Molecular Phylogenetics and Evolution 119 (2018) 128–143

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Evaluating methods for phylogenomic analyses, and a new phylogeny for a major frog clade (Hyloidea) based on 2214 loci

Jeffrey W. Streicher[a,b,*], Elizabeth C. Miller[a], Pablo C. Guerrero[c,d], Claudio Correa[d], Juan C. Ortiz[d], Andrew J. Crawford[e], Marcio R. Pie[f], John J. Wiens[a]

[a] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA
[b] Department of Life Sciences, The Natural History Museum, London SW7 5BD, UK
[c] Institute of Ecology and Biodiversity, Faculty of Sciences, University of Chile, 780-0024 Santiago, Chile
[d] Facultad de Ciencias Naturales & Oceanográficas, Universidad de Concepción, Concepción, Chile
[e] Department of Biological Sciences, Universidad de los Andes, A.A. 4976 Bogotá, Colombia
[f] Departamento de Zoologia, Universidade Federal do Paraná, Curitiba, Paraná, Brazil

ARTICLE INFO

Keywords:
Amphibia
Anura
Biogeography
Naive binning
Phylogenomics
Statistical binning

ABSTRACT

Phylogenomic approaches offer a wealth of data, but a bewildering diversity of methodological choices. These choices can strongly affect the resulting topologies. Here, we explore two controversial approaches (binning genes into "supergenes" and inclusion of only rapidly evolving sites), using new data from hyloid frogs. Hyloid frogs encompass ~53% of frog species, including true toads (Bufonidae), glassfrogs (Centrolenidae), poison frogs (Dendrobatidae), and treefrogs (Hylidae). Many hyloid families are well-established, but relationships among these families have remained difficult to resolve. We generated a dataset of ultraconserved elements (UCEs) for 50 ingroup species, including 18 of 19 hyloid families and up to 2214 loci spanning > 800,000 aligned base pairs. We evaluated these two general approaches (binning, rapid sites only) based primarily on their ability to recover and strongly support well-established clades. Data were analyzed using concatenated likelihood and coalescent species-tree methods (NJst, ASTRAL). Binning strongly affected inferred relationships, whereas use of only rapidly evolving sites did not (indicating ~87% of the data contributed little information). The optimal approaches for maximizing recovery and support of well-established clades were concatenated likelihood analysis and the use of a limited number of naive bins (statistical binning gave more problematic results). These two optimal approaches converged on similar relationships among hyloid families, and resolved them with generally strong support. The relationships found were very different from most previous estimates of hyloid phylogeny, and a new classification is proposed. The new phylogeny also suggests an intriguing biogeographical scenario, in which hyloids originated in southern South America before radiating throughout the world.

1. Introduction

Phylogenomic research is now generating massive datasets that can be used to address difficult phylogenetic problems. However, these datasets raise many questions about how the data should be analyzed. For example, should concatenated or coalescent-based (species-tree) analyses be preferred? If coalescent methods are used, which approaches are best? What if the properties of the data do not allow the use of the preferred method (e.g. because of too many genes, too many taxa, or too much missing data)? Should the data primarily determine the choice of methods, or should the choice of methods primarily determine what data are included?

Here, we address three major questions. First, what are the effects of binning on phylogenomic analyses? This approach involves combining sets of genes into bins or "supergenes." These supergenes are intended to provide better estimates of species trees when gene trees are poorly estimated. These supergenes are intended (Bayzid and Warnow, 2013). Simulations suggest that this approach can either improve phylogenetic accuracy (Bayzid and Warnow, 2013) or worsen it (Liu and Edwards, 2015; Liu et al., 2015; but see Springer and Gatesy, 2016), relative to unbinned analyses. There are also many potential approaches to binning, such as naive binning (with different possible numbers of bins) and statistical binning (using compatibility analyses to determine the optimal number of bins; Mirabab et al., 2014b). Second, can accuracy

* Corresponding author at: Department of Life Sciences, The Natural History Museum, London SW7 5BD, UK.
E-mail address: j.streicher@nhm.ac.uk (J.W. Streicher).

https://doi.org/10.1016/j.ympev.2017.10.013
Received 29 May 2017; Received in revised form 21 October 2017; Accepted 22 October 2017
Available online 27 October 2017
1055-7903/

be improved by excluding slower-evolving sites? Recent studies have suggested that accuracy might be improved by including only the fastest evolving sites (e.g. Salichos and Rokas, 2013; Hosner et al., 2016). However, similar to binning, the benefits of this approach have also been disputed (Betancur et al., 2014; Simmons and Gatesy, 2016). Third, what combination of these two approaches optimizes accuracy (i.e. recovery of the true phylogeny)? Previous papers have explored these approaches separately, but it remains unclear what combination of these approaches might yield optimal results.

We evaluate the performance of these approaches using empirical data from frogs. Few empirical systems offer a known phylogeny with which accuracy can be directly evaluated. Nevertheless, those branches that are supported by both molecular and morphological evidence can potentially be used to compare the performance of different sampling and inference methods (e.g. Wiens and Tiu, 2012; Streicher et al., 2016). It is difficult to imagine scenarios by which both molecular and morphological data will be systematically misled to give identical, incorrect relationships, especially if groups are relatively well sampled taxonomically (i.e. no long-branch attraction). Furthermore, empirical data may offer important advantages for evaluating methods relative to simulated data given that empirical data are, by definition, realistic. As one example, phylogenomic datasets often contain some level of missing data, but this is often not incorporated in simulation studies, especially those not specifically focused on this issue. It is not entirely clear what a realistic distribution of missing data would be (e.g. largely random across taxa and genes or more concentrated in certain taxa or genes with particular properties?). This is especially true for phylogenomic data from ultraconserved elements (UCEs; Bejerano et al., 2004; Faircloth et al., 2012), for which many basic properties are still being explored (e.g. Hosner et al., 2016; Meiklejohn et al., 2016; Streicher et al., 2016). Of course, empirical analyses cannot and should not replace simulation studies of method performance. Nevertheless, empirical analyses of method performance offer an important complement to simulation studies, despite remaining relatively underutilized.

In this study, we focus on the phylogeny of hyloid frogs. Hyloid frogs include the majority of frog species (~3600 or ~53%: Pyron and Wiens, 2011; AmphibiaWeb, 2016; Feng et al., 2017). Hyloidea includes many well-known frog families, including the true toads (Bufonidae), treefrogs (Hylidae), glassfrogs (Centrolenidae), and poison frogs (Dendrobatidae). They are distributed globally (especially bufonids and hylids) and include most frog species found in the New World (e.g. Roelants et al., 2007; Pyron and Wiens, 2011; AmphibiaWeb, 2016). Along with Ranoidea, they are one of the two major clades of Neobatrachia, the clade which contains ~95% of all frog species (Ford and Cannatella, 1993; Roelants et al., 2007; Pyron and Wiens, 2011; AmphibiaWeb, 2016).

Many hyloid families are now well established by morphological and molecular data (see below), but relationships among hyloid families have been very difficult to resolve (Fig. 1; Darst and Cannatella, 2004; Frost et al., 2006; Roelants et al., 2007; Pyron and Wiens, 2011; Zhang et al., 2013; Pyron, 2014; Feng et al., 2017; Hutter et al., in press). For example, Pyron and Wiens (2011) conducted a supermatrix analysis of 2871 amphibian species (including 1357 hyloid species), based on likelihood analyses of 12 concatenated genes (Fig. 1d). Among hyloid families, almost no relationships had bootstrap proportions > 70%, except for the clade Terrarana (Brachycephalidae, Ceuthomantidae, Craugastoridae, and Eleutherodactylidae) and the clade uniting Allophrynidae and Centrolenidae. Pyron (2014) analyzed a very similar matrix and obtained very similar results. Other studies have addressed hyloid relationships but with less extensive sampling of taxa (e.g. Darst and Cannatella, 2004; Frost et al., 2006. Roelants et al., 2007; Zhang et al., 2013) and genes (Wiens, 2007, 2011). These studies typically yielded weak support for relationships among hyloid families, and extensive conflicts with other estimates (Fig. 1). In contrast, Feng et al. (2017) found strong support for relationships among a subset of hyloid families using 95 nuclear loci and coalescent analyses (Fig. 1g). However, when they included additional families based on less data, the relationships became weakly supported (Fig. 1h). In summary, relationships among hyloid frogs remain largely unresolved (Fig. 1i). This is unfortunate, especially since numerous studies have now utilized these large-scale estimates of hyloid frog phylogeny, including analyses of life-history evolution (Gomez-Mestre et al., 2012), species richness patterns (Pyron and Wiens, 2013; Hutter et al., in press), diversification (e.g. Roelants et al., 2007; De Lisle and Rowe, 2015; Moen and Wiens, 2017), and ecomorph evolution (Moen et al., 2016).

Here, we analyze relationships among hyloid frogs and empirically evaluate two controversial approaches for phylogenomic data (binning and use of fast sites only). We first generate a novel dataset of ultraconserved (UCE) loci for 50 hyloid species and 5 outgroup taxa (Table 1). We identify 10 clades that are traditionally recognized and are relatively well established by molecular and morphological data. We then evaluate the ability of binning and exclusion of slow-evolving sites to recover and to support these clades (and their support for other clades). We use binning in conjunction with coalescent-based species-tree methods designed for large-scale phylogenomic datasets (NJst: Liu and Yu, 2011; ASTRAL: Mirabab et al., 2014a; Mirabab and Warnow, 2015). Our analyses include naive binning along with weighted and unweighted statistical binning. We also compare these coalescent-based methods to maximum likelihood (ML) analyses of concatenated data, and ML analyses that either include all sites or only fast-evolving sites. We then use the best approach(es) identified by these analyses to infer higher-level phylogenetic relationships among hyloid frogs. Our results offer a strongly supported hypothesis for this important but phylogenetically problematic group.

2. Materials and methods

2.1. Taxon sampling

We sampled 50 species that collectively represent 18 of 19 hyloid families (following the taxonomy of Pyron and Wiens, 2011; AmphibiaWeb, 2016). We were unable to sample the geographically restricted South American family Ceuthomantidae (Heinicke et al., 2009). However, the placement of this family with other terraranan families is well-established (e.g. Heinicke et al., 2009; Pyron and Wiens, 2011; Feng et al., 2017; Hutter et al., in press). We also included representatives of most hyloid subfamilies including: Centroleninae and Hyalinobatrachinae (Centrolenidae), Craugastorinae and Holoadeninae (Craugastoridae), Eleutherodactylinae and Phyzelaphryninae (Eleutherodactylidae), Hylinae, Pelodryadinae, and Phyllomedusinae (Hylidae), Cryptobatrachinae and Hemiphractinae (Hemiphractidae), and Leptodactylinae and Leiuperinae (Leptodactylidae). Note that we differ from AmphibiaWeb (2016) in placing Strabomantidae within Craugastoridae (following Pyron and Wiens, 2011; Padial et al., 2014, and others). We used five non-hyloid taxa as outgroups: *Spea bombifrons* (Scaphiopodidae), *Gastrophryne carolinensis* (Microhylidae), *Rana catesbeiana* (Ranidae), *Calyptocephalella gayi* (Calyptocephalellidae), and *Notaden bennettii* (Myobatrachidae). Many previous analyses have placed Myobatrachidae and Calyptocephalellidae as closely related to Hyloidea, with Ranoidea (including Ranidae and Microhylidae) and Pelobatoidea (Scaphiopodidae) as more distant outgroups (e.g. Roelants et al., 2007; Pyron and Wiens, 2011; Pyron, 2014; Feng et al., 2017). A summary of taxon sampling is presented in Table 1, and voucher information is available in Table S1.

2.2. Targeted sequence capture of ultraconserved elements

We generated UCE data for hyloid anurans using the laboratory protocols described in Faircloth et al. (2012) and the same tetrapod probes as Streicher et al. (2016) and Streicher and Wiens (2016, 2017). These probes are available from http://www.ultraconserved.org (as a FASTA file named “Tetrapods-UCE-5kv1”) and target 5060 UCEs (using