly added taxa (*Ericabatrachus* and some *Petropedetes* species) from our Bayesian tree, used the strict consensus to compare this topology with that of the Pyron and Wiens [10] tree restricted to the common taxa, and used AU tests to compare the fit of these two trees to our and to Pyron and Wiens [10] data (under GTR + G) restricted to the subset of taxa.

## Availability of data

The datasets used for the analyses of this study are available in TreeBASE (Study Accession URL: http://purl.org/phylo/treebase/phylows/study/TB2:S15260?format=html).

New sequences produced in this work were uploaded in Genbank (accession numbers from KF938362- KF938372), more details are provided in Additional file 3. A list of sequences added to the original alignment of Pyron and Wiens [10] is provided in Additional file 4. Other additional results supporting the findings of this study can be found in Additional files 1, 2 and 5.

## Additional files

**Additional file 1: Summary of saturation plots for all the gene partitions assessed.**

**Additional file 2: Table summarizing positions for *Ericabatrachus* in the bootstrap trees of the large-scale reconstruction.**

**Additional file 3: Sequences produced for this study.**

**Additional file 4: Accession numbers for the sequences added to the backbone alignment used.** Only the sequences retrieved from GenBank are shown (not the ones sequenced for this study). For a list of the sequences produced for this study see Additional file 3.

**Additional file 5: Single gene analyses for the gene partitions that included *Ericabatrachus*.** Each tree shown is a summary of the position of *Ericabatrachus* from the majority rule extended consensus of 100 non-parametric bootstraps of a single gene partition.

### Abbreviations

*12S*: 12SrRNA; *16S*: 16SrRNA; *28S*: 28SrRNA; *H3A*: histone 3a; *RAG1*: recombinase activating protein 1; *cytb*: cytochrome b; *CXCR4*: C-X-C chemokine receptor type 4; *SLC8A1*: solute-carrier family 8 member 1; *POMC*: pro-opiomelanocortin; *RHOD*: rhodopsin; *SIA*: seventh-in-absentia; *SLC8A3*: solute-carrier family 8 member 3; *TYR*: tyrosinase; HKY85: Hasegawa-Kishino-Yano 85 model; G: Gamma; ML: Maximum Likelihood; GTR: General Time Reversible model; CAT: Bayesian site-heterogeneous model; AU: Approximately Unbiased; BR: Bootstrap replicates; p: p-value; MS222: Tricaine methanesulfonate; ZNHM-AAU: Zoological Natural History Museum of Addis Ababa University, Ethiopia.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

KS, SPL, DJG, MW and DP designed the study and KS and DP performed the analyses. SPL, DJG, MM, AAM, RK, FG, RdS, and SAS, participated in the collection of samples. SPL generated sequence data. KS, SPL, DJG, MW and DP wrote the paper. All authors read and approved the final manuscript.

### Acknowledgements

For comments, and invaluable support and advice in the phylogenetic analyses we thank Chris Creevey. KS thanks Misha Paturyan and Leanne Haggerty (NUIM) for support using the servers. We would like to thank Michael Geiser and Silvia Schwaller (UB) whose fieldwork in Harenna located the first of our *Ericabatrachus* specimens in Fute. In addition, for advice, help with fieldwork, permits for research and export, we thank (no particular order) Michael Barej, Mark-Oliver Rödel (ZMB) Hailu Tadesse (EWCA), Afework Bekele (AAU), Abebe Getahun (AAU) Zelealem Tefera, Isa (WWF), Daniel Tilaye (BMNP), Alastair Nelson, Anouska Kinahan and Thadaigh Baggallay (FZS), Addis Berhanu, Sisay Shewamene, Kifle Argaw (EWCA), Yeneneh Teka (EWCA), Birutesfa Yimer (EWCA), Red Jackal Tour Operators, the people of Rira, and our local guides in 2008–2009 Issa, Ahmed, Hussein, Hussein, and Mohammed. Fieldwork in 2008 and 2009 was supported in part by a National Geographic Expedition Grant (CRE Grant #8532–08). The following bodies also funded surveys and lab work: Ethiopian Wildlife Conservation Authority, University of Basel, European Union Synthesys project (AM), The Natural History Museum London (DJG, SPL), and Stipendienkommission für Nachwuchskräfte aus Entwicklungsländern, Basel (AM), the Swiss National Science Foundation (31003A-133067) to SPL, and RdS acknowledges support from NSF-DEB award 1144692. MW was supported in part by grant BB/K007440/1 from the UK Biotechnology and Biological Sciences Research Council. The work of KS and DP was funded by the Irish Research Council (IRC) and Science Foundation Ireland (SFI). All the phylogenetic analyses were carried out using the High Performance Computing (HPC) facilities at NUI Maynooth and the Irish Centre for High-End Computing (ICHEC).

### Author details

[1]Molecular Evolution and Bioinformatics Lab, National University of Ireland, Maynooth, Co. Kildare, Ireland. [2]School of Biological Sciences and School of Earth Sciences, Woodland Road, Bristol BS8 1UG, UK. [3]Department of Life Sciences, The Natural History Museum, Cromwell Road, London SW7 5BD, UK. [4]Ethiopian Wildlife Conservation Authority, P.O. Box 386, Addis Ababa, Ethiopia. [5]Department of Zoology, Oklahoma State University, 311 D Life Sciences West, Stillwater, OK 74078, USA. [6]Tropical Biodiversity section, MUSE - Museo delle Scienze di Trento, Viale del Lavoro e della Scienza 3, Trento 38123, Italy. [7]Department of Environmental Sciences, University of Basel, Biogeography Research Group, Basel 4056, Switzerland. [8]Zoology Department, Faculty of Science, Al-Azhar University, Assiut, Egypt. [9]Department of Biology, University of Richmond, Richmond, VA 23173, USA.

Received: 28 October 2013 Accepted: 25 February 2014
Published: 10 March 2014

### References

1. Hillman JC: **The Bale Mountains National Park area, southeast Ethiopia, and its management.** *Mt Res Dev* 1988, **8:**253–258.
2. Largen MJ: **Catalogue of the amphibians of Ethiopia, including a key for their identification.** *Trop Zool* 2001, **14:**307–402.
3. Largen MJ, Drewes RC: **A new genus and species of brevicpitine frog (Amphibia Anura Microhylidae) from high altitude in the mountains of Ethiopia.** *Trop Zool* 1989, **2:**13–30.
4. Gower DJ, Aberra RK, Schwaller S, Largen MJ, Collen B, Spawls S, Menegon M, Zimkus BM, De Sá R, Mengistu A, *et al*: **Long-term data for endemic frog genera reveal potential conservation crisis in the Bale Mountains.** *Ethiopia. Oryx* 2013, **47**(1):56–59.
5. Largen MJ: **A new genus and species of petropedatine frog (Amphibia Anura Ranidae) from high altitude in the mountains of Ethiopia.** *Trop Zool* 1991, **4:**139–152.
6. Frost DR, Grant T, Faivovich J, Bain RH, Haas A, Haddad CF, De Sá RO, Channing A, Wilkinson M, Donnellan SC: **The amphibian tree of life.** *Bull Am Mus Nat Hist* 2006, **297:**1–291.
7. Dubois A: **Notes sur la classification des Ranidae (Amphibiens Anoures).** *Bull Mens Soc Linn Lyon* 1992, **61:**305–352.
8. Dubois A: **Amphibia Mundi. 1.1. An ergotaxonomy of recent amphibians.** *Alytes* 2005, **23**(1–2):1–24.
9. Scott E: **A phylogeny of ranid frogs (Anura: Ranoidea: Ranidae), based on a simultaneous analysis of morphological and molecular data.** *Cladistics* 2005, **21**(6):507–574.
10. Pyron AR, Wiens JJ: **A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians.** *Mol Phylogenet Evol* 2011, **61**(2):543–583.
11. Gower DJ, Doherty-Bone TM, Aberra RK, Mengistu A, Schwaller S, Menegon M, De Sá R, Saber SA, Cunningham AA, Loader SP: **High prevalence of the amphibian chytrid fungus (Batrachochytrium dendrobatidis) across multiple taxa and localities in the highlands of Ethiopia.** *Herpetol J* 2012, **22:**225–233.
12. White F: **The Afromontane Region.** In *Biogeography and Ecology of Southern Africa*. 31st edition. Edited by Werger MJA. The Hague: Springer Netherlands; 1978:463–513.
13. Isaac NJB, Redding DW, Meredith HM, Safi K: **Phylogenetically-Informed Priorities for Amphibian Conservation.** *PLoS ONE* 2012, **7**(8):e43912.
14. Ericabatrachus baleensis: *IUCN Red List of threatened species. Version 2013.1*. http://www.iucnredlist.org/.
15. Blackburn D, Wake D: **Class Amphibia Gray, 1825.** *Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness Zootaxa* 2011, **3148:**39–55.
16. Grandison AGC: **The occurrence of Nectophrynoides (Anura: Bufonidae) in Ethiopia. A new concept of the genus with a description of a new species.** *Monit Zool Ital* 1978, **11:**119–172.
17. Loader SP, Ceccarelli FS, Menegon M, Howell KM, Kassahun R, Mengistu A, Saber SA, Gebresenbet F, De Sá R, Davenport TB, *et al*: **Persistence and stability of Eastern Afromontane forests: evidence from brevicipitid frogs.** *J Biogeogr* 2014. accepted.
18. Barej MF, Rödel MO, Loader SP, Menegon M, Gonwouo NL, Penner J, Gvoždík V, Günther R, Bell RC, Nagel P, *et al*: **Light shines through the spindrift – phylogeny of African Torrent Frogs (Amphibia, Anura, Petropedetidae).** *Mol Phylogenet Evol* 2014, **71:**261–273.
19. Roelants K, Gower DJ, Wilkinson M, Loader SP, Biju S, Guillaume K, Moriau L, Bossuyt F: **Global patterns of diversification in the history of modern amphibians.** *Proc Natl Acad Sci USA* 2007, **104**(3):887–892.
20. van der Meijden A, Vences M, Hoegg S, Meyer A: **A previously unrecognized radiation of ranid frogs in Southern Africa revealed by nuclear and mitochondrial DNA sequences.** *Mol Phylogenet Evol* 2005, **37**(3):674–685.
21. Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford MJ: **A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata.** *Proc R Soc B* 2011, **278**(1703):298–306.
22. Rota-Stabelli O, Lartillot N, Philippe H, Pisani D: **Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study.** *Syst Biol* 2013, **62**(1):121–133.
23. Sperling EA, Peterson KJ, Pisani D: **Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa.** *Mol Biol Evol* 2009, **26**(10):2261–2274.
24. Zhang P, Liang D, Mao R-L, Hillis DM, Wake DB, Cannatella DC: **Efficient Sequencing of Anuran mtDNAs and a Mitogenomic Exploration of the Phylogeny and Evolution of Frogs.** *Mol Biol Evol* 2013, **30**(8):1899–1915.
25. Miehe S, Miehe G: *Ericaceous forests and heathlands in the Bale Mountains of South Ethiopia*. Stiftung walderhaltung in Afrika and Bundesforschungsansalt für Forst- und Holzwirtschaft: Hamburg, Germany; 1994.
26. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28**(10):2731–2739.
27. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**(5):1792–1797.
28. Mallatt J, Craig CW, Yoder MJ: **Nearly complete rRNA genes assembled from across the metazoan animals: Effects of more taxa, a structure-based alignment, and paired-sites evolutionary models on phylogeny reconstruction.** *Mol Phylogenet Evol* 2010, **55**(1):1–17.
29. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ: **Jalview Version 2—a multiple sequence alignment editor and analysis workbench.** *Bioinformatics* 2009, **25**(9):1189–1191.
30. Fourment M, Gibbs MJ: **PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change.** *BMC Evol Biol* 2006, **6**(1):1.
31. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**(21):2688–2690.
32. Lanfear R, Calcott B, Ho SY, Guindon S: **PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses.** *Mol Biol Evol* 2012, **29**(6):1695–1701.
33. Lartillot N, Lepage T, Blanquart S: **PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating.** *Bioinformatics* 2009, **25**(17):2286–2288.
34. Letunic I, Bork P: **Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.** *Nucleic Acids Res* 2011, **39**(2):W475–W478.
35. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection.** *Bioinformatics* 2001, **17**(12):1246–1247.

doi:10.1186/1471-2148-14-44
**Cite this article as:** Siu-Ting *et al.*: **Evolutionary relationships of the Critically Endangered frog *Ericabatrachus baleensis* Largen, 1991 with notes on incorporating previously unsampled taxa into large-scale phylogenetic analyses.** *BMC Evolutionary Biology* 2014 **14**:44.

# Concatabominations: Identifying Unstable Taxa in Morphological Phylogenetics using a Heuristic Extension to Safe Taxonomic Reduction

| | |
|---|---|
| Journal: | *Systematic Biology* |
| Manuscript ID: | USYB-2013-187.R1 |
| Manuscript Type: | Point of View |
| Date Submitted by the Author: | 14-Feb-2014 |
| Complete List of Authors: | Siu-Ting, Karen; National University of Ireland, Maynooth, Biology; University of Bristol, School of Biological Sciences<br>Pisani, Davide; University of Bristol, School of Biological Sciences<br>Creevey, Christopher; Aberystwyth University, Institute of Biological, Environmental & Rural Sciences<br>Wilkinson, Mark; The Natural History Museum, Department of Zoology |
| Keywords: | Rogue taxa, consensus, resolution, taxon removal |
| | |

RUNNING HEAD: CONCATABOMINATIONS

TITLE: Concatabominations: Identifying Unstable Taxa in Morphological Phylogenetics using a Heuristic Extension to Safe Taxonomic Reduction

AUTHORS: KAREN SIU-TING[1,2*], DAVIDE PISANI[2], CHRISTOPHER J. CREEVEY[3] AND MARK WILKINSON[4]

[1] *Dept. of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland*

[2] *School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK*

[3] *Institute of Biological, Environmental & Rural Sciences, Aberystwyth University, Aberystwyth SY23 3FG, UK*

[4] *Department of Life Sciences, The Natural History Museum, London SW7 5BD, UK*

** Corresponding author*

*Correspondence to be sent to:*

*School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK*

*E-mail: agalychnica@gmail.com*

KEYWORDS: Rogue taxa, consensus, resolution, taxon removal

For a variety of reasons, some phylogenetic datasets are replete with missing entries. Attitudes towards abundant missing data, specifically concerns over its potential to mislead or confound phylogenetic inferences, are varied. Thus there is a current debate on the impact of missing entries upon the accuracy of phylogenetic inferences (Wiens 2006; Lemmon et al. 2009; Philippe et al. 2011; Wiens and Morrill 2011; Roure et al. 2013). Perhaps less controversial is that individual taxa may sometimes be relatively phylogenetically unstable by virtue of limited data and extensive missing data (e.g. Wilkinson 1996; Sanderson and Shaffer 2002; Wiens 2003; Wilkinson 2003). Wilkinson (1995) developed an approach for diagnosing taxon instability due to missing data *a priori* termed safe taxonomic reduction (STR). STR allows the identification of "rogue" taxa that can be removed from a dataset safe in the knowledge that their removal will not impact upon the interrelationships that will be inferred among the remaining taxa under the parsimony criterion. The potential benefits of such deletion are reductions in numbers of optimal trees and run times and better resolved consensus summaries.

STR has been fairly widely used, mainly by palaeontologists confronted with relatively incomplete fossil taxa (see Anquetin 2012; Graf 2012; McDonald 2012; for some recent examples), but also in the context of the matrix representation with parsimony (Baum 1992; Ragan 1992) approach to supertree construction (e.g. Cardillo et al. 2004). Nonetheless STR is not always as effective as one might hope (e.g. Mannion et al. 2013). Here we present a simple heuristic method for identifying potentially unstable taxa that may be useful in cases where STR does not succeed in ameliorating all the problems caused by missing data. We illustrate the approach through application to the saurischian data of Gauthier (1986) which was previously

used to illustrate STR and thus is particularly appropriate for demonstrating the ability of the new method to achieve more than STR alone.

## The Method

STR is based on the understanding that if the character states of a leaf (OTU, terminal, tip) *w* are a subset of those of a second leaf *x* (such that *w* and *x* have a pairwise-dissimilarity or p-distance of zero) then (1) there exists at least one most parsimonious tree (MPT) in which leaves *w* and *x* are a cherry (sister or adjacent taxa), and (2) removing leaf *w* will not alter the combinations of character states present in the data, the length of most parsimonious trees (MPTs) or relationships inferred among the remaining taxa (Wilkinson 1995). If *w* is similarly potentially related to multiple other leaves (e.g. to *x*, *y*, *z*, etc.) there will be multiple optimal trees that differ only in the placement of *w* with *x* or with *y* or with *z* and so on. In such cases, removing *w*, which adds nothing to a parsimony analysis, can be helpful in reducing numbers of equally optimal trees and improving resolution of strict consensus trees. Figure 1 gives a classification of the sorts of relations that can pertain between pairs of taxa with p-distances of zero.

Sometimes missing (*qua* limited) data seem to be a problem, as evidenced by large numbers of equally optimal trees and poorly resolved consensus trees, but STR is of limited help. In such cases there may be many pairs of leaves with p-distances of zero but, because of the distribution of missing entries, the character states of neither are a proper subset of those of the other (category D, Fig. 1). Wilkinson (1995) called such pairs of leaves "potential taxonomic equivalents that are asymmetric both ways" (we will call them D pairs) and recognised that in contrast to the other categories of taxonomic equivalence the deletion of either member of the D pair cannot be guaranteed to be safe *a priori*. The new method we propose augments STR with a ranking of taxa intended to reflect the potential for their deletion to be safe, to substantially reduce numbers of MPTs, and to improve the resolution of strict consensus trees. Unlike STR the method is a heuristic in that the removal of candidate unstable leaves identified *a priori* by the method may not be safe, although it is not difficult to check this *a posteriori*.

The idea behind the new method is very simple. Given any D pair we can ask whether "forcing" these leaves together into a cherry on a parsimony tree would necessitate some homoplasy that is not already evident in the data. If it does not then it seems plausible that the two leaves could go together in some MPT. If one of these leaves has such a relation with many other leaves it seems plausible that this leaf will be unstable in phylogenetic analyses, which may therefore benefit from its removal.

Our approach to determining whether homoplasy is increased by forcing leaves to go together makes use of compatibility methods (e.g. Meacham and Estabrook 1985). Two characters are compatible if there is some tree on which they can both fit without any extra steps (homoplasy) and simulations have shown that compatibility decreases as homoplasy increases both for whole matrices (O'Keefe and Wagner 2001) and individual characters (Wagner 2012). We count the total number of character pairs in the data that are incompatible (Le Quesne 1969) and use this as a proxy estimate of homoplasy in the original data. We then combine the data for a D pair of leaves to make what we call a "*concatabomination*" (Fig. 2), add this construct to the original data and recalculate the pairwise incompatibility. We repeat the latter for each D pair in turn. For each leaf, we define D* as the number of times that leaf contributes to a concatabomination that does not appear to increase homoplasy (i.e. does not increase the number of pairwise character incompatibilities) in the data. We

also define, for each leaf, ABC as the number of taxonomic equivalences of that leaf in the STR categories A, B or C (each of which identifies scope for *a priori* safe deletion). Taxa can be ranked based on these individual scores or their sum.

Another way of thinking about this approach is to consider that whereas no individual characters provide evidence against the hypothesis that members of a given D pair are actually the same taxon it is possible that combining their data will reveal incompatibilities (homoplasy) that provide an argument that these leaves do not belong together. Consider a data set in which all pairs of characters are incompatible. In that case adding a concatabomination can never increase the pairwise incompatibility in the matrix irrespective of whether it would entail additional homoplasy or not. In such a case D* would be maximal for any leaves that contribute to any D pair and provides no basis for discriminating among them. Where the leaves can be ranked based on the sum of their D* and ABC scores we envisage users safely deleting any high ranked taxa for which ABC is non-zero and then experimentally deleting the taxa with highest D* (or D* + ABC) score to investigate whether this has beneficial impacts (i.e. reduction in numbers of optimal trees, increase in resolution of the strict consensus) while simultaneously checking that the deletion is safe. Removing a taxon is safe precisely when its inclusion or exclusion has no impact upon the inferred relationships of the remaining taxa, i.e., when sets of MPTs inferred with the taxon excluded or with the taxon included but subsequently pruned are identical. If tree length is insensitive to the inclusion/exclusion of a taxon this is also a good, though not infallible, indicator that it can be safely deleted (see Wilkinson 1995).

The new method has been implemented into a "*concatabominations pipeline*" in combination with STR that is available at http://code.google.com/p/concatabominations/. The pipeline uses the Jeffery and Wilkinson's STR software PerlEQ v.1.0 (http://www.molekularesystematik.uni-oldenburg.de/en/34011.html) to find all taxonomic equivalences and Simon Harris's program COMPASS (http://research.ncl.ac.uk/microbial_eukaryotes/downloads.html) to calculate incompatibility scores. The pipeline tallies the taxonomic equivalences, creates and analyses the concatabominations for every D pair and outputs D* and ABC scores of taxa together into a file that can be loaded into Cytoscape (Shannon et al. 2003) to provide a manipulable graphical representation of the results.

### AN EMPIRICAL EXAMPLE

We use Gauthier's (1986) morphological cladistic data for saurischians to illustrate the concatabomination approach in practice. This dataset is a much cited example of the problems of missing data in palaeontological phylogenetics (e.g., Wilkinson 1995; Kearney 2002; Norell and Wheeler 2003), having been previously used to illustrate STR (Wilkinson 1995), and comprising 17 taxa and 84 binary characters with 41% of the entries missing. Missing entries are not randomly distributed in these data but are especially concentrated in some particularly incomplete fossils taxa. Reanalysed with Paup v.4.0b10 (Swofford 2003) with branches collapsed when their maximum lengths are zero, we obtain 832,902 MPTs of 98 steps, the strict consensus of which (Fig. 3a) is disappointingly poorly resolved (with just three splits). Applied to this data set, STR identifies four taxa (*Hulsanpes, Liliensternus, Procompsognathus* and *Saurornitholestes*) that can be safely deleted *a priori*. Their deletion results in a substantial reduction in the number of MPTs (to 197, without any change in tree length) and an increase in the resolution (two additional splits) of their corresponding strict consensus tree (Fig. 3b). Note however that this improvement of the strict consensus can be obtained through the deletion of just *Hulsanpes* and *Saurornitholestes.* Although deletions of *Liliensternus* and/or*, Procompsognathus* are both safe and reduce the number of MPTs they are not effective at increasing the resolution of the corresponding strict consensus.

Table 1 shows the data obtained from the concatabominations pipeline and Figure 4a provides a graphical representation of the same in Cytoscape with vertices representing leaves and edges connecting pairs with either (1) taxonomic equivalences in categories A, B or C (which support safe deletion rules) or (2) concatabominations that do not increase the pairwise incompatibility of the data. The two leaves with the highest D* (*Hulsanpes* and *Sauronitholestes*) scores are also identified by traditional STR as taxa that can be safely deleted. Deletion of *Hulsanpes* alone reduces the number of MPTs for the remaining data to 45,654 without affecting tree length but does not improve (increase the number of splits in) the corresponding strict consensus. The further deletion of *Saurornitholestes* further reduces the number of MPTs to 2,758 and is sufficient to produce all the increased resolution of the consensus (from three to five splits) that can be achieved using traditional STR alone.

Beyond this the two approaches differ. Whereas STR identifies two additional taxa (*Procompsognathus* and *Liliensternus*) that can also be safely deleted, ranking based on D* scores prompts the experimental deletion of *Coelurus*. As already noted, the deletion of *Procompsognathus* and *Liliensternus* reduces the number of MPTs (to 197) but does not further improve the strict consensus. In contrast, deletion of *Coelurus* reduces the number of MPTs to 322 and improves the resolution of the corresponding strict consensus tree by adding an additional split (Fig. 3c). Deletion of *Coelurus* does not change MPT length and the sets of trees produced from the data after its deletion are identical to the trees produced with it included but from which it has been pruned. Thus we can be confident that the deletion of *Coelurus* is safe although it was not identified *a priori* as such by traditional STR.

We find using a graphical representation of the concatabominations pipeline output (Fig. 4), in which the degree of each vertex (leaf) represents the sum of the D* and ABC scores, to be very useful for visualising the potential equivalence relations among the taxa and especially useful in showing how these change with the successive removal of taxa (Fig. 4b-d). Disconnected components in the graph also help identify independent sets of taxonomic equivalents (e.g., the small set including *Procompsognathus* and *Lilienstemus* and the main set that contains *Hulsanpes* and *Saurornitholestes*). Rather than deleting taxa in the order suggested by the initial ranking of their scores, it makes more sense to recalculate the scores and re-rank the taxa after each deletion and this is perhaps most easily accomplished in Cytoscape. Note that after the deletion of *Coelurus* (Fig. 4d) all the taxa that were previously connected in the main set are now unconnected indicating no further potential taxonomic equivalence among those taxa.

The analysis can stop at this point because although additional safe deletions may be possible they cannot be expected to lead to sufficiently reduced numbers of MPTs such as to lead to additional splits in the corresponding strict consensus. Hence we find, *a posteriori*, that the deletions of two other taxa (*Ornitholestes* and *Microvenator*) are also safe but do not lead to any improvements of the strict consensus and are therefore quite unnecessary.

## Discussion

Since its introduction, STR has been adopted, with varying degrees of success, by many phylogenetic palaeontologists as a means of identifying relatively unstable

rogue taxa that can obfuscate what analyses of the data can tells us about phylogenetic relationships of other relatively more stable taxa. It has also been applied in some supertree studies that employ matrix representations (pseudocharacter encodings) of input trees. One undoubted attraction of STR is that a taxon is deleted *a priori* only if we are certain that this deletion cannot impact upon the relationships inferred among the remaining taxa. Thus it is not like throwing away data that could have an impact on the result and is consistent with a “total evidence” philosophy.

Taxon deletion is safe whenever the sets of trees produced by (1) excluding the taxon from the data and (2) pruning it from MPTs inferred with it included are identical. In any particular case there may be useful safe taxon deletions that are not identified *a priori* using STR. Our concatabomination approach is motivated by the desire to extend or augment STR by discovering these. It is a heuristic for identifying candidate rogue taxa, the deletion of which can only be confirmed as safe *a posteriori*. It is worth noting that even the “safe” removal of taxa might impact upon branch length estimation in parametric, model-based phylogenetics and that in stratocladistics (Fischer 2008) deleting potential equivalents would be counterproductive if they are from different time intervals.

The example dataset we used to illustrate the approach served also in the development of STR and might be considered fairly well studied and understood. Thus we were surprised when application of the concatabomination approach to these data led to such a clear cut improvement over what was achievable with STR alone. The example nicely illustrates how the approach can successfully lead to additional safe taxon deletions that improve the resolution of the strict consensus tree and our understanding of what phylogenetic hypotheses are supported by the parsimonious interpretation of the data. Although the approach is heuristic, we expect that highly ranked taxa that it identifies in practice will be the ones that most likely can be safely deleted while usefully reducing the number of MPTs.

We find the graphical representation of the results, with each taxon a vertex and edges representing potential equivalence, and the manipulation it enables to be particularly helpful. As highly connected, potentially unstable, taxa are deleted any changes in the degree of the remaining vertices and of their relative rankings will be apparent. Natural stopping points for experimental deletion are when formerly connected clusters of taxa completely separate or when connected taxa cannot be safely deleted or their safe deletion does not improve the consensus.

Recently, there has been growing interest in the detection of rogue taxa in large-scale phylogenetics mostly using purely *a posteriori* approaches (Aberer and Stamatakis 2011; Pattengale et al. 2011). Concatabominations, which sits somewhat between the pure *a priori* approach of STR and purely *a posteriori* approaches such as leaf stability (Thorley and Wilkinson 1999) or reduced consensus (Wilkinson 1994) offers another approach to this problem. That this approach can be applied to matrix representations of trees highlights its potential in diagnosing the often serious problem of ineffective overlap in broad phylogenomic (multi-gene) studies and in supertree construction (Wilkinson and Cotton 2006, Sanderson et al. 2011).

## Funding

This work was supported by the Biotechnology and Biological Sciences Research Council grant BB/K007440/1 to M.W.; and by an EMBARK scholarship from the Irish Research Council awarded to K.S.

## Acknowledgements

We thank F. Anderson, M. Charleston and P. Wagner for very helpful comments. The authors thank J. McInerney, W. Akanni and members of the Bioinformatics Lab at NUI, Maynooth for constructive discussions during the development of this method. The authors would also like to thank L. Haggerty, V. Paturyan and D. Gardner for technical help using the servers. Analyses were carried out using the computing facilities at the High-Performance Computing Centre – NUI Maynooth, Ireland and University of Bristol, UK.

## References

Aberer A.J., Stamatakis A. 2011.A simple and accurate method for rogue taxon identification. IEEE International Conference on Bioinformatics and Biomedicine; Atlanta (GA), IEEE, p. 118-122.

Anquetin J. 2012. Reassessment of the phylogenetic interrelationships of basal turtles (Testudinata). J. Syst. Palaeontol. 10:3-45.

Baum B. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3-10.

Cardillo M., Bininda ‐ Emonds R., Boakes E., Purvis A. 2004. A species ‐ level phylogenetic supertree of marsupials. J. Zool. (Lond.) 264:11-31.

Fisher D.C. 2008. Stratocladistics: integrating temporal data and character data in phylogenetic inference. Annu. Rev. Ecol. Evol. Syst. 39:365-385.

Gauthier J.A. 1986. Saurischian monophyly and the origin of birds. Mem. Calif. Acad. Sci. 8:1-47.

Graf J. 2012. A new Early Cretaceous coelacanth from Texas. Hist. Biol. 24:441-452.

Kearney M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Syst. Biol. 51:369-381.

Le Quesne W.J. 1969. A method of selection of characters in numerical taxonomy. Syst. Biol. 18:201-205.

Lemmon A.R., Brown J.M., Stanger-Hall K., Lemmon E.M. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58:130-145.

Mannion P.D., Upchurch P., Barnes R.N., Mateus O. 2013. Osteology of the Late Jurassic Portuguese sauropod dinosaur Lusotitan atalaiensis (Macronaria) and the evolutionary history of basal titanosauriforms. Zool. J. Linn. Soc.

McDonald A.T. 2012. Phylogeny of basal iguanodonts (Dinosauria: Ornithischia): an update. PloS one 7:e36745.

Meacham C.A., Estabrook G.F. 1985. Compatibility methods in systematics. Annu. Rev. Ecol. Syst. 16:431-446.

Norell M.A., Wheeler W.C. 2003. Missing Entry Replacement Data Analysis: A Replacement Approach to Dealing with Missing Data in Paleontological and Total Evidence Data Sets. J. Vert. Paleontol. 23:275-283.

O'Keefe F.R., Wagner P.J. 2001. Inferring and testing hypotheses of correlated character evolution using character compatibility. Syst.Biol. 50:657-675.

Pattengale N., Aberer A., Swenson K., Stamatakis A., Moret B. 2011.Uncovering Hidden Phylogenetic Consensus in Large Data Sets. IEEE/ACM Transactions on Computational Biology and Bioinformatics; IEEE, p. 902-911

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Ragan M. 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol. 1:53-58.

Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30:197-214.

Sanderson M.J., Shaffer H.B. 2002. Troubleshooting molecular phylogenetic analyses. Annu. Rev. Ecol. Syst. 33:49-72.

Sanderson M.J., McMahon M.M., Steel M. 2011. Terraces in phylogenetic tree space. Science. 333:448-450.

Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13:2498-2504.

Swofford D. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Thorley J.L., Wilkinson M. 1999. Testing the phylogenetic stability of early tetrapods. J. Theor. Biol. 200:343-344.

Wagner P.J. 2012. Modelling rate distributions using character compatibility: implications for morphological evolution among fossil invertebrates. Biol. Lett. 8:143-146.

Wiens J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52:528-538.

Wiens J. 2006. Missing data and the design of phylogenetic analyses. J. Biomed. Inform. 39:34-42.

Wiens J., Morrill M. 2011. Missing Data in Phylogenetic Analysis: Reconciling Results from Simulations and Empirical Data. Syst. Biol. 60:719-731.

Wilkinson M. 1994. Common Cladistic Information and its Consensus Representation: Reduced Adams and Reduced Cladistic Consensus Trees and Profiles. Syst. Biol. 43:343-368.

Wilkinson M. 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst. Biol. 44:501-514.

Wilkinson M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. Mol. Biol. Evol. 13:437-444.

Wilkinson M. 2003. Missing entries and multiple trees: Instability, relationships, and support in parsimony analysis. J. Vert. Paleontol. 23:311-323.

Wilkinson M., Cotton J.A. 2006. Supertree Methods for Building the Tree of Life: Divide-and-Conquer Approaches to Large Phylogenetic Problems. In: Hodkinson T., Parnell J. editors. Reconstructing the Tree of Life: Taxonomy and Systematics of Species Rich Taxa. Florida: CRC Press. p. 61-75.

Table 1. Results from the concatabominations pipeline analysis of the Gauthier (1986) dataset showing numbers of D* and ABC scores as well as the percentage of missing entries and abbreviations (Abb.) of taxon names used in the Figures.

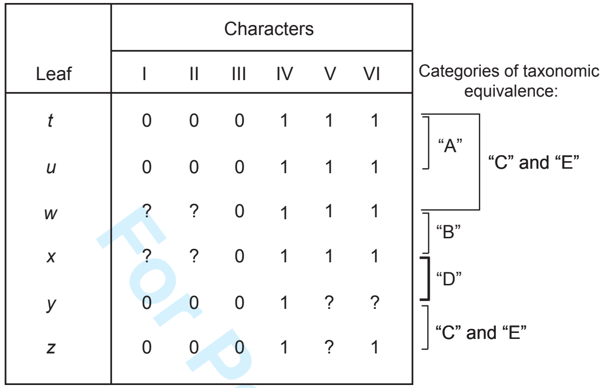
| | | % Missing | | | |
|---|---|---|---|---|---|
| Taxon | Abb. | entries | D* | ABC | Total |
| *Hulsanpes* | Hul | 81 | 7 | 2 | 9 |
| *Saurornitholestes* | Sas | 72 | 7 | 1 | 8 |
| *Coelurus* | Coe | 72 | 5 | 0 | 5 |
| *Ornitholestes* | Ors | 40 | 3 | 0 | 3 |
| *Compsognathus* | Com | 38 | 3 | 0 | 3 |
| *Microvenator* | Mic | 67 | 3 | 0 | 3 |
| Ceratosauria | Cer | 0 | 0 | 2 | 2 |
| Deinonychosauria | Dei | 6 | 0 | 2 | 2 |
| Caenagnathidae | Cae | 33 | 2 | 0 | 2 |
| Elmisauridae | Elm | 54 | 2 | 0 | 2 |
| *Procompsognathus* | Pro | 64 | 1 | 1 | 2 |
| *Liliensternus* | Lil | 48 | 1 | 1 | 2 |
| Ornithomimidae | Orm | 8 | 0 | 1 | 1 |
| Ornithischia | Orn | 0 | 0 | 0 | 0 |
| Sauropodomorpha | Sau | 0 | 0 | 0 | 0 |
| Carnosauria | Car | 2 | 0 | 0 | 0 |
| Avialae | Avi | 4 | 0 | 0 | 0 |

Figure 1. Hypothetical character data illustrating relations of taxonomic equivalence among pairs of taxa (after Wilkinson 1995) and the categories given in STR. Leaves *t* and *u*, which have no missing data and identical character states, are denoted actual equivalents (category A), all the other pairs have some missing data and are denoted potential equivalents. Leaves *w* and *x* have identical character data and are denoted symmetric potential equivalents (category B), all the other possible pairs (except *t* and *u*, *w* and *x*) are asymmetric potential equivalents. Leaves *x* and *y* are asymmetric potential equivalents both ways (category D), pairs *y* and *z*, and *t* and *w* are asymmetric all one way (categories C and E).

Figure 2. Producing a concatabomination (*x*+*y*) for a D pair of taxa with asymmetric potential equivalence both ways. Arrows show how the concatabomination leads to a composite taxon with missing data of each original taxon replaced where possible by data from its pair. In other words, the concatabomination of a D pair is a taxon comprising the union of the character states of the D pair.

Figure 3. Strict consensus trees of MPTs for the saurischian data of Gauthier (1986) or subsets thereof showing the increase in resolution obtained by deleting taxa. a) the complete dataset (no deletions); b) after safe deletion of four taxa identified by STR; c) after deleting the highest ranked taxa identified by the Concatabominations pipeline. For abbreviations used in the trees, refer to Table 1.

Figure 4. Taxonomic equivalences inferred from the concatabominations pipeline visualised in a network with all taxa (a) and with the successive deletions of *Hulsanpes* (Hul) (b), *Saurornitholestes* (Sas) (c) and *Coelurus* (Coe) (d). Vertices represent taxa and the edges represent the type of taxonomic equivalence shared between the taxa. Vertex size is scaled to represent the amount of taxonomic equivalences a taxon has, where the bigger the vertex the more equivalences it has, hence more unstable (see scale at the bottom of figure). Types of equivalence among nodes is represented by dashed lines (types C and E) and solid lines (type D). For a complete list of abbreviations used for the taxa names refer to Table 1.



| Leaf | Characters | | | | | |
|---|---|---|---|---|---|---|
| | I | II | III | IV | V | VI |
| *t* | 0 | 0 | 0 | 1 | 1 | 1 |
| *u* | 0 | 0 | 0 | 1 | 1 | 1 |
| *w* | ? | ? | 0 | 1 | 1 | 1 |
| *x* | ? | ? | 0 | 1 | 1 | 1 |
| *y* | 0 | 0 | 0 | 1 | ? | ? |
| *z* | 0 | 0 | 0 | 1 | ? | 1 |

Categories of taxonomic equivalence:

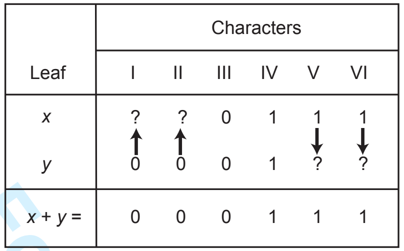
"A"

"C" and "E"

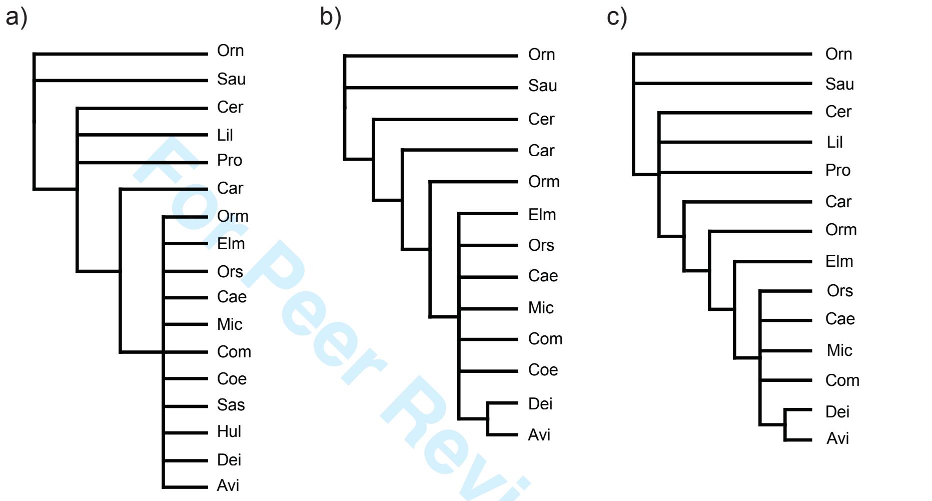
"B"

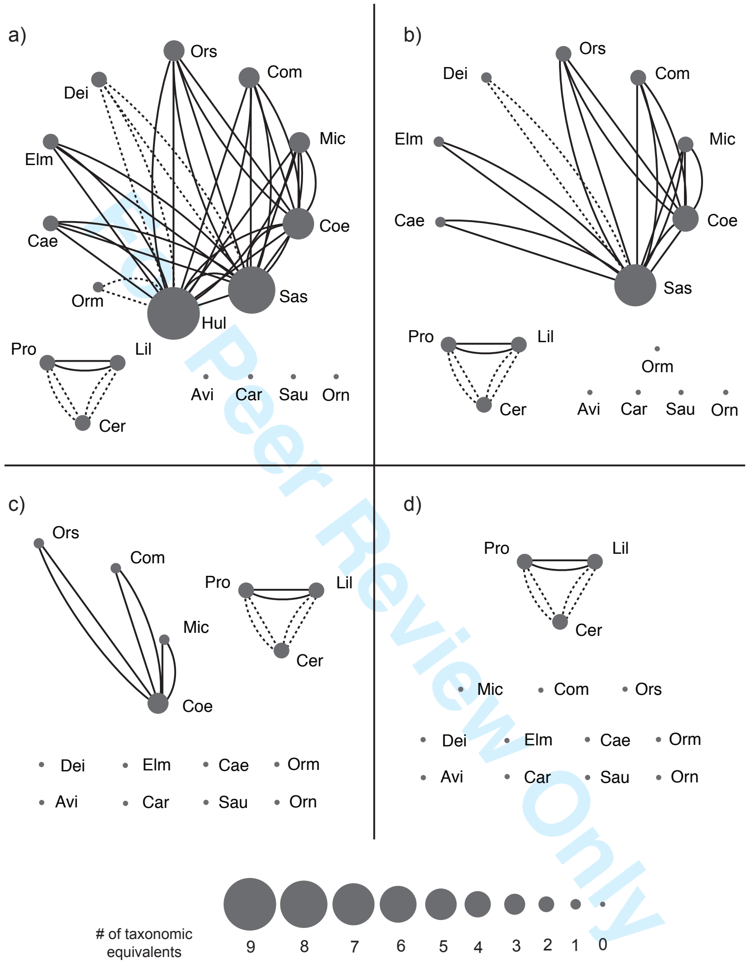
"D"

"C" and "E"



| Leaf | Characters | | | | | |
|---|---|---|---|---|---|---|
| | I | II | III | IV | V | VI |
| *x* | ? | ? | 0 | 1 | 1 | 1 |
| | ↑ | ↑ | | | ↓ | ↓ |
| *y* | 0 | 0 | 0 | 1 | ? | ? |
| *x* + *y* = | 0 | 0 | 0 | 1 | 1 | 1 |


a)
Orn
Sau
Cer
Lil
Pro
Car
Orm
Elm
Ors
Cae
Mic
Com
Coe
Sas
Hul
Dei
Avi
b)
Orn
Sau
Cer
Car
Orm
Elm
Ors
Cae
Mic
Com
Coe
Dei
Avi
c)
Orn
Sau
Cer
Lil
Pro
Car
Orm
Elm
Ors
Cae
Mic
Com
Dei
Avi


a)
Ors
Com
Dei
Mic
Elm
Coe
Cae
Sas
Orm
Hul
Pro
Lil
Avi
Car
Sau
Orn
Cer
b)
Ors
Dei
Com
Mic
Elm
Cae
Coe
Sas
Pro
Lil
Orm
Cer
Avi
Car
Sau
Orn
c)
Ors
Com
Pro
Lil
Mic
Cer
Coe
Dei
Elm
Cae
Orm
Avi
Car
Sau
Orn
d)
Pro
Lil
Cer
Mic
Com
Ors
Dei
Elm
Cae
Orm
Avi
Car
Sau
Orn
# of taxonomic equivalents
9
8
7
6
5
4
3
2
1
0