12-1-2016

Sequence capture and next-generation sequencing of ultraconserved elements in a large-genome salamander

Catherine E. Newman
Louisiana State University

Christopher C. Austin
Louisiana State University

Recommended Citation

Newman, C., & Austin, C. (2016). Sequence capture and next-generation sequencing of ultraconserved elements in a large-genome salamander. *Molecular Ecology, 25* (24), 6162-6174. https://doi.org/10.1111/mec.13909

Received Date : 06-Jun-2016
Revised Date : 26-Oct-2016
Accepted Date : 01-Nov-2016
Article type : Original Article

Sequence capture and next-generation sequencing of ultraconserved elements in a large-genome salamander

Catherine E. Newman[1,2,3], Christopher C. Austin[1,2]

[1] Museum of Natural Science, Louisiana State University, 119 Foster Hall, Baton Rouge, Louisiana 70803, USA
[2] Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, USA
[3] Corresponding author. Email: newma014@gmail.com, Fax: +1-225-578-3075

Keywords: phylogeography; *Plethodon serratus*; population genetics; species delimitation

Running title: UCEs for salamander phylogeography

Abstract

Amidst the rapid advancement in next-generation sequencing (NGS) technology over the last few years, salamanders have been left behind. Salamanders have enormous genomes – up to 40 times the size of the human genome – and this poses challenges to generating NGS data sets of quality and quantity similar to those of other vertebrates. However, optimization of laboratory protocols is time-consuming and often cost prohibitive, and continued omission of salamanders from novel phylogeographic research is detrimental to species facing decline. Here, we use a

salamander endemic to the southeastern US, *Plethodon serratus*, to test the utility of an established protocol for sequence capture of ultraconserved elements (UCEs) in resolving intraspecific phylogeographic relationships and delimiting cryptic species. Without modifying the standard laboratory protocol, we generated a data set consisting of over 600 million reads for 85 *P. serratus* samples. Species delimitation analyses support recognition of seven species within *P. serratus* sensu lato, and all phylogenetic relationships among the seven species are fully resolved under a coalescent model. Results also corroborate previous data suggesting non-monophyly of the Ouachita and Louisiana regions. Our results demonstrate that established UCE protocols can successfully be used in phylogeographic studies of salamander species, providing a powerful tool for future research on evolutionary history of amphibians and other organisms with large genomes.

Introduction

Salamanders have the largest genomes of any tetrapod, ranging from 9.9 gigabases (Gb) to 118 Gb (Animal Genome Size Database, http://www.genomesize.com). The median salamander genome size is 29 Gb, or roughly 8 times the size of the human genome. For comparison, the following are median genome sizes of other tetrapod groups: Anura (4.6 Gb), Aves (1.3 Gb), Mammalia (3.0 Gb), Reptilia (2.1 Gb). Because of cost and computational limits imposed by large genome sizes and the highly repetitive elements that make them large, no salamander genome has been fully sequenced to date. Salamander genome size is primarily due to an unusually large amount of highly repetitive transposable DNA (Sun *et al.* 2012), which poses challenges for next-generation sequencing (NGS) methods and genome assembly.

Repetitive sequences reduce the final sequence coverage of on-target reads by hybridizing to each other during the library preparation step in the laboratory protocol. In sequence-capture methods of NGS, DNA is randomly sheared, and probes targeting specific sequences are subsequently hybridized to the fragments. But DNA fragments are frequently longer than the probes, leaving single-stranded end(s) after hybridization. Because of the high density of repetitive sequences in the fragment pool, if these single-stranded ends correspond to repetitive regions, there is a higher probability that they will hybridize to other repetitive sequences and carry them through to sequencing (Hodges *et al.* 2009). This increases the number of off-target reads that will ultimately be discarded in downstream data analysis. A recent exon-

capture study with *Ambystoma* salamanders found that increasing the amount of species-specific c_0t-*1* blocker – added during library preparation to help prevent repetitive elements from hybridizing to the targeted sequences – increased the number of unique reads mapping to targets by decreasing the number of PCR duplicates (McCartney-Melstad *et al.* in press). While these results are promising, implementing these methods requires time-consuming modification of already well-established laboratory protocols. No study has used proven sequence-capture protocols to generate a phylogenetic or phylogeographic data set for a large number of individuals of a single salamander species.

The rationales for including multiple loci in studies of intraspecific systematics and evolutionary biogeography have been well established and include attaining the statistical power to resolve discordance among independent loci and estimation of demographic parameters (Edwards 2009; Knowles 2009). However, the field of systematics continues to face hurdles to adopting NGS methods for several reasons, including the focus on non-model organisms, the need for sequencing large numbers of samples, and the lack of guidance for choosing the most appropriate library preparation methods for particular questions (McCormack *et al.* 2013). This is particularly true for research on salamanders, whose large genomes make it especially difficult to design effective methods of genome reduction that would yield adequate sequencing coverage. Ultraconserved elements (UCEs), or regions of the genome that are highly conserved across deep time scales, are promising targets for sequence capture in salamanders because they do not include repetitive elements, and probe design does not require a genome from the species of study (Faircloth *et al.* 2012).

Salamanders are vital to our understanding of biogeography in the southeastern United States (US) – a region extremely rich in salamander biodiversity. The Appalachians alone are home to 76 species, or 14% of the global salamander diversity, and nearly half (35 species) of the species occurring in the Appalachians are endemic (Stuart *et al.* 2004; Gratwicke 2008). However, because amphibians are sensitive to even small changes in habitat, many species are suffering declines. The threats to amphibians are numerous and include habitat loss and contamination, overexploitation, climate change, and infectious disease (Lips *et al.* 2008). The foundation for research on amphibian declines is a solid understanding of the interplay of factors that shape amphibian geographic distributions in the region. Knowledge of the structure and

patterns of genetic variation within species is an essential component of understanding the historical and ecological processes driving distributions and, ultimately, species formation and decline.

In the Southeast, at least 14 of 144 amphibian species have disjunct distributions comprised of two or more isolated regions (Dorcas & Gibbons 2008; Mitchell & Gibbons 2010). Of those 14 species, 13 are salamanders. Because species with disjunct distributions are underrepresented in the literature, little is known about the phylogeographic history of these species. In the last few years, NGS technologies have facilitated large-scale genomic studies of model and non-model organisms, generating multilocus data sets comprised of hundreds to thousands of loci. But despite their importance as indicators of environmental health, salamanders have been omitted from these innovative studies in large part due to their enormous genomes, which present significant challenges to NGS methods (Keinath *et al.* 2015). Salamanders of the Southeast are therefore uniquely set to simultaneously serve as a proof-of-concept for using NGS methods with organisms with large genomes and serve as a case study for evolutionary studies of species with disjunct distributions.

The salamander *Plethodon serratus* is one of the 14 species in the Southeast with a disjunct geographic distribution. *Plethodon serratus* occurs in four isolated regions in the southeastern US (Fig. 1): central Louisiana, the Ouachita Mountains, the Ozark Mountains, and the Appalachian Mountains and Piedmont Province (Conant & Collins 1998). Within the Appalachian/Piedmont region, a separate subspecies in the Piedmont was historically recognized (*P. cinereus polycentratus*) based on morphology. This subspecies also included a sample from the Valley and Ridge province of the Appalachians in northwestern Georgia (Highton & Grobman 1956, Highton & Webster 1976). For simplicity and consistency, we continue to refer to the easternmost region of the *P. serratus* range in its entirety as "Appalachian region," and we refer to the Appalachian Mountain populations excluding Valley/Ridge as "Appalachian Highland."

In Louisiana, *P. serratus* is listed as Critically Imperiled because only three populations are known to exist in the state. *Plethodon serratus* is a fully terrestrial, mid-elevation species; throughout most of its range, the species is found on slopes between 100-800 m, but the sites in Louisiana are lower elevation, 40-100 m. *Plethodon serratus* has been recorded at elevations as

high as 1,700 m in the Appalachians (Huheey & Stupka 1967), where it is restricted to higher elevations in areas where the congener *P. ventralis* is present at lower elevations (Highton 1971). Our previous study on the phylogeographic relationships within *P. serratus* used traditional Sanger sequencing of one mitochondrial gene, five protein-coding nuclear genes, and one anonymous nuclear locus (Newman & Austin 2015). The study clarified relationships among populations and geographic regions to some extent but was unable to fully resolve the relationships among the Louisiana and Ouachita populations.

The genome size of *P. serratus* ranges from 19-24 gigabases (Gb; median = 20 Gb) (unpublished data), or slightly below the 29 Gb median for salamanders. The genome size of *P. cinereus*, a close relative of *P. serratus*, has been reported in several studies, ranging from 20-26 Gb (Mizuno & Macgregor 1974; Olmo 1974; Horner & Macgregor 1983; Sessions & Larson 1987; Licht & Lowcock 1991; Mueller *et al.* 2008).

The goals of the current study were twofold: (1) to determine whether an established sequence capture method with UCEs is an appropriate and effective method for generating substantial multilocus data sets for intraspecific phylogeography of large-genome organisms, and (2) to resolve the phylogenetic and population genetic relationships of *P. serratus* and test models of species delimitation.

Materials and Methods

Sequence capture of ultraconserved elements

We sampled 94 individuals of *P. serratus* and two individuals of one of its closest relatives *P. cinereus* (Fisher-Reid & Wiens 2011). Samples were from 27 localities distributed across the entire species range. Tissue samples were collected by hand or obtained from museums as part of a previous study (Newman & Austin 2015). For most samples, we used genomic DNA extracted as part of the previous study (Newman & Austin 2015), and we extracted genomic DNA from the remainder of samples following the same protocol. DNA extracts containing between 0.5 μg and 19 μg (average: 6 μg) of DNA at a concentration between 16 ng/μL and 270 ng/μL (average: 100 ng/μL) were sent to RAPiD Genomics (Gainesville, Florida) for library preparation, sequence capture, and sequencing. Samples were barcoded using standard Illumina TruSeq adapters with a unique 8 bp index for each individual. Our probe set

consisted of 2,064 probes targeting 1,745 UCE loci, a subset of the full tetrapods UCE probe set of 5,472 probes (see Faircloth *et al.* 2012 and http://ultraconserved.org for probe development). We searched the 5,472 probes in the full set against the only amphibian whole genome sequence database currently available, *Xenopus tropicalis*, using the NCBI nucleotide BLAST tool (http://ncbi.nlm.nih.gov) and retained the 2,064 probes that matched a segment of the *X. tropicalis* genome with an identity of >85% and a sequence length ≥100 bp. Enriched libraries were sequenced in a 100 bp paired-end run on a single lane of an Illumina HiSeq 2500. We received demultiplexed raw reads from RAPiD Genomics.

We filtered demultiplexed reads using a custom pipeline, Illumiprocessor (http://github.com/faircloth-lab/illumiprocessor), that incorporates Trimmomatic (Bolger *et al.* 2014) to remove adapter sequences, low quality ends, and ambiguous bases. Reads were assembled *de novo* using Trinity v.2.0.6 (Grabherr *et al.* 2011) in the software package Phyluce v.1.5 (Faircloth 2016). We also used Phyluce to filter assembled contigs for enriched UCE loci and generate sequence alignments for each locus using MAFFT v.7.130b (Katoh & Standley 2013). We completely removed from further analyses 9 individuals for which ≤15% of loci were successfully enriched. The remaining 87 individuals comprised the "all samples" set. To compare the effects of the amount of missing data versus number of loci on phylogeny estimation, we utilized two sets of individuals (= samples) (see Fig. 2): all individuals (87 individuals, "all samples" data set) and individuals with ≥1,000 loci (60 individuals, "1k samples" data set). The 1k samples data set included representatives from all major geographic areas and all major clades of the mitochondrial phylogeny generated in a previous study (Newman & Austin 2015). For both sets of individuals, we generated two sets of alignments, allowing 20% or 40% of the individuals to have missing data for each locus. For each of the four data sets, we determined the number of parsimony informative sites using Phyluce.

Likelihood analyses of concatenated loci

For each of the four data sets, we generated a concatenated alignment of all loci. We conducted a maximum-likelihood (ML) analysis of each concatenated alignment in RAxML v.8.2.0 (Stamatakis 2014), partitioning each alignment by locus and assigning each partition a GTR-GAMMA model of evolution. Nodal support was assessed with 1,000 rapid bootstrap

pseudoreplicates. Preliminary runs of normal (non-rapid) bootstraps yielded results qualitatively identical to the rapid bootstrap analyses (data not shown); rapid bootstraps were thus used with the full data sets to minimize computational time.

Cluster analyses

To generate SNP data sets for cluster analyses, we used a custom pipeline (http://github.com/mgharvey/seqcap_pop) that incorporates several programs as follows. Reads for each of the 87 individuals were mapped back to an index of consensus contigs in BWA v.0.7.8 (Li & Durbin 2009). We used SAMtools v.0.1.19 (Li *et al.* 2009), Picard (http://broadinstitute.github.io/picard), and Phyluce to generate BAM pileups, mark PCR duplicates, and prepare files for next steps. Using GATK (McKenna *et al.* 2010; DePristo *et al.* 2011; Van der Auwera *et al.* 2013), we called SNPs and indels, filtered low quality variant calls (QUAL <30.0), and phased SNPs. We then used custom Python scripts to convert VCF files of phased SNPs to formats suitable for downstream analyses. Because treating all SNPs as independent loci regardless of actual linkage may sometimes mislead clustering algorithms, we generated three SNP data sets: all SNPs from each locus, one random SNP from each locus, and the first SNP from each locus.

We assessed population structure across the species range using two approaches. First, we ran a Bayesian clustering analysis in Structure v.2.3.4 (Pritchard *et al.* 2000; Falush *et al.* 2003), implementing an admixture model (Pritchard *et al.* 2000), assuming correlated allele frequencies (Falush *et al.* 2003), and using sampling locality as prior information. For each *K* (number of clusters) from 1 to 10, we ran 20 iterations of 100,000 generations after a burn-in of 10,000 generations. We determined the best estimate of *K* by assessing the rate of change in log likelihood values between successive values of *K* (Evanno *et al.* 2005) through the Structure Harvester web server (Earl & vonHoldt 2011). We then combined all iterations of the best *K* in CLUMPP (Jakobsson & Rosenberg 2007) under a greedy algorithm to determine the most likely set of cluster membership coefficients.

Because Structure requires some level of subjectivity to determine the most appropriate value of *K* (Meirmans 2015), we also performed a cluster analysis in BAPS v.6.0 (Corander *et al.* 2003, 2008; Corander & Marttinen 2006). Unlike Structure, which uses an MCMC-based

algorithm, BAPS implements a stochastic optimization algorithm, which dramatically reduces computational time, especially for large data sets. We ran a "clustering of individuals" analysis for each of the three data sets (all SNPs, random SNPs, first SNPs), followed by an "admixture of individuals" analysis. For each clustering run, the maximum *K* was set at 10 and 15, running 10 iterations for each maximum *K*, for a total of 20 runs per data set. The admixture analysis estimates ancestry coefficients for each individual, assigning each individual to one of the *K* clusters from the cluster analysis. We set the admixture analysis to only recognize clusters with ≥2 individuals. The admixture analysis was run for 100 iterations, with 200 reference individuals per cluster.

Bayesian species delimitation

We used a Bayesian approach implemented in BPP v.3.2 (Yang & Rannala 2010, 2014) to simultaneously delimit species and infer the species phylogeny under a coalescent framework. BPP gives posterior probabilities for various models of numbers of species and for various species trees. We *a priori* assigned individuals to species/taxa according to the *K*=7 clustering scheme estimated by BAPS (see Results). To minimize the amount of missing data, we used the 1k samples data set with each alignment allowing up to 20% of individuals to have missing data. Analyses were run with two data sets: one data set consisting of the 70 most informative loci, and a second data set consisting of 70 loci randomly selected from the full set of informative loci.

We ran the analysis under six combinations of priors for ancestral population size (θ) and divergence time at the root of the species tree (τ) (Table 1). For each set of priors, four independent runs with different starting seeds were completed, each with a burn-in of 10,000 iterations and sampling every five iterations for a total of 200,000 iterations. Results from the independent runs were compared to ensure convergence. Because the runs with 70 random loci did not converge after 200,000 iterations (see Results), we ran the same analyses with that data set for an additional 300,000 iterations (total: 500,000).

Species tree inference in a coalescent framework

We jointly estimated gene trees and species trees in a coalescent framework using *BEAST (Bouckaert *et al.* 2014). As in the BPP analysis, we used the 1k samples data set with a maximum of 20% missing individuals per alignment, and we also used the same set of individuals and *a priori* species/taxa assignment as in BPP, with the exception of adding the outgroup taxon *P. cinereus*. Because running *BEAST with all loci is computationally intensive enough to be infeasible, we ran the analysis with nine sets of loci selected from all loci that include at least one member of the outgroup (*P. cinereus*) and each of the seven *a priori* species: the 20 most informative, the 50 most informative, the 70 most informative, the 100 most informative, and five sets of 70 loci randomly selected from the full set of informative loci (one of the five sets of random loci was the same data set used in the BPP analysis). The best-fit model of sequence evolution for each locus was estimated in jModelTest v.2.1.4 (Posada 2008). To reduce computational time, we followed Smith, et al. (Smith *et al.* 2014) in minimizing the number of parameters to be estimated by assigning an HKY model of evolution to loci assigned by jModelTest to a GTR model. We set all HKY models to use empirical base frequencies. We applied a strict molecular clock with rate fixed at 1.0 and a Yule species tree prior with a linear-with-constant-root population size model. For each data set, we ran 1 billion generations, sampling every 10,000 generations. We assessed MCMC convergence and determined appropriate burn-in by examining likelihood traces in Tracer v.1.6 (Rambaut & Drummond 2007) and ensuring all ESS values were ≥200.

Results

After quality filtering with Illumiprocessor, we obtained a total of 600 million reads for 85 *P. serratus* and 2 *P. cinereus* samples (1.6-16.5 million reads per sample), for a total of 58.5 billion base pairs. For 9 of the original 94 *P. serratus* samples, fewer than 15% of loci were successfully sequenced, and we excluded those individuals from further analysis (see Table S1). We assembled reads for the remaining 87 samples into an average of 16,176 contigs per sample (range: 4,332-35,038). Between 256-1,335 contigs per individual (average: 1,043) were aligned to UCE probes, with trimmed alignment length averaging 628 bp (range: 112-1562). Average read depth per contig per individual ranged from 2.2-42.6 times, and average read depth across all individuals was 14.7 times. The 1,517 alignments contained varying numbers of individuals

(average: 60, range: 3-79), and no alignment contained all 87 individuals. Five alignments contained no variation, and 58 alignments contained no parsimony informative sites. The remaining 1,459 informative loci contained an average of 9.5 parsimony informative sites (range: 1-79) (Fig. S1). As expected, the frequency of variant bases tended to be low in the center of the alignments – the conserved regions – and increased in the flanking regions (Fig. S2). After filtering out alignments exceeding the maximum allowed number of individuals with missing data (20%, 40%), the final four data sets contained between 321-1,387 alignments with 36-69 individuals per alignment (Table 2).

Likelihood analyses of concatenated loci

Phylogenies of the four concatenated data sets (Fig. 3) were consistent with our expectations based on our previous study with mitochondrial and nuclear data (Newman & Austin 2015). All four analyses recovered, with strong nodal support (bootstrap ≥ 99), the Appalachian Highland and Valley/Ridge salamanders as sister to each other (forming the clade we refer to as "Appalachian"), the Appalachian clade as sister to all other *P. serratus*, and the Ozark salamanders as sister to the Ouachita and Louisiana *P. serratus*. The Louisiana region and the Ouachita region were both always non-monophyletic. Three of the analyses recovered a clade containing eastern Ouachita salamanders that was sister to the clade containing Louisiana and western + central Ouachita, with moderate to high support (bootstrap = 79-100). Within the Louisiana + western/central Ouachita clade, Sicily Island (Louisiana) was sister to a strongly supported (bootstrap ≥ 98) clade containing the Kisatchie (Kisatchie National Forest, Louisiana) and western/central Ouachita salamanders.

Within the Kisatchie + western/central Ouachita clade, topology and nodal support varied across data sets. All analyses recovered Kisatchie as sister to the western + central Ouachita *P. serratus*. In the trees from the all samples data sets, the western + central Ouachita clade always excluded the single individual from Fodderstock Mtn. (that individual, OMNH 41642, was not present in the 1k data set). The individual from Fodderstock Mtn. instead was always placed sister to a clade containing central + remainder of western Ouachita (bootstrap: 74, 96). The central Ouachita populations (Fourche Mtn., Buck Knob) consistently fell out in a strongly supported clade (bootstrap ≥ 95). In contrast, western Ouachita populations were never

monophyletic. Nodal support for clades containing western Ouachita populations (Beavers Bend, Rich Mtn., Kiamichi Mtn., Winding Stair Mtn., Black Fork Mtn., Foran Gap, Iron Mtn.) varied across data sets.

Cluster analyses

Structure and BAPS consistently recovered a *K* of 7 across all data sets (Fig. 3). Both algorithms recovered the Appalachian Highlands + Valley/Ridge, Ozarks, Sicily Island, and Kisatchie as distinct clusters. The Ouachita region contained three clusters: an eastern cluster consisting of South Fourche, Ouachita Trail, Petit Jean Mtn., Mount Nebo, Caddo Gap, and County Rd. 240; a central cluster consisting of Fourche Mtn. and Buck Knob; and a western cluster consisting of all of the remaining populations. Both Structure and BAPS assigned all individuals from the Mount Nebo and Fodderstock Mtn. populations to more than one cluster. In addition, Structure, but not BAPS, assigned all individuals from Beavers Bend, Iron Mtn., and Caddo Gap to multiple clusters.

Species delimitation

We used the *K*=7 clustering scheme to *a priori* assign individuals to species for the BPP analysis. All BPP analyses converged on a model of seven species (posterior probability [*P*] = 1.0); no run collapsed any two or more species. Of the runs with the 70 most informative loci, only runs with small ancestral population size (theta = 2,2000) converged on the same species tree topology recovered by *BEAST (see Results below) with *P* >0.98 (Table 1, Fig. 4). All runs with the random loci data set failed to converge on a single topology with high support, even after an additional 300,000 iterations.

Species tree reconstruction

Species tree topologies (Fig. 4) were consistent across the data sets of 20, 50, and 70 most informative loci and concordant with the concatenated ML phylogenies. The 100 most informative loci data set and all random loci data sets failed to converge after 1 billion generations. Increasing the number of loci in the *BEAST analysis increased nodal support for the remainder of the clades, as follows. With the 20 most informative loci, the only clade within

P. serratus with strong support was western + central Ouachitas ($P = 0.99$). With the 50 most informative loci, all nodes were strongly supported ($P \geq 0.9$) except the node placing Sicily Island sister to Kisatchie + western/central Ouachitas ($P = 0.60$), leaving the relationship among the Louisiana and Ouachita lineages unresolved. With the 70 most informative loci, all nodes were resolved with strong support, placing Sicily Island as sister to Kisatchie + western/central Ouachitas with $P = 0.92$, and the eastern Ouachitas as sister to Sicily Island + (Kisatchie + west/central Ouachitas) ($P = 1.0$).

Because all of the random loci data sets failed to converge, we examined the average number of parsimony informative sites in each data set and ran an analysis of variance (ANOVA) comparing the number of informative sites in each 70-locus data set (most informative and random). The average number of informative sites in the data sets with the most informative loci ranged from 19.25-27.05, whereas the average number of informative sites in the random loci data sets ranged from 6.93-8.26. The random loci data sets all had significantly fewer informative sites than the data set of 70 most informative loci ($p < 0.001$). This result likely explains the lack of convergence in the *BEAST and BPP analyses using the random loci data sets. For comparison, the six nuclear loci in our previously published Sanger data (Newman & Austin 2015) set had 9-30 informative sites.

Discussion

Studies of amphibian phylogenetics and phylogeography have largely failed to adopt new methods using next-generation sequencing technology (but see McCartney-Melstad *et al.* in press; O'Neill *et al.* 2012; Barrow *et al.* 2014; Wielstra *et al.* 2014; Peloso *et al.* 2016). This is especially true for salamanders, with genome sizes ranging from ~14-120 Gb (Sun *et al.* 2012), an order of magnitude greater than the size of the human genome. Large genome size limits not only the cost feasibility of whole genome sequencing, but also the methods of genome reduction in library preparation that would potentially yield sufficient depth of sequence coverage from massively parallel sequencing. Our study is one of the first to use UCEs to generate a genome-scale, population level data set for an amphibian species to examine intraspecific phylogenetic relationships. One of the most exciting conclusions from our study is that the standard UCE protocol with a salamander species yielded not only a large amount of high quality data, but also

data with enough variation to fully resolve presumed intraspecific relationships. While other NGS methods are more commonly used to delimit cryptic species (e.g., Rittmeyer & Austin 2015), UCEs specifically target highly conserved regions of the genome and thus are most often applied to deeper timescales. However, UCEs have been previously demonstrated to be useful in addressing intraspecific questions in Neotropical bird species (Smith *et al.* 2014). But the question remained of whether or not the UCE data for a large-genome species would yield enough coverage depth and number of high quality reads to resolve relationships among populations that previous data showed to be very closely related.

The delimitation of species and topology of phylogenetic relationships among populations and species were consistent across multiple data set configurations with varying levels of missing data. Species trees generated from the 20 and 50 most informative loci left some relationships unresolved, whereas the species tree from the 70 most informative loci was fully resolved, with all nodes strongly supported. Interestingly, the topology and nodal support of our previously published species tree generated from six nuclear loci (Newman & Austin 2015) matched that of the tree from 50 UCE loci in this study, which left relationships among the Ouachita and Louisiana populations unresolved. Our results indicate a trade-off between increasing the number of loci and decreasing the average information content per locus, as the *BEAST run with the 100 most informative loci never converged. Contrary to the concern that UCEs are generally not as variable as nuclear loci commonly included in Sanger data sets, our most informative UCE loci were more variable than our Sanger loci, and it is likely that the large increase in number of loci along with the small increase in average number of informative sites per locus underlies the successful generation of a fully resolved species tree with UCE loci. Again we emphasize that these data were obtained without any laboratory testing or protocol modifications and without a reference genome.

Our results strongly support seven genetically distinct species within *P. serratus* sensu lato, corresponding to distinct geographic areas: the Ozarks, the Appalachians, Sicily Island, Kisatchie, and three apparently allopatric regions in the Ouachitas. Results also strongly support the non-sister relationship of the two Louisiana sites and the non-monophyly of the Ouachita populations, confirming results from our previous study that analyzed both mitochondrial and nuclear genes (Newman & Austin 2015). In contrast, an earlier study that included only

388 mitochondrial genes (Thesing *et al.* 2015) found the Louisiana populations to be sister to each
389 other. That study also showed the Appalachians to be contained in a clade that included Ozarks +
390 eastern Ouachitas and excluding Louisiana + western Ouachitas. The discordance between the
391 topology shown in (Thesing *et al.* 2015) and our study may be due, at least in part, to
392 mitochondrial gene duplication and rearrangement, as is known to occur in some plethodontid
393 salamanders (Chong & Mueller 2013). Our analyses also do not support recognition of the
394 Valley/Ridge/Piedmont populations as a species distinct from the Appalachian Highlands.
395 Further analyses should include samples from the Piedmont province to confirm this conclusion.
396 In the concatenated ML analyses, the lower nodal support for the western Ouachitas clade
397 in the phylogenies from the all samples data set is potentially due to a larger number of
398 individuals of mixed ancestry between the western Ouachitas, central Ouachitas, and Kisatchie.
399 Structure plots suggest considerable gene flow between Beavers Bend – which is the
400 southernmost population in the Ouachitas – and Kisatchie, and between Fodderstock and Iron
401 Mtns. and the central Ouachitas, so it would not be surprising that phylogenetic analyses are
402 unable to resolve relationships among those individuals with strong support. The Fodderstock
403 Mtn. individual in particular caused difficulty, always falling outside the western + central
404 Ouachitas clade in the concatenated analyses. In the Structure analysis, this individual was
405 assigned to both the central and western Ouachita clusters with nearly equal posterior
406 probabilities (PP = 0.397 and 0.443, respectively). For the species tree analyses, admixed
407 individuals were excluded, and we were able to obtain a fully resolved phylogeny. It is possible
408 that inclusion of data from additional samples from the Fodderstock Mtn. locality would better
409 resolve the delineation between the western and central Ouachitas clade; however, it is also
410 possible that the salamanders at that locality are true hybrids of the two lineages.
411 As noted in a previous study (Thesing *et al.* 2015), the location of lineage breaks in the
412 Ouachitas appears to correspond to the divide between eastward- and westward-draining rivers.
413 Populations in the eastern Ouachita clade are located along the eastward-draining Arkansas and
414 Ouachita River systems, whereas most of the populations in the western Ouachita clade are
415 located along the westward-draining Red River system. However, three western populations
416 (Rich Mtn., Black Fork Mtn., and Foran Gap) are located along eastern-draining rivers, as are the
417 central Ouachita populations. Because *P. serratus* is a direct developing terrestrial salamander

without an aquatic larval stage, its population structure is likely to be less strongly influenced by drainages. A similar pattern of association with river drainages in the Ouachitas was seen in the congeneric species *P. caddoensis* (Shepard & Burbrink 2011).

Of particular interest is the non-sister relationship of the two Louisiana sites and the apparent isolation of Sicily Island relative to Kisatchie. Despite its name, Sicily Island is not an actual island. Rather, the hills that comprise Sicily Island are surrounded by floodplain, giving it the appearance of an island on topographic maps (Fig. 5). While it is possible that the site may become temporarily surrounded by water during major flooding events, it is unknown how often such events occur or how the temporary isolation might affect *P. serratus* vagility. Perhaps more likely is that the floodplain surrounding Sicily Island inhibits movement of *P. serratus*, as the Louisiana salamanders are generally found on slopes. It is possible that the Sicily Island population is a relict that became geographically isolated soon after the colonization of Louisiana from the Ouachitas, while the Kisatchie population was able to maintain gene flow with the western Ouachitas for a substantially longer period of time. Ongoing gene flow between the Kisatchie and western Ouachitas is unlikely, given the vast geographic distance separating the regions and the unsuitable habitat in that intervening area.

Taxonomy and conservation

The tree-based analyses in this study suggest *P. serratus* sensu lato is comprised of seven major genetic lineages, and BPP analyses strongly support recognition of those lineages as distinct species. In addition, average pairwise mitochondrial sequence divergences among the seven lineages are similar to divergences among currently recognized *Plethodon* species (Highton *et al.* 2012), and the lineages occupy distinct environmental space (Newman & Austin 2015). The lineages are allopatric, separated by unsuitable habitat, with little to no opportunity for intergradation. Based on a general lineage species concept (De Queiroz 2007), we argue that the entirety of our data support elevating the six previously unrecognized lineages to species and recircumscribing *P. serratus* sensu stricto to only the lineage that includes the type locality (western Ouachitas, Rich Mtn.). Formal diagnoses and descriptions are forthcoming in a separate publication.

Plethodon serratus is listed as Critically Imperiled by the Louisiana Department of Wildlife and Fisheries. It is known from only three localities in the state: Sicily Island Hills WMA, the Longleaf Vista Outlook in the Kisatchie National Forest, and a recently discovered second Kisatchie locality 9 km straight-line distance away from the historical Longleaf Vista Outlook site (Newman & Austin 2015). Our results from UCE data analysis confirm the isolation and genetic uniqueness of the Sicily Island population in particular and highlight the need for extensive genetic study of other amphibian species in the region with similar geographic distributions, especially species that have ranges that span large geographic areas but also have small isolated populations. *Plethodon serratus* has historically been locally abundant in all regions except Louisiana, and consequently, its lack of recognition as a species of conservation concern above the state level gives a false sense of security. We emphasize that the loss of either of the two Louisiana populations would result in the loss of a substantial amount of genetic diversity within *P. serratus*. The Louisiana populations are by far the southernmost populations of *P. serratus*, and their habitat is the warmest part of the species range. Ecological niche modeling results in our previous study (Newman & Austin 2015) suggested that *P. serratus* responds to warming temperatures with range contraction, so we underscore the need for continued protection of the Louisiana *P. serratus* in the future.

Acknowledgements

We thank Benjamin D. Thesing for providing the majority of specimens included in our study. We also thank the following institutions for tissue samples: Museum of Vertebrate Zoology, Sam Noble Oklahoma Museum of Natural History, Sternberg Museum of Natural History, University of Alabama Herpetology Collection, and Louisiana State University Museum of Natural Science. In addition, we would also like to thank Brant Faircloth and Michael Harvey for bioinformatics help and comments, and Eric Rittmeyer and Jeffrey Weinell for assistance with collecting in the field. Finally, we thank the joint lab meeting group led by Jeremy Brown for extensive discussion leading to improvement of the manuscript. All tissue collection by the authors was conducted under the following state collecting permits for Louisiana: Scientific Collecting Permits LNHP-13-036 and LNHP-14-010 and Wildlife Division Special Use Permit to Conduct Research on WMAs #WL-Research-2013-05. Collecting by the authors was

conducted in accordance with IACUC protocol 13-060 approved at Louisiana State University.

This work was funded by National Science Foundation grants DEB 1405665 to CEN and CCA and DEB 1146033 to CCA.

References

Van der Auwera GA, Carneiro MO, Hartl C *et al.* (2013) From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, **43**.

Barrow LN, Ralicki HF, Emme SA, Lemmon EM (2014) Species tree estimation of North American chorus frogs (Hylidae: Pseudacris) with parallel tagged amplicon sequencing. *Molecular Phylogenetics and Evolution*, **75**, 78–90.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 1–7.

Bouckaert R, Heled J, Kühnert D *et al.* (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. (A Prlic, Ed,). *PLoS Computational Biology*, **10**, e1003537.

Chong RA, Mueller RL (2013) Low metabolic rates in salamanders are correlated with weak selective constraints on mitochondrial genes. *Evolution*, **67**, 894–9.

Conant R, Collins JT (1998) *A Field Guide to Reptiles and Amphibians: Eastern and Central North America*. Houghton Mifflin Harcourt.

Corander J, Marttinen P (2006) Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology*, **15**, 2833–43.

Corander J, Marttinen P, Sirén J, Tang J (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics*, **9**, 539.

Corander J, Waldmann P, Sillanpaa MJ (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–8.

Dorcas ME, Gibbons JW (2008) *Frogs and Toads of the Southeast*. University of Georgia Press, Athens, GA.

Earl DA, vonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Edwards SV (2009) Is a new and general theory of molecular systematics emerging? *Evolution*, **63**, 1–19.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–20.

Faircloth BC (2016) PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, **32**, 786–788.

Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–26.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fisher-Reid MC, Wiens JJ (2011) What are the consequences of combining nuclear and mitochondrial data for phylogenetic analysis? Lessons from *Plethodon* salamanders and 13 other vertebrate clades. *BMC Evolutionary Biology*, **11**, 300.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–52.

Gratwicke B (2008) Proceedings of the Appalachian Salamander Conservation Workshop. In: *IUCN/SSC Conservation Breeding Specialist Group*. Apple Valley, MN.

Highton R (1971) Distributional interactions among eastern North American salamanders of the genus *Plethodon*. In: *The Distributional History of the Biota of the Southern Appalachians: Vertebrates* (eds Holt PC, Paterson RA, Hubbard JP), pp. 139–189. Virginia Polytechnic Institute and State University, Blacksburg, VA.

Highton R, Grobman AB (1956) Two new salamanders of the genus *Plethodon* from the southeastern United States. *Herpetologica*, **12**, 185-188.

Highton R, Hastings AP, Palmer C *et al.* (2012) Concurrent speciation in the eastern woodland salamanders (Genus *Plethodon*): DNA sequences of the complete albumin nuclear and partial mitochondrial 12s genes. *Molecular Phylogenetics and Evolution*, **63**, 278–290.

Highton R, Webster TP (1976) Geographic protein variation and divergence in populations of the salamander *Plethodon cinereus*. *Evolution*, **30**, 33-45.

Hodges E, Rooks M, Xuan Z *et al.* (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protocols*, **4**, 960–974.

Horner HA, Macgregor HC (1983) C value and cell volume: their significance in the evolution and development of amphibians. *Journal of Cell Science*, **63**, 135–146.

Huheey J, Stupka A (1967) *Amphibians and Reptiles of the Great Smokey Mountains National Park*. University of Tennessee Press, Knoxville, TN.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–6.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–80.

Keinath MC, Timoshevskiy VA, Timoshevskaya NY *et al.* (2015) Initial characterization of the large genome of the salamander *Ambystoma mexicanum* using shotgun and laser capture chromosome sequencing. *Scientific Reports*, **5**, 1–13.

Knowles LL (2009) Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 593–612.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–60.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and

SAMtools. *Bioinformatics*, **25**, 2078–9.

Licht LE, Lowcock LA (1991) Genome size and metabolic rate in salamanders. *Comparative Biochemistry and Physiology -- Part B*, **100**, 83–92.

Lips KR, Diffendorfer J, Mendelson JR, Sears MW (2008) Riding the wave: reconciling the roles of disease and climate change in amphibian declines. *PLoS Biology*, **6**, e72.

McCartney-Melstad E, Mount GG, Shaffer HB (2016) *Exon capture optimization in large-genome amphibians*. *Molecular Ecology Resources*, **16**, 1084-1094.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–303.

Meirmans PG (2015) Seven common mistakes in population genetics and how to avoid them over confidence need diff analyses diff scales needed may not be able to do patterns errors. *Molecular Ecology*, 3223–3231.

Mitchell JC, Gibbons JW (2010) *Salamanders of the Southeast*. University of Georgia Press, Athens, GA.

Mizuno S, Macgregor HC (1974) Chromosomes, DNA sequences, and evolution in salamanders of the genus *Plethodon*. *Chromosoma*, **48**, 239–296.

Mueller RL, Gregory TR, Gregory SM, Hsieh A, Boore JL (2008) Genome size, cell size, and the evolution of enucleated erythrocytes in attenuate salamanders. *Zoology*, **111**, 218–30.

Newman CE, Austin CC (2015) Thriving in the cold: glacial expansion and post-glacial contraction of a temperate terrestrial salamander (*Plethodon serratus*). *Plos One*, **10**, e0130131.

O'Neill EM, Schwartz R, Bullock CT *et al.* (2012) Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander

(*Ambystoma tigrinum*) species complex. *Molecular Ecology*, **22**, 111-129.

Olmo E (1974) Further data on the genome size in the urodeles. *Bollettino Di Zoologia*, **41**, 29–33.

Peloso PLV, Frost DR, Richards SJ *et al.* (2016) The impact of anchored phylogenomics and taxon sampling on phylogenetic inference in narrow-mouthed frogs (Anura, Microhylidae). *Cladistics*, **32**, 113–140.

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253–6.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

De Queiroz K (2007) Species concepts and species delimitation. *Systematic Biology*, **56**, 879–86.

Rambaut A, Drummond AJ (2007) Tracer v.1.4.

Rittmeyer EN, Austin CC (2015) Combined next-generation sequencing and morphology reveal fine-scale speciation in Crocodile Skinks (Squamata: Scincidae: Tribolonotus). *Molecular Ecology*, **24**, 466–83.

Sessions SK, Larson A (1987) Developmental correlates of genome size in plethodontid salamanders and their implications for genome evolution. *Evolution*, **41**, 1239–1251.

Shepard DB, Burbrink FT (2011) Local-scale environmental variation generates highly divergent lineages associated with stream drainages in a terrestrial salamander, *Plethodon caddoensis*. *Molecular Phylogenetics and Evolution*, **59**, 399–411.

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT (2014) Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Systematic Biology*, **63**, 83–95.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–3.

Stuart S, Chanson J, Cox N, Young B (2004) Status and trends of amphibian declines and extinctions worldwide. *Science*, **306**, 1783–1786.

Sun C, Shepard DB, Chong RA *et al.* (2012) LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biology and Evolution*, **4**, 168–83.

Thesing BD, Noyes RD, Starkey DE, Shepard DB (2015) Pleistocene climatic fluctuations explain the disjunct distribution and complex phylogeographic structure of the southern red-backed salamander, *Plethodon serratus*. *Evolutionary Ecology*, **30**, 89-104.

Wielstra B, Duijm E, Lagler P *et al.* (2014) Parallel tagged amplicon sequencing of transcriptome-based genetic markers for *Triturus* newts with the Ion Torrent next-generation sequencing platform. *Molecular Ecology Resources*, **14**, 1080–9.

Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, **107**, 9264–9269.

Yang Z, Rannala B (2014) Unguided species delimitation using DNA sequence data from multiple Loci. *Molecular Biology and Evolution*, **31**, 3125–35.

Data Accessibility

Raw sequence read data and contig assemblies are available from the NCBI Sequence Read Archive and Targeted Locus Database, BioProject PRJNA345492. BioSample accession numbers: SAMN05867912-SAMN05868007. Sequence alignments, SNP data, and other supplementary files are available from DRYAD (http://datadryad.org): doi:10.5061/dryad.43t60.

Author Contributions

Conceived and designed the experiments: CEN, CCA. Performed the experiments: CEN. Analyzed the data: CEN, CCA. Contributed reagents, materials, analysis tools: CEN, CCA. Wrote the paper: CEN, CCA.

Figure Legends

Fig. 1. Map of *P. serratus* populations included in this study. Colors correspond to clusters in

Fig. 3. KNF: Kisatchie National Forest, SI: Sicily Island. Range map (gray): NatureServe, IUCN (http://natureserve.org) Inset: Ouachitas, with elevation overlaid onto map (US Geological Survey, http://www.nationalmap.gov).

Fig. 2. Graphical depiction of data sets and analyses. Numbers in parentheses indicate number of loci. Inds: individuals.

Fig. 3. Maximum likelihood phylogeny with corresponding cluster bar plots. Depicted phylogeny is for all-samples (40%) data set. Black circles indicate bootstrap support ≥75 across all four data sets. Otherwise, nodal support is indicated as follows: all-samples (40%) / all-samples (20%) / 1k (40%) / 1k (20%). Colors correspond to Fig. 1.

Fig. 4. Species phylogeny from *BEAST. For each clade, posterior probabilities (PP) are shown for each data set as follows (number of loci): 20/50/70. Asterisks indicate PP ≥ 0.90. Photograph: *P. serratus* (LSUMZ 98343) from Kisatchie Bayou.

Fig. 5. Topographical map of Sicily Island Hills WMA and surrounding region. Black line: Louisiana and Mississippi state boundary, light gray lines: parish boundaries. Catahoula Parish is highlighted by white line. Data source: US Geological Survey (http://www.nationalmap.gov).

Supplementary Information: Figure Legends

Fig. S1. Histograms of variable sites and informative sites. Top row: all samples, all loci. Bottom row: 1K-samples taxon set, 20% missing locus set.

Fig. S2. Smilograms of frequency of variant bases relative to position in alignment. Top row: all-samples taxon set. Bottom row: 1K-samples taxon set.