Rapture-ready darters: choice of reference genome and genotyping method (whole-genome or sequence capture) influence population genomic inference in *Etheostoma*

Brendan N. Reid[1*], Rachel L. Moran[2], Christopher J. Kopack[3], Sarah W. Fitzpatrick[1,4]

[1]Kellogg Biological Station, Michigan State University, Hickory Corners, MI, USA

[2]Department of Evolution, Ecology, and Behavior, University of Minnesota, Saint Paul, MN, USA

[3]Department of Biology, Colorado State University, Fort Collins, CO, USA

[4]Department of Integrative Biology, Michigan State University, East Lansing, MI, USA

[*]Corresponding author: reidbre1@msu.edu

Abstract

Researchers studying non-model organisms have an increasing number of methods available for generating genomic data. However, the applicability of different methods across species, as well as the effect of reference genome choice on population genomic inference, are still difficult to predict in many cases. We evaluated the impact of data type (whole-genome vs. reduced representation) and reference genome choice on data quality and on population genomic and phylogenomic inference across several species of darters (subfamily Etheostomatinae), a highly diverse radiation of freshwater fish. We generated a high-quality reference genome and developed a hybrid RADseq/sequence capture (Rapture) protocol for the Arkansas darter (*Etheostoma cragini*). Rapture data from 1900 individuals spanning four darter species showed recovery of most loci across darter species at high depth and consistent estimates of heterozygosity regardless of reference genome choice. Loci with baits spanning both sides of the restriction enzyme cut site performed especially well across species. For low-coverage whole-genome data, choice of reference genome affected read depth and inferred heterozygosity. For similar amounts of sequence data, Rapture performed better at identifying fine-scale genetic structure compared to whole-genome sequencing. Rapture loci also recovered an accurate phylogeny for the study species and demonstrated high phylogenetic informativeness across the evolutionary history of the genus *Etheostoma*. Low cost and high cross-species effectiveness regardless of reference genome suggest that Rapture and similar sequence capture methods may be worthwhile choices for studies of diverse species radiations.

Keywords: phylogeography, species radiation, bait design, heterospecific genome

Introduction

The advent of high-throughput sequencing (HTS) technology has enabled biologists to generate genome-scale molecular data from a variety of organisms, creating new opportunities for conservation genetics (Shafer et al. 2015), phylogenetics (Lemmon and Lemmon 2013, McCormack et al. 2013), and molecular ecology (Ekblom & Gallindo 2011). As the capacity for HTS has increased, however, repositories of sequence data have become increasingly biased toward sequences from a minority of model organisms (David et al. 2019). Although non-model organisms represent fruitful study systems for answering basic questions in biology (Russell et al. 2017), deciding on appropriate methods for generating and handling genomic data for non-model species remains a challenge.

Whole-genome sequencing may still remain out of reach for large-scale studies of non-model organisms, and as such reduced-representation approaches have grown popular as effective means for answering many questions (da Fonseca et al. 2016, Meek and Larson 2019). Sequence capture or targeted sequence enrichment methods represent an attractive method for generating repeatable, high-coverage sequence data (Grover et al. 2012, Harvey et al. 2016). A hybrid method that uses restriction-associated DNA sequencing (RADseq) combined with targeted enrichment of a user-defined subset of hundreds to thousands of RAD loci, termed 'Rapture' (Ali et al. 2016) has great potential as a rapid and efficient method for generating repeatable high-throughput genomic data at low cost and high efficiency. Rapture assays have so far been developed and applied to salmon (Ali et al. 2016), Tasmanian devils (Margres et al. 2018), marine turtles (Komoroske et al. 2019), frogs (Peek et al. 2019), and sea lampreys (Sard et al. 2020). The application of Rapture has mainly focused on population genomics within species,

58 although Rapture loci developed for one species have been shown to be useful for studying
59 hybridization among closely related species (Peek et al. 2019) and across species within slowly-
60 evolving lineages (Komoroske et al. 2019).
61
62 For both whole-genome and reduced-representation sequencing, high-quality reference genomes
63 can be used to improve genotype calling accuracy, inference of demographic history, and
64 identification of loci under selection (Manel et al. 2015, Brandies et al. 2019). For studies of non-
65 model species, however, reference genomes may not be available for the particular species of
66 interest. Assembling HTS data to heterospecific genomes of related species is a potential option
67 when such genomes are available. However, simulation studies indicate that even small
68 divergences (0.15% to 2%) between the heterospecific reference genomes and the conspecific
69 genome of the species of interest can increase errors in polymorphism calling and in estimates of
70 genetic diversity, particularly when read depths are low (Nevado et al. 2014). Still, the practice
71 of assembling short reads to a reference genome from a closely related species is common, and
72 other empirical studies have concluded that congeneric or confamilial reference genomes may be
73 suitable for SNP discovery, at least in groups with highly conserved genomes (Galla et al. 2019).
74
75 Using a conspecific reference genome in every situation is ideal but likely infeasible, especially
76 when studying highly diverse species radiations. Applying HTS to the study of diverse species
77 radiations will be particularly useful for understanding the effects of environmental context on
78 genome evolution and identifying links between genetic variation and adaptive traits. Indeed,
79 whole genome sequencing as well as reduced representation sequencing of adaptive radiations
80 has uncovered signatures of change in genome structure and selection in African cichlid fish

81 (Brawand et al. 2014) and specific genetic loci associated with beak and body size variation in
82 Darwin's finches (Chaves et al. 2016). However, the cost of generating separate reference
83 genomes for each species may be prohibitive, and making population genomic comparisons
84 among species often necessitates assembling data to a single reference genome (as in Chaves et
85 al. 2016). If using heterospecific reference genomes is unavoidable in studies of diverse species
86 radiations, it is important to quantify the biases that using these genomes will create when
87 working with different types of data.
88
89 Darters (subfamily Etheostomatinae) represent a species radiation with great potential for
90 illuminating the biotic and abiotic mechanisms that generate biological diversity. Darters are one
91 of the most diverse clades of freshwater fish in North America, consisting of approximately 250
92 currently described species that likely shared a common ancestor between 30 and 40 million
93 years ago (Near et al. 2011). Darters exhibit sexually dimorphic coloration that varies
94 substantially among species, and sexual isolation based on divergent sexual selection has likely
95 contributed to diversification in this group (Mendelson 2003, Moran et al. 2017, Moran and
96 Fuller 2018a, Moran and Fuller 2018b). Postzygotic barriers between many sympatric species are
97 not complete and hybridization is common, leading to gene tree discord and detectable signatures
98 of ancient and contemporary introgression (Bossu & Near 2013; Moran et al 2017; Moran et al.
99 2018). Darters are dispersal-limited and often restricted to small headwater streams, and as such
100 allopatric diversification due to physical isolation also plays a large role in their diversification
101 (Near and Benard 2004, Hollingsworth and Near 2009). In addition to driving diversification,
102 physical isolation and micro-endemicity, as well as habitat degradation, have created

103 conservation issues for many darter species, and a substantial proportion of darter species
104 diversity is currently considered threatened or endangered (Jelks et al. 2008).
105
106 HTS has great potential for providing insight into the forces controlling diversification in darters
107 as well as for landscape and conservation genomics. Darter research to date has been
108 characterized by a patchwork of molecular methods, making the comparison and integration of
109 data from different studies difficult. Most previous phylogenetic work in darters has focused on
110 Sanger sequencing of a small number of mitochondrial and nuclear genes (Near et al. 2011),
111 while conservation genetics, landscape genetics, and molecular ecology studies have mainly used
112 microsatellite markers developed for single species but with some applicability across the clade
113 (Tonnis 2006, Khudamrongsawat et al. 2007, Switzer et al. 2008, Gabel et al. 2008, Hudman et
114 al. 2008, Saarinen and Austin 2010). Recent work has begun to incorporate HTS methods,
115 employing single-digest RADseq (Moran et al. 2018, MacGuigan et al. 2019, Moran et al. 2020)
116 and double-digest RADseq (ddRAD, Moran et al. 2017, George 2018) to investigate phylogeny,
117 phylogeography, and reproductive barriers among species. While ddRAD and RADseq represent
118 a huge leap forward in terms of the amount of data generated, these methods often increase the
119 number of loci genotyped at the expense of missing data and low coverage (MacGuigan et al.
120 2019). As such, there is currently no published method for reproducibly generating data for a
121 single consistent set of loci distributed across the genome for darters. Furthermore, a reference
122 genome assembly has only recently become available for a single darter species (the
123 orangethroat darter *Etheostoma spectabile*; Moran et al. 2020).
124

125 Here, we describe an efficient and inexpensive Rapture-based method for reliably and repeatably
126 genotyping thousands of loci in darters. This method is based on a capture bait set developed
127 from RADseq data for Arkansas darters (*Etheostoma cragini*), a species of conservation concern
128 found in the Arkansas River and nearby drainages within the Great Plains. Previous work in this
129 species has used microsatellite markers to examine factors influencing population structure and
130 genetic diversity in the western portion of their range (Fitzpatrick et al. 2014). The capture bait
131 set targets over 2000 loci and includes both putatively neutrally-evolving loci as well as loci
132 showing some evidence of selection across this species' range. We assess two different tiling
133 schemes for these baits, targeting either one or both flanking regions adjacent to a restriction cut
134 site. We assess the performance of this capture bait set in a large set ($n > 1600$) of individual
135 Arkansas darters as well as for individuals of three additional species in the genus *Etheostoma*.
136 We assess the effects of aligning to either the heterospecific *E. spectabile* genome (which likely
137 diverged from *E. cragini* approximately 29 million years ago; Kelly et al. 2015) or to a novel
138 conspecific *E. cragini* genome, and we also compare estimates of genetic diversity and
139 population structure from Rapture to estimates from low-coverage whole-genome sequencing
140 (WGS) data for a subset of *E. cragini* individuals. We ask the following questions to gauge the
141 performance and applicability of the method: 1) How often are loci sequenced using the Rapture
142 baits recovered at high coverage (>20x), and how many reads per individual are needed to attain
143 high coverage?; 2) How much diversity is present within the set of Rapture loci for both the
144 target species and for other darter species?; 3) Can the Rapture loci identify distinct population
145 units within *E. cragini*?; and 4) Do the Rapture loci recover known phylogenetic relationships
146 among and within species? We also demonstrate how the choice of data type (Rapture vs. WGS)

147 and reference genome (heterospecific vs conspecific) affects inference of population genetic
148 parameters and population structure.
149
150 **Methods**
151 *Sampling*
152 Dipnetting and electrofishing were used by Kansas Department of Wildlife, Parks, and Tourism
153 personnel to collect 2,374 *E. cragini* individuals at 216 sites throughout Kansas in 2015-2016.
154 Fin clips were taken from adults (>28mm) and whole specimens were collected for juveniles
155 (<28mm). Samples were stored in 100% ethanol, shipped to Michigan State University (MSU),
156 then stored in a freezer (-20°C) prior to analysis. In addition to the Kansas samples, whole *E.*
157 *cragini* specimens were collected from six sites in Arkansas by the Arkansas Fish and Game
158 Commission. Tissue samples and isolated DNA from *E. cragini* individuals collected by the
159 Colorado Department of Parks and Wildlife were also available from a previous study
160 (Fitzpatrick et al. 2014). Sample information is provided in Supporting Table 1.
161
162 To examine the efficacy of Rapture across darter species, we also obtained genetic samples from
163 three additional darter species: rainbow darters (*E. caeruleum*) collected for a separate
164 population genetic study in southwestern Michigan (Oliveira et al. 2020), *E. spectabile*
165 specimens collected in the Salt Fork of the Vermillion River, Illinois, and fantail darter (*E.*
166 *flabellare*) specimens collected in Fox Creek, Illinois.
167
168 *DNA extractions*

169 DNA from a pilot set of 52 *E. cragini* individuals sampled from seven sites across the species'
170 range was extracted using Qiagen DNeasy Blood & Tissue kits (Qiagen, Hilden, Germany).
171 These extractions were done with a 60 ul elution in Qiagen EB buffer and quantified using a
172 Qubit (Thermo Fisher Scientific, Waltham, MA, USA). For high-throughput extractions, we used
173 a KingFisher Flex DNA extraction system (Thermo Fisher Scientific) to extract DNA from 20
174 sets of 90 samples (1800 samples total). We included an overnight digestion step in which tissues
175 were lysed in a 96-well PCR plate at a constant temperature of 55º C on an Eppendorf
176 Mastercycler thermal cycler (Eppendorf, Hamburg, Germany), and we included 10 uL Proteinase
177 K solution, 10 uL enhancer solution, 100 uL Qiagen Buffer EB solution, and approximately 10
178 mg tissue in each digest. We then used the MagMax whole blood protocol for an input volume of
179 200 uL and a final elution volume of 60 uL. We quantified DNA yield from high-throughput
180 extractions using a PicoGreen assay (Thermo Fisher Scientific, Waltham, MA), with the six
181 wells left unused in each plate used for assay standards and a negative control. High-throughput
182 extractions included an additional 1635 *E. cragini* samples from Kansas, 60 *E. cragini* samples
183 from Arkansas, 20 *E. cragini* samples from Colorado, and all seven *E. spectabile* and eight *E.*
184 *flabellare* samples (*E. caeruleum* samples were already extracted). DNA extracted for this study
185 from *E. cragini* covered a total of 232 collection sites (*n* = 2-10 per site; Supporting Table 1).
186 Nearly all of these samples yielded high-quality DNA and were included in the Rapture
187 genotyping analyses described below.
188
189 *Pilot RADseq library preparation & Illumina sequencing*
190 Using the pilot set of 52 *E. cragini* samples, two initial RADseq libraries (consisting of 24
191 samples and 28 samples respectively) were prepared and submitted to the MSU core genomics

192 facility for sequencing. We used the 'BestRad' protocol following Ali *et al.* 2016. Briefly,
193 genomic DNA (100 ng) from each sample was digested with a restriction enzyme (Sbfl-HF) and
194 indexed with a biotinylated RAD adapter. Pooled DNA was sheared to 500 bp fragments using a
195 Covaris sonicator (Covaris, Woburn, MA, USA). Shearing efficiency was evaluated with a
196 fragment analyzer. Dynabeads M-280 streptavidin magnetic beads (Thermo Fisher Scientific,
197 Waltham MA, USA) were used to physically isolate the RAD-tagged DNA fragments. The DNA
198 was then eluted in TE buffer and used in NEBNext Ultra DNA Library Prep Kit for Illumina
199 (New England Biosciences, Ipswich, MA, USA) with no modifications. The two libraries were
200 each sequenced with paired-end 150 bp reads on an Illumina HiSeq 4000 in separate lanes.
201
202 *Bioinformatic pipeline for pilot dataset*
203 As the BestRad protocol can result in sequences with barcodes on either the forward or reverse
204 reads, we used a Python script (Flip2BeRad, https://github.com/tylerhether/Flip2BeRAD) to flip
205 sequences with barcodes on the reverse read. We then filtered out potential PCR clones and
206 demultiplexed sequences using the clonefilter and process_radtags commands in Stacks v. 2.4
207 (Catchen et al. 2013; Rochette et al. 2019). Using the demultiplexed forward reads, we identified
208 loci containing single-nucleotide polymorphisms (SNPs) in ipyrad (Eaton and Overcast 2020).
209 Reads were filtered using ipyrad's default quality thresholds and mapped to an early draft
210 version of the *E. spectabile* genome. We retained an initial set of candidate loci that were
211 genotyped in ≥75% of the pilot set of 52 *E. cragini* individuals and that contained SNPs with a
212 minor allele frequency (MAF) > 0.05. Additionally, we only retained loci with SNPs that were
213 called in at least two *E. cragini* individuals, which imposed an additional floor on MAF and
214 removed SNPs called in only one individual due to sequencing error. We created a FASTA file

215 for all loci that passed these allele frequency filters and aligned these sequences to the draft *E.*
216 *spectabile* genome using bwa v.0.7.17 (Li and Durbin 2009). As bait capture is optimally
217 efficient as long as sequences are >95% similar to baits (Arbor Biosciences, personal
218 communication), any sequences that exhibited <95% similarity or aligned to multiple locations
219 on the *E. spectabile* draft genome were removed from the candidate set. Because the *E.*
220 *spectabile* draft genome contained many small scaffolds, we also removed any loci that were
221 located on scaffolds smaller than 10kb, as it would be difficult to determine whether these loci
222 were adjacent to any other loci in the final chromosome-level genome assembly.
223
224 To identify population clustering within the pilot samples, we conducted an exploratory PCA
225 using the r package adegenet (Jombart 2008, Jombart & Ahmed 2011) (Supporting Figure 1).
226 This preliminary analysis indicated that pilot samples clustered into three distinct groups. To
227 identify potential signatures of selection in this initial set of candidate loci, we used the program
228 BayeScan (Foll and Gaggiotti 2008), which takes a Bayesian approach to identify outlier loci
229 with higher or lower F_{ST} values than expected by chance given population structure. We
230 conducted an initial analysis using all populations. After this analysis showed a high average F_{ST}
231 and an overabundance of lower-than-expected F_{ST} outliers, we re-ran the analysis using only
232 populations in the mainstem Arkansas river (Supporting Figure 2). We used a false discovery
233 rate of 0.05 to identify outlier SNPs in both datasets.
234
235 *Bait design*
236 From the candidate set of RAD loci, we identified three different categories of potential baits to
237 be used as targets for Rapture: (1) short loci (n=3,176), in which ipyrad identified a locus

238 containing at least one SNP that was located on one side of the restriction cut site only; (2) long
239 loci (n=249), consisting of paired loci that both contained a SNP and were located on either side
240 of the cut site; and (3) outlier loci (n=29) identified by Bayescan. Long loci were initially chosen
241 to assess stretches of homozygosity or as potentially more phylogenetically informative blocks of
242 sequence. We obtained BED coordinates for all target loci on the *E. spectabile* draft genome and
243 provided these coordinates and the draft genome to Arbor Biosciences (Ann Arbor, MI, USA).
244 Arbor Biosciences designed and produced a set of 4,966 80-bp baits to capture all long and
245 outlier target RAD loci, as well as 1,841 of the short target RAD loci, for a total of 2,119 Rapture
246 loci. While most previous Rapture study designs have used 120-bp baits (Ali et a. 2016,
247 Komoroske et al. 2019, Peek et al. 2019), we used 80-bp baits tiled in an overlapping manner
248 along the target loci to increase capture efficiency (as in Sard 2020). For short and outlier loci,
249 two baits were tiled along each locus (starting at the restriction site), meaning that approximately
250 40 bps in the center of each locus were covered by two baits and the regions flanking this central
251 region were only covered by one bait. For the long loci, five baits were tiled across both regions
252 flanking the restriction site (Figure 1), meaning that a much longer region (approximately 160
253 bp) was covered by more than one bait.
254
255 *E. cragini whole genomes*
256 To compare population genetic statistics generated with Rapture to those generated using WGS,
257 we produced a reference genome for *E. cragini* and conducted low-coverage whole-genome
258 resequencing. We submitted *E. cragini* muscle tissue from two young-of-year fish of unknown
259 sex raised at the John W. Mumma Native Aquatic Species Restoration Facility to Dovetail
260 Genomics (Scotts Valley, CA, USA) to produce a high-quality reference genome for this species.

261 Dovetail performed Illumina shotgun library preparation, paired-end 2x150 sequencing on an
262 Illumina HiSeq X, and *de novo* assembly in Meraculous (Chapman et al. 2011) using a kmer size
263 of 79. The assembly was refined using Chicago and Hi-C libraries, and scaffolds were
264 constructed using HiRise (Putnam et al. 2016).
265
266 We submitted isolated DNA from 24 *E. cragini* samples for low-coverage whole-genome
267 resequencing at the MSU Core Genomics center. Samples were chosen to include several
268 individuals in each of several population clusters identified by Rapture (see below). We used the
269 Illumina Coverage calculator
270 (https://support.illumina.com/downloads/sequencing_coverage_calculator.html) to estimate the
271 amount of sequencing needed to achieve ≥5x coverage based on genome size and an estimate of
272 20% duplicate sequences. These samples were submitted in two batches of 12, and each batch
273 also contained four samples from another fish with a similar genome size (*Gambusia affinis*). As
274 such, we used 75% of a lane of sequencing for each batch of 12 samples (1/16th of a lane for each
275 sample). Due to maintenance problems at MSU, the initial batch of sequencing produced fewer
276 reads than expected. MSU sent the first batch of samples to the University of Michigan genomics
277 core for additional sequencing and sent the second batch of samples to the Illumina FastTrack
278 Sequencing Service Center for sequencing. All sequencing was performed on an Illumina HiSeq
279 4000.
280
281 We used FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to assess
282 sequencing quality across individuals. We used BWA v. 0.7.17-r1188 (Li and Durbin 2009) to
283 align sequences to either the conspecific *E. cragini* genome or the heterospecific *E. spectabile*

284 genome. We used samtools v.1.9 (Li et al. 2009) to filter out low-quality sequences and
285 improperly paired reads, remove duplicates, and compute average coverage over the whole
286 genome and over all covered sites for alignment to either the conspecific or heterospecific
287 genomes.
288
289 *Rapture library preparation, sequencing, data processing pipeline, and quality control*
290 We used the BestRAD protocol described above along with a sequence capture step that
291 incorporated the Rapture bait sequences to conduct reduced-representation library preparation for
292 1900 individuals (1,855 *E. cragini*, 28 *E. caeruleum*, 8 *E. flabellare* and 9 *E. spectabile*). We
293 aimed for a target DNA mass of 200 ng in 10 uL for the starting material in each reaction. For
294 DNA samples with concentrations of 15-20ng/uL of DNA, we used 10uL total DNA. For
295 samples with concentrations <15 ng/uL, we used a ThermoSavant DNA120 Speedvac (Thermo
296 Fisher, Waltham, MA, USA) to dry down a sample volume containing 200ng and then
297 resuspended in 10 uL 1x TE buffer. We performed library preparation in batches of four 96-well
298 plates (containing 95 samples and one 1X TE blank), using the BestRAD barcode sequences and
299 a plate-specific Illumina adapter for each plate. After BestRAD library preparation, we pooled all
300 four plates and performed sequence capture using the protocol provided by Arbor Biosciences.
301 Briefly, this involved performing a hybridization step at 65ºC for at least 16 hours, isolating bait-
302 target hybrids using streptavidin-coated magnetic beads and washing to remove non-target DNA,
303 and performing PCR amplification of captured DNA for sequencing. We submitted these
304 libraries for sequencing at the MSU Genomics Core facility in five batches of 380 samples each
305 using paired-end 2x150 bp reads on an Illumina HiSeq 4000, using a single lane of sequencing
306 for each batch. For each batch, we altered the number of cycles used for PCR amplification of

307 libraries during the library preparation and the sequence capture steps in order to ensure a high
308 enough concentration for sequence capture and sequencing, respectively. We used 12 cycles
309 during library preparation for all libraries except libraries used in batch 3, where 11 cycles were
310 used. For PCR amplification during sequence capture we used 12 cycles in the first 2 batches and
311 11 cycles in the three subsequent batches. We used the steps described above for BestRAD to
312 process the raw data, and we used BWA to align reads to both the *E. spectabile* (v.
313 UIUC_Espe_1.0, downloaded from NCBI; Moran et al. 2020) and the *E. cragini* reference
314 genomes, and used samtools v. 1.7 (Li et al. 2009) to remove improperly paired reads. We
315 generated two updated bed files by aligning the baits to each genome, merging all loci together
316 into a single file, and creating a buffer (+500 bp from the 3' end of the baits for short and outlier
317 loci, +/- 500 bp on either side of the baits for long loci). We filtered all BAM files using these
318 buffered regions before performing population genetic and phylogenetic analyses.
319
320 We evaluated data quality and potential differences caused by mapping to either a conspecific or
321 heterospecific reference genome using multiple metrics, including the proportion of clonal reads
322 per library preparation, the proportion of reads mapping to either reference genome per
323 individual sample, and the proportion of mapped reads that overlapped the buffered Rapture loci
324 per individual sample. We evaluated potential batch effects by comparing these metrics across
325 Rapture batches. We also estimated individual-level coverage of Rapture loci using bedtools. We
326 assessed two metrics of coverage: (1) the number of reads with any overlap for each buffered
327 locus; and (2) per-base coverage of each buffered locus for a subset of individuals to examine
328 how read depth changed with distance from the restriction site for different types of loci (long vs
329 short).

330

331 *Population genomics, population structure, and selection*

332 We used ANGSD v.0.928 (Korneliussen et al. 2014) to calculate genotype likelihoods for single-
333 nucleotide polymorphisms (SNPs) based on aligned BAM files for Rapture and low-coverage
334 WGS data. As we only had WGS data from *E. cragini*, we excluded the other species from this
335 analysis. To create subsetted datasets with comparable numbers of total reads for comparing
336 population genetic inferences between Rapture and WGS, we randomly selected 570 individuals
337 (corresponding to the number of individuals sequenced on 1.5 lanes) from the Rapture dataset.
338 We assumed bi-allelic SNPs when calculating genotype likelihoods and estimated major and
339 minor allele frequencies for all SNPs. We used sites for which ANGSD detected a SNP with a *p*-
340 value of $< 1x\ 10^{-6}$, and we discarded SNPs with a minor allele frequency < 0.05 and SNPs which
341 were genotyped in <50% of individuals (after Komoroske et al. 2018). We then used the program
342 PCAngsd v.0.981 (Meisner and Albrechtsen 2018) to conduct downstream population genomic
343 analyses. We first conducted principal component analyses (PCA) and calculated genotype
344 probabilities on each of six datasets (WGS, full Rapture, and subsetted Rapture sequence sets,
345 with each set aligned either the *E. spectabile* or *E. cragini* genome), with the optimum number of
346 principal components determined by PCAngsd using a minimum average partial (MAP) test. We
347 examined PCA results to evaluate evidence for batch effects in PCAngsd analyses. To obtain
348 estimates of heterozygosity, we used PCAngsd to call genotypes using a probability threshold of
349 0.9. We compared matched individual heterozygosity values between data types (WGS or
350 Rapture) and between data aligned to either the *E. spectabile* or the *E. cragini* reference genome.
351 We also used PCAngsd to estimate individual admixture proportions for each individual and to
352 perform a PCA-based scans for loci potentially under selection (i.e. loci exhibiting greater

353 differentiation along PCs than expected by drift; Galinsky et al. 2016). We calculated *p*-values
354 for the test statistics generated by PCAngsd selection scan using a one-tailed chi-squared test
355 with one degree of freedom. To account for the large number of tests conducted for the selection
356 scan, we set a conservative significance threshold for each dataset (Rapture and WGS) of 0.05
357 divided by the number of SNPs in the dataset. We compared within-species population structure
358 and selection scan results along the first principal component axis between data types and
359 reference genomes for all *E. cragini* individuals.
360
361 *Phylogenetics and phylogenetic informativeness*
362 We compiled filtered BAM files aligned to the *E. cragini* genome from a subset 56 individuals
363 (two individuals from *E. spectabile* and *E.flabellare*, two individuals from each of two sites for
364 *E. caeruleum*, and two individuals from each of 19 sites covering the full distribution of *E.*
365 *cragini*) and ran the ref_map.pl script in Stacks using the "populations: phylip_var_all" option
366 and default parameter values to call SNPs and output PHYLIP-formatted concatenated multiple
367 sequence alignments for each individual. We then used IQTREE (Nguyen et al. 2015) to
368 construct a phylogenetic tree of all sequences. We used the default maximum likelihood model
369 selection and tree search methods in IQTREE with 1000 bootstraps to calculate support values.
370
371 We converted this tree into a time-calibrated ultrametric tree using the R package ape (Paradis
372 and Schliep 2019). We set estimated branching times for three splits based on a published study
373 of darter evolution (Kelly et al. 2012) using the makeChronosCalib function to calibrate ranges
374 of potential branching times for three interspecific splits. We set the root of the tree, identified
375 here as the common ancestor of the clades *Oligocephalus*, *Psychromaster*, and *Catonotus*, to 24-

376 34 million years ago. We also set the root of *Oligocephalus* (corresponding to the *E. caeruleum* –
377 *E. spectabile* split in our tree) to 17.5-27.5 million years ago, and the common ancestor of
378 *Psychromaster* and *Catonotus* (corresponding the *E. flabellare* – *E. cragini* split in our tree) to
379 16.5 – 26.5 million years ago. We then used the function chronos to construct a time-calibrated
380 tree under three clock models (correlated, relaxed, and discrete). The correlated model had the
381 highest likelihood and we used the tree calibrated using this model in all further analyses. We
382 plotted the time-calibrated tree in ape and plotted the tips of the tree in space using the R package
383 phytools (Revell 2012).
384
385 To calculate phylogenetic informativeness, we created separate PHYLIP files for each set of loci
386 (short, long, and outlier) and concatenated all three into a single dataset and exported as a Nexus
387 file using ape. We then input this alignment and the time-calibrated tree into the PhyDesign web
388 interface (López-Giráldez and Townsend 2011). We examined the inferred net phylogenetic
389 informativeness for each set of baits over the time period covered by the phylogeny (30 million
390 years ago – present).
391
392 **Results**
393 *E. cragini whole genomes*
394 The *E. cragini* genome was similar in terms of contiguity and completeness compared to other
395 published percid reference genomes, although it was smaller (643 Mb vs. greater than 850 Mb in
396 all other percid genomes), contained less repetitive content, and exhibited a number of
397 chromosomal rearrangements, especially relative to *E. spectabile* (Supporting Information 1,
398 Supporting Figure 3). Coverage for resequenced *E. cragini* individuals varied based on the

399 number of reads generated and the reference genome used. Shotgun sequencing for low-coverage
400 WGS generated between 20.5 – 37.8 million read pairs per individual. Between 8.1%-15.5% of
401 sequences were duplicates. Average read depth for covered sites (i.e. all sites with at least 1x
402 coverage) and for all sites in each genome increased with the number of reads (Figure 2; $r > 0.99$
403 in all cases). Average read depth and depth of covered sites were highest when reads were
404 aligned to the *E. cragini* genome and were almost identical, indicating that nearly all sites in the
405 *E. cragini* genome assembly were covered at least 1x. Read depth progressively decreased by
406 approximately 20% for covered sites and by approximately 44% for all sites when reads were
407 aligned to the *E. spectabile* genome (Figure 2). While some of the decline in coverage over all
408 sites may be attributable to the 30% greater length of the *E. spectabile* assembly (which is
409 suggestive of a reduction in genome size for *E. cragini* relative to *E. spectabile*), lower coverage
410 at sites with at least 1x coverage (presumably present in both genomes) also suggests loss of
411 sequencing information resulting from poor alignment to the heterospecific reference genome.
412
413 *Rapture quality control and coverage across species*
414 Between 15.6% and 38.38% of total reads were identified as clones, and the proportion of reads
415 identified as clones decreased in later Rapture batches (Supporting Figure 4). Most of the
416 samples (96%) sequenced using Rapture generated >10,000 read pairs, and 93.7% generated
417 >100,000 read pairs. For *E. cragini* samples with >10,000 read pairs, a high proportion of reads
418 (generally >90%) mapped to the *E. cragini* reference genome. There were batch effects in
419 proportion of reads mapping to the reference, with a higher proportion mapping in earlier
420 Rapture batches. Approximately 90% of reads from *E. caeruleum* samples and approximately
421 80% of *E. spectabile* and *E. flabellare* reads mapped to *E. cragini* genome, with lower mapping

422 success for these samples compared to *E. cragini* sequenced in the same batch (Supporting
423 Figure 5a). The proportion of reads aligning to Rapture loci ranged from 46% to 70% and also
424 displayed batch effects, as well as a decrease in reads mapping to Rapture loci with total read
425 number (Supporting Figure 5a). Alignment of *E. caeruleum* and *E. spectabile* reads to the
426 Rapture loci displayed similar patterns to *E. cragini* reads from the same batch, while the
427 proportion of *E. flabellare* sequences aligning to the Rapture loci was distinctly lower compared
428 to *E. cragini* from the same batch (Supporting Figure 5a). A lower proportion of *E. cragini* and
429 *E. flabellare* sequences mapped to the heterospecific *E. spectabile* genome across batches, while
430 a higher proportion of *E. caeruleum* and *E. spectabile* reads mapped to this reference genome
431 (Supporting Figure 5b). The proportion of mapped reads aligning to the Rapture loci, however,
432 was generally highly similar across species (Supporting Figure 5b).
433
434 The number of Rapture loci covered increased with number of reads for a given individual and
435 tended to reach an asymptote above 10,000-100,000 reads (Figure 3, Supporting Figure 6). The
436 maximum number of loci covered varied between species and between types of loci. For *E.*
437 *cragini*, nearly all of the 2,119 Rapture loci were covered at each read depth. For the other
438 species, a maximum of 1,700-1,800 of the Rapture loci were covered (Figure 3). The reduction
439 in covered loci mainly came from a loss of short loci, of which only ~1,500 of 1,841 (~80%)
440 were covered. A higher proportion of long loci (88%-95%) were sequenced at high coverage,
441 and almost all of the outlier loci were sequenced at high coverage as well (Figure 3). Coverage
442 for Rapture loci was nearly identical when the heterospecific *E. spectabile* reference genome was
443 used for alignment (Supporting Figure 6). Per-base read depth was high for the portion of each
444 locus covered by the capture baits for both long and short loci, representing large numbers of

445 forward reads starting from the cut site overlapping the same region, although short loci had
446 lower read depth beyond the capture baits compared to long loci (Supporting Figure 7).
447
448 *Polymorphism and heterozygosity*
449 For the full Rapture dataset, there were 8,694 SNPs for the alignment to *E. cragini* and 10,495
450 SNPs for the alignment to *E. spectabile* across all 2,119 Rapture loci after filtering, indicating the
451 presence of multiple SNPs per locus. The number of SNPs detected was similar for the subsetted
452 Rapture datasets (8,581 SNPs for the alignment to *E. cragini* and 10,339 SNPs for the alignment
453 to *E. spectabile*). For the WGS dataset, there were 5,759,437 SNPs for the alignment to *E.*
454 *cragini* and 14,020,671 SNPs for the alignment to *E. spectabile*. Individual SNP heterozygosities
455 were highly correlated across datasets – however, estimated heterozygosities were higher for the
456 WGS datasets, and heterozygosity was higher for the *E. spectabile* WGS dataset than the *E.*
457 *cragini* WGS dataset (Figure 4).
458
459 *Population structure and selection*
460 We compared the results of population structure and selection in analyses for *E. cragini* among
461 different datasets (full and subsetted Rapture datasets versus WGS) to evaluate how data type
462 and reference genome affected downstream population genetic inferences. In contrast to batch
463 effects on mapping and alignment to Rapture loci, we did not see strong evidence for batch
464 effects in PCA-based analyses, and samples tended to cluster strongly by metapopulation rather
465 than by batch (Supporting Figure 8). The admixture analysis in PCAngsd indicated that the best
466 population delineation included 16 different populations for the both the full and subsetted
467 Rapture datasets aligned to either reference genome (Figure 5a, Supporting Figure 9a-c). For

468 WGS data, however, PCAngsd found 3 populations for the data aligned to the *E. cragini*
469 reference, and 2 populations for the data aligned to the *E. spectabile* reference (Figure 5b,
470 Supporting Figure 5d). The populations resolved for Rapture datasets broadly corresponded to
471 major river drainages. The populations resolved for WGS lumped together populations in the
472 major northern and southern drainages.
473
474 For the both the Rapture and WGS datasets, PCAngsd did not identify any loci with significant
475 evidence for selection when aligned to either reference genome after correction for the large
476 number of tests (Supporting Figure 10).
477
478 *Phylogeny*
479 Maximum likelihood phylogenetic analyses indicated that the Rapture loci were capable of
480 resolving phylogenetic relationships with fairly strong support. The ML analysis produced
481 100% bootstrap support for correctly grouping *E. spectabile* with *E. caeruleum* and for grouping
482 *E. flabellare* with *E. cragini*, as well as for grouping all individuals within their respective
483 species (Figure 6a). Several deep phylogenetic splits (approximately 2.5-6 million years old)
484 within *E. cragini* also received high support, and individuals within sites and within drainages
485 were often grouped together with high support. Within *E. cragini*, populations showed a nested
486 phylogeographic structure, with Arkansas populations basal to all populations to the east, and
487 populations in east Kansas basal to populations further west. There was also a strongly supported
488 split between populations in the mainstem Arkansas River and its tributaries and populations in
489 drainages to the south of the mainstem Arkansas River (Figure 6b).
490

Per-site phylogenetic informativeness profiles for the three categories of loci showed similar overall patterns from 30 million years ago to 2-3 million years ago, with a slightly convex but relatively stable informativeness profile over time (Supporting Figure 11). Informativeness dropped rapidly from 2 million years ago to the present for the long and short loci, but outlier loci exhibited a secondary peak from 1-2 million years ago for the outlier loci followed by a steep decline. Long loci tended to have lower per-site phylogenetic informativeness than short or outlier loci.

Discussion

There are a number of common questions any researcher involved in the design and implementation of a population genomic or phylogenomic study in a non-model organism will have to address. These include: how many loci and how many individuals do I need to include? Should I sequence loci over the entire genome or should I use sequence capture to target a smaller number of loci at high depth? Should I generate a reference genome for my species or will I be able to use a reference genome from a closely related species, and how will this choice affect the interpretation of my data? Will one methodology work equally well across all target populations and species? And how cost-effective are these alternative methods? All of these questions are perhaps even more relevant for projects aimed at diverse species radiations, as such projects by their nature encompass a number of closely related species. Based on the work described here, we discuss how these questions can be addressed and which methods are most appropriate for different applications.

514 *To Rapture or not to Rapture (and how to Rapture)*

515

516 A number of sequence capture methods exist, ranging from anchored probes (Lemmon et al.
517 2012) and ultraconserved elements (Faircloth et al. 2012) developed for use across a wide variety
518 of taxa, to more focused methods that develop and use a bait set for a single species (Margres et
519 al. 2018). Previous work with the Rapture method in marine turtles demonstrated that baits
520 developed for a single species work well in related species that diverged tens of millions of years
521 ago (Komoroske et al. 2019), and we confirm in this work that Rapture loci developed for a
522 single darter species can also be used in other species from the same group. Rapture loci were
523 recovered with highest coverage from the target species (*E. cragini*) but a majority of loci were
524 recovered from all four species. Long loci spanning both sides of the restriction site were
525 recovered with higher frequency than loci that did not span the restriction site (short loci) across
526 species, and we obtained higher coverage in regions flanking the RAD locus for long loci as
527 well. This is possibly because of a greater possibility of bait capture for more dissimilar
528 sequences with more baits per locus (5 baits for long loci compared to 2 baits for short loci). We
529 also used 80-bp tiled baits as opposed to 120-bp baits used in previous Rapture studies, which
530 may have improved the likelihood of capture as well. These results suggest overall that using loci
531 that span both sides of the restriction site and using 80-bp tiled baits will likely lead to the most
532 consistent recovery of Rapture loci across related species. Incorporating multiple reference
533 genomes or creating pseudo-reference genomes from pilot RAD data for other species of interest
534 may be useful in designing bait sets that will function best across species radiations.

535

536 One of the goals of a Rapture approach is to consistently genotype a large number of loci for a
537 large number of samples. Processing a very large number of samples will necessitate splitting
538 these samples into batches, and batch effects are known to plague some HTS analyses (Leigh et
539 al. 2018, Lambert et al. 2019). The genotyping described here was conducted in 5 batches over a
540 time period of approximately 8 months, and we did find evidence for batch effects in some
541 aspects of Rapture data generation and analysis, specifically in the proportion of clonal reads and
542 the proportion of reads mapping to reference genomes and Rapture loci. The decrease in clonal
543 reads likely resulted from using fewer PCR cycles in the final amplification step in later batches
544 after we had determined that fewer cycles were needed to produce an adequate DNA
545 concentration. Batch effects related to reference genome mapping could potentially be a
546 downstream effect of the change in clone frequency, or they could also be related to somewhat
547 reduced efficacy of the capture reaction over time (possibly due to freezing and thawing of
548 reagents over time, although baits were aliquoted to minimize freeze-thaw cycles). PCA results,
549 however, suggest that these batch effects did not have much downstream effect on the
550 interpretation of the data, as on PC axes explaining most of the variation in the data samples did
551 not cluster by batch, and rather clustered strongly by metapopulation. Additionally, high overall
552 coverage across Rapture loci likely alleviates problems of nonrandom data missingness across
553 batches and allelic dropout commonly seen in traditional RADseq data (Malinsky et al. 2018).
554 Our conclusion is that batch effects should not strongly affect downstream analyses of Rapture
555 data collected over multiple lab preparations and sequencing lanes.
556
557 *Conspecific vs heterospecific reference genomes*
558

559 For capture-based reduced-representation genomic methods, the choice of reference genome
560 impacts both the initial phase of SNP discovery and development of capture baits as well as the
561 analysis phase. Due to the currently limited availability of reference genomes and the time and
562 cost required to sequence a novel genome, researchers initiating a sequence capture project may
563 need to use a heterospecific genome for the initial SNP discovery and bait design steps, as we did
564 here. Studies in birds have suggested that heterospecific reference genomes can be useful for
565 SNP discovery (Galla et al. 2019), although strongly conserved genome structure across bird
566 taxa (Ellegren 2010) may increase the utility of heterospecific reference genomes for SNP
567 discovery in this group. In this study, using a heterospecific reference genome for the initial
568 design phase resulted in baits that successfully captured polymorphic RAD loci in our target
569 species (*E. cragini*) and in congeneric species. This may reflect conserved chromosome number
570 and large regions of synteny among genomes for species belonging to this group (Supporting
571 Information 1). However, comparison of the *E. cragini* and *E. spectabile* genomes indicated
572 substantial changes in genome size and organization among species within *Etheostoma*,
573 suggesting that genome structure has evolved substantially over the approximately 30 million-
574 year history of the genus. Genome structure and karyotype may in some cases vary widely within
575 species radiations (e.g. Vershinina and Lukhtanov 2016), and as such caution should still be
576 exercised when using heterospecific reference genomes in SNP discovery and bait design. As
577 more eukaryotic genomes become available, adopting a pan-genome approach, used in the past
578 to identify core regions common to prokaryotic genomes within specified taxonomic groups
579 (Vernikos et al. 2015)), could become an attractive alternative to using a single reference
580 genome. This approach may be particularly appealing to researchers working with species

581 radiations, as targeting genomic regions that are conserved throughout a radiation should
582 increase the utility of the capture bait set across species.
583
584 Choice of reference genome will also impact the downstream analysis and interpretation of both
585 targeted sequence capture and WGS data. Although mapping sequence reads generated from one
586 species to the genome of a closely related species is still common practice, the effects of
587 mapping reads to a heterospecific genome versus a conspecific genome are still relatively
588 understudied. Galla et al. (2019) mapped RADseq and low-coverage WGS data to either a
589 conspecific, congeneric, confamilial, or conordinal genome and found a decreasing alignment
590 rate with increasing phylogenetic distance, as well as less consistency in estimates of genetic
591 diversity when reads were mapped to a more distantly-related genome. Our WGS results
592 generally agree with these findings. Mapping reads generated from low-coverage WGS of *E.*
593 *cragini* individuals to the *E. cragini* reference genome was generally more successful than
594 mapping to the *E. spectabile* genome. Lower read depth and allelic dropout could contribute to
595 different estimates of genetic diversity from conspecific and heterospecific reference genomes.
596 However, even though we inferred population structure using genotype likelihoods, which
597 should mitigate the effects of lower read depth associated with mapping to a heterospecific
598 reference genome (Nevado et al. 2014), admixture results were also affected by the choice of
599 reference genome. The existence of multiple rearrangements among darter genomes observed in
600 this study and others (Moran et al. 2020) could aggravate the effects of using a heterospecific
601 reference genome. For our Rapture dataset, however, the effects of mapping to a heterospecific
602 reference were much reduced, and downstream inferences regarding diversity and population
603 structure were similar, regardless of which reference genome was used for mapping. This

604 suggests that sequence capture may reduce biases associated with the absence of a closely-
605 related reference genome, possibly because the RAD loci targeted by sequence capture in this
606 case were designed by alignment to a heterospecific reference genome and thus were fairly
607 conserved across genomes.
608
609 *Effects of reference genome choice and data type on population genomic inferences*
610
611 Different data types can be differentially suited to different analyses. We found that Rapture was
612 much better at identifying fine-scale population structure than WGS. This is likely partially due
613 to the much greater spatial coverage and the greater number of individuals we were able to
614 sequence via this method with a similar amount of sequencing effort. Higher coverage overall for
615 the Rapture data may also alleviate allelic dropout and decreased sensitivity for calling
616 heterozygotes associated with low-coverage WGS. However, as PCAngsd uses genotype
617 likelihoods rather than called genotypes, this issue may not have strongly affected these analyses.
618
619 Previous work has asserted that WGS is typically better suited to detecting evidence of selection
620 using genome scan methods than RAD-based approaches, which target relatively small portions
621 of the genome (Lowry et al. 2016). For selection scan methods to be accurate, however, they
622 must take into account variation in allele frequencies due to neutral processes, such as change in
623 population size over time or spatial population structure (de Villemereuil et al. 2014). This may
624 be particularly important in species with limited dispersal capability, such as freshwater fish
625 (Shurin et al. 2009). We found little evidence for selection in any of the Rapture datasets, even
626 for loci that showed some evidence of selection in the pilot dataset. As the pilot dataset did not

627 delineate many of the fine-scale populations indicated by the full or subsetted Rapture datasets,
628 however, we believe it is very plausible that the outlier loci identified in these preliminary
629 analyses represent loci that were differentiated due to neutral population within broad-scale
630 groupings rather than selection. We found little evidence of selection for the WGS datasets as
631 well. This may be due again to the lower number of individuals sampled for WGS, and less
632 accurate estimation of population structure may also have confounded detection of loci under
633 selection using WGS. Future analyses of selection using low-coverage WGS data may also
634 benefit from the application of methods that estimate linkage disequilibrium based on genotype
635 likelihoods and prune closely linked SNP loci (Fox et al. 2019), which would reduce overall SNP
636 density but also potentially reduce false positives and increase power by lowering thresholds for
637 accurately detecting loci under selection while accounting for multiple comparisons.
638
639 RADseq-based methods can detect selection if marker density is high relative to the size of
640 linkage disequilibrium blocks (Catchen et al. 2017), and Rapture workflows designed to detect
641 selection with these factors in mind may be comparable to WGS. Alternatively, Rapture methods
642 can also include loci with known *a priori* effects on fitness (such as loci associated with disease
643 susceptibility). While our Rapture loci identified *a priori* as under selection by genome scans in
644 the pilot dataset did not show strong evidence of selection in the larger datasets, Rapture panels
645 designed to include high marker density as well as immune-associated loci constituted an
646 effective means identifying loci associated with survival in female Tasmanian devils
647 (*Sarcophilus harrisii*) with a transmissible cancer (Margres et al. 2018).
648
649 *Phylogenetic informativeness of Rapture loci*

650

651 Sequence capture strategies targeting ultra-conserved elements (UCEs) and protein-coding genes
652 have been evaluated in the past for percomorph fishes (Gilbert et al. 2015). UCE flanks and
653 protein-coding genes in general showed great utility for resolving deeper split but a loss of
654 phylogenetic signal for more recent epochs, with per-locus phylogenetic informativeness for
655 UCE flanks and protein-coding genes peaking between 20-40 million years ago and exhibiting
656 rapid decline from 20 million years ago to the present. The phylogenetic informativeness of
657 Rapture loci was fairly constant over time and potentially more useful for examining relatively
658 recent splits between closely related species. However, phylogenetic informativeness for Rapture
659 declined rapidly for very recent epochs. This is also potentially reflected in support values
660 estimated here for relationships within the *E. cragini*. We obtained 100% bootstrap support for
661 older splits between populations in Arkansas versus populations further east, as well as high
662 support for a sister relationship between populations in eastern Kansas and all other populations
663 to the west and a split dating to approximately 3 million years ago between populations in
664 drainages associated with the mainstem Arkansas river and populations in drainages to the south
665 of the Arkansas River. For more recent splits, support values were overall fairly high (95-100%)
666 but much lower for some nodes, indicating ambiguous support for some relationships. This likely
667 represents both a true lack of phylogenetic informativeness (i.e. substation rates too low to allow
668 for reliably distinguishing among alternative relationships) as well as potentially other
669 confounding factors, such as gene flow and maintenance of ancestral polymorphisms.
670 Alternative methods that incorporate gene flow and demographic modeling (Jackson et al. 2017,
671 Scott et al. 2018) could allow for more reliable inferences for recently diverged populations.
672

673 *Costs and benefits of different sequencing methods*

674

675 With limited funding, cost will always be a consideration. Rapture has a somewhat costly initial
676 investment but is still highly cost-effective ($13.42 sample including bait design and production
677 for 1,900 individuals in our study, or <$10 per sample if baits are already available; Table 1) in
678 terms of cost per sample when compared to either BestRAD or low-coverage WGS. Given these
679 low costs, Rapture is a very attractive method for conducting future work in the darter system,
680 especially when extensive individual-level and spatial sampling are important components of the
681 project design. The data produced by Rapture can be supplemented by low-coverage WGS if this
682 is needed for the study, and the relatively high cost per individual of WGS in this study (~$275
683 per sample) could potentially be reduced by using poolseq (Schlötterer et al. 2014).

684

685 Overall, the Rapture method outlined here represents a potentially powerful methodology for
686 phylogenomics and population genomics, both in darters and in diverse radiations of non-model
687 organisms more generally. We have also shown several potential pitfalls associated with using
688 heterospecific genomes . While targeted sequence capture seems to mitigate some of these
689 pitfalls, choosing to use a heterospecific reference genome still has consequences that should be
690 carefully considered during study design. As more reference genomes and sequence capture
691 methods become available, Rapture will become an increasingly attractive option, especially
692 when large sample sizes, extensive spatial coverage, or high read depth are important.

References

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). Rad capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics*, *202*(2), 389–400. https://doi.org/10.1534/genetics.115.183665

Bossu, C. M., & Near, T. J. (2013). Characterization of a contemporaneous hybrid zone between two darter species (*Etheostoma bison* and *E. caeruleum*) in the Buffalo River System. *Genetica*, *141*(1–3), 75–88. https://doi.org/10.1007/s10709-013-9707-8

Brandies, P., Peel, E., Hogg, C.J, Belov, K. (2019). The value of reference genomes in the conservation of threatened species. *Genes*, 10(11), 846.

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., … Di Palma, F. (2015). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, *513*(7518), 375–381. https://doi.org/10.1038/nature13726

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources*, *17*(3), 362–365. https://doi.org/10.1111/1755-0998.12669

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Chapman, J. A., Ho, I., Sunkara, S., Luo, S., Schroth, G. P., & Rokhsar, D. S. (2011). Meraculous: De novo genome assembly with short paired-end reads. *PLoS ONE*, *6*(8). https://doi.org/10.1371/journal.pone.0023501

Chaves, J. A., Cooper, E. A., Hendry, A. P., Podos, J., De León, L. F., Raeymaekers, J. A. M., … Uy, J. A. C. (2016). Genomic variation at the tips of the adaptive radiation of Darwin's finches. *Molecular Ecology*, *25*(21), 5282–5295. https://doi.org/10.1111/mec.13743

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., … Pereira, R. J. (2016). Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, *30*, 3–13. https://doi.org/10.1016/j.margen.2016.04.012

David, K. T., Wilson, A. E., & Halanych, K. M. (2019). Sequencing disparity in the genomic era. *Molecular Biology and Evolution*, *36*(8), 1624–1627. https://doi.org/10.1093/molbev/msz117

de Villemereuil, P., Frichot, E., Bazin, E., François, O., & Gaggiotti, O.E. (2014). Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, *23*(8), 2006-2019.

Eaton, D. A. R., & Overcast, I. (2020). ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz966

Ekblom, R., & Galindo, J. (2011, July 8). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, Vol. 107, pp. 1–15. https://doi.org/10.1038/hdy.2010.152

Ellegren, H. (2010). Evolutionary stasis: the stable chromosomes of birds. *Trends in Ecology and Evolution, 25(5)*, 283-291.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Systematic Biology*, Vol. 61, pp. 717–726. https://doi.org/10.2307/41677973

Fitzpatrick, S. W., Crockett, H., & Funk, W. C. (2014). Water availability strongly impacts population genetic patterns of an imperiled Great Plains endemic fish. *Conservation Genetics*, *15*(4), 771–788. https://doi.org/10.1007/s10592-014-0577-0

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*, *180*(2), 977–993. https://doi.org/10.1534/genetics.108.092221

Fox, E.A., Wright, A.E., Fumagalli, M., & Vieira, F.G. (2019). *ngsLD*: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics, 35(19)*, 3855-3856.

Gabel, J. M., Dakin, E. E., Freeman, B. J., & Porter, B. A. (2008). Isolation and identification of eight microsatellite loci in the Cherokee darter (*Etheostoma scotti*) and their variability in other members of the genera *Etheostoma*, *Ammocrypta*, and *Percina*. *Molecular Ecology Resources*, *8*(1), 149–151. https://doi.org/10.1111/j.1471-8286.2007.01903.x

Galinsky, K. J., Bhatia, G., Loh, P. R., Georgiev, S., Mukherjee, S., Patterson, N. J., & Price, A. L. (2016). Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *American Journal of Human Genetics*, *98*(3), 456–472. https://doi.org/10.1016/j.ajhg.2015.12.022

Galla, S. J., Forsdick, N. J., Brown, L., Hoeppner, M., Knapp, M., Maloney, R. F., … Steeves, T. E. (2018). Reference Genomes from Distantly Related Species Can Be Used for Discovery of Single Nucleotide Polymorphisms to Inform Conservation Management. *Genes*, *10*(1), 9. https://doi.org/10.3390/genes10010009

George, M. (2018). Phylogeny of the Orangethroat Darter (*Etheostoma spectabile*) species complex in the Ozark Highlands of Arkansas. *Animal Science Undergraduate Honors Theses*. Retrieved from https://scholarworks.uark.edu/anscuht/22

Gilbert, P. S., Chang, J., Pan, C., Sobel, E. M., Sinsheimer, J. S., Faircloth, B. C., & Alfaro, M. E. (2015). Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. *Molecular Phylogenetics and Evolution*, *92*, 140–146. https://doi.org/10.1016/j.ympev.2015.05.027

Grover, C. E., Salmon, A., & Wendel, J. F. (2012). Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany*, *99*(2), 312–319. https://doi.org/10.3732/ajb.1100323

Harvey, M., Smith, B., Glenn, T., … B. F.-S., & 2016, undefined. (n.d.). Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Academic.Oup.Com*. Retrieved from https://academic.oup.com/sysbio/article-abstract/65/5/910/2223860

Hoffberg, S. L., Kieran, T. J., Catchen, J. M., Devault, A., Faircloth, B. C., Mauricio, R., & Glenn, T. C. (2016). RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Molecular Ecology Resources*, *16*(5), 1264–1278. https://doi.org/10.1111/1755-0998.12566

Hollingsworth Jr, P. R., & Near, T. J. (2009). Temporal patterns of diversification and microendemism in eastern highland endemic barcheek darters (Percidae: Etheostomatinae). *Evolution*, *63*(1), 228–243. https://doi.org/10.1111/j.1558-5646.2008.00531.x

Hudman, S. P., Grose, M. J., Landis, J. B., Skalski, G. T., & Wiley, E. O. (2008). Twenty-three microsatellite DNA loci for population genetic studies and parentage assignment in orangethroat darter, *Etheostoma spectabile*. *Molecular Ecology Resources*, *8*(6), 1483–1485. https://doi.org/10.1111/j.1755-0998.2008.02312.x

Jackson, N. D., Morales, A. E., Carstens, B. C., & O'Meara, B. C. (2017). PHRAPL: Phylogeographic Inference Using Approximate Likelihoods. *Systematic Biology*, *66*(6), 1045–1053. https://doi.org/10.1093/sysbio/syx001

Jelks, H. L., Walsh, S. J., Burkhead, N. M., Contreras-Balderas, S., Diaz-Pardo, E., Hendrickson, D. A., … Warren, M. L. (2008). Conservation Status of Imperiled North American Freshwater and Diadromous Fishes. *Fisheries*, *33*(8), 372–407. https://doi.org/10.1577/1548-8446-33.8.372

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24.11: 1403-1405.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics, 27(21), 3070-3071.

Kelly, N. B., Near, T. J., & Alonzo, S. H. (2012). Diversification of egg-deposition behaviours and the evolution of male parental care in darters (Teleostei: Percidae: Etheostomatinae).

Journal of Evolutionary Biology, *25*(5), 836–846. https://doi.org/10.1111/j.1420-9101.2012.02473.x

Khudamrongsawat, J., Heath, L. S., Heath, H. E., & Harris, P. M. (2007). Microsatellite DNA primers for the endangered vermilion darter, *Etheostoma chermocki*, and cross-species amplification in other darters (Percidae: *Etheostoma*). *Molecular Ecology Notes*, *7*(5), 811–813. https://doi.org/10.1111/j.1471-8286.2007.01712.x

Komoroske, L. M., Miller, M. R., O'Rourke, S. M., Stewart, K. R., Jensen, M. P., & Dutton, P. H. (2019). A versatile Rapture (RAD-Capture) platform for genotyping marine turtles. *Molecular Ecology Resources*, *19*(2), 497–511. https://doi.org/10.1111/1755-0998.12980

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Lambert, S.M., Streicher, J.W., Fisher-Reid, M.C., de la Cruz, F.R.M., Martínez-Méndez, N., García-Vázquez, U.O., Montes de Oca, A.N., & Wiens, J.J. (2019). Inferring introgression using RADseq and *D*FOIL: Power and pitfalls revealed in a case study of spiny lizards (*Sceloporus*). *Molecular Ecology Resources, 19*, 818–837.

Lápez-Giráldez, F., & Townsend, J. P. (2011). PhyDesign: An online application for profiling phylogenetic informativeness. *BMC Evolutionary Biology*, *11*(1), 152. https://doi.org/10.1186/1471-2148-11-152

Leigh, D.M., Lischer, H.E.L., & Keller, L.F. (2018). Batch effects in a multiyear sequencing study: False biological trends due to changes in read lengths. *Molecular Ecology Resources*, *18*, 778-788.

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. *Syst. Biol*, *61*(5), 727–744. https://doi.org/10.1093/sysbio/sys049

Lemmon, E. M., & Lemmon, A. R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 99–121. https://doi.org/10.1146/annurev-ecolsys-110512-135822

Li, H., Bioinformatics, R. D.-, & 2009, undefined. (n.d.). Fast and accurate short read alignment with Burrows–Wheeler transform. *Oxford University Press*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: an evaluation of the utility of restriction site-associated DNA

sequencing for genome scans of adaptation. *Molecular Ecology Resources*, *17*(2), 142–152. https://doi.org/10.1111/1755-0998.12635

MacGuigan, D. J., & Near, T. J. (2019). Phylogenomic Signatures of Ancient Introgression in a Rogue Lineage of Darters (Teleostei: Percidae). *Systematic Biology*, *68*(2), 329–346. https://doi.org/10.1093/sysbio/syy074

Malinsky, M., Trucchi, E., Lawson, D.J., & Falush, D. (2018). RADpainter and fineRADstructure: population inference from RADseq Data. *Molecular Biology and Evolution*, *35(5),* 1284–1290.

Manel, S., Perrier, C., Pratlong, M., Abi-Rached, L., Paganini, J., Pontarotti, P., & Aurelle, D. (2016). Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Molecular Ecology*, *25*(1), 170–184. https://doi.org/10.1111/mec.13468

Margres, M. J., Jones, M. E., Epstein, B., Kerlin, D. H., Comte, S., Fox, S., … Storfer, A. (2018). Large-effect loci affect survival in Tasmanian devils (*Sarcophilus harrisii*) infected with a transmissible cancer. *Molecular Ecology*, *27*(21), 4189–4199. https://doi.org/10.1111/mec.14853

Meek, M. H., & Larson, W. A. (2019). The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era. *Molecular Ecology Resources*, *19*(4), 795–803. https://doi.org/10.1111/1755-0998.12998

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, *210*(2), 719–731. https://doi.org/10.1534/genetics.118.301336

Mendelson, T. C. (2003). Sexual isolation evolves faster than hybrid inviability in a diverse and sexually dimorphic genus of fish (Percidae: *Etheostoma*). *Evolution*, *57*(2), 317–327. https://doi.org/10.1111/j.0014-3820.2003.tb00266.x

Moran, R. L., Catchen, J. M., & Fuller, R. C. (2020). Genomic resources for darters (Percidae: Etheostominae) provide insight into postzygotic barriers implicated in speciation. *Molecular Biology and Evolution*, *37*(3), 711–729. https://doi.org/10.1093/molbev/msz260

Moran, R. L., & Fuller, R. C. (2018). Male-driven reproductive and agonistic character displacement in darters and its implications for speciation in allopatry. *Current Zoology*, *64*(1), 101–113. https://doi.org/10.1093/cz/zox069

Moran, R. L., & Fuller, R. C. Agonistic character displacement of genetically based male colour patterns across darters. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1884). https://doi.org/10.1098/rspb.2018.1248

Moran, R. L., Zhou, M., Catchen, J. M., & Fuller, R. C. (2018). Hybridization and postzygotic isolation promote reinforcement of male mating preferences in a diverse group of fishes with traditional sex roles. *Ecology and Evolution*, *8*(18), 9282–9294. https://doi.org/10.1002/ece3.4434

Moran, R. L., Zhou, M., Catchen, J. M., & Fuller, R. C. (2017). Male and female contributions to behavioral isolation in darters as a function of genetic distance and color distance. *Evolution*, *71*(10), 2428–2444. https://doi.org/10.1111/evo.13321

Near, T. J., & Benard, M. F. (2004). Rapid allopatric speciation in logperch darters (Percidae: *Percina*). *Evolution*, *58*(12), 2798–2808. https://doi.org/10.1111/j.0014-3820.2004.tb01631.x

Near, T. J., Bossu, C. M., Bradburd, G. S., Carlson, R. L., Harrington, R. C., Hollingsworth, P. R., … Etnier, D. A. (2011). Phylogeny and temporal diversification of darters (Percidae: Etheostomatinae). *Systematic Biology*, *60*(5), 565–595. https://doi.org/10.1093/sysbio/syr052

Nevado, B., Ramos-Onsins, S. E., & Perez-Enciso, M. (2014). Resequencing studies of nonmodel organisms using closely related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics. *Molecular Ecology*, *23*(7), 1764–1779. https://doi.org/10.1111/mec.12693

Nguinkal, J. A., Brunner, R. M., Verleih, M., Rebl, A., de los Ríos-Pérez, L., Schäfer, N., … Goldammer, T. (2019). The First Highly Contiguous Genome Assembly of Pikeperch (*Sander lucioperca*), an Emerging Aquaculture Species in Europe. *Genes*, *10*(9), 708. https://doi.org/10.3390/genes10090708

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Oliveira, D.A., Reid, B.R., Fitzpatrick, S. W. (2020). Genome-wide diversity and habitat underlie fine-scale phenotypic differentiation in the rainbow darter (*Etheostoma caeruleum*). *Evolutionary Applications*, *accepted*.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Peek, R. A., Bedwell, M., O'Rourke, S. M., Goldberg, C., Wengert, G. M., & Miller, M. R. (2019). Hybridization between two parapatric ranid frog species in the northern Sierra Nevada, California, USA. *Molecular Ecology*, *28*(20), 4636–4647. https://doi.org/10.1111/mec.15236

Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., … Green, R. E. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research*, *26*(3), 342–350. https://doi.org/10.1101/gr.193474.115

Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*(2), 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737–4754. https://doi.org/10.1111/mec.15253

Russell, J. J., Theriot, J. A., Sood, P., Marshall, W. F., Landweber, L. F., Fritz-Laylin, L., … Brunet, A. (2017). Non-model model organisms. *BMC Biology*, *15*(1), 1–31. https://doi.org/10.1186/s12915-017-0391-5

Saarinen, E. V., & Austin, J. D. (2010). When Technology Meets Conservation: Increased Microsatellite Marker Production Using 454 Genome Sequencing on the Endangered Okaloosa Darter (*Etheostoma okaloosae*). *Journal of Heredity*, *101*(6), 784–788. https://doi.org/10.1093/jhered/esq080

Sard, N. M., Smith, S. R., Homola, J. J., Kanefsky, J., Bravener, G., Adams, J. V., … Scribner, K. T. (2020). RAPTURE (RAD capture) panel facilitates analyses characterizing sea lamprey reproductive ecology and movement dynamics. *Ecology and Evolution*, *10*(3), 1469–1488. https://doi.org/10.1002/ece3.6001

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014, November 25). Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, Vol. 15, pp. 749–763. https://doi.org/10.1038/nrg3803

Scott, P. A., Glenn, T. C., & Rissler, L. J. (2018). Resolving taxonomic turbulence and uncovering cryptic diversity in the musk turtles (*Sternotherus*) using robust demographic modeling. *Molecular Phylogenetics and Evolution*, *120*, 1–15. https://doi.org/10.1016/j.ympev.2017.11.008

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., … Zieliński, P. (2015, February 1). Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, Vol. 30, pp. 78–87. https://doi.org/10.1016/j.tree.2014.11.009

Shurin, J.B., Cottenie, K., & Hillebrand, H. (2009). Spatial autocorrelation and dispersal limitation in freshwater organisms. *Community Ecology, 159*, 151-159.

Switzer, J. F., Welsh, S. A., & King, T. L. (2008). Microsatellite DNA primers for the candy darter, *Etheostoma osburni* and variegate darter, *Etheostoma variatum*, and cross-species

amplification in other darters (Percidae). *Molecular Ecology Resources*, *8*(2), 335–338. https://doi.org/10.1111/j.1471-8286.2007.01946.x

Tonnis, B. D. (2006). Microsatellite DNA markers for the rainbow darter, *Etheostoma caeruleum* (Percidae), and their potential utility for other darter species. *Molecular Ecology Notes*, *6*(1), 230–232. https://doi.org/10.1111/j.1471-8286.2005.01203.x

Vernikos, G., Medini, D., Riley, D.R, & Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology*, 23, 148-154.

Vershinina, A.O., & Lukhtanov, V.A. (2017). Evolutionary mechanisms of runaway chromosome number change in *Agrodiaetus* butterflies. *Scientific Reports, 7*, 8199.

Acknowledgments

This work was funded by grant #SCTF911C from the Colorado Parks and Wildlife Species Conservation Trust Fund and grant # E-30-R-1 from the Kansas Department of Wildlife, Parks, and Tourism. We would like to thank Harry Crockett (Colorado Parks and Wildlife Agency), Jordan Hoffmeier and Mark van Scoyoc (Kansas Department of Wildlife, Parks, and Tourism), and Brian Wagner (Arkansas Game and Fish Commission) for providing samples. Rainbow darters were collected under a scientific collector's permit issued by the Michigan Department of Natural Resources, and collection of *E. spectabile* and *E. flabellare* was approved by the Illinois Department of Natural Resources under Scientific Collecting Permit A15.4035.
We thank M. Meek and D. Oliveira for providing helpful feedback on our manuscript. This is W.K. Kellogg Biological Station publication number xx.

Data accessibility statement

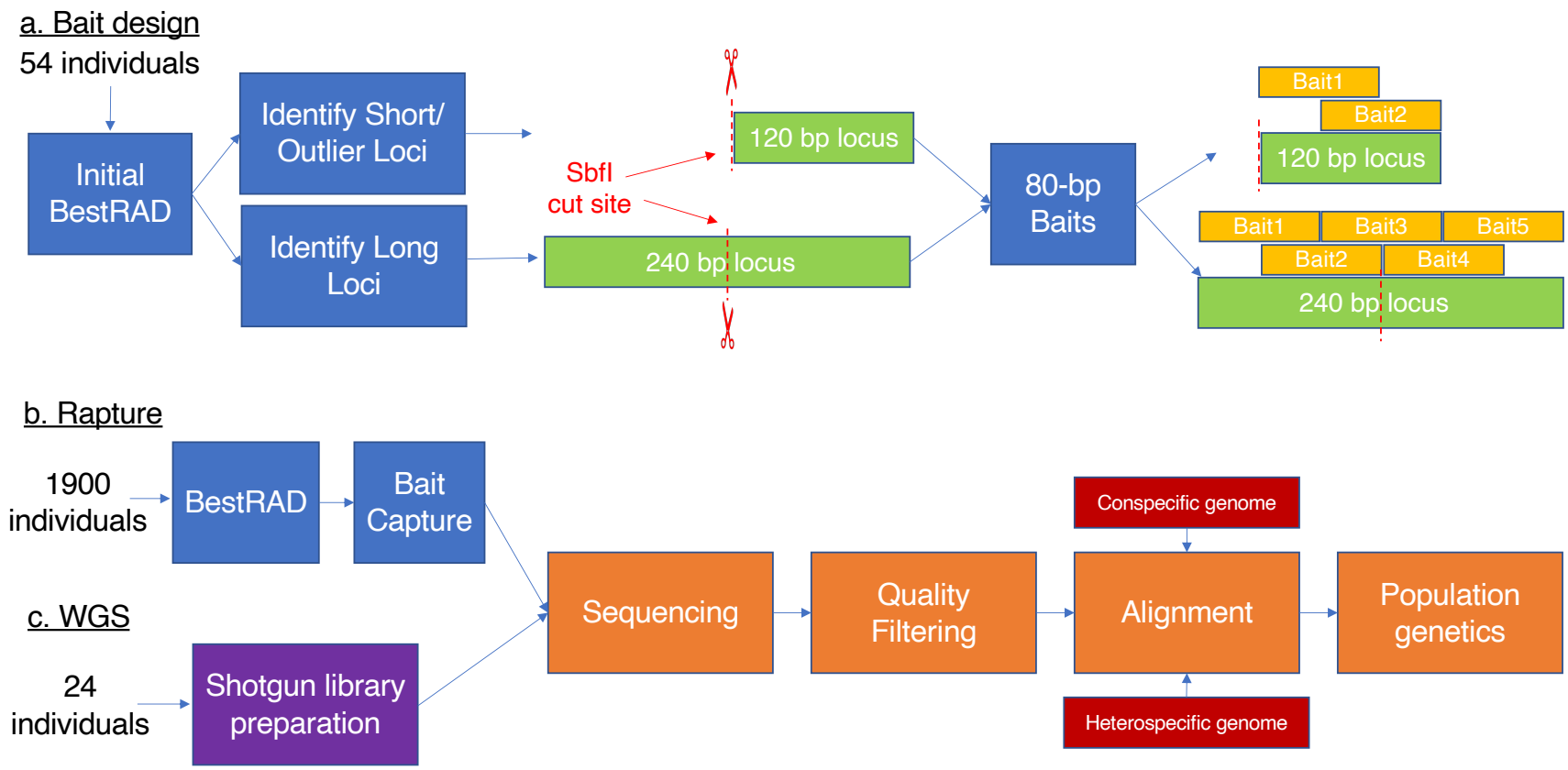
The *E. cragini* Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAAVJE000000000. The version described in this paper is version JAAVJE010000000. Short-read data have been uploaded to the NCBI as BioProject PRJNA611833. Analysis scripts and capture bait sequences are available at https://github.com/nerdbrained/darter_rapture.

964
965 **Table 1.** Cost of Rapture and WGS methods used in this study.

Method	# samples	Pilot dataset	Baits	Library preparation	Sequencing	Total cost	Cost/sample	Cost/sample after bait design
BestRAD (pilot)	52	N/A	N/A	$733.00	$2,661.00	$3,394.00	$65.27	
Rapture	1900	$3,394.00	$3,600.00	$5,190.00	$13,305.00	$25,489.00	$13.42	$9.73
WGS	24	N/A	N/A	$2,616.00	$3,991.50	$6,607.50	$275.31	

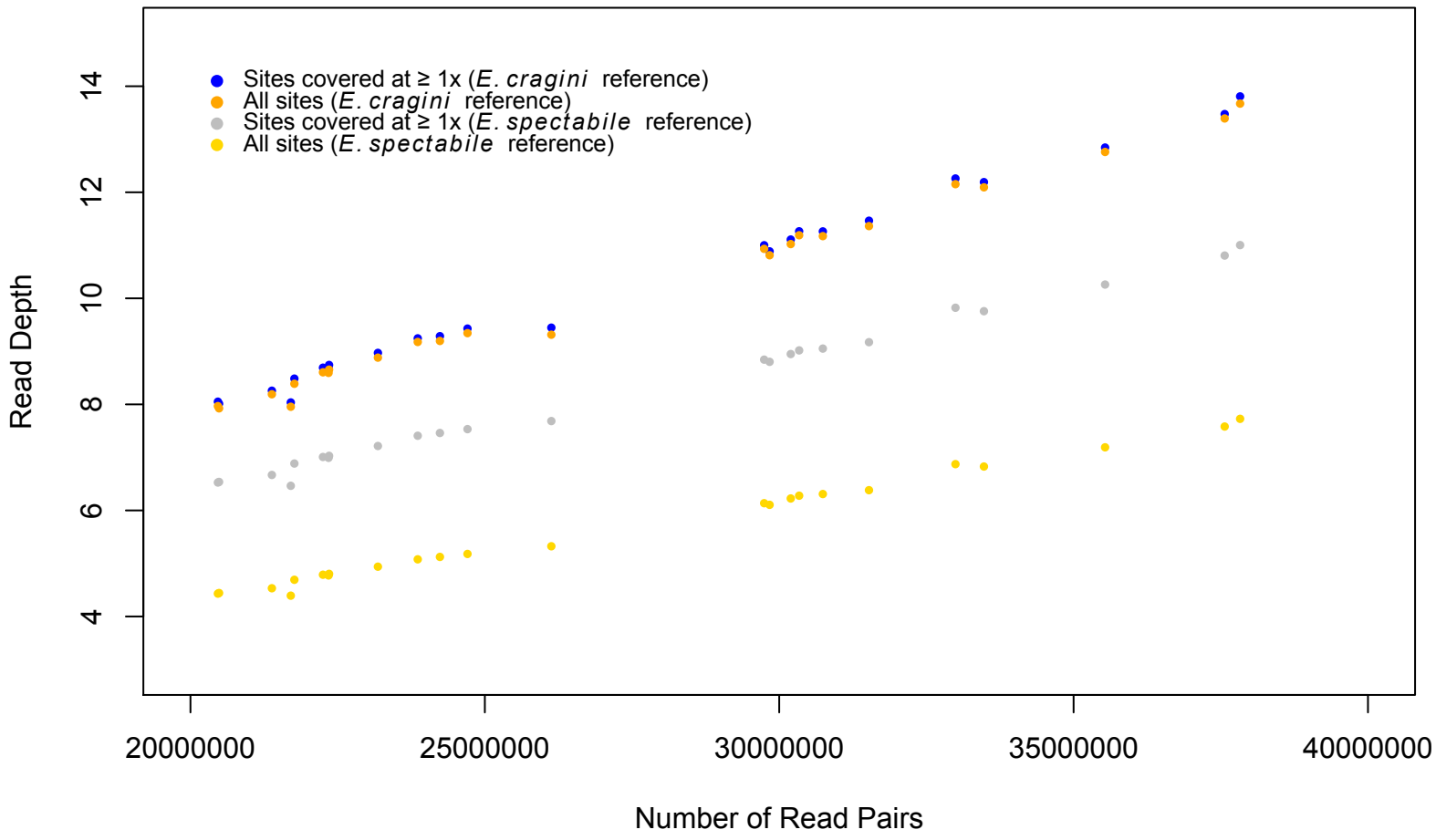
966

967 Figure 1. Flow chart showing procedures used to design *E. cragini* Rapture baits from BestRAD RADseq data, test the Rapture baits
968 using a subsequent round of BestRAD RADseq, and compare the results of population genomic analyses using Rapture versus WGS
969 data.
970
971



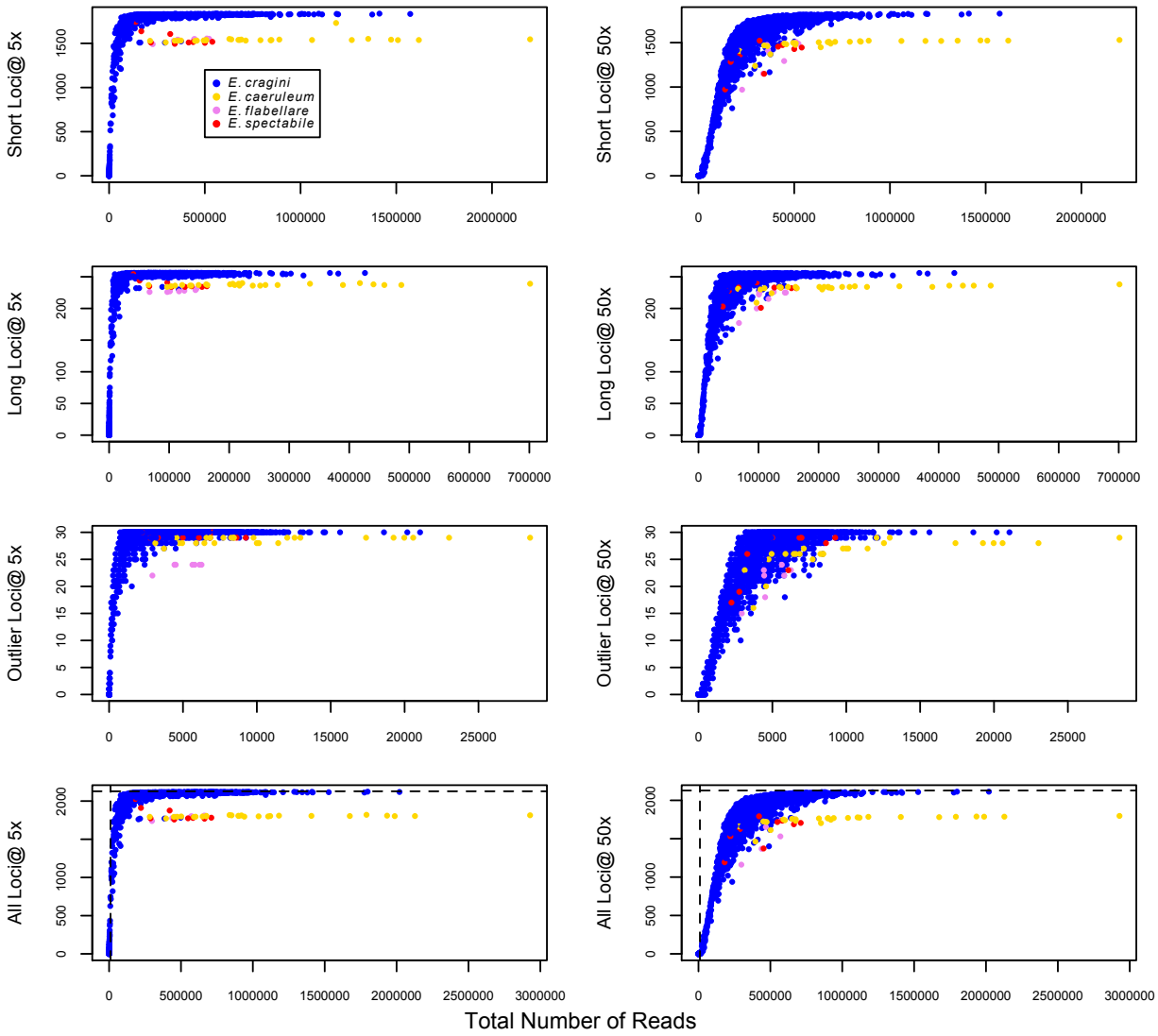
972

973 Figure 2. Read depth and number of read pairs for whole genome sequencing data. Average depth is shown for either a subset of sites
974 with at least 1x coverage or for all sites in the genome, with alignment to either the conspecific *E. cragini* genome or the
975 heterospecific genome of a closely related species (*E. spectabile*).



976

977 Figure 3. Coverage for Rapture loci (either all loci combined or short, long, and outlier loci taken
978 separately) mapped to the *E. cragini* genome across four *Etheostoma* species.
979
980



981

Figure 4. Comparisons of heterozygosity estimates across different datasets. For each comparison, the first estimate is plotted on the x-axis and the second is plotted on the y-axis. A 1:1 dotted line (expected for complete agreement across datasets) is also shown.

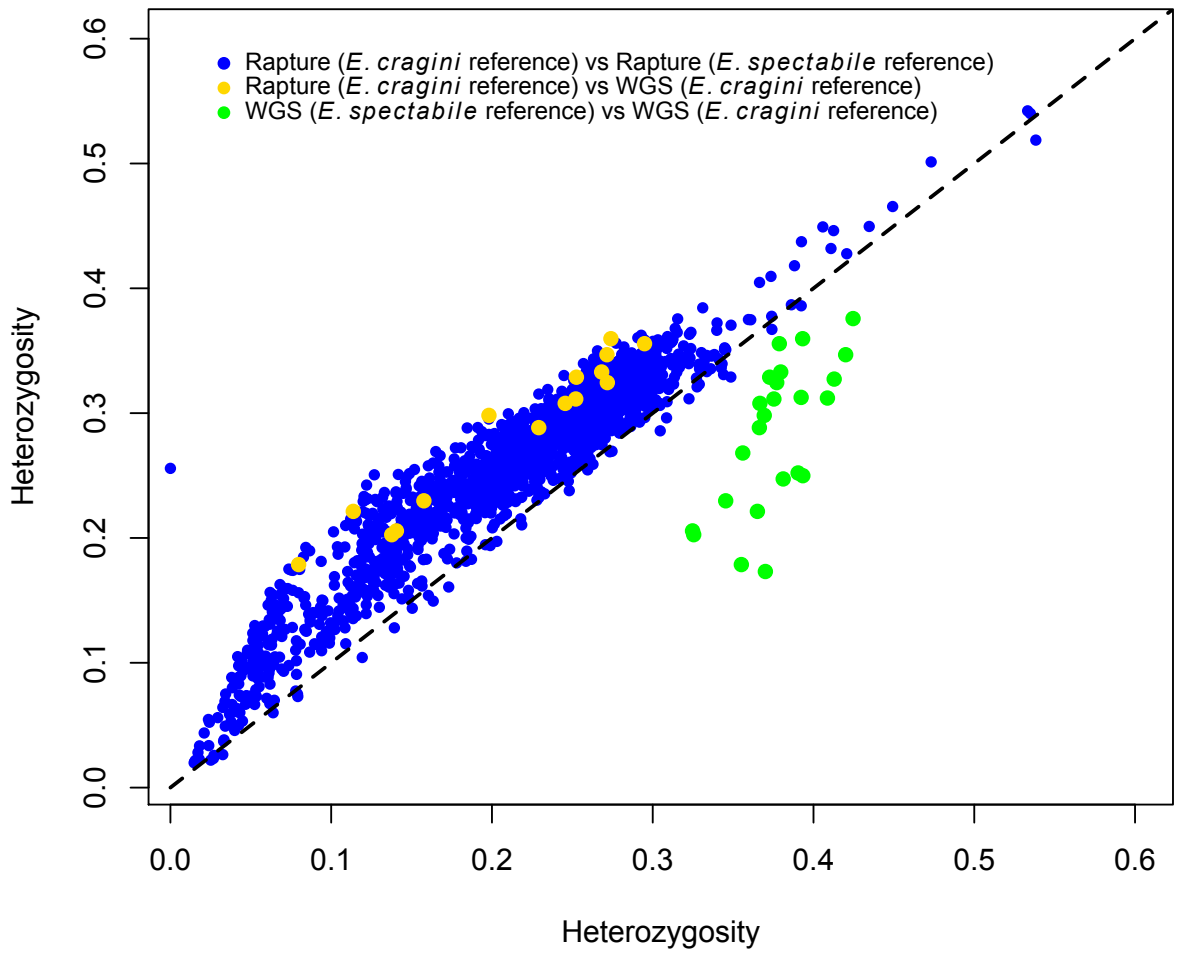
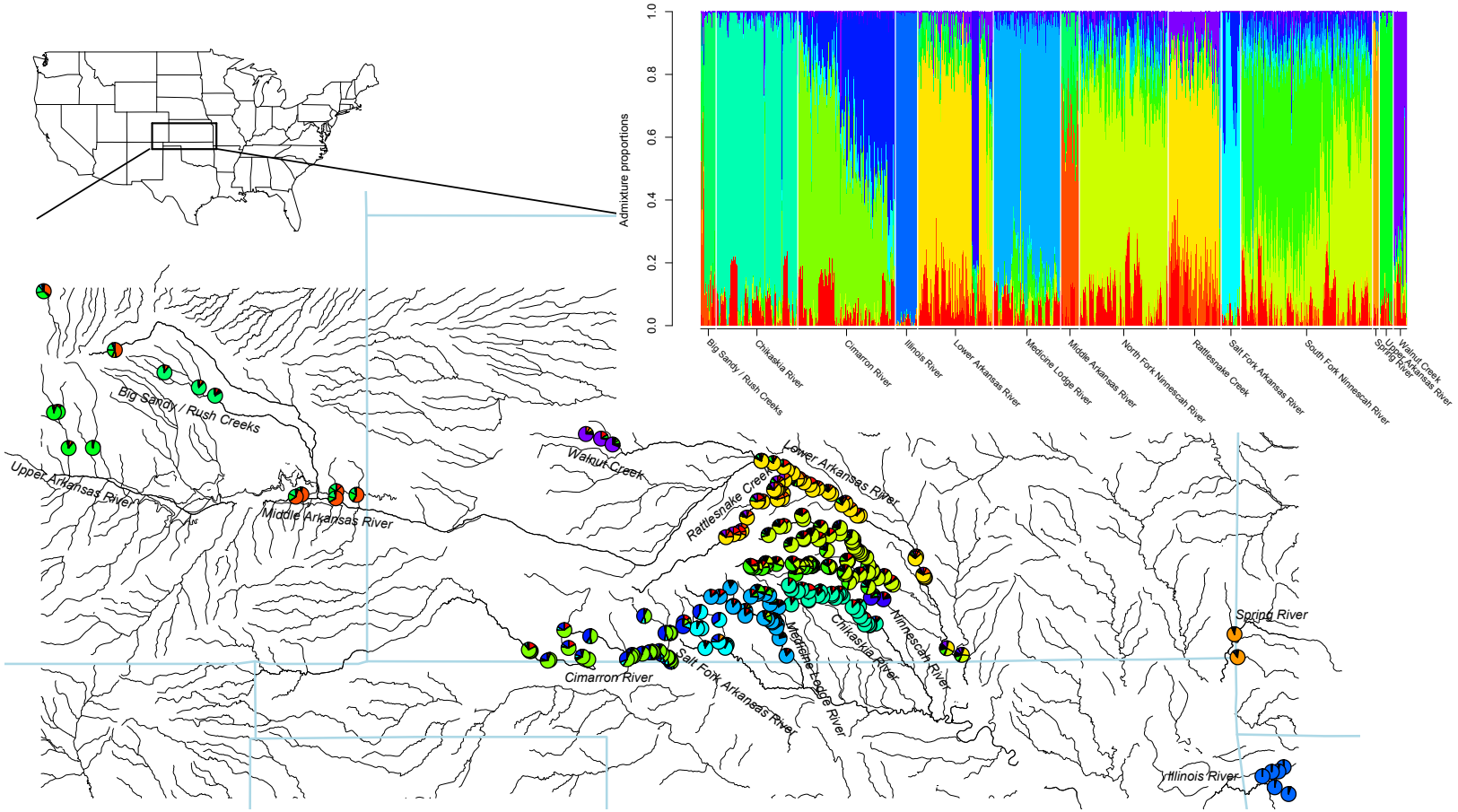


Figure 5. Admixture plots and mapped ancestry proportions. Each line in barplots represents an individual, and colors represent proportion of ancestry for each individual assigned to a given population. For the maps, pie charts represent either ancestry proportions aggregated for all individuals at a given site (for Rapture data) or admixture proportions for a single individual (for WGS data). Text on barplots indicates drainage of origin.

5a. Rapture loci, full dataset, *E. cragini* reference.



5b. WGS data, *E. cragini* reference

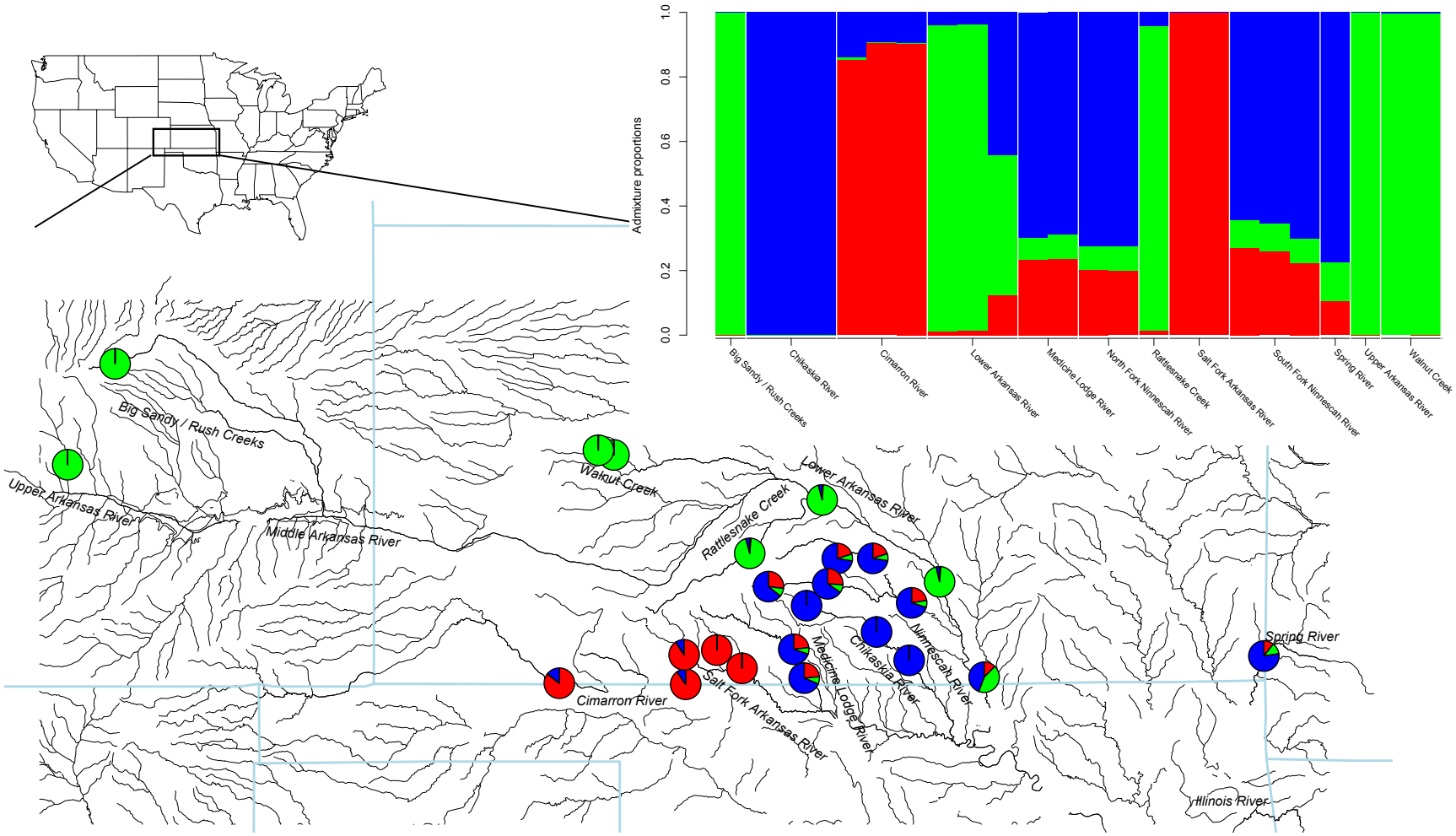
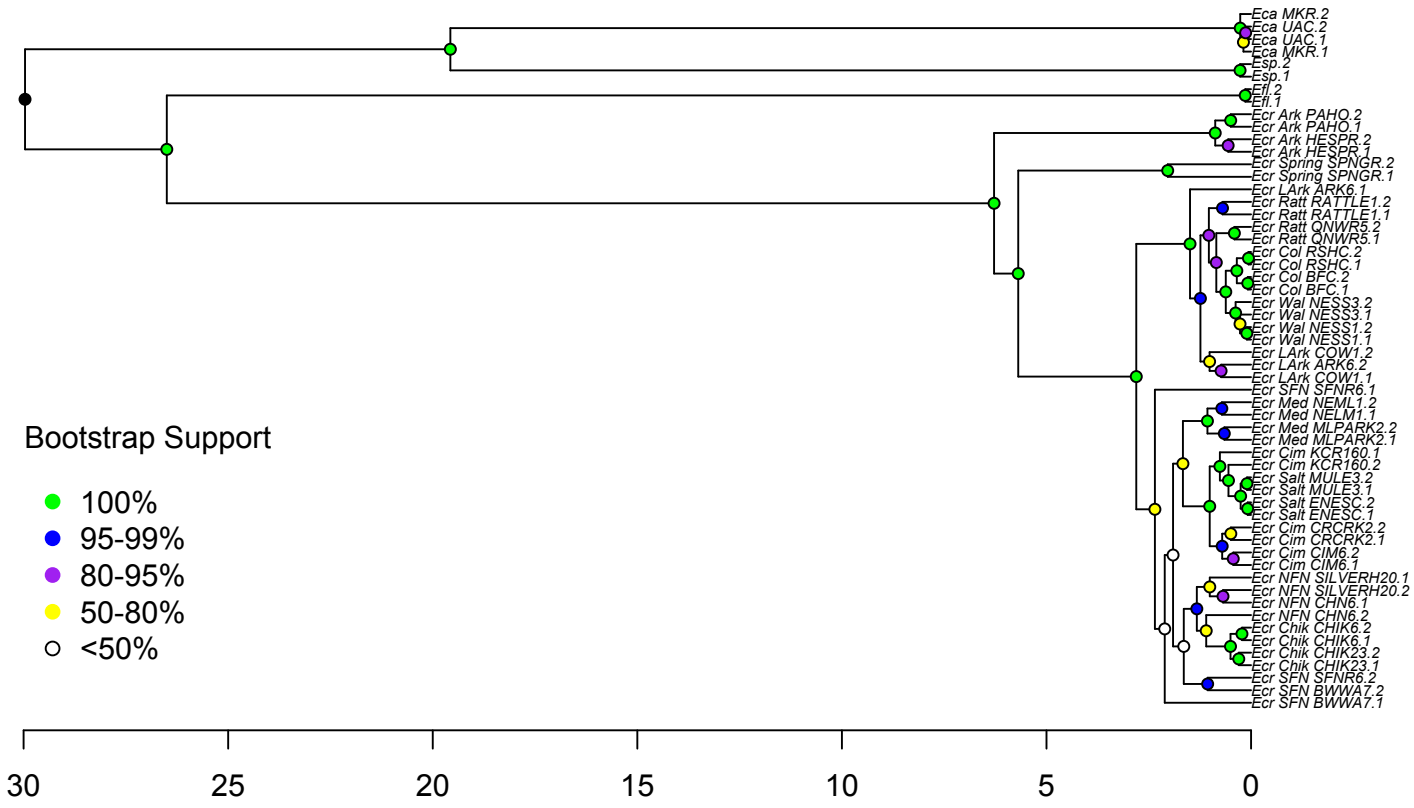
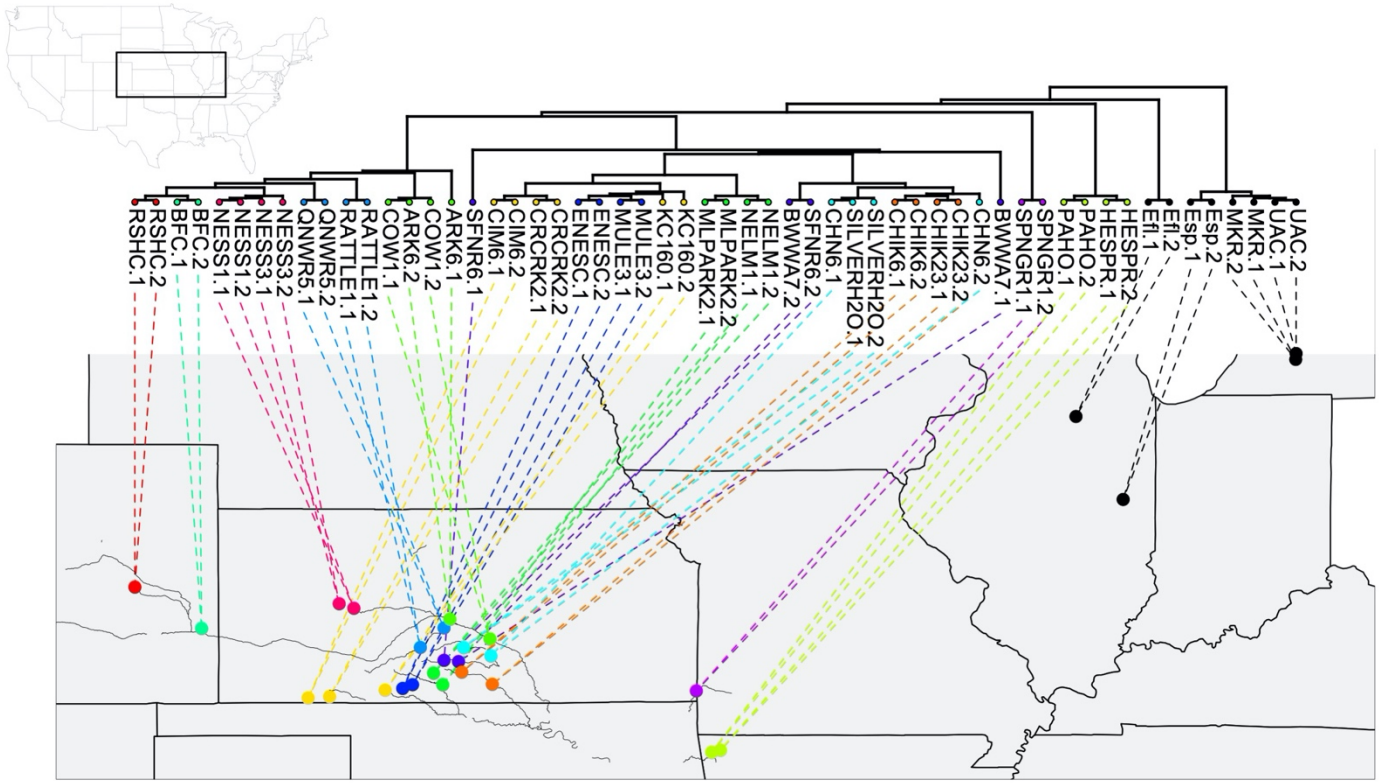


Figure 6a. Time-calibrated maximum likelihood phylogeny for Rapture data. Eca = *E. caeruleum*, Esp = *E. spectabile*, Efl = *E. flabellare*, Ecr = *E. cragini*. Node labels for *E. caeruleum* individuals include a site identifier, and node labels for *E. cragini* individuals include a metapopulation identifier followed by a site identifier. Time (on the x-axis) is expressed in millions of years ago.



6b. Phylogeny plotted in space.