



OPEN

Genome-wide SNPs redefines species boundaries and conservation units in the freshwater mussel genus *Cyprogenia* of North America

Kyung Seok Kim & Kevin J. Roe

Detailed information on species delineation and population genetic structure is a prerequisite for designing effective restoration and conservation strategies for imperiled organisms. Phylogenomic and population genomic analyses based on genome-wide double digest restriction-site associated DNA sequencing (ddRAD-Seq) data has identified three allopatric lineages in the North American freshwater mussel genus *Cyprogenia*. *Cyprogenia stegaria* is restricted to the Eastern Highlands and displays little genetic structuring within this region. However, two allopatric lineages of *C. aberti* in the Ozark and Ouachita highlands exhibit substantial levels (mean uncorrected $F_{ST} = 0.368$) of genetic differentiation and each warrants recognition as a distinct evolutionary lineage. Lineages of *Cyprogenia* in the Ouachita and Ozark highlands are further subdivided reflecting structuring at the level of river systems. Species tree inference and species delimitation in a Bayesian framework using single nucleotide polymorphisms (SNP) data supported results from phylogenetic analyses, and supports three species of *Cyprogenia* over the currently recognized two species. A comparison of SNPs generated from both destructively and non-destructively collected samples revealed no significant difference in the SNP error rate, quality and amount of ddRAD sequence reads, indicating that nondestructive or trace samples can be effectively utilized to generate SNP data for organisms for which destructive sampling is not permitted.

Precise information on the delineation of evolutionary lineages, their genetic composition and population sub-structure is imperative for guiding population restoration and the long-term preservation of biodiversity¹. However, uncertainty about the accurate delimitation of species and the identification of distinct populations within species remain major challenges to the conservation of imperiled species that are also poorly studied from a taxonomic perspective^{1–3}.

Freshwater mussels (Bivalvia: Unionidae), are one of the most endangered faunas in the world⁴. Estimates are that 65% of the North American mussel fauna are imperiled⁵, and additional extinctions (up to 50% of extant species) are projected over the next century⁶. Freshwater mussels are also important components of aquatic ecosystems and provide important ecosystem services in the systems they inhabit^{7–9}. The life cycle of North American freshwater mussels includes a larval stage, called a glochidium, that must attach to a host fish to complete their development¹⁰. The larval dispersal distance, and therefore gene flow, of freshwater mussels is believed to depend largely on the dispersal capacity of their host fishes^{11,12}. It has been shown that some freshwater mussel species utilize different host fishes in different river systems^{13,14}, and research has shown that mussel larvae transform to juveniles at higher rates on sympatric populations of host-fish species^{13,15}, indicating that the relationship between obligate parasitic mussel larvae and host fishes may be reinforced by co-evolutionary forces.

The freshwater mussel genus *Cyprogenia* inhabits high-gradient streams within the Mississippi faunal province which includes the Eastern, Ozark, and Ouachita highland regions of North America¹⁶. Currently, two species of *Cyprogenia* are recognized: the Fanshell *Cyprogenia stegaria* (Rafinesque 1820), which is listed as a federally endangered species (USFWS 1990) and is found east of the Mississippi River in tributaries of the Ohio River, and the Western Fanshell *Cyprogenia aberti* (Conrad 1850), which is found west of the Mississippi River in the

Department of Natural Resource Ecology and Management, Iowa State University, Ames, IA, USA. email: kjroe@iastate.edu

Arkansas, White, Black, and Ouachita rivers^{17,18}. Populations of *C. aberti* are declining and have been severely fragmented due to habitat loss and degradation by human disturbance¹⁹, and Harris et al.¹⁸ recommended that the status of *C. aberti* in Arkansas be changed from Threatened to Endangered. In many mussel species, the glochidia larvae are packaged, along with unfertilized eggs, into structures called conglutinates that resemble worms and facilitate host infection. The mature glochidia are almost completely transparent, and the color of the conglutinate lure results from the pigmentation of included unfertilized eggs^{13,14}. In *Cyprogenia* species, the colors of conglutinates observed to date include brown, red, and white.

Previous phylogenetic and population genetic research on *Cyprogenia* have provided conflicting answers regarding the validity of the two nominal species. Phylogenetic research based on mitochondrial DNA (mtDNA)^{20,21} indicated that the two species of *Cyprogenia* were not reciprocally monophyletic, and that the divergent mitochondrial clades recovered in analyses were associated with the color of the conglutinate lures used by mussels to attract and infest host fishes²². In contrast, population genetic research using microsatellite markers^{21,23} supported the recognition of both *C. aberti* and *C. stegaria*, as allopatric species inhabiting the Western and Eastern highland regions respectively. Additionally, Chong et al.²³ further advanced the hypothesis originally proposed by Serb & Barnhart²² that the mitochondrial lineages of *Cyprogenia* were associated with conglutinate color and these divergent mtDNA lineages were maintained by negative frequency-dependent selection by host fish. Despite the above findings, the taxonomic status of these species while clearer was still not fully resolved, and *C. aberti* has been assessed as Data Deficient due to these ongoing taxonomic issues (<http://www.iucnredlist.org/details/6182/0>).

Non-genomic markers are still the most commonly applied tool to address conservation genetic questions for freshwater mussels, however, the low information content and the potential non-neutrality of mtDNA markers, and costs and scoring issues associated with microsatellite markers have limited their usefulness. Recently, genomic data has been shown to be a valuable tool to address these challenges^{24,25}. Application of population genomic tools, including reduced representation with restriction site associated markers allows for the harnessing of high-throughput next-generation sequencing to yield thousands of polymorphic genetic markers to address the conservation and management of this group. Genotyping-by-sequencing (GBS) or double digest restriction-site associated DNA sequencing (ddRAD) are reduced representation approaches which have generated highly informative single nucleotide polymorphisms (SNPs) for population genomic inference^{24,26–29} and phylogenomic research^{25,27,30–41} for numerous species across a broad range of taxa.

The availability of a large number of genome-wide SNP markers using the ddRAD-Seq approach has enabled high-density SNP discovery, genotyping, and genetic mapping at high resolution, and has provided solutions to unresolved questions concerning population genetic structure, species delineation, and to inform the management and conservation of a number of high-profile species of conservation concern^{28,35,42}. Despite these advances, perceptions about the necessity of high quality starting material for the development of genomic libraries may have precluded the application of these techniques for endangered taxa because collecting permit restrictions only permit non-destructive sampling, which can yield small quantities of genomic DNA, are otherwise contaminated with non-target species or otherwise degraded⁴³.

Here, we employed genome-wide ddRAD-Seq data to address species delimitation in *Cyprogenia*, document the genetic structuring, and attempt to further elucidate the relationship between conglutinate color and mtDNA lineages. We also explore and discuss possible barriers inhibiting gene flow and genetic connectivity across the range of *Cyprogenia*, the timing of population changes, as well as forces driving genetic divergence among mussel populations within the highlands. Finally, we used this opportunity to examine the utility of reduced representation approaches when only non-destructive samples are available as source material for genomic library creation by comparing the quality of results obtained from samples collected using a non-destructive method (cytology brushes) with those collected from mantle tissue biopsies.

Results

STACKS parameter exploration identified a combination of assembly parameters ($-m = 3$, $-M = 2$, $-n = 3$, $-max_locus_stacks = 3$, with SNP model) as the best performing parameters (details in “Supplementary Information”), and this parameter setting was chosen for de novo ddRADseq assembly and SNP discovery, and for downstream genetic analyses. In total, 309,727 catalog loci were generated and 14,176 RAD loci passed sample/population constraints ($-p 2$, $-r 0.60$), where only loci that were present in 50% of four species-groups (*C. stegaria*_Eastern, *C. aberti*_Ozark, and *C. aberti*_Ouachita, and *Dromus dromas*) were retained. Total of 7,243 SNPs appeared to have a single SNP per locus with the mean depth of coverage of 7.24 (SD = 5.19).

Genetic diversity of *Cyprogenia*. A total of 7243 SNPs were used to estimate population genetic parameters for four pre-defined mussel groups (*C. stegaria*_Eastern, *C. aberti*_Ozark, *C. aberti*_Ouachita, and *D. dromas*) as well as for ten rivers throughout the range of *Cyprogenia*. Across all three species, measures of genetic diversity were the highest in *C. aberti*_Ozark, followed, in decreasing order, by *C. stegaria*_Eastern, *C. aberti*_Ouachita and *D. dromas* (Table 1). Among rivers, genetic diversity was lowest in specimens from the Saline River and highest in the Black River (Table 1). Among highland regions, mean level of genetic diversity was the lowest in *C. aberti*_Ouachita, followed by *C. stegaria*_Eastern and *C. aberti*_Ozark. Genetic diversity estimates (H_o and H_e) were significantly different ($P = 0.001$; based on 999 permutation using OSx-statistic) among the three highlands regions.

Genetic differentiation (F_{ST}) among rivers within each highland region was lowest in *C. aberti*_Ozark ($F_{ST} = 0.0733$), followed by *C. stegaria*_Eastern ($F_{ST} = 0.013$) and *C. aberti*_Ouachita ($F_{ST} = 0.143$) (Table 1). Pairwise genetic differentiation (F_{ST}) among pairs of all ten rivers throughout the range of *Cyprogenia* showed similar patterns of differentiation for both raw and corrected data. Corrected F_{ST} values are lower than uncorrected values

Species/highland	River	Size	Na	Neff	Ho	He	F_{IS}	F_{ST}
<i>C. stegaria</i> /Eastern	Clinch	8	1.334	1.193	0.107	0.131	0.178	
	Green	10	1.409	1.221	0.121	0.144	0.162	
	Licking	24	1.462	1.225	0.120	0.141	0.153	
	Salt	9	1.393	1.215	0.122	0.142	0.142	
	All (Mean)	51	1.400	1.214	0.118	0.140	0.159	0.013
<i>C. aberti</i> /Ozark	Black	28	1.547	1.252	0.126	0.161	0.218	
	Spring	27	1.529	1.240	0.123	0.152	0.186	
	St. Francis	16	1.447	1.219	0.117	0.140	0.169	
	All (Mean)	71	1.508	1.237	0.122	0.151	0.191	0.073
<i>C. aberti</i> /Ouachita	Caddo	5	1.245	1.155	0.097	0.104	0.064	
	Ouachita	6	1.273	1.164	0.096	0.112	0.145	
	Saline	13	1.288	1.155	0.076	0.101	0.250	
	All (mean)	24	1.269	1.158	0.090	0.106	0.153	0.143
<i>D. dromas</i>	Clinch	8	1.131	1.076	0.031	0.051	0.393	

Table 1. Measure of genetic Diversity and inbreeding coefficient for populations from rivers, Highlands and species. A total of 154 individuals except two geographically isolated individuals (Ozark_105 and Ozark_107) were included in this analysis. *Na* mean number of alleles, *Neff* effective number of alleles (The number of alleles in a population, weighted for their frequencies), *Ho* observed heterozygosity, *He* expected heterozygosity assuming Hardy–Weinberg equilibrium, F_{IS} inbreeding coefficient, F_{ST} uncorrected for missing data.

Highland	River	<i>C. stegaria_Eastern</i>				<i>C. aberti_Ozark</i>			<i>C. aberti_Ouachita</i>		
		Clinch	Green	Licking	Salt	Black	Spring	St.Francis	Caddo	Ouachita	Saline
Eastern	Clinch	–	0.026	0.026	0.019	0.197	0.259	0.244	0.271	0.248	0.215
	Green	0.051	–	0.003	0.006	0.241	0.304	0.287	0.307	0.286	0.248
	Licking	0.050	0.007	–	0.005	0.238	0.298	0.283	0.301	0.283	0.250
	Salt	0.059	0.012	0.012	–	0.212	0.275	0.257	0.278	0.259	0.226
Ozark	Black	0.506	0.508	0.513	0.508	–	0.018	0.030	0.134	0.119	0.098
	Spring	0.535	0.537	0.539	0.537	0.041	–	0.024	0.174	0.160	0.140
	St.Francis	0.563	0.558	0.554	0.561	0.106	0.073	–	0.173	0.155	0.136
Ouachita	Caddo	0.612	0.595	0.580	0.605	0.308	0.354	0.402	–	0.007	0.036
	Ouachita	0.609	0.594	0.582	0.604	0.311	0.357	0.400	0.046	–	0.030
	Saline	0.632	0.619	0.603	0.630	0.342	0.397	0.443	0.206	0.177	–

Table 2. Pairwise F_{ST} 's among 10 rivers throughout distribution range of genus *Cyprogenia* based on 7243 SNP loci. F_{ST} values from raw data below diagonal and F_{ST} values from corrected data above diagonal. A total of 146 individuals except two geographically isolated individuals (Ozark_105 and Ozark_107) were included in this analysis. Significance was tested by genic differentiation for each population pair from exact G test using default Markov chain parameters (Dememorisation: 10,000, Batches: 100, Iterations per batch: 5000). Bold Italicized indicates highly significant F_{ST} value. Corrected data refers to data that missing data were replaced with randomly drawn alleles based on the overall allele frequencies. Significance level was adjusted for multiple testing using Bonferroni correction.

due to the use of averages to impute missing data (required for PCA analysis). The lowest and nonsignificant F_{ST} ($F_{ST}=0.007$) was observed between the Licking and the Green rivers, and the highest ($F_{ST}=0.630$, $P=0.001$) was found between the Salt and the Saline rivers for the uncorrected data. Most pairwise comparisons were statistically significant after correction for multiple tests (Table 2). In general, high genetic differentiation was observed between the sampling locations of *C. stegaria* and *C. aberti*, but very high F_{ST} values were also observed between *C. aberti* from the Ouachita and Ozark highland regions.

Analysis of molecular variance (AMOVA) according to hierarchical groupings revealed a high degree of variance among the three highland regions, accounting for 21.6% of total variation ($F_{CT}=0.216$, $P=0.001$; using an Infinite Allele Model (IAM) and 999 permutations), which may largely reflect species divergence (Table 3). Significant but little genetic differentiation was observed among populations within regions ($F_{SC}=0.020$, $P=0.001$).

The estimated value for F_{ST} between mussels grouped by conglutinate color revealed low level of genetic differentiation ($F_{ST}=0.0377$), indicating no evidence for population subdivision based on this phenotype. This difference is smaller than the values among rivers (Ozark: $F_{ST}=0.0733$; Ouachita: $F_{ST}=0.143$) within highland regions.

Source of variation	%var	F-stat	F-value	c.i.2.5%	c.i.97.5%	P-value
Three highlands	(Ozark, Ouachita, and Eastern)					
Within individual	0.668	F_{IT}	0.332	0.323	0.341	–
Among individual	0.100	F_{IS}	0.131	0.125	0.136	0.001
Among population	0.016	F_{SC}	0.020	0.019	0.022	0.001
Among highlands	0.216	F_{CT}	0.216	0.209	0.223	0.001
Among population without grouping	–	F_{ST}	0.179	0.174	0.185	0.001
Two highlands	(Ozark and Ouachita)					
Within individual	0.790	F_{IT}	0.210	0.198	0.222	–
Among individual	0.100	F_{IS}	0.113	0.104	0.121	0.001
Among population	0.027	F_{SC}	0.030	0.027	0.033	0.001
Among highlands	0.082	F_{CT}	0.082	0.074	0.091	0.107
Among population without grouping	–	F_{ST}	0.069	0.064	0.075	0.001

Table 3. Summary of hierarchical AMOVA analysis for *Cyprogenia* from the highland regions based on 7243 SNP loci. A total of 146 individuals except two geographically isolated individuals (Ozark_105 and Ozark_107) were included in this analysis. AMOVA was conducted separately according to hierarchical groupings. Three Highlands correspond to the Ozark, Ouachita, and Eastern highland regions, consisting of two *Cyprogenia* species. Two Highlands correspond to Ozark and Ouachita Highlands, regions consisting of same species, *C. aberti*. Missing data are replaced with randomly drawn alleles based on the overall allele frequencies. Significance was tested using 999 permutations.

Population structure of *Cyprogenia*. A PCA along PC1 vs. PC2 for all 156 specimens showed a clear segregation of samples into three clusters (Fig. 1A). Further subdivision was observed along PC1 & PC3 and PC2 & PC3, in which *C. aberti* were further divided into two separate groups, *C. aberti*_Ouachita and *C. aberti*_Ozark. Specimens of *C. aberti*_Ozark were grouped with those of *D. dromas* for PC1 vs. PC3, whereas *C. aberti*_Ozark was grouped with *C. stegaria* for PC2 vs. PC3.

Additionally, a PCA excluding *D. dromas* samples displayed a clear separation into the three highland regions representing two *Cyprogenia* species (Fig. 1B). A PCA for 97 *C. aberti* specimens alone from Ozark and Ouachita highlands clearly distinguished specimens from the two highland regions, and specimens from the Fall and Spring rivers, KS (Fig. 1C). Finally, Discriminant Analysis of Principal Components (DAPC) for *C. aberti* specimens further subdivided *C. aberti* specimens into three discrete groups, 1: Caddo and Ouachita rivers in the Ouachita Highlands, 2: Saline River in the Ouachita Highlands, 3. Black, Spring, and St. Francis rivers in the Ozark Highlands (Fig. 1D).

STRUCTURE identified $K = 3$ as the optimal number of clusters for the *Cyprogenia* samples indicating little evidence of admixture among the three highland regions (Fig. 2 and left bar plot in Fig. 3). Further exploration of each cluster using STRUCTURE resulted in the identification of substructure within the Ouachita and Eastern highlands ($K = 2$), and the Ozark Highlands ($K = 3$). Based on these results, at least five genetically distinct clusters are identifiable within *C. aberti*, that correspond to the 1. Ouachita/Caddo rivers AR, 2. Saline River, AR (Ouachita Highlands), 3. Black/Spring rivers, AR, 4. St. Francis River, MO, and 5. Fall/Spring rivers, KS (Ozark Highlands) (right bar plot in Fig. 3). Interestingly, two specimens (Ozark105 and Ozark107) sampled from the Spring River and the Fall River in Kansas shared genetic composition indicative of both the Ozark and Ouachita Highlands (Fig. 3). Most of *C. stegaria* specimens in the Eastern Highland did not show any genetic structuring, but some specimens from the Clinch River had a genetic composition distinct from the other rivers (Fig. 3).

Phylogenetic relationships among *Cyprogenia*. Maximum likelihood (ML) trees constructed using different filtering criteria produced nearly identical topologies, therefore for downstream analyses we used a data set in which SNPs were present in at least 60% of samples (Fig. S4). This same criterion was used for the Bayesian analysis. Both ML and Bayesian trees identified four distinct clades, *C. stegaria* (Clade A), *C. aberti* (Clades B, C) and the outgroup, *D. dromas* (Clade D). The trees also indicated at least 6 subclades within *C. aberti*, all with strong bootstrap and posterior supports (A, B-1, B-2, B-3, C-1, C-2), reflecting river systems (Fig. 3).

Specimens from the Ouachita region were divided into two subclades, Clade C-1 (Caddo and Ouachita rivers, AR), and Clade C-2 (Saline River, AR). In the Ozark region, specimens from the St. Francis River, AR formed an internal clade within the Ozark clades. Notably, two individuals (Ozark 105 and Ozark 107), from the Spring and Fall rivers, in the Ozark region formed a clade (Clade B-3) sister to all other *C. aberti* clades. These individuals are also detected as a distinct cluster in the STRUCTURE analysis. *C. stegaria* specimens from the Eastern Highland region did not show clear pattern of geographic clustering (Fig. 3).

ML and Bayesian trees for *C. aberti* specimens with pre-defined conglutinate lure color displayed clear allopatric clades that corresponded to geographic regions (Ozark and Ouachita) and not with the color of the conglutinate lures (Fig. S5).

Species tree and species delimitation of *Cyprogenia*. Species trees showed clear separation between *C. aberti* and *C. stegaria*. *Cyprogenia aberti* was further separated into two monophyletic groups reflecting highland regions (Fig. 4). Trees were generated based on two different sampling criteria, random sampling of indi-

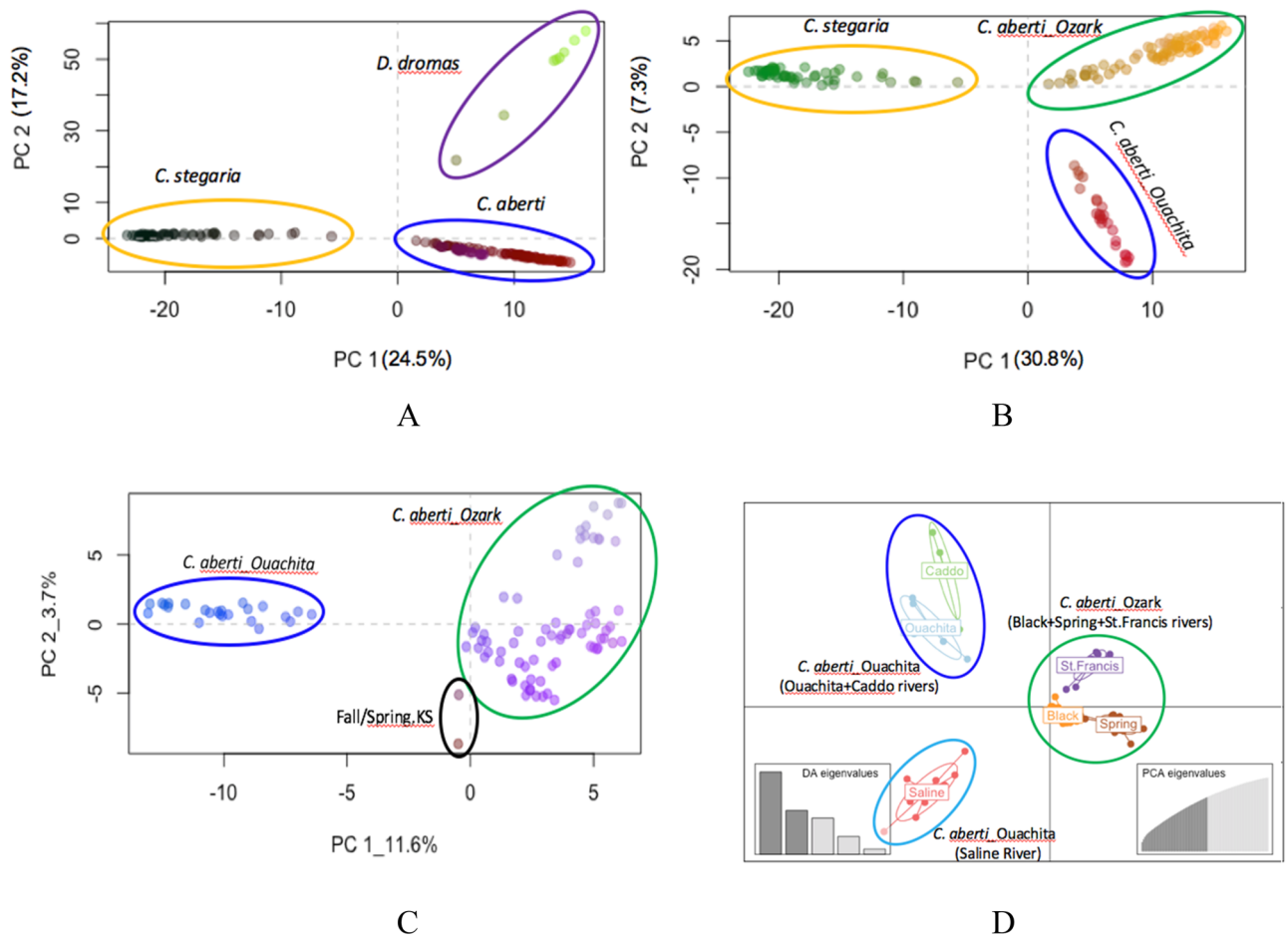


Figure 1. Results of the Principal Component Analyses (PCA) and Discriminant Analysis of Principal Components (DAPC) in R. In this Figure, genetic diversity is represented in two ways: by the distances (further away = more genetically divergent), and by the colors (more divergent colors = more genetically divergent). (A) PCA for entire 156 samples representing three freshwater mussel species (*Cyprogenia* + *Dromus*), (B) PCA for 148 samples representing genus *Cyprogenia*, (C) PCA for 97 samples representing *Cyprogenia aberti*. (D) DAPC for 97 samples representing *Cyprogenia aberti* (Number of PCA axes retained = 50, and number of PCA axes retained = 5).

individuals or by selecting the individual samples with the highest coverage (smallest number of missing SNPs). A tree selected based on coverage (Fig. 4A) revealed that, in the Ozark Highlands, the Black River was basal to the Spring and St. Francis rivers, and in the Ouachita Highlands, the Saline River diverged prior to Caddo and Ouachita rivers. These results indicated support for the results of the ML and Bayesian phylogenetic analysis.

The species delimitation results for the six *Cyprogenia* models are summarized in Table 4. The top-ranked model (Model2) included three species. In this model, *C. aberti* is divided into two species, Ozark Highlands (Black, St. Francis, Spring rivers) and Ouachita Highlands (Caddo, Ouachita, Saline rivers). Model 2 has the largest MLE value and received greater supported than the model (Model 1) that includes two *Cyprogenia* species, (current taxonomy model). Among models, the BF in support for Model 2 was decisive compared to other models (Table 4). The current taxonomy model (Model 1) was ranked 4th among models and was more supported than two paraphyletic models (Model 3 and Model 4). It is notable that a paraphyletic model (Model 3), where half the samples of the Ozark and Ouachita Highlands were intermixed with each other, was the least supported among models.

Demographic changes in *Cyprogenia*. The past demographic changes of each of the sampling locations (five demographic models) were explored using Approximate Bayesian Computation (ABC) approach. The best-fit demographic model was selected by comparison of five models and the resultant estimations of current effective population sizes (N_e) for each sampling site are summarized in Table 5. All *Cyprogenia* populations showed either model 2 (DEC) or model 4 (INCDEC), or both models as the best fit model, depending on estimation method (direct or logistic regression) (Table 5), indicating a population decline after the Last Glacial Maximum (LGM), as supported by the best-fit demographic model (Table 5).

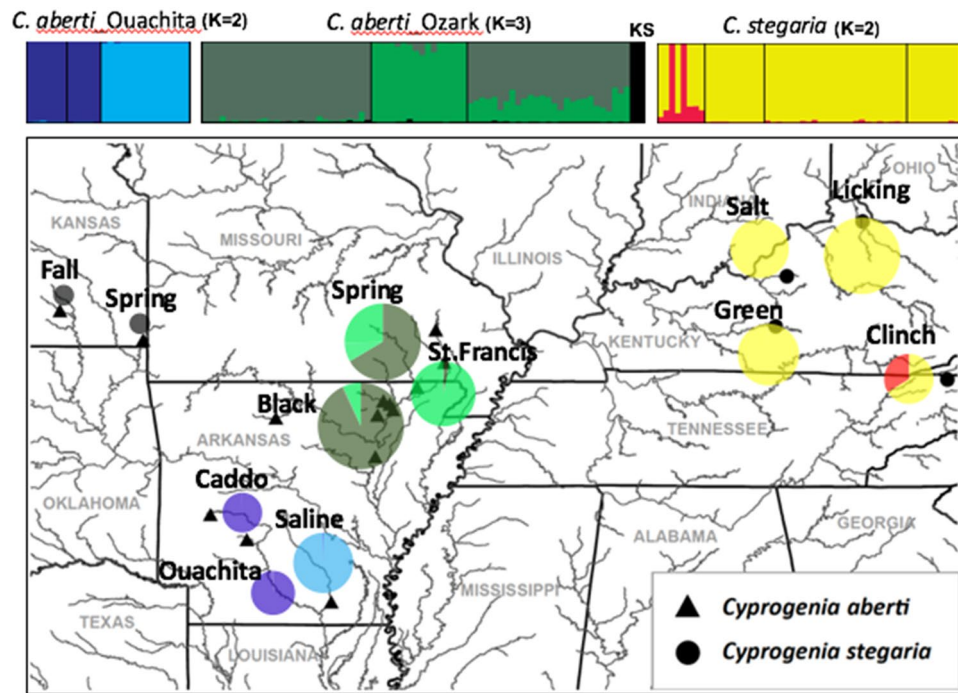


Figure 2. Geographic locations and pie charts of membership of each sampled population inferred by STRUCTURE analysis of *Cyprogenia*. The STRUCTURE plots show the posterior probability for individual assignments of samples to different genetic clusters based on the results of re-analysis of the original $K=3$ clusters (Fig. S1). The optimal number of genetic clusters within Highlands was $K=2$ for Ouachita, $K=3$ for Ozark, and $K=2$ for Eastern. Pie charts indicate proportions of membership to clusters inferred by STRUCTURE analysis for each Highland (see text for details). Circle size of pie charts is proportional to sample size. Sample sites for *Cyprogenia aberti* are marked as triangles and *Cyprogenia stegaria* as circles. Base map reproduced from Chong et al.²³.

Quality comparison between genetic materials using non-destructive and destructive methods. Genetic samples collected using cytology brushes showed a lower level of sample retention, compared to those from tissue samples. Forty six out of 88 nondestructive samples (52.3%) passed the quality filtering steps of the STACKS pipeline, whereas 110 out of 126 destructive samples (87.3%) passed the steps, which is statistically significant at the threshold of 0.05 (Fisher exact test statistic value = 0.022). However, after filtering, the amount (in estimates of mean \pm SD) of ddRAD sequence reads (i.e. number in low quality, no radtags, retained and total loci) do not appear to differ between genetic samples using either method (Fig. S6). There was no significant difference in SNP error rates between duplicated genetic samples from nondestructive and destructive methods [SNP error rate: four duplicated cytology brush samples = 0.049 ± 0.014 (SD) and four duplicated mantle biopsy samples = 0.056 ± 0.011 (SD)]. Lastly, ddRAD sequences of randomly selected destructive (Fig. S7A: CSF_21, 3,023,365 loci) and non-destructive (cytology brush) (Fig. S7B: CSF_19, 3,881,449 loci) samples showed similar amount and pattern in the percentage of ddRAD sequences mapped to four genome sequences and PhiX sequences.

Discussion

Understanding the population subdivision and genetic differentiation within a species can serve as a basis for defining units for conservation management⁴⁴ and can inform management decisions to conserve species⁴⁵. Detailed information on species delimitation, population genetic structure and genetic composition of populations at both regional and fine scales together will help establish the long-term management and design tailored conservation plans for freshwater mussel species that are among the most endangered faunas in North America. The present genome-wide SNP data provides improved resolution on species delimitation of *Cyprogenia* and also allowed detection of unrecognized population genetic structure at the regional scale.

Cyprogenia specimens were genetically more similar within regions than between different regions (Table 3). A biogeographic pattern reflecting highland regions has been also observed in other freshwater organisms, e.g. rainbow darter, rosyface shiner, streamline chub, etc., that are endemic to the highland regions of North America^{46–48}. It has been proposed that Pleistocene glaciations have played an important role in determining faunal composition in the highlands of North America, in which a series of glacial cycles fragmented the region and its associated fauna^{48–50}. Formation of the three major highland regions, the Ozark and Ouachita highlands west of the Mississippi River, and the Eastern Highlands east of the Mississippi River, from the contiguous Central Highlands of North America, during the Pleistocene has been discussed elsewhere⁵¹. Several case studies^{47,52–55} on population genetic structure for fish species in the North American Highlands evidenced diminished gene

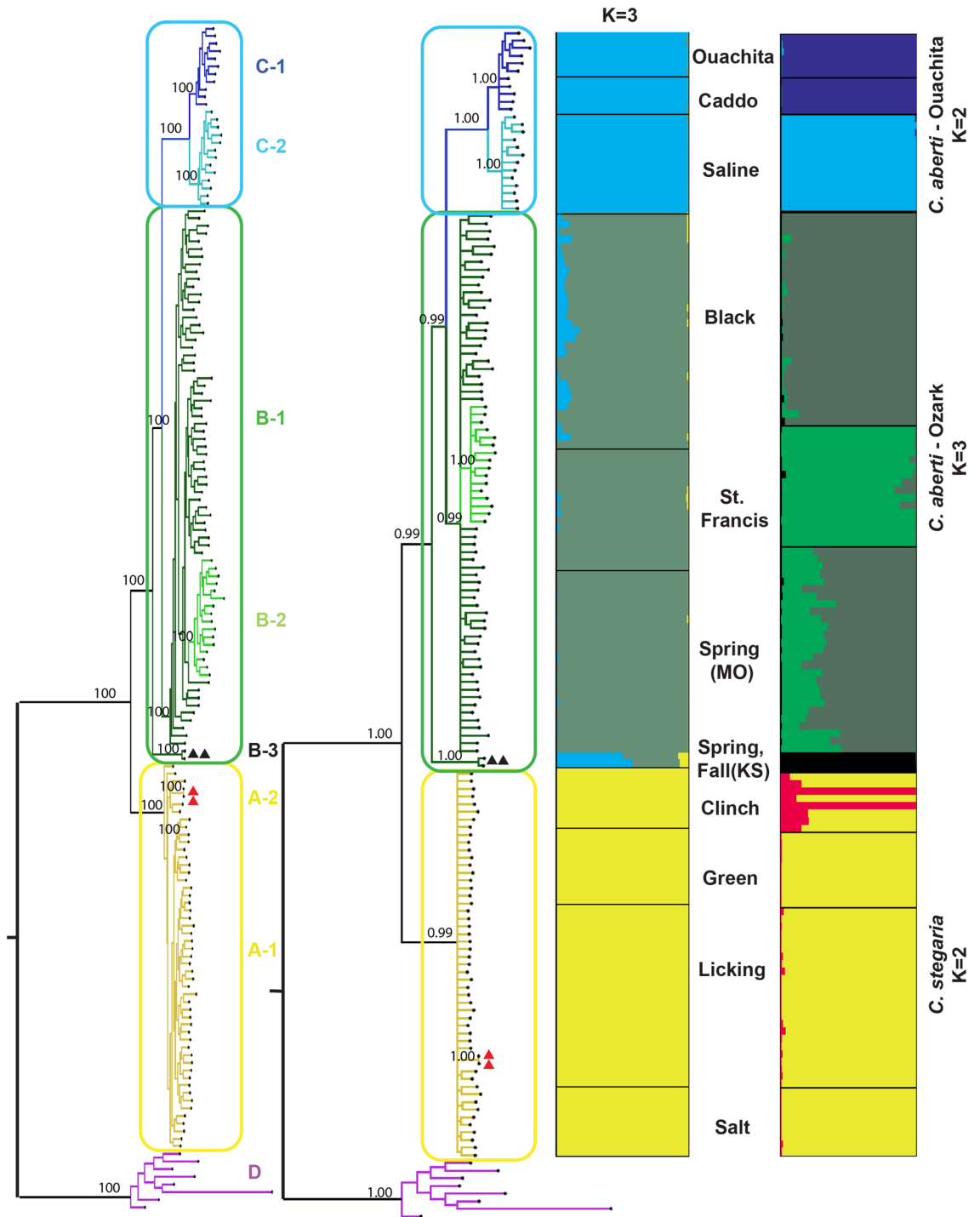


Figure 3. Phylogenetic relationships and population genetic composition of freshwater mussel species. Maximum likelihood (ML) tree (left) and Bayesian tree (right) were constructed based on 4395 informative sites from concatenated sequences of 9673 SNPs (see text for details). Bootstrap support values (ML) and posterior probability (Bayesian) respectively were provided at the node of major clade. The STRUCTURE plots show the posterior probability for individual assignments of samples to different genetic clusters. The main plot (left) shows the results for the optimal number of genetic clusters ($K = 3$) for 148 *Cyprogenia* specimens. The three smaller plots to the right show the results of re-analyses of samples from each Highland. The optimal number of genetic clusters was $K = 2$ for Ouachita, $K = 3$ for Ozark, and $K = 2$ for Eastern. Colors of clusters in the phylogenetic trees correspond to colors in the STRUCTURE plots and Fig. 2. Red triangles: Clinch river, TN (Eastern15 and Eastern18 in Table S1), Black triangles: Spring and Fall rivers, KS (Ozark105 and Ozark107 in Table S1).

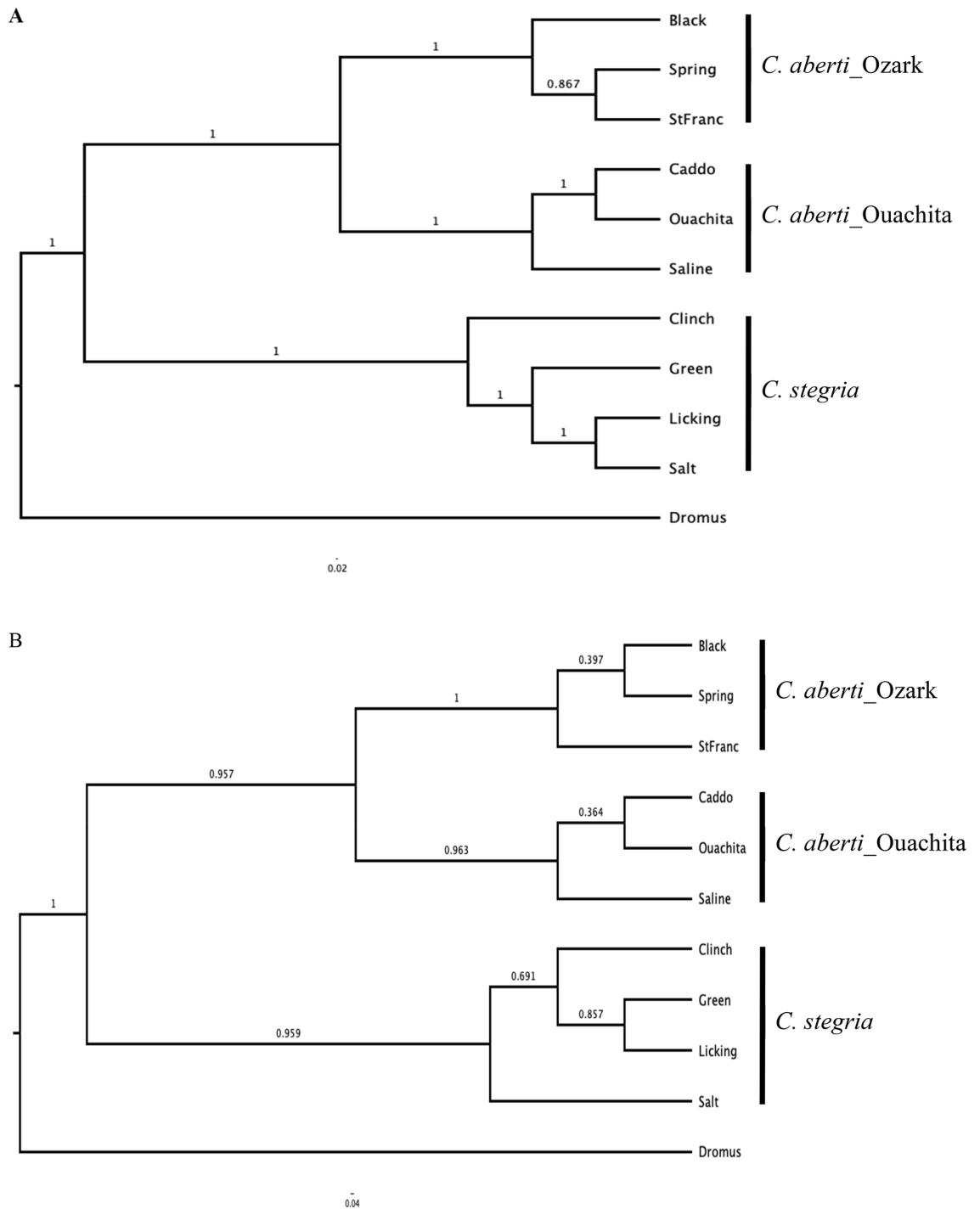


Figure 4. Species trees generated with Maximum clade credibility (MCC) and Median heights option in TreeAnnotator and with Transform branches (proportional) option in FigTree. Posterior probabilities are shown on the branch of each node. Total of 40 samples for 10 sampling sites (four samples per each sampling site) for *Cyprogenia* species were aligned and selected based on by selecting the individual samples with the highest coverage (smallest number of missing SNPs) (A) or by random sampling of individuals (B) using a custom Python script.

flow within and between these same highland regions. The results of the ABC simulations provide support for existing biogeographic scenarios within the Ozark and Ouachita regions. For example, the origin of the Ouachita River population via the admixture of the Caddo and Saline rivers is not inconsistent with hypotheses for the evolution of these drainages from their earlier pre-glacial condition^{51,56}. Our scenario for the divergence of the St. Francis River also fits well with current hypotheses for the history of this river⁵¹, and although the divergence

Model	Classification	No. species	MLE	Rank	BF	Detailed classification
1	Current taxonomy	2	-55,853	4	-	Two <i>Cyprogenia</i> species: <i>C. aberti</i> and <i>C. stegaria</i>
2	Split <i>C. aberti</i>	3	-53,201	1	-5304	<i>C. aberti</i> was split into Ozark (Black, St. Francis, Spring) and Ouachita (Caddo, Ouachita, Saline)
3	Mix <i>C. aberti</i>	3	-56,827	6	1949	Half of Ozark and Ouachita Highlands were intermixed
4	Reassign <i>C. aberti</i>	3	-56,323	5	940	Black from Ozark was moved to Ouachita, and Saline from Ouachita was moved to Ozark
5	Reassign <i>C. aberti</i>	3	-54,742	2	-2222	Saline from Ouachita was moved to Ozark
6	Reassign <i>C. aberti</i>	3	-55,285	3	-1136	Black from Ozark was moved to Ouachita

Table 4. Path sampling results for the six species delimitation models for *Cyprogenia* species shown in Fig. 2. All Bayes factor (BF) calculations are made against the current taxonomy model (Model 1). Therefore, positive BF values indicate support for current taxonomy model, and negative BF values indicate support for alternative model. The BF scale is as follows: $0 < \text{BF} < 2$ is not worth more than a bare mention, $2 < \text{BF} < 6$ is positive evidence, $6 < \text{BF} < 10$ is strong support, and $\text{BF} > 10$ is decisive. Current and alternative species delimitation models were analyzed with the path sampling steps of 36, MCMC sample length of 100,000 for each path sampling step, alpha of 0.3, burnInPercentage = 10, and preBurnin of 10,000.

Highlands	Location/sample	Best scenario direct (logistic)	Parameter	Median	Quantile 5%	Quantile 95%
Ozark	Black	DEC (DEC)	<i>Ne</i>	4.70×10^2	1.62×10^2	9.73×10^2
			<i>tt</i>	1.97×10^2	1.13×10^2	4.19×10^2
	Spring	DEC (DEC)	<i>Ne</i>	7.78×10^2	2.62×10^2	1.60×10^3
			<i>tt</i>	2.35×10^2	1.19×10^2	5.63×10^2
	St. Francis	DEC (DEC)	<i>Ne</i>	2.00×10^3	6.45×10^2	3.93×10^3
			<i>tt</i>	1.32×10^3	3.16×10^2	3.24×10^3
Ouachita	Ouachita	DEC (DEC)	<i>Ne</i>	4.26×10^3	9.88×10^2	1.09×10^4
			<i>tt</i>	2.33×10^3	4.46×10^2	4.26×10^3
	Caddo	DEC (DEC)	<i>Ne</i>	3.79×10^3	8.56×10^2	1.07×10^4
			<i>tt</i>	2.30×10^3	4.45×10^2	4.24×10^3
	Saline	DEC (DEC)	<i>Ne</i>	3.29×10^3	9.04×10^2	7.31×10^3
			<i>tt</i>	1.29×10^3	2.39×10^2	3.85×10^3
Eastern	Salt	INCDEC (INCDEC)	<i>Ne</i>	1.55×10^3	3.97×10^2	4.26×10^3
			<i>t1</i>	3.82×10^2	1.30×10^2	1.52×10^3
			<i>t2</i>	3.62×10^3	2.42×10^2	4.42×10^3
	Green	INCDEC (INCDEC)	<i>Ne</i>	3.78×10^2	1.10×10^2	1.29×10^3
			<i>t1</i>	1.60×10^2	1.06×10^2	6.00×10^2
			<i>t2</i>	3.90×10^3	2.54×10^3	4.46×10^3
	Clinch	DEC (INCDEC)	<i>Ne</i>	2.35×10^3	6.93×10^2	5.16×10^3
			<i>tt</i>	6.50×10^2	1.63×10^2	2.89×10^3
	Licking	INCDEC (INCDEC)	<i>Ne</i>	2.23×10^2	8.51×10	5.49×10^2
			<i>t1</i>	1.22×10^2	1.03×10^2	2.06×10^2
			<i>t2</i>	4.03×10^3	2.62×10^3	4.47×10^3

Table 5. Best fit demographic model and posterior distribution of the current effective population sizes (*Ne*) and generation time (*T*) (Median and 95% credible interval) for each *Cyprogenia* sample site using ABC simulation implemented in DIYABC. Five demographic models proposed by by Cabrera and Palsbøll⁸¹ are evaluated. Model 1: CON (constant population size), Model 2: DEC (a single instantaneous decrease in population size), Model 3: INC (a single instantaneous increase in population size), Model 4: INCDEC (a single instantaneous increase followed by a single instantaneous decrease in population size), Model 5: DECINC (a single instantaneous decrease followed by a single instantaneous increase in population size). Time is in number of generations assuming a generation time of 5 years for *Cyprogenia*; *tt* denotes a unique event after the LGM, *t1* and *t2* denote an event after and before the LGM. Detailed prior model parameterization for five demographic models was provided in “Supplementary Information”. No mutation model parameterization was required for SNPs.

scenario for the Black and Spring rivers runs counter to our understanding of the origins of these rivers, it could accurately reflect the origin of the *Cyprogenia* populations that inhabit them.

Given the relationship between freshwater mussel larvae and host fishes, the reduced compatibility of freshwater mussels with allopatric populations of host fishes as compared to sympatric host fishes may serve to limit gene flow between mussel populations in different rivers^{13,57}. Other factors such as anthropogenic disturbance (resource extraction, impoundments, etc.) and natural barriers can limit dispersal of host fishes within and between regions. These barriers may further reinforce genetic differences between host fish populations by inhibiting gene flow among Highlands, leading to population subdivision within and between highlands.

Phylogenetic trees derived from our genome-wide ddRAD-Seq SNP data identified three monophyletic groups in the genus *Cyprogenia*. *Cyprogenia stegaria* formed a distinct clade corresponding to the Eastern Highland Region, whereas *C. aberti* was divided into two major clades, which corresponded to the Ozark and Ouachita highlands of the Western Highland region (Fig. 2). Within the Ozark and Ouachita regions, *Cyprogenia* specimens were further subdivided into two distinct lineages delimited mainly by river systems. The three main genetic clusters in *Cyprogenia* identified by Bayesian model-based clustering analyses are consistent with the species delimitation analysis of *Cyprogenia* based on species trees in this study, and are in agreement with the results from the previous study using nuclear microsatellite markers²³. Importantly, the present genome-wide ddRAD-Seq dataset also recognized additional genetically distinct clusters within the range of *C. aberti* not observed by Chong et al.²³. These results indicate that the findings of host-fish compatibility trials, which indicated sympatric host fishes produced higher transformation rates of larvae to juvenile mussels than allopatric host fishes of the same species, may possibly have a genetic basis²².

Coalescent-based Bayesian species trees also indicated that *C. aberti* was comprised two clusters, Ozark and Ouachita, (Fig. 4). Species delimitation analysis revealed that a three species model was more supported than the two species model based on current taxonomy. Interestingly, the model, where samples of each highland region were intermixed (Model 3), showed the lowest support among the six competing species delimitation models (Table 4). Results based on species delimitation, phylogenetic analysis, population genetic structure, and PCA analyses, together support the recognition of three *Cyprogenia* species. However, more thorough and integrative understanding on geographical distribution, ecological and demographic characters, and host fishes use of the species, will help to address this issue.

Recognition of the three evolutionary lineages within *Cyprogenia* corresponding to the three highland regions supports continued recognition of *Cyprogenia stegaria* (Rafineque, 1820) as a distinct entity occurring east of the Mississippi River and occupying tributaries of the Ohio River basin. *Cyprogenia* that are found west of the Mississippi River, within the range of *Cyprogenia aberti* (Conrad, 1850) form two distinct lineages, one restricted to the Ozark region (White and St. Francis River systems) and the other in the Ouachita region (Ouachita and Saline river systems). Both of these entities are distinct enough to warrant recognition as separate species, however our current data set included only a small number of samples from the Spring and Fall rivers in Kansas, and we feel that additional samples from this area are needed to confirm this pattern. These samples formed a sister clade to the remaining western *C. aberti* populations and were also a distinct group in the Bayesian clustering analysis.

Since the Fall and Verdigris rivers are the type localities for the names *Unio aberti* Conrad, 1850 and *Unio popenoi* Call, 1855, the affinities of the Fall River populations and any extant populations in the Verdigris River are critical to the assignment to the name *Cyprogenia aberti* (Conrad, 1850). Because of the small number of samples from the Fall River and total lack of samples from the Verdigris River, additional taxonomic changes must await a more thorough geographic sampling of *Cyprogenia*.

Our Approximate Bayesian Computation (ABC) approach investigating demographic changes indicates that all *Cyprogenia* populations experienced a decline in size following the LGM (Table 5). Although some populations of *Cyprogenia* occur south of the maximal glacial extent, the impact of glaciation has been shown to still have impacts on population size and other demographic features in other taxa (e.g.⁵⁸). The inferred reduction in population size and concomitant loss of genetic diversity likely contributed to the genetic differentiation revealed between these two highland regions in this study.

In this study, the combination of population genomic and phylogenomic findings suggests that conservation plans for *Cyprogenia* should consider at least 5, and possibly 7 distinct evolutionarily lineages when developing management plans; 1. *C. stegaria* in the Licking, Salt, and Green rivers, KY), 2. *C. "aberti"* in the Black/Spring rivers, AR, 3. *C. "aberti"* in the St. Francis River, MO, 4. *C. "aberti"* in the Ouachita/Caddo rivers, AR, 5. *C. "aberti"* in the Saline River, AR. These recommendations are justified by the fact that these five lineages have distinct genetic compositions and high level of genetic differentiation from other lineages (Table 2, Fig. 3). Individuals in these lineages form separate groups in PCA and DAPC (Fig. 1), and are further characterized by the lack of ecological exchangeability as evidenced by regional mussels' compatibility to their local host fishes¹³. In addition to these, two additional regions, including *C. aberti* in the Spring and Fall rivers in Kansas and *C. stegaria* in the Clinch River in Tennessee warrant special attention as they appear to be genetically distinct entities, but have very limited geographic distributions.

The present ddRAD-Seq study employed genetic materials obtained by both nondestructive (cytology brush), and destructive (mantle biopsy) methods. Using cytology brushes (or swabs) is commonly used for extracting genomic DNA from endangered and threatened freshwater mussels. Despite the lower recovery rate of samples that passed the filtering steps of the STACKS pipeline in nondestructive genetic samples, our data showed that the quality and amount of ddRAD sequence reads were not significantly different between genetic samples using nondestructive and destructive methods. Our results therefore provide evidence that this nondestructive method can successfully be used for phylogenomic and population genomic studies using ddRAD-Seq analysis and can benefit conservation and management of endangered species such as freshwater mussels in North America.

Materials and methods

Mussel specimens and ddRAD-Seq library preparation. Genomic DNA for this study was derived from DNA extractions of specimens prepared for previous studies^{20,21,23}. *Cyprogenia aberti* specimens were collected from rivers in Arkansas, Missouri, and Kansas, which occur within the Ouachita (Ouachita, Caddo, and Saline rivers) and the Ozark (Fall, Spring, St. Francis, Black, Current, Buffalo, and Strawberry rivers) highland regions (Fig. 2, Table S1). *C. stegaria* specimens were collected from the Eastern Highland Region (Licking, Salt, and Green rivers in Kentucky and the Clinch River in Tennessee).

Samples were collected nondestructively using cytology brushes and stored in the buffer supplied with the Puregene Buccal Cell Kit (Qiagen, Hilden, Germany) and DNA was extracted following the manufacturer's instructions. Mantle biopsies were conducted by taking ~2 mm² tissue and storing it in 95% ethanol (see Chong et al.²³ for details). DNA was extracted from tissue samples using the DNeasy Blood & Tissue Kits (Qiagen, Hilden, Germany). For population genomic analyses, all *Cyprogenia* specimens were grouped according to their sampling locations, i.e., river. In addition, 14 *Dromus dromas* specimens were collected from the Clinch River, TN for this study and were used as an outgroup for phylogenetic analysis and for genetic comparison at the genus level. A total of 214 samples (200 *Cyprogenia sp.* + 14 *Dromus sp.*) were used in this study. Of these, 24 samples were randomly chosen as replicates to explore STACKS^{59,60} parameters for de novo assembly and SNP discovery (see below).

DNA libraries were prepared with modifications of Peterson et al.⁶¹. Briefly, a total of 200 ng of genomic DNA for each sample was normalized into 20 ng/μl, which was then digested using high fidelity restriction enzymes *Pst*I and *Msp*I (New England Biolabs). Resulting fragments were ligated to custom-made P1 containing sample-specific barcodes and forward primer annealing sites and P2 adapters with index and reverse primer annealing sites. Each individual was then PCR-amplified using Phusion Polymerase and following PCR conditions: 25 cycles at 98 °C for 20 s, 60 °C for 30 s and 72 °C for 40 s, with an initial denaturation step at 98 °C for 30 s and a final extension step at 72 °C for 12 min. PCR products were electrophoresed in 1.5% agarose gels and only samples with smeared PCR products were chosen for the next steps. Individually barcoded and indexed samples were then pooled into libraries and size-selected (325–500 bp) using a Blue Pippin (Sage Science) so that the remaining adapters and unbound primers/primer dimers can be removed. Each library was quantified and quality-checked using an Agilent BioAnalyzer 2100 (Agilent Technologies).

A total of 238 samples (214 samples + 24 replicates) were divided into 5 lanes (48 samples per lane, with one lane having 46 samples) and subjected to sequencing. Libraries were single-ended sequenced (100-bp) in separate flow cells on an Illumina HiSeq 3000 (Illumina Inc.) at the DNA Facility of Iowa State University. A simplified flowchart of our methods including ddRAD raw sequence reads, STACKS parameter selection, resultant datasets and data analyses is shown in Fig. S1.

ddRAD locus assembly, STACKS parameter selection, and sample/loci filtering. The STACKS software pipeline, version 1.46^{59,60} was used for demultiplexing, quality filtering, de novo locus assembly, and SNP discovery (see “Supplementary Information”). We evaluated a variety of parameter settings in STACKS according to the recommendations of Mastretta-Yanes et al.⁶² to determine which combination balanced an acceptable SNP error rate and the number of loci retained (Details in “Supplementary Information”). A total of 156 specimens (148 *Cyprogenia* + 8 *Dromus*) were selected for phylogenetic and population genetic analyses. In this study, we applied the following STACKS population constraints to the 156 specimens. In STACKS, *-p*: refers to the minimum number of samples that a locus must be present in for inclusion of the locus in phylogenetic analyses: (1) for the phylogenetic study, only loci that were present in >60% of all samples (i.e. 94 of 156) were retained, (2) for the population genetic study, only loci that were present in 50% of four species-groups (i.e. *-p* 2, here 2 of 4: *C. stegaria*_Eastern, *C. aberti*_Ozark, and *C. aberti*_Ouachita, and *Dromus dromas*), according to Chong et al.²³, and in >60% of all samples within each group (i.e. *-r* 0.6) were retained. Additionally, concatenated SNP data for all samples were also used for phylogenetic analysis of *C. aberti* specimens identified by conglutinate egg colors (14 red and 17 brown). Sample information for the 156 specimens used in this study is shown in Table S1. Raw ddRAD-seq data for each specimen are deposited into the NCBI SRA database (BioProject: PRJNA454895) under the accession numbers listed in Table S1.

Population genetic diversity and structure. The STACKS populations program (*-p* 2, *-r* 0.6) was used to retain only those loci that were present in at least two of the four species groups (*C. stegaria*_Eastern, *C. aberti*_Ozark, *C. aberti*_Ouachita, and *D. dromas*) and present in over 60% of all samples within each group. In total, 14,890 RAD loci were obtained and 7243 SNPs (consisting of only the first SNP per ddRAD-Seq locus) were recovered and used for subsequent population genetic analyses.

Measures of genetic diversity for each of ten sampling locations (rivers) were calculated using GENODIVE⁶³. Pairwise F_{ST} values to assess levels of genetic differentiation between locations were estimated following standard ANOVA⁶⁴ using GENEPOP v.4.7⁶⁵. Significance was tested through genetic differentiation for each population pair using the exact G test with default Markov chain parameters (Dememorisation: 10,000, Batches: 100, Iterations per batch: 5000). Significance level was adjusted for multiple tests using Bonferroni correction.

To explore the effect of missing data in each locus, we compared pairwise F_{ST} 's calculated for both raw data and corrected data in which missing data were replaced with randomly drawn alleles based on the overall allele frequencies using GENODIVE⁶³.

GENODIVE was further used to conduct a hierarchical Analysis of Molecular Variance (AMOVA)⁶⁶ to assess the partitioning of genetic variation among highlands, populations within highlands, individuals within populations and within individuals. Two separate AMOVA analyses were conducted (three highland regions vs. two highland regions). The comparison of three highland regions compared the Ozark, Ouachita, and Eastern

highlands, which mainly accounts for variance between *Cyprogenia aberti* and *C. stegaria*. Whereas the comparison of two highland regions (Ozark vs Ouachita highlands), account for variance within *C. aberti*. Analyses were carried out using an Infinite Allele Model (IAM) and 999 permutations to assess its significance, where missing data were replaced by randomly drawn alleles based on overall allele frequencies.

We examined the distributions of individual multilocus genotypes in multivariate space, to assess the geographic structuring among individuals. We carried out a principal component analysis (PCA) based on 7243 SNP loci across all samples and *Cyprogenia* samples only using the adegenet package⁶⁷ implemented in R⁶⁸. A scatter diagram was plotted based on factor scores along the first two PC axes accounting for the most variation. We also conducted a Discriminant Analysis of Principal Components (DAPC) using an interactive web interface for DAPC distributed with the package [using R command: `adegenetServer(what = "DAPC")`]. DAPC aims to provide an efficient description of genetic clusters using a few synthetic variables that are constructed as linear combinations of the original variables (alleles) which have the largest between-group variance and the smallest within-group variance⁶⁹. We used the default parameters except selecting the option of suggested number of PCA components.

To infer the most likely number of genetic clusters, we applied a Bayesian clustering approach implemented in STRUCTURE 2.3.4⁷⁰. The admixture ancestry model was used with correlated allele frequencies and no prior on population membership. The initial analysis of the complete dataset was conducted ten times for each K value (the number of inferred genetic clusters) from 1 to 10 with 200,000 Markov Chain Monte Carlo (MCMC) generations and a burn-in of 100,000. The optimal K value was inferred from ad hoc posterior probability models of $[\text{Pr}(X|K)]^{70}$ and $\text{delta}K^{71}$ using STRUCTURE HARVESTER⁷². To assess the presence of substructure within major clusters, subsequent analyses were conducted within each of main clusters, using the same parameters as described above. A STRUCTURE bar plot to infer the genetic composition of individuals was visualized using DISTRUCT v.1.1⁷³.

Phylogenetic analyses. To explore the effect of missing data, we compared data sets generated by four different filtering criteria using the STACKS population program. We varied this criterion in order of strictness (i.e., inclusion of loci shared by fewer to more samples); $-p\ 78$, $-p\ 94$, $-p\ 109$, and $-p\ 125$, which only retains loci present in at least 50%, 60%, 70% and 80% of the 156 samples respectively. Using the STACKS population program (`-phylip_var`) and the *D. dromas* specimens as an outgroup, these four data sets of concatenated SNPs were used to investigate evolutionary relationships among the 156 specimens. These four different filtering criteria respectively generated concatenated SNPs of 15,829, 9673, 4674, and 1283, each with parsimony-informative sites of 7440, 4395, 2014, and 511. Phylogenetic trees were constructed with Maximum likelihood (ML) analyses in IQtree 2.1.2^{74,75} with 1000 ultrafast bootstrap replicates (`-bb`), 10,000 iterations (`-nm`) and Ascertainment bias correction (`+ASC`) model. The TVM + F + ASC + G4 model of sequence evolution was selected as a best-fit model according to a Bayesian Information Criterion (BIC) model selection implemented in IQtree 2.1.2. ML trees constructed from different filtering criteria produced nearly identical topologies (Fig. S4), therefore we used a data set in which SNPs present in at least 60% of samples ($-p\ 94$) for downstream analyses.

We used MrBayes v 3.2⁷⁶ to construct phylogenetic trees using Bayesian inference (BI) with the GTR-I-GAMMA model (General Time Reversible model with a proportion of invariable sites and a gamma-shaped distribution of rates across sites), which includes a parameter for site heterogeneity. BI employed four simultaneous Monte Carlo Markov chains (one cold and three heated) with 1,000,000 generations and sampled every 500 generations. The first 25% of the data points were discarded as burn-in. We confirmed that both ML and Bayesian tree searches achieved convergence (ML tree: Bootstrap correlation coefficient of split occurrence frequencies = 0.99, Bayesian tree: PSRFs = 1.000592). The consensus trees from both ML and BI were illustrated using FigTree v 1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree/>).

Species tree and species delimitation. Species tree inference and species delimitation in a Bayesian framework using SNP data were investigated using SNAPP⁷⁷ and MODEL_SELECTION implemented in BEAST version 2⁷⁸. Species Trees were inferred directly from SNP markers by bypassing gene trees in a full coalescent analysis, using a polynomial-time algorithm that computes the likelihood of a species tree directly from the markers under a finite-sites model of mutation effectively integrating over all possible gene trees⁷⁷. Out of 148 samples for *Cyprogenia*, a total of 40 samples for ten sampling sites (four samples per each sampling site) were aligned and selected based on two sampling criteria, random sampling of individuals or by selecting the individual samples with the highest coverage (smallest number of missing SNPs) using a custom Python script. Coalescent-based species tree was inferred using XML file with following model parameters; with selecting "Cal mutation rates", unselecting "Include non-polymorphic sites" and selecting "Use Log Likelihood Correction" in Model Parameter, and the chain length of 5,000,000 with storing every 1000 in MCMC, and the rest of parameters were set to default.

For species delimitation exploration, candidate species delimitation models were explored using marginal likelihood estimation (MLE) of each competing species delimitation model, ranking models by marginal likelihood, and using Bayes factors⁷⁹ to assess support for model rankings. For species delimitation using genetic data and coalescent methods, we designed six candidate species delimitation models; Model 1: current taxonomy (two *Cyprogenia* species), Model 2: split *C. aberti* based on highland regions into two species (*C. aberti* was split into Ozark (Black, St. Francis, Spring) and Ouachita (Caddo, Ouachita, Saline)), Model 3: intermix half of two candidate *C. aberti* species from each other (Half of Ozark and Ouachita were equally intermixed), Model 4: reassign *C. aberti* (Black from Ozark was moved to Ouachita, and Saline from Ouachita was moved to Ozark), Model 5: reassign *C. aberti* (Saline from Ouachita was moved to Ozark), Model 6: reassign *C. aberti* (Black from Ozark was moved to Ouachita). For current and alternative species delimitation models, we edited the XML

file for marginal likelihood estimation, with the path sampling steps of 36, MCMC sample length of 100,000 for each path sampling step, alpha of 0.3, burnInPercentage = 10, and preBurnin of 10,000. Our path sampling parameters might be low, and a thorough analysis may require more path sampling steps and higher MCMC. However, given computational time and capacity, and similar results from various parameter sets applied, such parameter sets are reasonable.

Bayes factor (BF) calculations based on marginal likelihood estimation (MLE) of each competing species delimitation model are made against the current taxonomy model (Model 1), subtracting the MLE values for two models, and then multiplying the difference by two ($BF = 2 \times (\text{Model 1} - \text{alternative model})$). The strength of support from BF comparisons of competing models was evaluated using the framework of Kass and Raftery⁷⁹. The BF scale is as follows: $0 < BF < 2$ is not worth more than a bare mention, $2 < BF < 6$ is positive evidence, $6 < BF < 10$ is strong support, and $BF > 10$ is decisive. All Bayes factor (BF) calculations are made against the current taxonomy model (Model 1). Therefore, positive BF values indicate support for current taxonomy model, and negative BF values indicate support for alternative model.

Trace files generated by SNAPP were inspected and convergence of chain by checking Effective Sample Size (ESS) for various variables was evaluated using the program Tracer (<http://tree.bio.ed.ac.uk/software/tracer>). Posterior distribution of species trees was summarized to identify the topology with the best posterior support using TreeAnnotator implemented in BEAST version 2⁷⁸. The target tree file generated with Maximum clade credibility (MCC) and Median heights option in TreeAnnotator was displayed in FigTree.

Demographic changes in *Cyprogenia*. We investigated the demographic changes for each of the *Cyprogenia* sampling locations in the three highland regions using the Approximate Bayesian Computation (ABC) approach implemented in DIYABC 2.1⁸⁰. To evaluate the past demographic changes, we applied the methods of Cabrera and Palsbøll⁸¹. Five demographic models were evaluated. Model 1: CON (constant population size), Model 2: DEC (a single instantaneous decrease in population size), Model 3: INC (a single instantaneous increase in population size), Model 4: INCDEC (a single instantaneous increase followed by a single instantaneous decrease in population size), Model 5: DECINC (a single instantaneous decrease followed by a single instantaneous increase in population size). These five models were tested for each of sampling locations and the best fit model was selected by the model comparison implemented in the DIYABC 2.1. We followed the prior model parameterization designed by Cabrera and Palsbøll⁸¹ with the slight modification of generation time (T) and a uniform distribution with a range of 10–20,000 (N). Similar to Cabrera and Palsbøll⁸¹, the Last Glacial Maximum (LGM) of approximately 26,500–15,000 years ago was employed as the main cause for changes in population size of these species. In the case of the models DEC and INC, the change in population size occurred after the LGM (i.e., approximately 11,250 years ago), and in the case of the models INCDEC and DECINC, the changes in population size occurred just before and just after the LGM (i.e., approximately 22,500 and 11,250 years ago, respectively)⁸¹. We employed a generation time of 5 years for *Cyprogenia*, therefore a uniform prior distribution with a range of 100–2250 for T1 (after the LGM), and a range of 2250–4500 for T2 (before the LGM). These prior distributions are reasonable given the species' presumed census size (100–20,000 mussels) and generation time (5–9 years per generation)⁸².

To minimize effect of missing data in the analyses, we used only loci that were found in all populations using independent parameters of populations program in STACKS for each of three Highlands (-p 3 for 3 populations in Ozark Highland, and -p 3 for 3 populations in Ouachita Highland, and -p 4 for four populations in Eastern Highland). A vcf dataset generated from the Stacks populations command was changed to .snp file using python script (vcf2diyabc.py) implemented in the DIYABC program and monomorphic loci were further removed from the dataset so that only polymorphic loci were employed in the ABC analyses. No mutation model parameterization was required for SNPs. Models for demographic changes and their prior distributions were provided in the "Supplementary Information".

Genetic relationship among specimens identified by conglutinate lure color. The color of conglutinate lures (red and brown colored eggs) of included specimens were identified previously²³, and a total of 31 *C. aberti* specimens with known conglutinate lure color (14 with red conglutinates and 17 with brown conglutinates) were selected to examine genetic relationships between specimens. Concatenated SNPs obtained from a moderate criterion (-p 94) described earlier were used for phylogenetic analyses. Maximum likelihood (ML) and Bayesian trees were constructed to examine phylogenetic relationships for total of 33 specimens, 31 *C. aberti* specimens, and two *C. stegeria* specimens used as outgroups. The TVM + I + G4 model of sequence evolution was selected according to a Bayesian Information Criterion (BIC) implemented in IQtree.

Quality comparison between genetic materials using non-destructive and destructive methods. To investigate the utility of non-destructive methods as source material for genomic library creation, we compared the results obtained from samples collected using cytology brushes and mantle-tissue biopsies. First, we compared the retention rate of samples from each method after the aforementioned filtering steps. Then, we compared the quality and amount of ddRAD sequence reads between genetic samples using both methods. In addition, we estimated and compared differences in SNP error rates between replicated samples from each sample type. Lastly, we mapped and compared ddRAD fastq sequences from all samples to four genome sequences (bivalve, human, yeast and a bacterium) and PhiX sequences using FastQ_Screen (https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/). Genome sequences of a bivalve, *Mytilus galloprovincialis*, are available from GenBank (accession: LNJA000000000). Genome sequences of human (*Homo sapiens*), and yeast (*Saccharomyces cerevisiae*) are available from ftp://ftp.ensembl.org/pub/current_fasta. Genome sequences of a

bacterium (*Escherichia coli*) are available from GenBank (accession: U00096.3). PhiX sequences are available from Refseq accession NC_001422.1.

Data availability

Raw ddRAD-seq data for each sample are deposited in the NCBI SRA database (BioProject ID: PRJNA454895) under the individual SRA accession numbers listed in Table S1. Raw ddRAD-seq data are available in the same BioProject under the SRA Accession no. (SAMN09060972–SAMN09061127).

Received: 4 January 2021; Accepted: 10 May 2021

Published online: 24 May 2021

References

1. Frankham, R. Challenges and opportunities of genetic approaches to biological conservation. *Biol. Conserv.* **143**, 1919–1927 (2010).
2. Goldstein, P. Z., Desalle, R., Amato, G. & Vogler, A. P. Conservation genetics at the species boundary. *Conserv. Biol.* **14**, 120–131 (2000).
3. Isaac, N. J. B., Mallet, J. & Mace, G. M. Taxonomic inflation: Its influence on macroecology and conservation. *Trends Ecol. Evol.* **19**, 464–469 (2004).
4. Lydeard, C. *et al.* The global decline of nonmarine mollusks. *Bioscience* **54**, 321–330 (2004).
5. Haag, W. R. & Williams, J. D. Biodiversity on the brink: An assessment of conservation strategies for North American freshwater mussels. *Hydrobiologia* **735**, 45–60 (2014).
6. Ricciardi, A. & Rasmussen, J. Extinction rates of North American freshwater fauna. *Conserv. Biol.* **13**, 1220–1222 (1999).
7. Spooner, D. E. & Vaughn, C. C. Context-dependent effects of freshwater mussels on stream benthic communities. *Freshw. Biol.* **51**, 1016–1024 (2006).
8. Vaughn, C. C., Spooner, D. E. & Galbraith, H. S. Context-dependent species identity effects within a functional group of filter-feeding bivalves. *Ecology* **88**, 1654–1662 (2007).
9. Vaughn, C. C., Nichols, S. J. & Spooner, D. E. Community and foodweb ecology of freshwater mussels. *J. N. Am. Benthol. Soc.* **27**, 409–423 (2008).
10. McMahon, R. F. *Ecology and Classification of North American Freshwater Invertebrates* (Academic Press, 1991).
11. Watters, G. T. Unionids, fishes, and the species-area curve. *J. Biogeogr.* **19**, 481–490 (1992).
12. Haag, W. R. & Warren, M. L. Host fishes and reproductive biology of 6 freshwater mussel species from the Mobile Basin, USA. *J. N. Am. Benthol. Soc.* **16**, 576–585 (1997).
13. Eckert, N. L. Reproductive biology and host requirement differences among isolated populations of *Cyprogenia aberti* (Conrad, 1850). *MS Thesis, Southwest Missouri State University, Springfield* (2003).
14. Barnhart, M. C., Haag, W. R. & Roston, W. N. Adaptations to host infection and larval parasitism in Unionoida. *J. N. Am. Benthol. Soc.* **27**, 370–394 (2008).
15. Rogers, S. O., Watson, B. T. & Neves, R. J. Life history and population biology of the endangered tan riffleshell (*Epioblasma florentina walkeri*) (Bivalvia: Unionidae). *J. N. Am. Benthol. Soc.* **20**, 582–594 (2001).
16. Burr, B. M. & Mayden, R. L. Phylogenetics and North American freshwater fishes. In: *Systematics, Historical Ecology, and North American Freshwater Fishes*. (Stanford University Press, 1992).
17. Oesch, R. D. *Missouri Naiades: A Guide to the Mussels of Missouri* (Missouri Department of Conservation, 1995).
18. Harris, J. L. *et al.* Unionoida (Mollusca: Margaritiferidae, Unionidae) in Arkansas, third status review. *J. Ark. Acad. Sci.* **63**, 50–86 (2009).
19. Obermeyer, B. K. *Recovery plan for four freshwater mussels in southeast Kansas: Neosho mucket (Lampsilis rafinesqueana), Ouachita kidneyshell (Ptychobranchius occidentalis), rabbitsfoot (Quadrula cylindrica cylindrica), and western fanshell (Cyprogenia aberti)*. Kansas Department of Parks and Wildlife (2000).
20. Serb, J. M. Discovery of genetically distinct sympatric lineages in the freshwater mussel *Cyprogenia aberti* (Bivalvia: Unionidae). *J. Molluscan Stud.* **72**, 425–434 (2006).
21. Grobler, J. P., Jones, J. W., Johnson, N. A., Neves, R. J. & Hallerman, E. M. Homogeneity at nuclear microsatellite loci masks mitochondrial haplotype diversity in the endangered fanshell pearlymussel (*Cyprogenia stegaria*). *J. Hered.* **102**, 196–206 (2011).
22. Serb, J. M. & Barnhart, M. C. Congruence and conflict between molecular and reproductive characters when assessing biological diversity in the Western Fanshell *Cyprogenia aberti* (Bivalvia, Unionidae) I. *Ann. Missouri Bot. Gard.* **95**, 248–261 (2008).
23. Chong, J. P., Harris, J. L. & Roe, K. J. Incongruence between mtDNA and nuclear data in the freshwater mussel genus *Cyprogenia* (Bivalvia: Unionidae) and its impact on species delineation. *Ecol. Evol.* **6**, 2439–2452 (2016).
24. Hohenlohe, P. A. *et al.* Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* **6**, e1000862 (2010).
25. Leaché, A. D., Fujita, M. K., Minin, V. N. & Bouckaert, R. R. Species delimitation using genome-wide SNP Data. *Syst. Biol.* **63**, 534–542 (2014).
26. Bruneaux, M. *et al.* Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach. *Mol. Ecol.* **22**, 565–582 (2013).
27. Wagner, C. *et al.* Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol. Ecol.* **22**, 787–798 (2013).
28. Larson, W. A. *et al.* Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evol. Appl.* **7**, 355–369 (2014).
29. Lee, S.-R., Jo, Y.-S., Park, C.-H., Friedman, J. M. & Olson, M. S. Population genomic analysis suggests strong influence of river network on spatial distribution of genetic variation in invasive saltcedar across the southwestern United States. *Mol. Ecol.* **27**, 636–646 (2017).
30. Massatti, R., Reznicek, A. A. & Knowles, L. L. Utilizing RADseq data for phylogenetic analysis of challenging taxonomic groups: A case study in carex sect. Racemosae. *Am. J. Bot.* **103**, 337–347 (2016).
31. Razkin, O. *et al.* Species limits, interspecific hybridization and phylogeny in the cryptic land snail complex *Pyramidula*: The power of RADseq data. *Mol. Phylogenet. Evol.* <https://doi.org/10.1016/j.ympev.2016.05.002> (2016).
32. Rubin, B. E. R., Ree, R. H. & Moreau, C. S. Inferring phylogenies from RAD sequence data. *PLoS ONE* **7**, e33394 (2012).
33. Takahashi, T., Nagata, N. & Sota, T. Application of RAD-based phylogenetics to complex relationships among variously related taxa in a species flock. *Mol. Phylogenet. Evol.* <https://doi.org/10.1016/j.ympev.2014.07.016> (2014).
34. Boucher, F. C., Casazza, G., Szövényi, P. & Conti, E. Sequence capture using RAD probes clarifies phylogenetic relationships and species boundaries in *Primula* sect. *Auricula*. *Mol. Phylogenet. Evol.* <https://doi.org/10.1016/j.ympev.2016.08.003> (2016).
35. Combosch, D. J., Lemer, S., Ward, P. D., Landman, N. H. & Giribet, G. Genomic signatures of evolution in *Nautilus*—An endangered living fossil. *Mol. Ecol.* **26**, 5923–5938 (2017).
36. Cruaud, A. *et al.* Empirical assessment of RAD sequencing for interspecific phylogeny. *Mol. Biol. Evol.* **31**, 1272–1274 (2014).

37. Eaton, D. A. R. & Ree, R. H. Inferring phylogeny and introgression using RADseq data: An example from flowering plants (Pedicularis: Orobanchaceae). *Syst. Biol.* **62**, 689–706 (2013).
38. Emerson, K. J. *et al.* Resolving postglacial phylogeography using high-throughput sequencing. *Proc. Natl. Acad. Sci. USA* **107**, 16196–16200 (2010).
39. Herrera, S. & Shank, T. M. RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. *Mol. Phylogenet. Evol.* **100**, 70–79 (2016).
40. Hipp, A. L. *et al.* A framework phylogeny of the American Oak Clade based on sequenced RAD data. *PLoS ONE* **9**, e93975 (2014).
41. Jones, J. C., Fan, S., Franchini, P., Scharl, M. & Meyer, A. The evolutionary history of Xiphophorus fish and their sexually selected sword: A genome-wide approach using restriction site-associated DNA sequencing. *Mol. Ecol.* **22**, 2986–3001 (2013).
42. Funk, W. C. *et al.* Adaptive divergence despite strong genetic drift: Genomic analysis of the evolutionary mechanisms causing genetic differentiation in the island fox (*Urocyon littoralis*). *Mol. Ecol.* **25**, 2176–2194 (2016).
43. Taberlet, P. & Luikart, G. Non-invasive genetic sampling and individual identification. *Biol. J. Linn. Soc.* **68**, 41–55 (1999).
44. Palsbøll, P. J., Bérubé, M. & Allendorf, F. W. Identification of management units using population genetic data. *Trends Ecol. Evol.* **22**, 11–16 (2007).
45. Gibbs, J., Jr. Hunter, M. & Sterling, E. Population genetics: Diversity within versus diversity among populations. In: *Problem-Solving in Conservation Biology and Wildlife Management: Exercises for Class, Field, and Laboratory* 29–35 (Blackwell Publishing Ltd., 2008). <https://doi.org/10.1002/9781444319576.ch4>.
46. Berendzen, P. B., Simons, A. M., Wood, R. M., Dowling, T. E. & Secor, C. L. Recovering cryptic diversity and ancient drainage patterns in eastern North America: Historical biogeography of the *Notropis rubellus* species group (Teleostei: Cypriniformes). *Mol. Phylogenet. Evol.* **46**, 721–737 (2008).
47. Ray, J. M., Wood, R. M. & Simons, A. M. Phylogeography and post-glacial colonization patterns of the rainbow darter, *Etheostoma caeruleum* (Teleostei: Percidae). *J. Biogeogr.* **33**, 1550–1558 (2006).
48. Strange, R. M. & Burr, B. M. Intraspecific phylogeography of North American highland fishes: A test of the pleistocene vicariance hypothesis. *Evolution (N. Y.)* **51**, 885–897 (1997).
49. Pflieger, W. L. A distributional study of missouri fishes. *Univ. Kans. Publ. Mus. Nat. Hist.* **20** (1971).
50. Thornbury, W. D. Regional geomorphology of the United States. *J. Geol.* **73**, 815–816 (1965).
51. Mayden, R. Vicariance biogeography, parsimony, and evolution in North American freshwater fishes. *Syst. Zool.* **37**, 329–355 (1988).
52. Echelle, A. A., Echelle, A. F., Smith, M. H. & Hill, L. G. Analysis of genic continuity in a headwater fish, *Etheostoma radiosum* (Percidae). *Copeia* **1975**, 197–204 (1975).
53. Haponski, A. E., Bollin, T. L., Jedlicka, M. A. & Stepien, C. A. Landscape genetic patterns of the rainbow darter *Etheostoma caeruleum*: A catchment analysis of mitochondrial DNA sequences and nuclear microsatellites. *J. Fish Biol.* **75**, 2244–2268 (2010).
54. Turner, T. F. & Trexler, J. C. Ecological and historical associations of gene flow in darters (Teleostei: Percidae). *Evolution (N. Y.)* **52**, 1781–1801 (1998).
55. Turner, T. F., Trexler, J. C., Kuhn, D. N. & Robison, H. W. Life-history variation and comparative phylogeography of darters (Pisces: Percidae) from the North American Central Highlands. *Evolution (N. Y.)* **50**, 2023–2036 (1996).
56. Cross, F., Mayden, R. & Stewart, J. *Fishes in the Western Mississippi Drainage. The Zoogeography of North American Freshwater Fishes* (Wiley, 1986).
57. Barnhart, M. C. *Reproduction and Fish Host of the Western Fanshell, Cyprogenia aberti (Conrad 1850)* (Kansas Department of Wildlife and Parks, 1997).
58. Inoue, K., Monroe, E. M., Elderkin, C. L. & Berg, D. J. Phylogeographic and population genetic analyses reveal Pleistocene isolation followed by high gene flow in a wide ranging, but endangered, freshwater mussel. *Heredity (Edinb.)* **112**, 282–290 (2014).
59. Catchen, J. M., Amores, A., Hohenlohe, P. A., Cresko, W. A. & Postlethwait, J. H. Stacks: Building and genotyping Loci de novo from short-read sequences. *G3 Genes Genomes Genet.* **1**, 171–182 (2011).
60. Catchen, J. M., Hohenlohe, P. A., Bassham, S., Amores, A. & Cresko, W. A. Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–3140 (2013).
61. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. & Hoekstra, H. E. Double digest RADseq: An inexpensive method for De Novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7**, e37135 (2012).
62. Mastretta-Yanes, A. *et al.* Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Mol. Ecol. Resour.* **15**, 28–41 (2015).
63. Meirmans, P. G. & Van Tienderen, P. H. Genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. *Mol. Ecol. Notes* **4**, 792–794 (2004).
64. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution (N. Y.)* **38**, 1358–1370 (1984).
65. Raymond, M. & Rousset, F. GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. *J. Hered.* **86**, 248–249 (1995).
66. Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491 (1992).
67. Jombart, T. adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).
68. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at <https://www.R-project.org/> (2018).
69. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94 (2010).
70. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
71. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software structure: A simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).
72. Earl, D. A. & Vonholdt, B. M. Structure Harvester: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**, 359–361 (2012).
73. Rosenberg, N. A. distruct: A program for the graphical display of population structure. *Mol. Ecol. Notes* **4**, 137–138 (2004).
74. Minh, B., Nguyen, M.-A. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
75. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
76. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
77. Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A. & Roychoudhury, A. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* **29**, 1917–1932 (2012).
78. Bouckaert, R. *et al.* BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, 1–6 (2014).
79. Kass, R. E. & Raftery, A. E. kass1995BayesFactors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).

80. Cornuet, J.-M. *et al.* DIYABC v2.0: A software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics* **30**, 1187–1189 (2014).
81. Cabrera, A. A. & Palsbøll, P. J. Inferring past demographic changes from contemporary genetic data: A simulation-based evaluation of the ABC methods implemented in diyabc. *Mol. Ecol. Resour.* **17**, e94–e110 (2017).
82. Jones, J. W. & Neves, R. J. Life history and propagation of the endangered fanshell pearlymussel, *Cyprogenia stegaria* Rafinesque (Bivalvia: Unionidae) The University of Chicago Press on behalf of the Society for F. J. N. *Am. Benthol. Soc.* **21**, 76–88 (2002).

Acknowledgements

We would like to acknowledge J. Serb, P. Grobler, J. Jones and J. P. Chong for sharing DNA samples and John Harris for collecting samples. We are also grateful for the valuable comments of Jeanne Serb, John Harris and other anonymous reviewers for providing their helpful comments and edits on this manuscript.

Author contributions

K.R. conceived the study and K.K. conducted laboratory work and analyzed the data. K.K. and K.R. drafted and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-90325-0>.

Correspondence and requests for materials should be addressed to K.J.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021