A synthetic phylogeny of freshwater crayfish: insights for conservation

Christopher L. Owen1, Heather Bracken-Grissom2, David Stern1 and Keith A. Crandall1,3

1Computational Biology Institute, George Washington University, Ashburn, VA 20132, USA
2Department of Biological Sciences, Florida International University, Biscayne Bay Campus, North Miami, FL 33181, USA
3Department of Invertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013, USA

Phylogenetic systematics is heading for a renaissance where we shift from considering our phylogenetic estimates as a static image in a published paper and taxonomies as a hardcopy checklist to treating both the phylogenetic estimate and dynamic taxonomies as metadata for further analyses. The Open Tree of Life project (opentreeoflife.org) is developing synthesis tools for harnessing the power of phylogenetic inference and robust taxonomy to develop a synthetic tree of life. We capitalize on this approach to estimate a synthesis tree for the freshwater crayfish. The crayfish make an exceptional group to demonstrate the utility of the synthesis approach, as there recently have been a number of phylogenetic studies on the crayfishes along with a robust underlying taxonomic framework. Importantly, the crayfish have also been extensively assessed by an IUCN Red List team and therefore have accurate and up-to-date area and conservation status data available for analysis within a phylogenetic context. Here, we develop a synthesis phylogeny for the world’s freshwater crayfish and examine the phylogenetic distribution of threat. We also estimate a molecular phylogeny based on all available GenBank crayfish sequences and use this tree to estimate divergence times and test for divergence rate variation. Finally, we conduct EDGE and HEDGE analyses and identify a number of species of freshwater crayfish of highest priority in conservation efforts.

1. Introduction

The freshwater ecosystems represent only 0.8% of the Earth’s surface, but house nearly 6% of all described species and are under severe pressure from multiple impacts, including: overexploitation, water pollution, flow modification, destruction or degradation of habitat, and invasion by exotic species [1]. Freshwater ecosystems in areas such as the southeastern United States house a highly diverse array of fauna that exhibits high levels of endemism [2]. The fragmented nature of these habitats both drives speciation (resulting in exceptional biodiversity) and results in high susceptibility to habitat destruction and limited dispersal capability. The combination of these factors promotes an accelerated extinction rate (on the order of 4% per decade) on par with extinction rates in tropical rainforests [3]. Thus, freshwater habitats are critical to biodiversity, but are at extremely high risk and therefore are in need of conservation efforts.

Crayfish are an important component of these endangered ecosystems and represent an opportunity to capitalize on our knowledge of their diversity (taxonomic and phylogenetic) to help assess relative conservation priorities for freshwater ecosystems as well as conservation priorities for these endangered species themselves [4,5]. Crayfish play a central ecological role in many freshwater ecosystems and provide an important economic and cultural role in many communities [6,7]. Indeed, they have been categorized as keystone species in stream communities both based on consumer activity [8] as well as directly through predation and indirectly through sediment bioturbation and increasing organic
matter processing rates [9]. Even in terrestrial systems, the freshwater crayfish play a significant role as ecological engineers [10,11]. Unfortunately, both the freshwater ecosystems and the crayfish from around the world are under severe pressure and should be high priorities for conservation efforts.

Crayfish are a highly endangered component of these freshwater ecosystems with over 30% of the world’s described species considered endangered and at risk of extinction [12]. The extant representatives are composed of over 600 species distributed taxonomically across three families (Parastacidae, Cambaridae and Astacidae) and 30 genera [13,14]. They are distributed in temperate areas across the globe and are on all continents except Antarctica and continental Africa (although there is an endemic genus in Madagascar). The centre of diversity for the Southern Hemisphere family, Parastacidae, is in southeast Australia and Tasmania; whereas, the centre of diversity for the species rich Cambaridae is in the southeastern United States [13]. Crayfish represent an excellent candidate for phylogenetic synthesis because of their robust underlying taxonomy [15,16] and extensive phylogenetic work across the group (table 1 and references therein). Importantly, they have also been the focus of a recent International Union for Conservation of Nature (IUCN) Red List assessment and have been thoroughly assessed using Red List criteria for endangerment [12].

In this study, we bring together recent crayfish phylogenetic studies with this underlying taxonomic framework to estimate a synthesis tree for all the freshwater crayfish (taxonomy + phylogeny). With this synthesis tree, we map IUCN endangered species status and test for associations with phylogenetic clades and taxonomic groups. Additionally, we estimate a crayfish phylogram using GenBank sequence data from across the freshwater crayfish to obtain branch length estimates and anchor these with divergence time estimates calibrated with extensive fossil data [17]. Using the resulting chronogram, we estimate divergence and extinction rates across the freshwater crayfish. Finally, by combining this phylogenetic information with phylogenetic diversity (PD) calculations and endangered status, we identify crayfish species that are especially evolutionarily distinct and globally endangered (EDGE analysis) [36]. Therefore, our study effectively demonstrates the power of combining robust taxonomy with synthetic phylogeny to aid conservation assessment based on PD and endangerment assessment.

2. Material and methods

(a) Phylogenetic analyses and synthesis

(i) Synthetic tree estimation

Phylogenetic synthesis is the merging of multiple sources of phylogenetic information with an underlying taxonomy. Thus, the generation of synthetic trees differs from supertree approaches [37] both conceptually as well as practically. Supertrees treat multiple phylogenetic estimates as the ‘data’ for a new phylogenetic analysis resulting in the supertree without consideration of the underlying taxonomy. Synthetic trees are the graphical uniting of multiple estimates of phylogeny without re-estimation and therefore they do not suffer from signal enhancement, where novel relationships can appear in the supertree that are not present in the input source trees [38]. Instead, conflict can be visualized and traced back to the source trees without novel relationships being generated from conflicting source trees [39]. Because the synthetic tree approach uses a taxonomy as the underlying backbone structure, conflicts in taxonomy can also be quickly identified.

(ii) Crayfish phylogeny with branch lengths

In order to calculate PD measures, we require a tree with branch lengths. Phylogeny branch lengths are estimates of genetic
change along the branch of a phylogeny, usually specified in units of expected number of substitutions per site. The synthesis tree does not include branch length information; currently, therefore, we generated a crayfish phylogram using PHLAWD [43] to obtain a phylogeny with branch lengths that contains all species with sequence data from GenBank. We removed intraspecific sequences and tried to retain only species because the PHLAWD method does not distinguish between species and intraspecific variants. This includes those new species that authors of previous crayfish studies identified as potential species, but may not be formally described. The loci assembled included portions of the three mitochondrial genes (12S, 16S and cytochrome c oxidase subunit I (COI)) and portions of two nuclear genes (18S and 28S). Sequences for COI were checked for complete open reading frame and GenBank sequence identifiers that included ‘-like’ to prevent nuclear copies of the COI gene from being included [44]. All loci were aligned using MAFFT 7.130b [45], while poorly aligned regions were identified and removed using cBlocks [46] according to the least stringent settings. Two lobsters, Homarus americanus and Eupomatias occidentalis, from two superfamilies within the Astacidae were used as outgroup taxa [17]. An optimal partitioning scheme and models of sequence evolution were determined using PARTITIONFINDER v. 1.1.1 [47] according to the Bayesian Information Criterion (BIC) with the nucleotide alignment divided into seven subsets: 12S, 16S, 18S, 28S, COI 1st pos., COI 2nd pos. and COI 3rd pos. A maximum-likelihood (ML) phylogeny was estimated in GARLI 2.0 [48] according to the models and partitions identified in the PARTITIONFINDER analysis. Multiple searches from random starting trees were conducted to ensure searches were not being trapped in local optima. Branch support was assessed using 100 non-parametric bootstrap replicates [49].

(ii) Crayfish chronogram
A chronogram was estimated using penalized likelihood in r8s [50] with the best ML tree. Node calibrations included six fossils used in previous crayfish chronogram estimation studies [21,22]. These six calibration points spanned from Mid Triassic (approx. 225 million years ago (Ma)) [51], through Late Jurassic (approx. 145 Ma) [52], Early Cretaceous (approx. 135 Ma) [53], to the Eocene (approx. 40 Ma) [54], providing a variety of calibration points throughout the phylogeny [55]. The optimal smoothing parameter was chosen based on a cross-validation procedure [56].

(b) Diversification rates through time
To determine whether this group radiated through time at a constant rate, we first assessed whether branching times fit a pure-birth Yule model using the Monte Carlo Constant Rates Test [57]. Studies have shown that this test is sensitive to non-random sampling and missing taxa [58–60]; therefore, we chose to account for missing taxa by using birth–death chronogram simulations [58]. We simulated constant rate birth–death chronograms using CORSIM [58] assuming birth–death chronogram simulations [58–60]; therefore, we chose to account for missing taxa. At the end of a cycle, the largest change in rate is recorded with the likelihood of the model. This process is continued for additional cycles with each cycle adding one more diversification rate change to the model, while conditioning on the rate changes previously identified in earlier iterations. A χ²-test was used to compare alternative models with different numbers of rate changes with three d.f. We set the maximum number of diversification rate changes to four, while estimating a birth–death model every 1 Myr.

(c) Conservation status, phylogenetic diversity, and EDGE and HEDGE analyses
Crayfish conservation priorities were assigned first by designating IUCN Red List status to each species retrieved from the IUCN Red List of Threatened Species v. 2013.2 [12,64]. Next, we conducted an Evolutionarily Distinct, Globally Endangered (EDGE) analysis [36] and Heightened Evolutionary Distinctiveness and Globally Endangered (HEDGE) analysis [65]. The EDGE analysis ranks species according to their evolutionary distinctness by measuring the length of the branches leading to the tip taxa weighted by the number of descendants from each node, thereby calculating a probability that a species may go extinct. This calculation requires a probability of extinction which we assigned using a numerical designation associated with the IUCN Red List category [36,66]: Least Concern = 0.025, Near Threatened = 0.05, Vulnerable = 0.1, Endangered = 0.2 and Critically Endangered = 0.4. The HEDGE analysis ranks species based on their expected contribution to PD; therefore, the metric aims to preserve species that contribute the most PD. The HEDGE calculation is an extension of probabilistic PD where the probability of extinction can reach 0 [67,68]. Both EDGE and HEDGE metrics were estimated in Mesquite [69] using the Tuatara module [70].

In addition to estimating EDGE and HEDGE, we measured the PD of crayfish in each of eight terrestrial Freshwater Animal Diversity Assessment (FADA: http://fada.biodiversity.be) recognized ecozones [71] to examine broad geographical patterns of diversity and endangerment within the context of pre-established freshwater ecozones. We used the IUCN Red List [61] to gather range data and code the presence or absence of FADA ecozones for each of the 382 species in the chronogram. This information was used to estimate the phylogenetic species variability (PSV), which is a measure of PD that is independent of species richness [72]. PSV is a metric that evaluates a hypothetical independent neutral trait forward in time along a phylogeny to quantify how shared evolutionary history decreases the variance of the hypothetical trait [72]. As applied here, we are examining the variance in branch lengths as a proxy of relatedness within FADA ecozones. It is important to note that we also estimate PSV variance because our phylogeny is not clocklike; therefore, there is variation around the single PSV value. We chose to use PSV as measure of diversity and relatedness due to the variability in the number of species across regions. PSV estimates can range from 0 to 1 with 0 indicating species within a FADA ecozone are closely related and 1 indicating species within a FADA ecozone are distantly related. We calculated PSV and Faith’s PD [73] in the R package picante [74] using the chronogram without the outgroup; picante calculates PD for each FADA ecozone by summing the total length of all branches connecting species within a FADA zone on the unrooted phylogeny.

3. Results
(a) Phylogenetic synthesis
Construction of the synthetic tree and curation of the taxonomy yielded 719 terminal taxa (figure 1). The terminal taxa consist of all 590 described species of freshwater crayfish [12] and multiple representatives of species with unique GenBank identifiers from population level data. Of the 719 taxa, 387 (60% of all described species) are unique species represented with publicly available sequence data. The genus Procambarus is the largest genus with the lowest number of species with available sequence data (31 available/178 species described).
(b) Molecular phylogeny
The final nucleotide alignment contained 5259 bp of nucleotides after gBlocks trimming. The BIC favoured each partition subset as the best partitioning scheme. The ML tree strongly supported sister clades consisting of the Northern Hemisphere and Southern Hemisphere taxa (bootstrap value (BS) = 100); however, branch support only favoured monophyly for one family (figure 2). Parastacidae is strongly supported (BS = 98), while Cambaridae and Astacidae are paraphyletic (figure 2). The monophyly of Southern Hemisphere genera is mostly supported, but no support is present for the relationships among the genera (figure 2). This is similar to the Northern Hemisphere genera, but in the Northern Hemisphere the genera containing many species (e.g. Orconectes, Procambarus, Cambarus) are not supported as being monophyletic (figure 2).

(c) Diversification rates through time
The Monte Carlo Constant Rates test rejected the Yule model with a single rate of diversification through time. A three-rate birth–death model fitted the branching times the best according to the $\chi^2$-test with changes in diversification rate occurring at 20 Ma and 4 Ma (figure 3). The death parameter increased towards the present, while the birth parameter decreases and then increases at the present. Finally, the turnover rate (death/birth) is the greatest for the rate nearest the present.

(d) Conservation status, phylogenetic diversity, and EDGE and HEDGE analyses
The PD and PSV calculations of FADA ecozones differed when compared with one another owing to variation in species...
Figure 2. Chronogram estimated in MrBayes using the ML phylogeny. Node calibrations identical to Breinholt et al. [22] and Toon et al. [21]. Bootstrap support greater than 70% from 100 non-parametric bootstrap replicates shown.
parameter estimates are given for each birth–death model.

Australia (e.g. endangered, with the bulk of these species being located in
scores include species that are critically endangered and
all species. The 10 species with the highest EDGE and HEDGE
(figure 4). Calculations of EDGE ranged from 35.54 to 0.36,
species with IUCN values in our molecular phylogeny
heterogeneity (non-clocklike) among the seven taxa.

It should be noted that the PSV variance is the greatest for
more closely related, ultimately lowering the PSV value.
The low PSV can be attributed to shorter branches, thus being
diverse (there is only the Malagasy genus with seven species).

The EDGE and HEDGE scores were calculated for all
species with IUCN values and associated statistics to address conservation
questions in order to focus species-based conservation efforts on preserving the greatest amount of PD [36,65,66,75]. Species that have few close relatives have a greater responsibility for evolutionary history [73]. Thus, the combination of phylogenetic information with endangered status allows for conservation biologists to make informed decisions about conservation priorities that incorporate threat and evolutionary history and processes [76]. Despite its predictive value, no comprehensive phylogeny has been proposed for the crayfish. Crayfish suffer from severe habitat loss, and as a result a large percentage of the species are endangered [12]. A phylogenetically informed conservation plan would benefit this group of organisms immensely to prioritize species-based conservation efforts and provide overarching protection for diverse freshwater ecosystems. Here we synthesize all taxonomic and phylogenetic information to define priorities given the current available information. Finally, we describe the limitations of findings and what is needed to achieve a robust framework to inform conservation priorities in the future.

4. Discussion
Recently, phylogenies have been used in combination with IUCN values and associated statistics to address conservation questions in order to focus species-based conservation efforts on preserving the greatest amount of PD [36,65,66,75]. Species that have few close relatives have a greater responsibility for evolutionary history [73]. Thus, the combination of phylogenetic information with endangered status allows for conservation biologists to make informed decisions about conservation priorities that incorporate threat and evolutionary history and processes [76]. Despite its predictive value, no comprehensive phylogeny has been proposed for the crayfish. Crayfish suffer from severe habitat loss, and as a result a large percentage of the species are endangered [12]. A phylogenetically informed conservation plan would benefit this group of organisms immensely to prioritize species-based conservation efforts and provide overarching protection for diverse freshwater ecosystems. Here we synthesize all taxonomic and phylogenetic information to define priorities given the current available information. Finally, we describe the limitations of findings and what is needed to achieve a robust framework to inform conservation priorities in the future.

(a) Taxonomy
Taxonomy is the foundation of conservation, and without sound taxonomy conservation priorities can become misleading [76–78]. Our phylogenetic analyses, in addition to previous published phylogenies, show crayfish taxonomy is in need of review and curation. For example, both our molecular phylogeny and synthetic tree confirm multiple genera are paraphyletic (figures 1 and 2). This is concordant with previous molecular studies of North American fauna where species are not monophyletic [20] and genera are not monophyletic [18]. In addition, our dataset only supports one of three families as monophyletic (figures 1 and 2).

Two issues may be causing the paraphyly and polyphyly of families and genera. First, traditional molecular loci used in crayfish systematics may not be the best choice to resolve the relationships being estimated (discussed in the molecular data section below). The second potential cause of conflict

<table>
<thead>
<tr>
<th>ecozone</th>
<th>PSV</th>
<th>PSV variance</th>
<th>PD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australasian</td>
<td>0.72</td>
<td>1.07 x 10^{-5}</td>
<td>6025.27</td>
</tr>
<tr>
<td>Afrotropical</td>
<td>0.16</td>
<td>7.40 x 10^{-3}</td>
<td>266.54</td>
</tr>
<tr>
<td>Neotropical</td>
<td>0.59</td>
<td>2.49 x 10^{-6}</td>
<td>11026.91</td>
</tr>
<tr>
<td>Neotropical</td>
<td>0.74</td>
<td>2.19 x 10^{-4}</td>
<td>1407.77</td>
</tr>
<tr>
<td>Paleartic</td>
<td>0.63</td>
<td>8.53 x 10^{-4}</td>
<td>861.7</td>
</tr>
</tbody>
</table>

Figure 3. Log-lineage through time plot of the ML chronogram. Vertical dashed lines indicate timing of the $\chi^2$ favoured changes in diversification rate. ML parameter estimates are given for each birth–death model.
between taxonomy and molecular phylogeny may be that the morphological characters used to diagnose species are not informative about species relationships at higher taxonomic levels. The majority of taxonomic conflict occurs among the North American fauna. This group of crayfish forms a clade that most likely resulted from a recent rapid radiation, which is portrayed on a phylogeny as short branches near the terminal nodes (figure 2). Quick bursts of cladogenesis leave few morphological synapomorphies, making it difficult to establish relationships among species due to the short amount of time sister taxa were in isolation or gene flow was limited. We suggest a major reappraisal of North American crayfish taxonomy with the aid of a molecular phylogeny that capitalizes on loci from throughout the genome (see suggestions below).

The Southern Hemisphere taxonomy is not exempt from problems. The centre of diversity for the Southern Hemisphere is Australia, where the phylogeny consists of long terminal branches relative to the North American fauna [21]. The long terminal branches reflect low diversification rates in recent years, which may have resulted from the desiccation of Australia starting with the formation of the Antarctic Circumpolar Current during the Miocene and the formation of the Antarctic ice sheets and glacial cycles [79]. With the long terminal branches and increased time since speciation, the genera form monophyletic groups, unlike the Northern Hemisphere crayfish. However, published phylogenies have shown paraplectic relationships for genera such as *Euastacus* [29] and *Engaeus* [24]. Although there is limited conflict between phylogeny and taxonomy of Southern Hemisphere taxa, most of these discrepancies represent undescribed species rather than true taxonomy conflict.

As part of this study, we revised the existing taxonomy that forms the basis of the synthetic tree. Although we removed non-recognized species names from the list of all species in use in morphological and molecular studies, a larger effort is needed to describe the undescribed material and formally re-describe higher taxa according to new molecular, highly supported, phylogenetic hypotheses. This level of curation will be difficult as the number of alpha taxonomists has declined greatly in the molecular and genomics era [80], but it is desperately needed for conservation efforts.

**Figure 4.** Bivariate plot of the largest EDGE and HEDGE scores for taxa included in the molecular phylogeny. Calculations were performed in the *Mesquite* module.

(b) Crayfish phylogeny

Our effort here to reconstruct the crayfish phylogeny with publicly available sequence data shows the immense historical effort needed to reconstruct the evolutionary relationships. Despite this great effort, many generic and intergeneric relationships remain uncertain. This can be seen in our synthetic tree and supermatrix we assembled for this study (figures 1 and 2, respectively). We have sampled approximately 60% of all crayfish species. Unfortunately, from a conservation priority standpoint, our confidence in conservation priorities estimated in this study is limited by the percentage of sampled species. The EDGE and HEDGE calculations all rely on a fully resolved and sampled phylogeny with branch lengths. Our phylogeny contains 60% of the known species; therefore, missing 40% of the known species may drastically overestimate EDGE and HEDGE calculations owing to the missing taxa not shortening the edge lengths of terminal taxa. EDGE and HEDGE analyses have primarily focused on the mammals, where there are estimates of a complete phylogeny. Although the mammal phylogeny is nearly complete, researchers have dealt with missing taxa by placing them as best they can on the phylogeny [81]. While we can place these taxa according to taxonomy (and have done so in our synthetic tree, figure 1), we do not have branch length information for these data which is critical to the EDGE/HEDGE calculations. Furthermore, many of the genera in the Northern Hemisphere crayfish are paraplectic, making the placement of species based on taxonomy suspect. Therefore, our EDGE and HEDGE estimates should be taken lightly until a more robust phylogeny and taxonomy are assembled (work in progress).
Traditionally in crayfish phylogenetics, five core genes are sequenced: COI, 12S, 16S, 18S and 28S [82]. Although providing a well-supported backbone for many groups (figure 2), this set of loci in combination has failed here to resolve many of the relationships within and among genera and families. We are at a time in molecular systematics where genomic data are being generated at substantial pace and systematics can begin to look across the genome for loci with a modest amount of work. Three techniques are currently being used to target single-copy nuclear genes across the nuclear genome: highly conserved regions [83], ultraconserved regions [84] and transcrip- toms [85]. Each of these methods produces hundreds to thousands of loci informative across time scales. The crayfish phylogeny would benefit greatly from a study that targets one of these sources for loci to estimate and confirm the phylo- genetic backbone and add support to the relationships within and among genera.

(c) Diversification rates and extinction
Recent advances in comparative methods have allowed researchers to estimate extinction from phylogenies [86], although some are sceptical [87]. Our diversification analyses show diversification rates were not constant through time (figure 3). In fact, the best-fit birth–death model for our data supports three different time-dependent diversification rates. The timing of changes in diversification are fairly congruent with the timing of the formation of the Antarctic Circumpolar Current and the formation of ice sheets [88] approximately 20 Ma, while the most recent change in diversification rate corresponds to recent glaciation in the Northern Hemisphere [89]. Future crayfish divergence time estimation should attempt estimation in a Bayesian framework. Unfortunately, getting convergence of parameter estimates for a large tree is difficult; therefore, we have relied on point estimates here as an approximation and recognize the variance associated with these dates.

Although we give a general picture of historical diversifi- cation and extinction here on a geological time scale, future studies can leverage a more complete tree and other compara- tive methods to obtain diversification and extinction rates on a local time scale. For example, new binary and multi-state character models of diversification rates exist [90,91]. We envision using these models in conjunction with environmental data layers to look at diversification and extinction rates of taxa associated with certain types of habitat and natu- ral history characteristics. Analyses that combine ecology, morphology, geography and phylogeny can provide power- ful correlative evidence with high extinction rates that are associated with particular characters.

Acknowledgements. We thank the Royal Society and the bioGENESIS working group for organizing and supporting the ‘Phylogeny, extinction risks and conservation’ workshop at the Royal Society. We are grateful to the George Washington University Colonial One high perfor- mance computing cluster for providing analysis capacity for this work. We thank Joseph Brown for his help with the synthesis analy- ses. We thank two anonymous reviewers and the associate editor for providing helpful comments to improve our manuscript.

Funding statement. This work was supported by the US NSF grant no. DEB 13-01820.

References


