# Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations

Thomas C. Nelson[1,2,*] and William A. Cresko[1,†]

[1]Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403
[2]Current address: Division of Biological Sciences, University of Montana, Missoula, MT 59812
*thom.nelson@mso.umt.edu
[†]To whom correspondence should be addressed: wcresko@uoregon.edu

Running title: Ancient genomic variation fuels recent adaptation
Keywords: adaptation, evolutionary genomics, speciation islands, RAD-seq
Abstract word count: 254
Total word count: 5867

ABSTRACT

Adaptation in the wild often involves standing genetic variation (SGV), which allows rapid responses to selection on ecological timescales. However, we still know little about how the evolutionary histories and genomic distributions of SGV influence local adaptation in natural populations. Here, we address this knowledge gap using the threespine stickleback fish (*Gasterosteus aculeatus*) as a model. We extend the popular restriction site-associated DNA sequencing (RAD-seq) method to produce phased haplotypes approaching 700 base pairs (bp) in length at each of over 50,000 loci across the stickleback genome. Parallel adaptation in two geographically isolated freshwater pond populations consistently involved fixation of haplotypes that are identical-by-descent. In these same genomic regions, sequence divergence between marine and freshwater stickleback, as measured by $d_{XY}$, reaches ten-fold higher than background levels and structures genomic variation into distinct marine and freshwater haplogroups. By combining this dataset with a *de novo* genome assembly of a related species, the ninespine stickleback (*Pungitius pungitius*), we find that this habitat-associated divergent variation averages six million years old, nearly twice the genome-wide average. The genomic variation that is involved in recent and rapid local adaptation in stickleback has actually been evolving throughout the 15-million-year history since the two species lineages split. This long history of genomic divergence has maintained large genomic regions of ancient ancestry that include multiple chromosomal inversions and extensive linked variation. These discoveries of ancient genetic variation spread broadly across the genome in stickleback demonstrate how selection on ecological timescales is a result of genome evolution over geological timescales, and *vice versa*.

IMPACT STATEMENT

Adaptation to changing environments requires a source of genetic variation. When environments change quickly, species often rely on variation that is already present – so-called standing genetic variation – because new adaptive mutations are just too rare. The threespine stickleback, a small fish species living throughout the

Northern Hemisphere, is well-known for its ability to rapidly adapt to new environments. Populations living in coastal oceans are heavily armored with bony plates and spines that protect them from predators. These marine populations have repeatedly invaded and adapted to freshwater environments, losing much of their armor and changing in shape, size, color, and behavior.

Adaptation to freshwater environments can occur in mere decades and probably involves lots of standing genetic variation. Indeed, one of the clearest examples we have of adaptation from standing genetic variation comes from a gene, *eda*, that controls the shifts in armor plating. This discovery involved two surprises that continue to shape our understanding of the genetics of adaptation. First, freshwater stickleback from across the Northern Hemisphere share the same version, or allele, of this gene. Second, the 'marine' and 'freshwater' alleles arose millions of years ago, even though the freshwater populations studied arose much more recently. While it has been hypothesized that other genes in the stickleback genome may share these patterns, large-scale surveys of genomic variation have been unable to test this prediction directly.

Here, we use new sequencing technologies to survey DNA sequence variation across the stickleback genome for patterns like those at the *eda* gene. We find that *every* region of the genome associated with marine-freshwater genetic differences shares this pattern to some degree. Moreover, many of these regions are as old or older than *eda*, stretching back over 10 million years in the past and perhaps even predating the species we now call the threespine stickleback. We conclude that natural selection has maintained this variation over geological timescales and that the same alleles we observe in freshwater stickleback today are the same as those under selection in ancient, now-extinct freshwater habitats. Our findings highlight the need to understand evolution on macroevolutionary timescales to understand and predict adaptation happening in the present day.

## INTRODUCTION

The mode and tempo of adaptive evolution depend on the sources of genetic variation affecting fitness (Wright 1932; Orr 2005). While new mutation is the ultimate origin of all genetic variation, recent studies of adaptation in the wild have documented adaptive genetic variation that was either segregating in the ancestral population as standing genetic variation (SGV)(Barrett & Schluter 2008; Domingues *et al.* 2012; Schrider & Kern 2017), or introgressed from a separate population or species (Huerta-Sánchez *et al.* 2014; Fontaine *et al.* 2015). The use of SGV during evolution appears particularly important when dramatic responses to selection occur on ecological timescales, in dozens of generations or fewer (Barrett & Schluter 2008). When environments change rapidly, SGV can propel rapid evolution in ecologically relevant traits even in populations of long-lived organisms like Darwin's finches (Grant & Grant 2002), monkeyflowers (Wright *et al.* 2013), and threespine stickleback fish (Colosimo *et al.* 2005).

The contribution of SGV to rapid divergence has important consequences for our understanding of evolutionary genetics. Existing genetic variants have evolutionary histories that are often unknown, but which may none-the-less have significant impacts on subsequent adaptation (Kirkpatrick & Barton 2006; Wright *et al.* 2013). The abundance, genomic distribution, and fitness effects (Charlesworth *et al.* 1993; Colosimo *et al.* 2005; Kirkpatrick & Barton 2006; Linnen *et al.* 2009; Stankowski & Streisfeld 2015) of SGV are themselves the products of evolution, and their unknown history raises fascinating questions for the genetics of adaptation in the wild. When did adaptive variants originally arise? How are they structured, across both geography and the genome? Which evolutionary forces shaped their current distribution and how might this evolutionary history channel future evolutionary change?

Answers to these questions are critical for our understanding of the importance of SGV in nature, as well as our ability to predict the paths available to adaptation on ecological timescales (Wright *et al.* 2013). Biologists are beginning to probe evolutionary histories of SGV using genome-wide sequence variation across multiple individuals in numerous populations (Pease *et al.* 2016), but this level of inference has

been unavailable for most natural systems because of methodological limitations that remove phase information (e.g. pool-seq: Schlotterer *et al.* 2014) or produce very short reads (e.g. RAD-seq: Davey *et al.* 2011). Here, we investigate the structure and evolutionary history of divergent SGV by modifying the original sheared RAD-seq method to generate ~700 bp haplotypes at tens of thousands of loci sampled across the stickleback genome. This approach allows us to accurately measure sequence variation and estimate divergence times across the genome. By collecting more detailed sequence information at each RAD locus, this approach also provides more accurate estimates of polymorphism and divergence at each locus, and with far smaller sample sizes, compared to traditional short-read methods (Nei 1987 chapters 10 and 13; Wakeley 2009; Cruickshank & Hahn 2014).

SGV has long been postulated to be critical to adaptation in stickleback, and several recent population genomic studies have supported this hypothesis (Hohenlohe *et al.* 2010; Jones *et al.* 2012; Roesti *et al.* 2015; Samuk *et al.* 2017). Marine stickleback have repeatedly colonized freshwater lakes and streams (Bell & Foster 1994b; Jones *et al.* 2012; Wund *et al.* 2016), and adaptive divergence in isolated freshwater habitats is highly parallel at the phenotypic (Colosimo *et al.* 2004; Cresko *et al.* 2004a) and genomic levels (Hohenlohe *et al.* 2010; Jones *et al.* 2012; but see Stuart *et al.* 2017). In addition, analyses of haplotype variation at the genes *eda* (Colosimo *et al.* 2005; Roesti *et al.* 2014) and *atp1a1* (Roesti *et al.* 2014) present two clear results: separate freshwater populations share common 'freshwater' haplotypes that are identical-by-descent (IBD), and sequence divergence between the major marine and freshwater haplogroups suggests their ancient origins – perhaps over two million years ago in the case of *eda* (Colosimo *et al.* 2005). While intriguing, it is not clear whether the deep evolutionary histories of these loci are outliers or representative of more widespread ancient history across the genome. Furthermore, although recent population genomic studies have made important contributions to identifying that SGV across the genome, the short reads employed limit the accuracy of genealogical inference across the genome.

136 To address fundamental questions of genealogical relationships and molecular
137 evolution in stickleback, we utilize the new RAD-seq haplotyping approach to assay
138 genome-wide variation associated with adaptive divergence in two young freshwater
139 ponds, which formed during the end-Pleistocene glacial retreat (c. 12,000 years ago:
140 Francis *et al.* 1986; Cresko *et al.* 2004a, Fig. 1). In addition, we generated a *de novo*
141 genome assembly of the sister taxon ninespine stickleback (*Pungitius pungitius*),
142 allowing us to estimate divergence times for genealogies across the genome. Our
143 results clearly demonstrate that the previous findings of deep evolutionary history based
144 upon candidate loci are not unique but in fact the rule. A suite of adaptive variation
145 structured into distinct marine and freshwater haplotypes that evolved over millions of
146 years forms the foundation of a deep pool of SGV that undergirds repeated and rapid
147 evolution in stickleback.
148
149 METHODS
150 *Sample collection*
151 Wild threespine stickleback were collected from Rabbit Slough (N 61.5595, W
152 149.2583), Boot Lake (N 61.7167, W 149.1167), and Bear Paw Lake (N 61.6139, W
153 149.7539). Rabbit Slough is an offshoot of the Knik Arm of Cook Inlet and is known to
154 be populated by anadromous populations of stickleback that are stereotypically oceanic
155 in phenotype and genotype (Cresko *et al.* 2004b; Hohenlohe *et al.* 2010). Boot Lake
156 and Bear Paw Lake are both shallow lakes formed during the end-Pleistocene glacial
157 retreat. Fish were collected in the summers of 2009 (Rabbit Slough), 2010 (Bear Paw
158 Lake), and 2014 (Boot Lake) using wire minnow traps and euthanized *in situ* with
159 Tricaine solution. Euthanized fish were immediately fixed in 95% ethanol and shipped to
160 the Cresko Laboratory at the University of Oregon (Eugene, OR, USA). DNA was
161 extracted from fin clips preserved in 95% ethanol using either Qiagen DNeasy spin
162 column extraction kits or Ampure magnetic beads (Beckman Coulter, Inc) following
163 manufacturer's instructions. Yields averaged 1-2 µg DNA per extraction (~30 mg
164 tissue). Treatment of animals followed protocols approved the University of Oregon
165 Institutional Animal Care and Use Committee (IACUC).

166

167 *Sequencing strategy and rationale*

168 We designed our sequencing to maximize detection of sequence variation and
169 divergence, with the ultimate goal being the estimation of absolute divergence times of
170 marine and freshwater haplogroups. Previous work by us and others using short
171 sequence reads provided clear evidence of changes in relative frequencies of alleles
172 across stickleback populations (Hohenlohe *et al.* 2010; Roesti *et al.* 2014; Lescak *et al.*
173 2015; Roesti *et al.* 2015), but could not sufficiently address questions of haplotype ages.
174 We therefore designed a RAD sequencing approach to (1) accurately estimate
175 sequence diversity within and divergence between threespine stickleback ecotypes and
176 (2) recover sufficient RAD loci that map unambiguously to an outgroup genome
177 sequence from the ninespine stickleback that we could confidently compare diversity
178 within threespine stickleback to divergence from the ninespine stickleback.

179 To achieve our aims, we designed a sequencing method to produce phased
180 haplotypes of ~700 bp at each RAD locus (Fig. 1B-D) and to sample the genome
181 densely enough to identify signatures of selection after the likely dropout of RAD loci
182 without clear homology in the ninespine stickleback genome. We used the single-digest,
183 sheared RAD approach to limit biases in our estimates of sequence diversity. RAD-seq
184 has known biases due to mutations in restriction sites causing allele dropout (Arnold *et*
185 *al.* 2013; Gautier *et al.* 2013), the potential for which increases with increasing
186 sequence divergence and leads to underestimates of genetic diversity. Diversity
187 estimates are, however, substantially more accurate with sheared RAD-seq compared
188 to other RAD-seq approaches (e.g. double-digest RAD-seq: Peterson *et al.* 2012).
189 Importantly for the coalescent analyses we present here, such allele dropout is unlikely
190 to affect estimates of overall divergence across the clade of alleles. When in the rare
191 cases it does, the bias is toward *underestimation* of the divergence age (Arnold *et al.*
192 2013), which would make our findings of deep divergence even more striking.

193 Our sequencing design facilitated accurate inference of sequence variation even
194 with smaller population samples than are typical among population genomic studies.
195 While allele frequency-based statistics like $F_{ST}$ have particularly high variance with small

196 sample sizes (Willing *et al.* 2012), our study is fortunate to be built upon numerous
197 properly powered, previous population genomic studies in stickleback including in these
198 populations. The genome-wide patterns of $F_{ST}$ we observed using our new approach
199 closely matched multiple previous studies (Hohenlohe *et al.* 2010; Jones *et al.* 2012,
200 Fig. 2A). Because of this extensive body of previous work, we relied on $F_{ST}$ only to draw
201 inference of larger genomic regions containing tens or hundreds of RAD loci. Instead,
202 as stated above the focus of this work is to extend these previous findings by
203 addressing the ages of allelic divergence. We therefore do not expect the higher
204 variance associated with smaller sample size to qualitatively influence our results.
205 Importantly, estimation of sequence diversity ($\pi$) (Nei 1987) and divergence ($d_{XY}$) (Nei
206 1987; Cruickshank & Hahn 2014) at a given locus improves greatly with increases in
207 sequence length. Using equations 10.9 and 13.83 from Nei (1987; Box 1 in Cruickshank
208 & Hahn 2014), the predicted sampling variances in both $\pi$ and $d_{XY}$ using 700 bp
209 sequences in five individuals are lower than those obtained using standard 100 bp
210 sequences at any sample size (Suppl. Fig. S1). Therefore, not only is this novel
211 application of RAD-seq ideally suited for our questions, our findings show that this
212 approach may significantly decrease the necessary sample size, and thus resource
213 expenditure, for many population genomic studies.
214
215 *Library preparation*
216 To identify sufficient sequence variation at a RAD locus, and to simplify
217 downstream sequence processing and analysis, we took advantage of longer
218 sequencing reads available on newer Illumina platforms and the phase information
219 captured by paired-end sequencing. We generated RAD libraries from these samples
220 using the single-digest sheared RAD protocol from Baird et al. (2008) with the following
221 specifications and adjustments: 1 µg of genomic DNA per fish was digested with the
222 restriction enzyme *PstI-HF* (New England Biolabs), followed by ligation to P1 Illumina
223 adaptors with 6 bp inline barcodes. Ligated samples were multiplexed and sheared by
224 sonication in a Bioruptor (Diagenode). To ensure that most of our paired-end reads
225 would overlap unambiguously and produce longer contiguous sequences, we selected a

226 narrow fragment size range of 425-475 bp. The remainder of the protocol was per Baird
227 et al. (Baird *et al.* 2008b). All fish were sequenced on an Illumina HiSeq 2500 using
228 paired-end 250 bp sequencing reads at the University of Oregon's Genomics and Cell
229 Characterization Core Facility (GC3F).
230

231 *Sequence processing*

232 Raw Illumina sequence reads were demultiplexed, cleaned, and processed
233 primarily using the Stacks v1.46 pipeline (Catchen *et al.* 2011; Catchen *et al.* 2013a).
234 Paired-end reads were demultiplexed with **`process_shortreads`** and cleaned using
235 **`process_radtags`** using default criteria (throughout this document, names of scripts,
236 programs, functions, and command-line arguments will appear in **`fixed-width`**
237 **`font`**). Overlapping read pairs were then merged with **`fastq-join`** (Aronesty 2011).
238 Pairs that failed to merge were removed from further analysis. To retain the majority of
239 the sequence data for analysis in Stacks and still maintain adequate contig lengths,
240 merged contigs were trimmed to 350 bp and all contigs shorter than 350 bp were
241 discarded. We aligned these contigs to the stickleback reference genome (Jones *et al.*
242 2012; Glazer *et al.* 2015) using **`bbmap`** v35.69 with the most sensitive alignment settings
243 ('**`vslow=t`**'; http://jgi.doe.gov/data-and-tools/bbtools/) and required that contigs mapped
244 uniquely to the reference. We then used the **`pstacks`**, **`cstacks`**, and **`sstacks`**
245 components of the Stacks pipeline to identify RAD-tags and call SNPs using the
246 maximum likelihood algorithm implemented in **`pstacks`**, create a catalog of RAD tags
247 across individuals, and match tags across individuals. All data were then passed
248 through the Stacks error correction module **`rxstacks`** to prune unlikely haplotypes. We
249 ran the Stacks component program **`populations`** on the final dataset to filter loci
250 genotyped in fewer than four individuals in each population and to create output files for
251 sequence analysis. We use the naming conventions of Baird et al. (2008a): A "RAD tag"
252 refers to sequence generated from a single end of a restriction site and the pair of RAD
253 tags sequenced at a restriction site comprises a "RAD locus" (Fig. 1D).

254 We used the program **phase** v2.1 (Stephens *et al.* 2001; Stephens & Scheet
255 2005) to phase pairs of RAD tags originating from the same restriction site. We coded
256 haplotypes present at each RAD tag, which often contain multiple SNPs, into multiallelic
257 genotypes. This both simplified and reduced computing time for the phasing process.
258 We also performed coalescent simulations to generate, 'cut', and re-phase haplotypes
259 to demonstrate the high accuracy of this method using sequences and sample sizes
260 similar to those in this study (Suppl. Fig. S2). Custom Python scripts automated this
261 process and are included as supplementary files. We required that each individual had
262 at least one sequenced haplotype at each tag for phasing to be attempted. If a sample
263 had called genotypes at only one tag in the pair, the sample was removed from further
264 analysis of that locus. The resultant phased haplotypes were used to generate
265 sequence alignments for import into BEAST.
266 We recovered a total of 236,787 RAD tags after filtering, mapping to 151,813 *PstI*
267 restriction sites. At 84,974 restriction sites, we recovered and successfully phased
268 adjacent RAD tags (169,948 RAD tags) into single RAD loci. RAD tags with no variable
269 sites were simply concatenated to the adjacent tag to form a single locus. We retained
270 these 84,974 RAD loci for our analysis. For population genetic analyses, inclusion of
271 singleton (i.e. unpaired) RAD tags did not qualitatively change our results. We chose to
272 restrict genealogical analyses to loci of uniform length and to use the same set of loci in
273 analyses of polymorphism and gene tree topologies.
274
275 *Ninespine stickleback genome assembly*
276 In order to estimate the $T_{MRCA}$ of threespine stickleback RAD alleles, we used the
277 ninespine stickleback (*Pungitius pungitius*) as an outgroup. RAD sequence analysis,
278 however, relies on the presence of homologous restriction sites among sampled
279 individuals and results in null alleles when mutations occur within a restriction
280 site(Arnold *et al.* 2013). Because this probability increases with greater evolutionary
281 distance among sampled sequences, we elected to use RAD-seq to only estimate
282 sequence variation within the threespine stickleback. We then generated a contig-level
283 *de novo* ninespine stickleback genome assembly from a single ninespine stickleback

284 individual from St. Lawrence Island, Alaska (collected by J. Postlethwait) using
285 DISCOVAR *de novo* revision 52488
286 (https://software.broadinstitute.org/software/discovar/blog/). We used this single
287 ninespine stickleback haplotype to estimate threespine-ninespine sequence divergence
288 and time calibrate coalescence times within the threespine stickleback. DISCOVAR de
289 novo requires a single shotgun library of paired-end 250-bp sequence reads from short-
290 insert-length DNA fragments. High molecular weight genomic DNA was extracted from
291 an ethanol-preserved fin clip by proteinase K digestion followed by DNA extraction with
292 Ampure magnetic beads. Purified genomic DNA was mechanically sheared by
293 sonication and size selected to a range of 200-800 bp by gel electrophoresis and
294 extraction. We selected this fragment range to agree with the recommendations for *de*
295 *novo* assembly using DISCOVAR *de novo*. This library was sequenced on a single lane
296 of an Illumina HiSeq2500 at the University of Oregon's Genomics and Cell
297 Characterization Core Facility (GC3F: https://gc3f.uoregon.edu/). We assembled the
298 draft ninespine stickleback genome using DISCOVAR de novo. Raw sequence read
299 pairs were first quality filtered and adaptor sequence contamination removed using the
300 program **process_shortreads**, which is included in the Stacks analysis pipeline
301 (Catchen *et al.* 2013b). We ran the genome assembly on the University of Oregon's
302 Applied Computational Instrument for Scientific Synthesis (ACISS: http://aciss-
303 computing.uoregon.edu).
304
305 *Alignment of RAD tags to the ninespine assembly*
306 We included the single ninespine stickleback haplotype into our sequence
307 analyses by aligning a single phased threespine stickleback RAD haplotype from each
308 locus to the ninespine genome assembly. For those that aligned uniquely (59,254 RAD
309 loci), we used a custom Python script to parse the alignment fields of the output BAM
310 file (Li *et al.* 2009) and reconstruct the ninespine haplotype by introducing threespine-
311 ninespine substitutions into the threespine RAD locus sequence. The final dataset
312 consists of 57,992 RAD loci that mapped to the 21 threespine stickleback chromosomes
313 and aligned uniquely to the ninespine assembly.

314

315 *Lineage sorting and time to the most recent common ancestor*

316 Allelic divergence can occur by multiple modes of lineage sorting during
317 adaptation. To identify patterns of lineage sorting associated with freshwater
318 colonization, we analyzed gene tree topologies at all RAD loci using BEAST v. 1.7
319 (Drummond & Rambaut 2007; Drummond *et al.* 2012). We chose BEAST because it co-
320 estimates tree topologies and node ages for sequenced genomic loci. BEAST does not
321 explicitly perform model selection, and this may affect divergence time estimates in
322 genomic regions under direct or indirect selection. However, other methods developed
323 to estimate the age of adaptive alleles model evolutionary scenarios that are likely not
324 relevant to the evolutionary histories we infer here. First, some models assume a recent
325 origin of an adaptive allele compared to adjacent genomic variation (Peter *et al.* 2012;
326 Ormond *et al.* 2016), which is the opposite of what we describe here, so that measures
327 of variation at linked sites and the decay of linkage disequilibrium can be used to
328 estimate when a sweep began. Selection in the stickleback populations we study likely
329 acted on SGV, as has been supported by previous studies, and we hypothesize that this
330 SGV may be quite old. Therefore, adaptive alleles already existed on distinct haplotype
331 backgrounds, which masks the differences between selected and linked neutral sites.

332 Second, a recent model developed to infer ages of standing genetic variants
333 assumes that the variant was evolving neutrally at some point during its trajectory
334 through a population (Peter *et al.* 2012). This assumption is unlikely for many of the loci
335 we detect here, except in the very distant past and for those loci that have evolved
336 recently arose in genomic regions already heavily influenced by selection. Rather, the
337 patterns of haplotype variation we observed in the genomic regions that differentiate
338 marine and freshwater populations reflect long-term maintenance and isolation of
339 separate haplogroups that mimics population structure and even speciation, with
340 selective sweeps being important but constituting a small minority of the time these
341 haplotypes have segregating in the stickleback metapopulation. For all of these reasons
342 we therefore chose to estimate tree topologies and divergence times with BEAST,
343 which makes minimal assumptions regarding specific evolutionary processes.

344 We used blanket parameters and priors for BEAST analyses across all RAD loci.
345 Markov chain Monte Carlo (MCMC) runs of 1,000,000 states were specified, and trees
346 logged every 100 states. We used a coalescent tree prior and the GTR+Γ substitution
347 model with four rate categories and uniform priors for all substitution rates. We identified
348 evidence of lineage sorting by using the program **`treeannotator`** v1.7.5 to select the
349 maximum clade credibility (MCC) tree for each RAD locus and the
350 **`is.monophyletic()`** function included in the R package 'ape' v3.0 (Paradis *et al.*
351 2004; Popescu *et al.* 2012). We determined for each MCC tree whether tips originating
352 from marine (RS) or freshwater (BL+BP) formed monophyletic clades.
353 To convert node ages estimated in BEAST into divergence times, in years, we
354 assumed a 15 million-year divergence time between threespine and ninespine
355 stickleback at each RAD locus (Aldenhoven *et al.* 2010). The $T_{MRCA}$ of all alleles in each
356 gene tree was set at 15 Mya and each node age of interest was converted into years
357 relative to the total height of the tree. Additionally, to use the ninespine stickleback as
358 an outgroup, we required that threespine stickleback haplotypes at a RAD locus were
359 monophyletic to the exclusion of the ninespine haplotype. Doing so reduced our
360 analysis to 49,672 RAD loci for analyses included in Fig. 4 of the main text. RAD loci not
361 showing this pattern of lineage sorting did not show evidence of a genome-wide
362 correlation with marine-freshwater divergence and thus do not impact the assertions in
363 the main text. We used medians of the posterior distributions as point estimates of
364 $T_{MRCA}$ for each RAD locus. Because of the somewhat limited information from any single
365 RAD locus, and because the facts of the genealogical process mean that the true $T_{MRCA}$
366 at any locus likely differs from the 15 My estimate (Kingman 1982a, b; Tajima 1983), we
367 do not rely heavily on $T_{MRCA}$ estimates at individual RAD loci. Rather, we use these
368 estimates to understand patterns of broad patterns of ancestry throughout the
369 threespine stickleback genome — spatially along chromosomes and genome-wide
370 patterns.
371 We determined $T_{MRCA}$ outlier genomic regions by permuting and kernel
372 smoothing the genomic distribution of $T_{MRCA}$ estimates using the same window sizes as
373 we present in the main text. Windows where the actual $T_{MRCA}$ exceeded 99.9% of

374 permuted windows were considered outliers. This method controls for the local density
375 of RAD loci (poorly sampled regions will have larger confidence bands) and the size of
376 the windows used.
377
378 *Sequence diversity and haplotype networks*
379 We quantified sequence diversity within and among populations and sequence
380 divergence between populations using R v3 (R Core Team 2016). We used the R
381 package 'ape' (Paradis *et al.* 2004) to compute pairwise distance matrices for all alleles
382 at each RAD locus and used these matrices to calculate the average pairwise
383 nucleotide distances, $\pi$, within and among populations along with $d_{XY}$, the average
384 pairwise distance between two sequences using only across-population comparisons
385 (Nei 1987). We also calculated the haplotype-based $F_{ST}$ from Hudson et al. (1992)
386 implemented in the R package 'PopGenome' v2.2.4 (Pfeifer *et al.* 2014). We used
387 permutation tests written in R to identify differences in variation within- and between-
388 habitat type at divergent RAD loci versus the genome-wide distributions. Mann-Whitney-
389 Wilcoxon tests implemented in R were used to identify variation in genome-wide
390 diversity among populations and habitat types.
391 We constructed haplotype networks of the RAD loci at *eda* and *atp1a1* using the
392 infinite sites model with the function **`haploNet()`** in the R package 'pegas' (Paradis
393 2010). The *atp1a1* network was constructed from from a RAD locus spanning exon 15
394 of *atp1a1* and including portions of introns 14 and 15 at (chr1:21,726,729-21,727,381
395 [BROAD S1, v89]; chr1: 26,258,117-26,257,465 [re-scaffolding from Glazer, *et al*
396 (2015)]). The *eda* network spans exon 2 and portions of introns 1 and 3 of *eda* (chr4:
397 12,808,396-12,809,030).
398
399 *Code availability*
400 Scripts used to phase RAD-tags, summarize gene trees, calculate population genetic
401 statistics, and produce figures and statistics presented in paper are available at
402 https://github.com/thomnelson/ancient-divergence. Scripts for processing raw sequence
403 data are available from the authors upon request.

404

405 *Data availability*

406 Raw sequence data supporting these findings are available on the Sequence Read
407 Archive at PRJNAXXXXXX. The final datasets needed to reproduce the figures and
408 statistics presented in the paper are available at https://github.com/thomnelson/ancient-
409 divergence.

410

411 RESULTS AND DISCUSSION

412

413 Parallel adaptation to freshwater environments has been a major theme of
414 stickleback evolutionary history (Bell & Foster 1994a). Stereotypical morphological
415 changes to, for example, bony armor(Colosimo *et al.* 2004) and craniofacial structures
416 (Kimmel *et al.* 2005) presumably reflect adaptation to similar selective regimes
417 (Reimchen 1994; Arnegard *et al.* 2014). These phenotypic changes are accompanied
418 by parallel genomic divergence (Hohenlohe *et al.* 2010; Jones *et al.* 2012), which
419 involves large regions spanning many megabases (Schluter & Conte 2009; Roesti *et al.*
420 2014), including multiple chromosomal inversions (Jones *et al.* 2012). The leading
421 hypothesis for the genetics of parallel divergence in stickleback posits that distinct
422 freshwater-adaptive haplotypes that are identical-by-descent (IBD) are shared among
423 freshwater populations due to historical gene flow between marine and freshwater
424 populations (Schluter & Conte 2009). We tested for the presence of these haplotypes
425 directly and at a genomic scale.

426

427 *Parallel divergence involves a shared suite of haplotypes genome-wide*

428 Our sequencing strategy produced 57,992 RAD loci, with 690 potential variable
429 sites each, present across the three threespine stickleback populations and aligned to
430 the ninespine stickleback genome assembly. These data comprise over 40 Mb of
431 sequence, or nearly 10% of the threespine stickleback genome (9.5% of 419 Mb
432 assigned to chromosomes) (Jones *et al.* 2012; Glazer *et al.* 2015). All loci we recovered
433 were polymorphic and we observed a median of seven segregating sites per locus

434 (range: 2-155, Suppl. Fig. S3, Suppl. Table 1). By including haplotypes from all three
435 populations in these genealogical analyses, we were able to jointly calculate population
436 genetic statistics ($F_{ST}$, $\pi$, $d_{XY}$) and identify patterns of identity-by-descent (IBD) among
437 populations, which we defined as haplotypes from two populations forming a
438 monophyletic group to the exclusion of the third population.

439 We find that parallel population genomic divergence in the two freshwater pond
440 populations consistently involved haplotypes that were identical-by-descent (IBD)
441 among both freshwater populations (Fig. 2). Background $F_{ST}$ between populations
442 ranged from 0.139-0.226, with genome-wide differentiation between the freshwater
443 populations BL and BP being highest ($F_{ST(RS-BL)}$ = 0.139, $F_{ST(RS-BP)}$ = 0.194, $F_{ST(BL-BP)}$ =
444 0.226; two-sided Mann-Whitney test for all pairwise comparisons: $p \leq 1x10^{-10}$). The
445 degree and genomic distribution of pairwise $F_{ST}$ between the BL, BP, and RS
446 populations were similar to those previously reported (Hohenlohe *et al.* 2010). This
447 similarity included marine-freshwater $F_{ST}$ outlier regions on chromosome 4 over a broad
448 span in which the *eda* gene is embedded (orange triangle in Fig. 2A), and three regions
449 now known to be associated with chromosomal inversions on chromosomes 1, 11, and
450 21 (yellow bars in Fig. 2; hereafter referred to as *inv1*, *inv11*, and *inv21*). The gene
451 *atp1a1* (green triangle in Fig. 2A) is contained within *inv1*. As expected, we found
452 distinct haplogroups associated with marine and freshwater habitats at both *eda* and
453 *atp1a1* (Fig. 3, insets).

454 Strikingly, this finding of habitat specific haplogroups was not at all unique to
455 these well studied genes or chromosomal inversions. The two isolated freshwater
456 populations shared IBD haplotypes within all common marine-freshwater $F_{ST}$ peaks
457 even though IBD was rare elsewhere (Fig. 2B). Furthermore, we observed a separate
458 clade of haplotypes representing the marine RS population at the majority (1129 of
459 2172, 52%) of RAD loci showing freshwater IBD. The result was a genome-wide pattern
460 of reciprocal monophyly between marine and freshwater haplotypes. Notably, this is the
461 same genealogical structure previously reported at *eda* (Colosimo *et al.* 2005; Roesti *et*
462 *al.* 2014) and *atp1a1* (Roesti *et al.* 2014), demonstrating that these loci are but a small
463 part of a genome-wide suite of genetic variation sharing similar habitat-specific

464 evolutionary histories, and the previous documentation of their genealogies was a
465 harbinger of a much more extensive pattern across the genome revealed here.
466 Hereafter, we refer collectively to this class of RAD loci as 'divergent loci'.
467
468 *Adaptive marine-freshwater sequence divergence involves ancient allelic origins*
469 Because the genealogical structure of divergence across the genome mirrors
470 that at *eda* and *atp1a1*, we asked whether levels of sequence variation and divergence
471 also showed consistent genomic patterns. At all RAD loci we therefore calculated $\pi$
472 within each population, as well as in the combined freshwater populations, and $d_{XY}$
473 between marine and freshwater habitat types. Genome-wide diversity was similar
474 across populations and habitat types (mean $\pi_{RS} = 0.0032$, $\pi_{BL} = 0.0034$, $\pi_{BP} = 0.0026$,
475 $\pi_{FW} = 0.0038$) and comparable to previous estimates (Hohenlohe *et al.* 2010). Likewise,
476 genome-wide $d_{XY}$ among habitat types was modest (0.0049) when compared to $\pi$
477 across all populations ($\pi = 0.0042$, two-sided Mann-Whitney test: $p \leq 1\text{x}10^{-10}$; Suppl.
478 Fig. S4). Among divergent loci, however, we observed reductions in diversity in both
479 habitats (mean $\pi_{RS\text{-divergent}} = 0.0012$, $\pi_{RS\text{-divergent}} = 0.0016$, two-sided permutation test: $p \leq$
480 $1\text{x}10^{-4}$, Fig. 3), indicating natural selection in both habitats. Sequence divergence
481 associated with reciprocal monophyly was striking, however, averaging nearly three
482 times the genome-wide mean (mean $d_{XY\text{-divergent}} = 0.0124$). This divergence ranged
483 more than an order of magnitude (0.0013–0.0442), from substantially lower than the
484 genome-wide average to ten times greater than the average. These findings indicate
485 that much of the genetic variation underlying adaptive divergence is vastly older than
486 the diverging freshwater populations themselves. Not only was adaptive variation
487 standing and structured by habitat, but it has been segregating and accumulating for
488 millennia.
489 These data clearly support the hypothesis of Schluter and Conte (2009) of
490 ancient haplotypes 'transported' among freshwater populations. Much of the divergence
491 we observed was ancient in origin, with levels of sequence divergence at some RAD
492 loci exceeding that observed at *eda* (Fig. 3, gold line) and suggestive of divergence
493 times of at least two million years ago (Colosimo *et al.* 2005). Our observation that

494 sequence variation was consistently reduced in both habitat types emphasizes that
495 alternative haplotypes at these loci are likely selected for in the marine population as
496 well as the freshwater. These alternative fitness optima — driven by divergent ecologies
497 — provide a favorable landscape for the maintenance of variation (Charlesworth *et al.*
498 1997; Lenormand 2002), but also lead to a more potent barrier to gene flow among
499 freshwater populations if there are fitness consequences in the marine habitat for
500 stickleback carrying freshwater-adaptive variation. Conditional fitness effects through
501 genetic interactions (i.e. dominance or epistasis: Otto & Bourguet 1999; Phillips 2008)
502 and genotype-by-environment interactions (McGuigan *et al.* 2011) could potentially
503 extend the residence time of freshwater haplotypes in the marine habitat. Future work
504 should consider the phenotypic effects of divergently adaptive variation in different
505 external environments (McGuigan *et al.* 2011; McCairns & Bernatchez 2012).
506 Adaptive divergence between marine and freshwater stickleback genomes is
507 likely ongoing, with recently derived alleles arising on already highly divergent genomic
508 backgrounds. We found reciprocal monophyly associated with a spectrum of sequence
509 divergence, including a substantial fraction of divergent loci (11.0%, 124/1129) with $d_{XY}$
510 below the genome-wide average. Thus, ongoing marine-freshwater ecological
511 divergence may continue to yield additional marine-freshwater genomic divergence.
512 Moreover, while this younger variation is shared between the freshwater populations in
513 this study, and localizes to genomic regions of divergence shared globally (Jones *et al.*
514 2012), some adaptive variants may be distributed only locally (e.g. limited to southern
515 Alaska or the eastern Pacific basin). Global surveys of shared variation have been
516 performed (Jones *et al.* 2012), but future work in this system should quantify the
517 distributions of locally or regionally limited genomic variation involved in ecological
518 divergence, because regional pools of variation may contribute substantially to
519 stickleback genomic and phenotypic diversity (Stuart *et al.* 2017).
520
521 *Habitat associated genomic divergence is as old as the threespine stickleback species*
522 Sequence divergence provides an important relative, but ultimately incomplete,
523 evolutionary timescale. To more directly compare the timescales of ecological

524 adaptation and genomic evolution, we translated patterns of sequence variation into the
525 time to the most recent common ancestor ($T_{MRCA}$) of allelic variation, in years. To do so,
526 we performed a *de novo* genome assembly of the ninespine stickleback (*Pungitius*
527 *pungitius*), a member of the Gasterosteidae that diverged from the threespine
528 stickleback lineage approximately 15 million years ago (MYA) (Aldenhoven *et al.* 2010)
529 (Fig. 4A, Suppl. Table 2). We then aligned our RAD dataset to this assembly and
530 estimated gene trees for each alignment with BEAST (Drummond *et al.* 2012), setting
531 divergence to the ninespine stickleback at 15 MYA (see Methods).
532 We find that the divergence of key marine and freshwater haplotypes has been
533 ongoing for millions of years and extends back to the split with the ninespine stickleback
534 lineage (Fig. 4B). Genome-wide variation averaged 4.1 MY old, and $T_{MRCA}$ for the vast
535 majority of RAD loci was under 5 MY old. In contrast, divergence times at habitat-
536 associated loci averaged 6.4 MYA and, amazingly, the most ancient 10% (118 of 1129)
537 are each estimated at over 10 MY old. This deep genomic divergence not only
538 underscores that local adaptation to marine and freshwater habitats has been occurring
539 throughout the history of the threespine stickleback lineage, for which there is evidence
540 in the fossil record going back 10 million years (Bell *et al.* 1985), but it also
541 demonstrates that at least some of the variation fueling those ancient events has
542 persisted until the present day. In some genomic regions, then, marine and freshwater
543 threespine stickleback are as divergent as threespine and ninespine stickleback, which
544 are classified into separate genera.
545 Adaptive divergence has impacted the history of the stickleback genome as a
546 whole (Fig. 4C). We identified 32.6 Mb, or 7.5% of the genome, as having elevated
547 $T_{MRCA}$ (gray boxes in Fig. 4C; two-sided permutation test of smoothed genomic
548 intervals, $p \leq 0.001$). Outside of the non-recombining portion of the sex chromosome
549 (chr. 19), the oldest regions of the stickleback genome were those enriched for
550 divergent loci. Patterns of ancient ancestry closely mirrored recent divergence in allele
551 frequencies (Fig. 2A) and it appears that historical and contemporary marine-freshwater
552 divergence has impacted ancestry across much of the length of some chromosomes.
553 Chromosome 4, for example, contains at least three broad peaks in $T_{MRCA}$ and a total of

554 5.9 Mb identified as genome-wide outliers (two-sided permutation test, $p \leq 0.001$). This
555 chromosome has been of particular interest because of its association with a number of
556 phenotypes (Colosimo *et al.* 2004; Miller *et al.* 2014), including fitness (Barrett *et al.*
557 2008). We found the major-effect armor plate locus *eda* comprised a local peak (mean
558 $T_{MRCA}$ = 6.4 MYA) nested within a large region of deep ancestry spanning 8.1 Mb.
559 Moreover, at least two other peaks distal to *eda,* centered at 21.4 Mb and 26.6 Mb,
560 were also several million years older than the genomic average at 6.8 MYA and 7.0
561 MYA, respectively.
562
563 *Long-term divergence maintains linked variation and promotes genomic structural*
564 *evolution*
565 Intriguingly, genomic regions of elevated $T_{MRCA}$ remained outliers even after
566 removing marine-freshwater relative divergence outlier loci (as measured by $F_{ST}$: Suppl.
567 Fig. S5). We estimated that 7.5% of the genome had increased $T_{MRCA}$ even though only
568 1.9% of RAD loci (1129 of 57,992) were classified as divergent based on marine-
569 freshwater reciprocal monophyly. When we removed these loci, along with loci with
570 elevated marine-freshwater $F_{ST}$ ($F_{ST} > 0.5$), many of the regions in which they resided
571 were still $T_{MRCA}$ outliers. It is possible that the remainder of this old variation is neutral
572 with respect to fitness. However, we identified divergence outliers based on only a
573 single axis of divergence: the marine-freshwater axis. Throughout the entire species
574 range, populations are locally experiencing multiple axes of divergence, including lake-
575 stream and benthic-limnetic axes (McKinnon & Rundle 2002), that often shares a
576 common genomic architecture (Deagle *et al.* 2012; Roesti *et al.* 2015). Our data may
577 indicate underlying similarities in selection regimes. Alternatively, this co-localized
578 ancient variation may represent the accumulation of adaptive divergence along multiple
579 axes in the same genomic regions, whether or not the underlying adaptive variants are
580 the same. Aspects of the genomic architecture, such as gene density or local
581 recombination rates, may in part govern where in the genome adaptive divergence can
582 occur (Roesti *et al.* 2013; Aeschbacher *et al.* 2017; Samuk *et al.* 2017). Multiple axes of

583 divergence may therefore act synergistically to maintain genomic variation across the
584 stickleback metapopulation.
585 Nevertheless, much of the ancient variation we observe may in fact itself be
586 neutral, having been maintained by close linkage to loci under divergent selection
587 between the marine and freshwater habitats (Charlesworth *et al.* 1997). Indeed, the
588 broadest peaks of $T_{MRCA}$ we observe occur in genomic regions with low rates of
589 recombination (Roesti *et al.* 2013; Glazer *et al.* 2015) in other stickleback populations,
590 which would extend the size of the linked region affected by divergent selection. On
591 ecological timescales, low recombination rates in stickleback are thought to promote
592 divergence by making locally adapted genomic regions resistant to gene flow (Roesti *et*
593 *al.* 2013). Our results potentially extend the inferred impact of recombination rate
594 variation on genomic variation to timescales that are 1000-fold longer, maintaining both
595 multimillion-year-old adaptive variation and large stores of linked genetic variation.
596 Future modeling efforts will be needed to explore the range of population genetic
597 parameter values (e.g. selection coefficients, migration rates, and recombination rates)
598 required to produce the extent of divergence we see here.
599 Lastly, our findings demonstrate that known chromosomal inversions maintain
600 globally distributed, multilocus haplotypes. The three chromosomal inversions known to
601 be associated with marine-freshwater divergence (Jones *et al.* 2012; Roesti *et al.* 2015)
602 (*inv1*, *inv11*, and *inv21*; yellow bars in Fig. 4C) all showed sharp spikes in $T_{MRCA}$.
603 Genomic signatures of these inversions are distributed throughout the species range,
604 including coastal marine-freshwater population pairs in the Pacific and Atlantic basins
605 (Jones *et al.* 2012) and inland lake-stream pairs in Switzerland (Roesti *et al.* 2015).
606 Despite our limited geographic sampling, our finding that all three of these inversions
607 are over six million years old is further evidence of single, ancient origins of each,
608 followed by their spread across the species range. Each inversion contained a high
609 density of divergent RAD loci (*inv1*: 64% of loci divergent; *inv11*: 60%; *inv21*: 71%) but
610 we also identified regions within these inversions in which haplotypes from marine or
611 freshwater habitats, or both, were not monophyletic. *inv1* and *inv11* both contained two
612 regions separated by loci in which neither habitat type was monophyletic; *inv21*, the

613 largest of the three, contained ten such regions. Additionally, $T_{MRCA}$ and $F_{ST}$ decreased
614 sharply to background levels outside of the inversions, demonstrating the potential for
615 gene flow and recombination to homogenize variation in these regions. We interpret this
616 as evidence that these inversions help maintain linkage disequilibrium among multiple
617 divergently adaptive variants in regions susceptible to homogenization (Kirkpatrick &
618 Barton 2006; Guerrero *et al.* 2012). The presence of these inversions in addition to
619 divergence in regions of generally low recombination (Glazer *et al.* 2015), therefore,
620 further supports the hypothesis that the recombinational landscape can influence where
621 in the genome adaptive divergence can occur (Roesti *et al.* 2013; Samuk *et al.* 2017)
622 and emphasizes the degree to which gene flow among divergently adapted stickleback
623 populations has impacted global genomic diversity.
624
625 *Conclusions*
626 Selection operating on two very different timescales — the ecological and the
627 geological — has shaped genomic patterns of SGV in the threespine stickleback. On
628 ecological timescales, selection drives phenotypic divergence in decades or millennia
629 by sorting SGV across geography and throughout the genome (Hendry *et al.* 2002;
630 Hohenlohe *et al.* 2010; Lescak *et al.* 2015; Roesti *et al.* 2015). Our findings show that
631 persistent ecological diversity and continual local adaptation of stickleback has set the
632 stage for long-term divergent selection and for the accumulation and maintenance of
633 adaptive variation over geological timescales. Some of the genetic variants fueling
634 contemporary, rapid adaptation may even have been present – and under selection –
635 since before the threespine-ninespine stickleback lineages split. The genomic
636 architecture of ecological adaptation in one focal population is therefore the product of
637 millions of years of evolution taking place in multiple populations, many of which are
638 now extinct. These findings underscore the need to understand macroevolutionary
639 patterns when studying microevolutionary processes, and vice versa.
640
641
642

ACKNOWLEDGEMENTS

We thank P. Phillips, M. Streisfeld, J. Postlethwait, K. Sterner for valuable input and lively discussion throughout this project. We also thank K. Alligood, E. Beck, S. Bassham, M. Chase, M. Currey, M. Hahn, L. Fishman, P. Ralph, C. Small, S. Stankowski, J. Willis, two anonymous reviewers, and members of the Cresko Lab and the Institute of Ecology and Evolution for advice and comments on previous versions of this manuscript. J. Postlethwait graciously donated ninespine stickleback tissue, collected under award XXXXXXXX. We acknowledge National Science Foundation awards NSF DEB 1501423 (WAC and TCN), NSF DEB 0949053 (WAC), and National Institutes of Health award NIH T32GM007413 (TCN).

AUTHOR CONTRIBUTIONS

TCN and WAC conceived of the project and designed sampling, sequencing, and analysis. TCN prepared sequencing libraries, wrote software, and performed data analysis. TCN and WAC wrote the paper.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

663 LITERATURE CITED
664
665
666 Aeschbacher, S., Selby, J.P., Willis, J.H. & Coop, G. (2017). Population-genomic inference of
667 the strength and timing of selection against gene flow. *P Natl Acad Sci USA*, 114, 7061-
668 7066.
669 Aldenhoven, J.T., Miller, M.A., Corneli, P.S. & Shapiro, M.D. (2010). Phylogeography of
670 ninespine sticklebacks (Pungitius pungitius) in North America: glacial refugia and the
671 origins of adaptive traits. *Mol Ecol*, 19, 4061-4076.
672 Arnegard, M.E., McGee, M.D., Matthews, B., Marchinko, K.B., Conte, G.L., Kabir, S. *et al.*
673 (2014). Genetics of ecological divergence during speciation. *Nature*, 511, 307-311.
674 Arnold, B., Corbett-Detig, R.B., Hartl, D. & Bomblies, K. (2013). RADseq underestimates
675 diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol*
676 *Ecol*, 22, 3179-3190.
677 Aronesty, E. (2011). ea-utils: Command-line tools for processing biological sequencing data.
678 Baird, N.a., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.a. *et al.* (2008a).
679 Rapid SNP discovery and genetic mapping using sequenced RAD markers. *Plos One*, 3,
680 1-7.
681 Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A. *et al.* (2008b).
682 Rapid SNP discovery and genetic mapping using sequenced RAD markers. *Plos One*, 3,
683 e3376.
684 Barrett, R.D., Rogers, S.M. & Schluter, D. (2008). Natural selection on a major armor gene in
685 threespine stickleback. *Science*, 322, 255-257.
686 Barrett, R.D.H. & Schluter, D. (2008). Adaptation from standing genetic variation. *TREE*, 23, 38-
687 44.
688 Bell, M.A., Baumgartner, J.V. & Olson, E.C. (1985). Patterns of Temporal Change in Single
689 Morphological Characters of a Miocene Stickleback Fish. *Paleobiology*, 11, 258-271.
690 Bell, M.A. & Foster, S.A. (1994a). Evolutionary inference: the value of viewing evolution through
691 stickleback-tinted glasses. In: *The Evolutionary Biology of the Threespine Stickleback*
692 (eds. Bell, MA & Foster, SA). Oxford University Press New York, pp. 472-486.
693 Bell, M.A. & Foster, S.A. (1994b). Introduction to the evolutionary biology of the threespine
694 stickleback. In: *The Evolutionary Biology of the Threespine Stickleback* (eds. Bell, MA &
695 Foster, SA). Oxford University Press New York, pp. 1-27.
696 Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A. & Cresko, W.A. (2013a). Stacks: an
697 analysis tool set for population genomics. *Mol Ecol*, 22, 3124-3140.
698 Catchen, J., Hohenlohe, P.a., Bassham, S., Amores, A. & Cresko, W.a. (2013b). Stacks: An
699 analysis tool set for population genomics. *Mol Ecol*, 22, 3124-3140.
700 Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W. & Postlethwait, J.H. (2011). Stacks:
701 building and genotyping Loci de novo from short-read sequences. *G3-Genes Genom*
702 *Genet*, 1, 171-182.
703 Charlesworth, B., Morgan, M.T. & Charlesworth, D. (1993). The Effect of Deleterious Mutations
704 on Neutral Molecular Variation. *Genetics*, 134, 1289-1303.
705 Charlesworth, B., Nordborg, M. & Charlesworth, D. (1997). The effects of local selection,
706 balanced polymorphism and background selection on equilibrium patterns of genetic
707 diversity in subdivided populations. *Genet Res*, 70, 155-174.
708 Colosimo, P.F., Hosemann, K.E., Balabhadra, S., Villarreal, G., Jr., Dickson, M., Grimwood, J.
709 *et al.* (2005). Widespread parallel evolution in sticklebacks by repeated fixation of
710 Ectodysplasin alleles. *Science*, 307, 1928-1933.

711 Colosimo, P.F., Peichel, C.L., Nereng, K., Blackman, B.K., Shapiro, M.D., Schluter, D. *et al.*
712 (2004). The genetic architecture of parallel armor plate reduction in threespine
713 sticklebacks. *PLoS Biol*, 2, 635-641.
714 Cresko, W.A., Amores, A., Wilson, C., Murphy, J., Currey, M., Phillips, P. *et al.* (2004a). Parallel
715 genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback
716 populations. *Proc Natl Acad Sci U S A*, 101, 6050-6055.
717 Cresko, W.a., Amores, A., Wilson, C., Murphy, J., Currey, M., Phillips, P. *et al.* (2004b). Parallel
718 genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback
719 populations. *P Natl Acad Sci USA*, 101, 6050-6055.
720 Cruickshank, T.E. & Hahn, M.W. (2014). Reanalysis suggests that genomic islands of
721 speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol*, 23, 3133-3157.
722 Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. & Blaxter, M.L. (2011).
723 Genome-wide genetic marker discovery and genotyping using next-generation
724 sequencing. *Nat Rev Genet*, 12, 499-510.
725 Deagle, B.E., Jones, F.C., Chan, Y.F., Absher, D.M., Kingsley, D.M. & Reimchen, T.E. (2012).
726 Population genomics of parallel phenotypic evolution in stickleback across stream-lake
727 ecological transitions. *Proc Roy Soc B-Biol Sci*, 279, 1277-1286.
728 Domingues, V.S., Poh, Y.P., Peterson, B.K., Pennings, P.S., Jensen, J.D. & Hoekstra, H.E.
729 (2012). Evidence of adaptation from ancestral variation in young populations of beach
730 mice. *Evolution*, 66, 3209-3223.
731 Drummond, A.J. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling
732 trees. *BMC Evol Biol*, 7, 214.
733 Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. (2012). Bayesian phylogenetics with
734 BEAUti and the BEAST 1.7. *Mol Biol Evol*, 29, 1969-1973.
735 Fontaine, M.C., Pease, J.B., Steele, A., Waterhouse, R.M., Neafsey, D.E., Sharakhov, I.V. *et al.*
736 (2015). Mosquito genomics. Extensive introgression in a malaria vector species complex
737 revealed by phylogenomics. *Science*, 347, 1258524.
738 Francis, R.C., Baumgartner, J.V., Havens, A.C. & Bell, M.A. (1986). Historical and ecological
739 sources of variation among lake populations of threespine sticklebacks, Gasterosteus
740 aculeatus, near Cook Inlet, Alaska. *Canad J Zool*, 64, 2257-2265.
741 Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhue, C., Pudlo, P. *et al.* (2013). The
742 effect of RAD allele dropout on the estimation of genetic variation within and between
743 populations. *Mol Ecol*, 22, 3165-3178.
744 Glazer, A.M., Killingbeck, E.E., Mitros, T., Rokhsar, D.S. & Miller, C.T. (2015). Genome
745 Assembly Improvement and Mapping Convergently Evolved Skeletal Traits in
746 Sticklebacks with Genotyping-by-Sequencing. *G3-Genes Genom Genet*, 5, 1463-1472.
747 Grant, P.R. & Grant, B.R. (2002). Unpredictable evolution in a 30-year study of Darwin's finches.
748 *Science*, 296, 707-711.
749 Guerrero, R.F., Rousset, F. & Kirkpatrick, M. (2012). Coalescent patterns for chromosomal
750 inversions in divergent populations. *Philos Trans Roy Soc B*, 367, 430-438.
751 Hendry, A.P., Taylor, E.B. & McPhail, J.D. (2002). Adaptive divergence and the balance
752 between selection and gene flow: lake and stream stickleback in the Misty system.
753 *Evolution*, 56, 1199-1216.
754 Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A. & Cresko, W.A. (2010).
755 Population genomics of parallel adaptation in threespine stickleback using sequenced
756 RAD tags. *Plos Genet*, 6, e1000862.
757 Hudson, R.R., Slatkin, M. & Maddison, W.P. (1992). Estimation of Levels of Gene Flow from
758 DNA-Sequence Data. *Genetics*, 132, 583-589.

759 Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B.M., Vinckenbosch, N. *et al.* (2014).
760 Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*,
761 512, 194-197.
762 Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J. *et al.* (2012). The
763 genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, 484, 55-61.
764 Kimmel, C.B., Ullmann, B., Walker, C., Wilson, C., Currey, M., Phillips, P.C. *et al.* (2005).
765 Evolution and development of facial bone morphology in threespine sticklebacks. *Proc*
766 *Natl Acad Sci U S A*, 102, 5791-5796.
767 Kingman, J.F.C. (1982a). The coalescent. *Stoch Proc Appl*, 13, 235-248.
768 Kingman, J.F.C. (1982b). On the genealogy of large populations. *J Appl Probab*, 19, 27-43.
769 Kirkpatrick, M. & Barton, N. (2006). Chromosome inversions, local adaptation and speciation.
770 *Genetics*, 173, 419-434.
771 Lenormand, T. (2002). Gene flow and the limits to natural selection. *TREE,* 17, 183-189.
772 Lescak, E.A., Bassham, S.L., Catchen, J., Gelmond, O., Sherbick, M.L., von Hippel, F.A. *et al.*
773 (2015). Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proc Natl*
774 *Acad Sci U S A*, 112, E7204-7212.
775 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. *et al.* (2009). The Sequence
776 alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, 2078-2079.
777 Linnen, C.R., Kingsley, E.P., Jensen, J.D. & Hoekstra, H.E. (2009). On the origin and spread of
778 an adaptive allele in deer mice. *Science*, 325, 1095-1098.
779 McCairns, R.J.S. & Bernatchez, L. (2012). Plasticity and heritability of morphological variation
780 within and between parapatric stickleback demes. *J Evol Biol*, 25, 1097-1112.
781 McGuigan, K., Nishimura, N., Currey, M., Hurwit, D. & Cresko, W.A. (2011). Cryptic genetic
782 variation and body size evolution in threespine stickleback. *Evolution*, 65, 1203-1211.
783 McKinnon, J.S. & Rundle, H.D. (2002). Speciation in nature:the threespine stickleback model
784 systems. *TREE*, 17, 480-488.
785 Miller, C.T., Glazer, A.M., Summers, B.R., Blackman, B.K., Norman, A.R., Shapiro, M.D. *et al.*
786 (2014). Modular Skeletal Evolution in Sticklebacks Is Controlled by Additive and
787 Clustered Quantitative Trait Loci. *Genetics*, 197, 405-420.
788 Nei, M. (1987). *Molecular evolutionary genetics*. Columbia university press.
789 Ormond, L., Foll, M., Ewing, G.B., Pfeifer, S.P. & Jensen, J.D. (2016). Inferring the age of a
790 fixed beneficial allele. *Mol Ecol*, 25, 157-169.
791 Orr, H.A. (2005). The genetic theory of adaptation: a brief history. *Nat Rev Genet*, 6, 119-127.
792 Otto, S.P. & Bourguet, D. (1999). Balanced polymorphisms and the evolution of dominance. *Am*
793 *Nat*, 153, 561-574.
794 Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular
795 approach. *Bioinformatics*, 26, 419-420.
796 Paradis, E., Claude, J. & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in
797 R language. *Bioinformatics*, 20, 289-290.
798 Pease, J.B., Haak, D.C., Hahn, M.W. & Moyle, L.C. (2016). Phylogenomics reveals three
799 sources of adaptive variation during a rapid radiation. *PLoS Biol*, 14, e1002379.
800 Peter, B.M., Huerta-Sanchez, E. & Nielsen, R. (2012). Distinguishing between Selective
801 Sweeps from Standing Variation and from a De Novo Mutation. *Plos Genet*, 8.
802 Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. & Hoekstra, H.E. (2012). Double Digest
803 RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model
804 and Non-Model Species. *Plos One*, 7.
805 Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S.E. & Lercher, M.J. (2014). PopGenome: An
806 Efficient Swiss Army Knife for Population Genomic Analyses in R. *Mol Biol Evol*, 31,
807 1929-1936.

808 Phillips, P.C. (2008). Epistasis--the essential role of gene interactions in the structure and
809 evolution of genetic systems. *Nat Rev Genet*, 9, 855-867.
810 Popescu, A.A., Huber, K.T. & Paradis, E. (2012). ape 3.0: New tools for distance-based
811 phylogenetics and evolutionary analysis in R. *Bioinformatics*, 28, 1536-1537.
812 Reimchen, T.E. (1994). Predators and morphological evolution in threespine stickleback. In: *The*
813 *Evolutionary Biology of the Threespine Stickleback* (eds. Bell, MA & Foster, SA). Oxford
814 University Press New York, pp. 240-276.
815 Roesti, M., Gavrilets, S., Hendry, A.P., Salzburger, W. & Berner, D. (2014). The genomic
816 signature of parallel adaptation from shared genetic variation. *Mol Ecol*, 23, 3944-3956.
817 Roesti, M., Kueng, B., Moser, D. & Berner, D. (2015). The genomics of ecological vicariance in
818 threespine stickleback fish. *Nat Commun*, 6, 8767.
819 Roesti, M., Moser, D. & Berner, D. (2013). Recombination in the threespine stickleback genome
820 - Patterns and consequences. *Mol Ecol*, 22, 3014-3027.
821 Samuk, K., Owens, G.L., Delmore, K.E., Miller, S., Rennison, D.J. & Schluter, D. (2017). Gene
822 flow and selection interact to promote adaptive divergence in regions of low
823 recombination. *Mol Ecol*, 26, 4378-4390.
824 Schlotterer, C., Tobler, R., Kofler, R. & Nolte, V. (2014). Sequencing pools of individuals-mining
825 genome-wide polymorphism data without big funding. *Nat Rev Genet*, 15, 749-763.
826 Schluter, D. & Conte, G.L. (2009). Genetics and ecological speciation. *Proc Natl Acad Sci U S*
827 *A*, 106 Suppl 1, 9955-9962.
828 Schrider, D.R. & Kern, A.D. (2017). Soft Sweeps Are the Dominant Mode of Adaptation in the
829 Human Genome. *Mol Biol Evol*, 34, 1863-1877.
830 Stankowski, S. & Streisfeld, M.A. (2015). Introgressive hybridization facilitates adaptive
831 divergence in a recent radiation of monkeyflowers. *Proc Roy Soc B-Biol Sci*, 282,
832 20151666.
833 Stephens, M. & Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype
834 inference and missing-data imputation. *Am J Hum Genet*, 76, 449-462.
835 Stephens, M., Smith, N.J. & Donnelly, P. (2001). A new statistical method for haplotype
836 reconstruction from population data. *Am J Hum Genet*, 68, 978-989.
837 Stuart, Y.E., Veen, T., Weber, J.N., Hanson, D., Ravinet, M., Lohman, B.K. *et al.* (2017).
838 Contrasting effects of environment and genetics generate a continuum of parallel
839 evolution. *Nat Ecol Evol*, 1, 0158.
840 Tajima, F. (1983). Evolutionary Relationship of DNA-Sequences in Finite Populations. *Genetics*,
841 105, 437-460.
842 R Core Team. (2016). R: A language and environment for statistical computing. R Foundation
843 for Statistical Computing Vienna, Austria.
844 Wakeley, J. (2009). *Coalescent Theory: An Introduction*. Harvard University.
845 Willing, E.M., Dreyer, C. & van Oosterhout, C. (2012). Estimates of Genetic Differentiation
846 Measured by F-ST Do Not Necessarily Require Large Sample Sizes When Using Many
847 SNP Markers. *Plos One*, 7.
848 Wright, K.M., Lloyd, D., Lowry, D.B., Macnair, M.R. & Willis, J.H. (2013). Indirect Evolution of
849 Hybrid Lethality Due to Linkage with Selected Locus in Mimulus guttatus. *PLoS Biol*, 11.
850 Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution.
851 *Proceedings of the Sixth International Congress on Genetics*, 1, 356-366.
852 Wund, M.A., Singh, O.D., Geiselman, A. & Bell, M.A. (2016). Morphological evolution of an
853 anadromous threespine stickleback population within one generation after reintroduction
854 to Cheney Lake, Alaska. *Evol Ecol Res*, 17, 203-224.
855
856

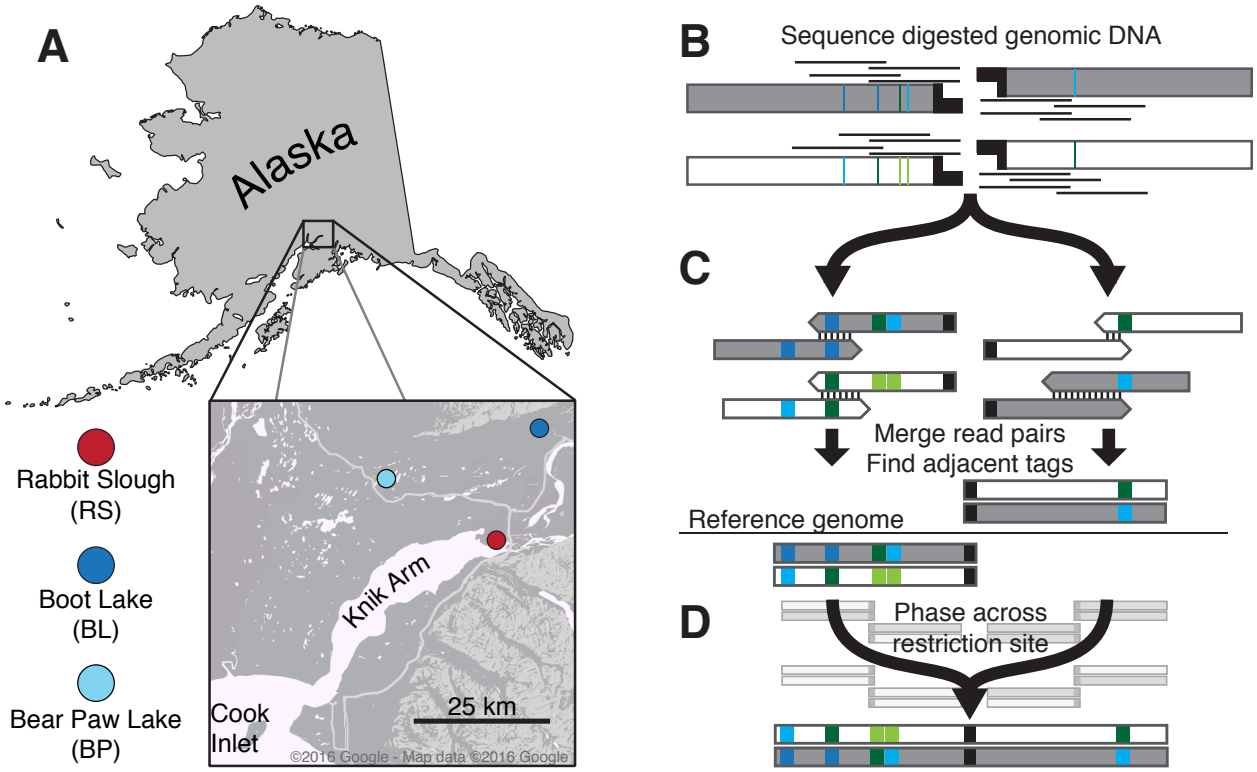



**Figure 1. Stickleback sampling and RAD sequencing to measure haplotype variation.** A) Threespine stickleback sampling locations in this study. Colors represent habitat type: red: marine; blue: freshwater. B-D: We modified the original RAD-seq protocol to generate local haplotypes. Colored bars represent polymorphic sites. For a detailed description of haplotype construction, see *Methods*. B) Overlapping paired-end reads are anchored to *PstI* restriction sites. C) Paired reads mapping to each halfsite are merged into contigs. Contigs mapping to the same restriction site are identified by alignment to the reference genome. D) Sequences from each half of a restriction site are phased to generate a single RAD locus. RAD tags in the background represent multiple genotypes used in phasing.

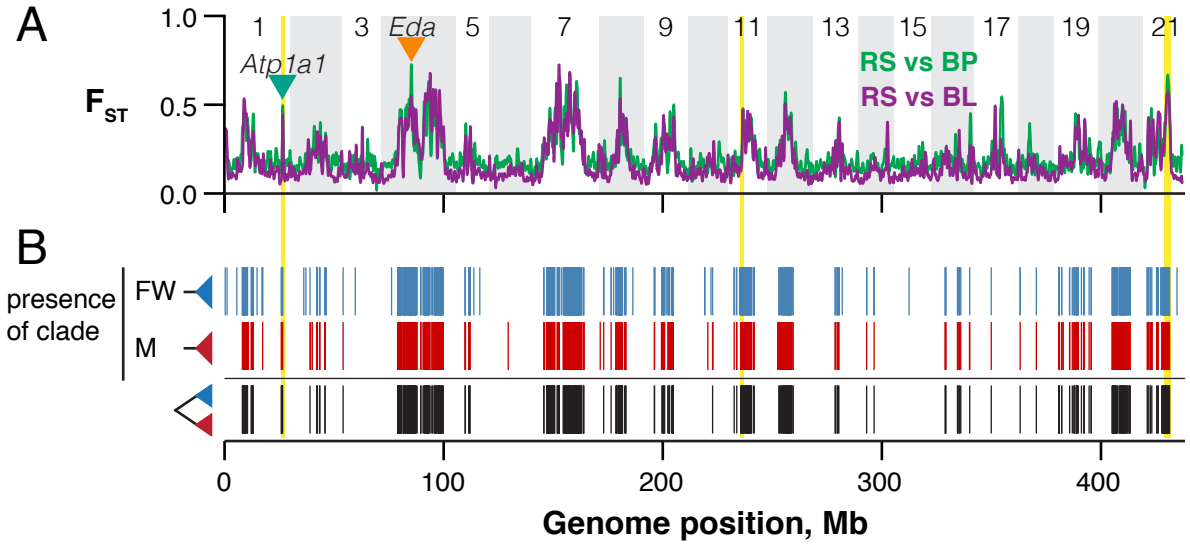



**Figure 2. The genealogical structure of parallel genomic divergence.** A) Genome-wide $F_{ST}$ for both marine-freshwater comparisons was kernel-smoothed using a normally distributed kernel with a window size of 500 kb. Inverted triangles indicate the locations of two genes known to show extensive marine-freshwater haplotype divergence, *Eda* and *Atp1a1*. Three chromosomal inversions are highlighted in yellow. B) Lineage sorting patterns were identified from maximum clade credibility trees for each RAD locus. Blue bars: haplotypes from both freshwater populations form a single monophyletic group; red: haplotypes from the marine population form a monophyletic group; black: A RAD locus is structured into reciprocally monophyletic marine and freshwater haplogroups.

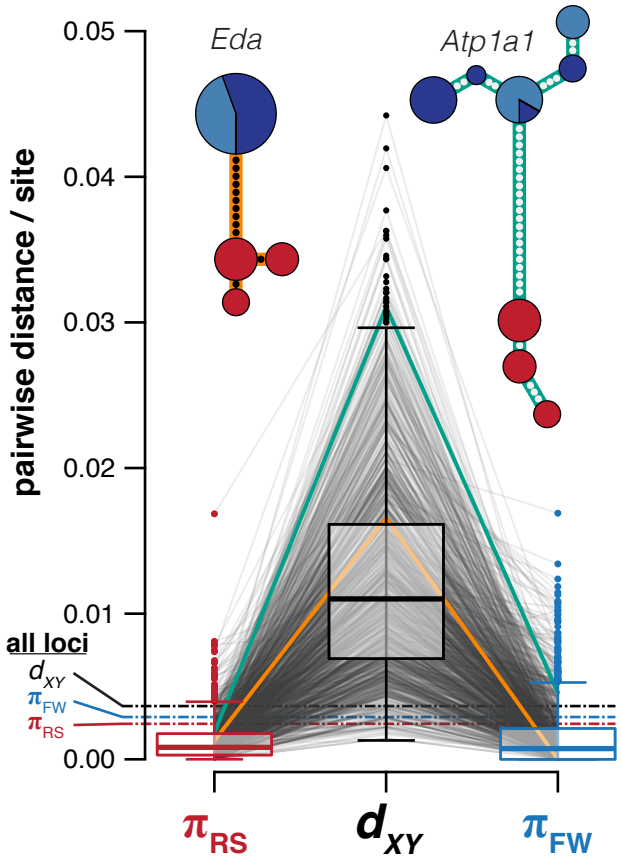



**Figure 3. Extensive sequence divergence between marine and freshwater haplogroups accompanies reciprocal monophyly.** For each reciprocally monophyletic RAD locus, we calculated sequence variation ($\pi$) within and sequence divergence between habitat types ($d_{XY}$). Each RAD locus is shown as a pair of lines connecting estimates of π and $d_{XY}$. Boxplots show distributions across all reciprocally monophyletic RAD loci: Boxes are upper and lower quartiles, including the median; whiskers extend to 1.5x interquartile range. Dashed lines are the genome-wide medians. Single RAD loci from within the transcribed regions of *Eda* and *Atp1a1* are shown as gold and green lines, respectively, and presented as haplotype networks. Dots represent mutational steps. Circle sizes indicate the number of haplotypes and colors indicate population of origin as in Figure 1. Each network = 29 haplotypes.

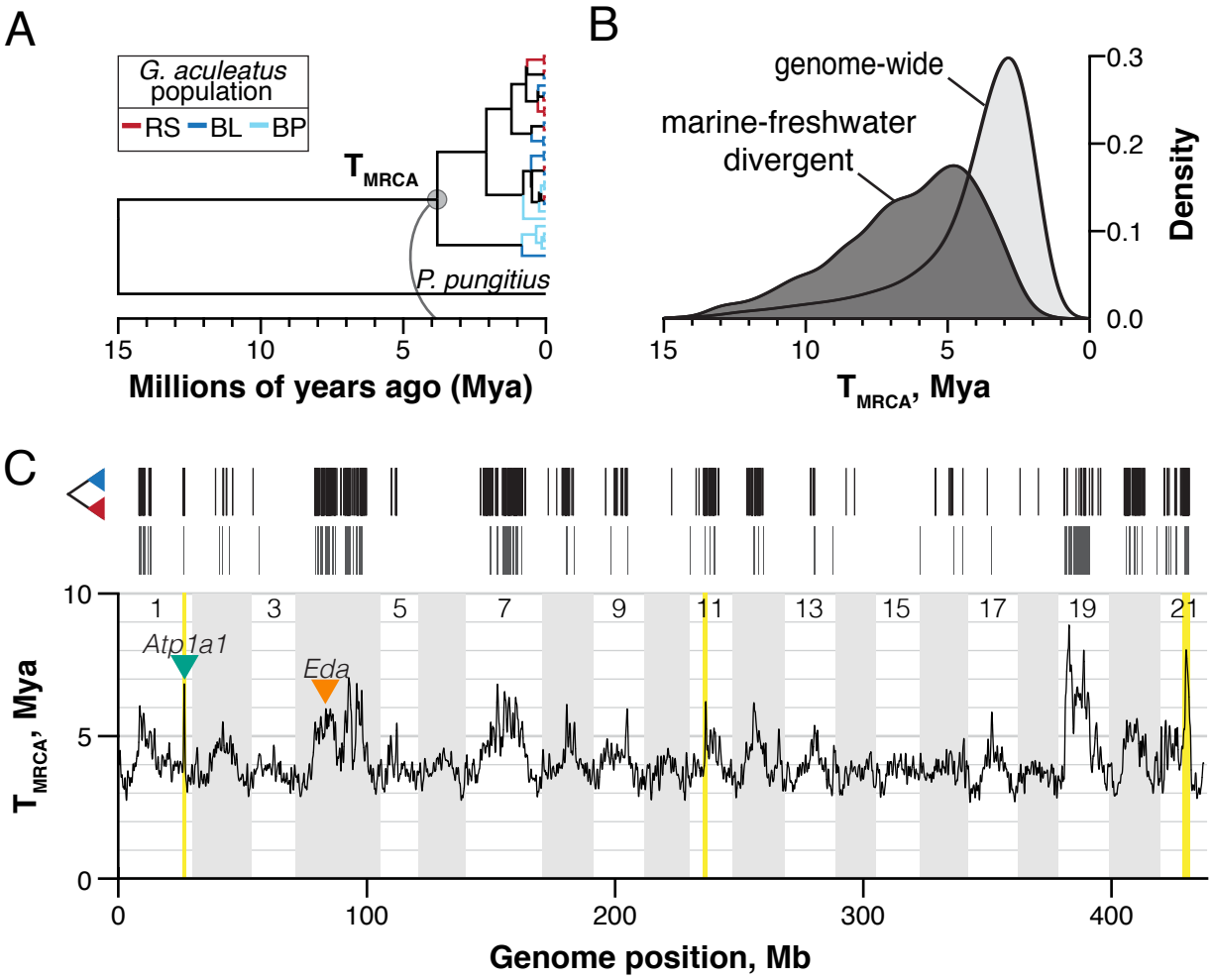



**Figure 4. Marine-freshwater divergence has evolved over millions of years, affecting large genomic regions.** We performed Bayesian estimation of the time to the most recent common ancestor ($T_{MRCA}$) of alleles at threespine stickleback RAD loci. We calibrated coalescence times within threespine stickleback by including a *de novo* genome assembly from the ninespine stickleback (*Pungitius pungitius*) and setting threespine-ninespine divergence at 15 million years ago. A) Maximum clade credibility RAD gene tree representative of the genome-wide average $T_{MRCA}$. Branches within threespine are colored by population of origin. B) Kernel-smoothed densities of $T_{MRCA}$ distributions for all RAD loci containing a monophyletic group of threespine stickleback alleles (light gray) and those structured into reciprocally monophyletic marine and freshwater haplogroups. C) The genomic distribution of reciprocally monophyletic RAD loci (black, as in Figure 2) is associated with increased $T_{MRCA}$ at a genomic scale. $T_{MRCA}$ outlier windows (those exceeding 99.9% of permuted genomic windows) are shown as gray bars. Genome-wide $T_{MRCA}$ was kernel-smoothed using a normally distributed kernel with a window size of 500 kb. Inverted triangles indicate the locations of *Eda* and *Atp1a1*. Three chromosomal inversions are highlighted in yellow.

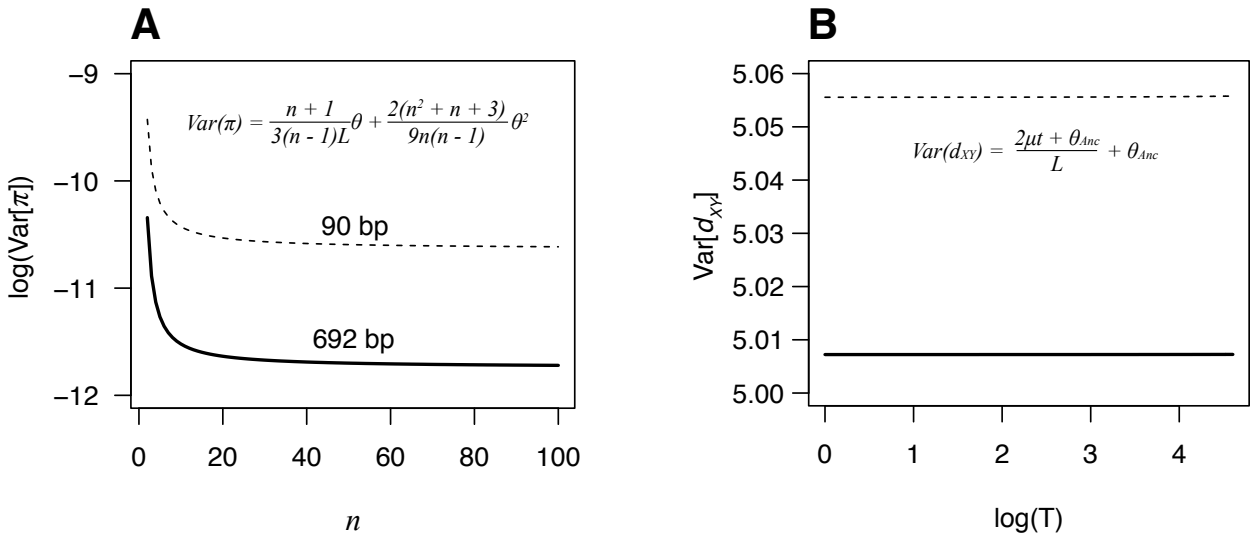



**Figure S1. Longer sequences reduce variance in estimates of sequence diversity and divergence.** A: Variance in π as a function of the number of chromosomes sampled, using sequence lengths typical of RAD-seq experiments (90 bp) and those in this study (692 bp). Variance was calculated using equation 10.9 in Nei (1987). Right: Variance in $d_{XY}$ (using the equation in box 1 in Cruickshank and Hahn (2014)) as a function of (log-scaled) divergence time of two populations. The change in variance as function of divergence time is dwarfed by the difference in variance obtained with different sequence lengths. In both panels, $n$ = sequences sampled; $L$ = length of sequence sampled; $\theta = 4N\mu$; $\theta_{Anc} = \theta$ in the population ancestral to those sampled; $\mu$ = mutation rate per nucleotide; $t$ = time since population split.

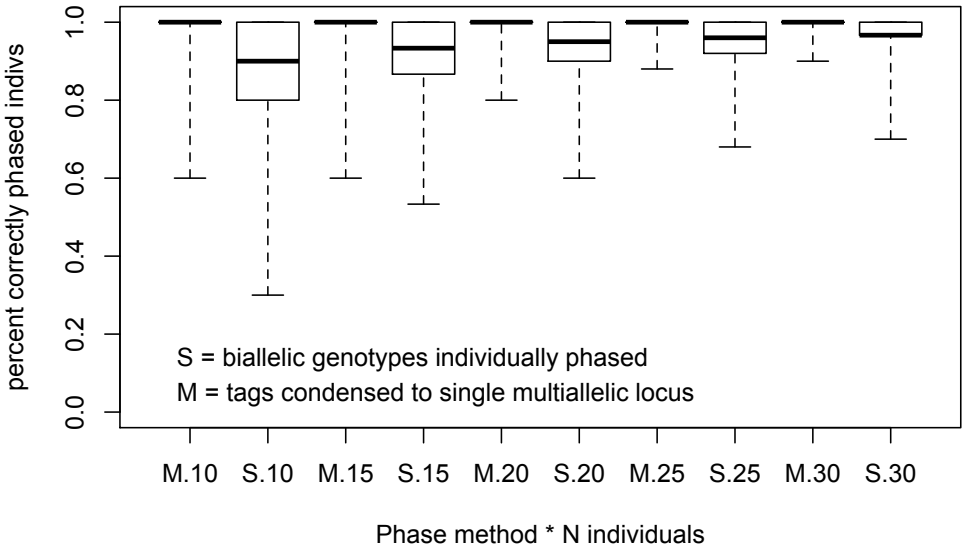
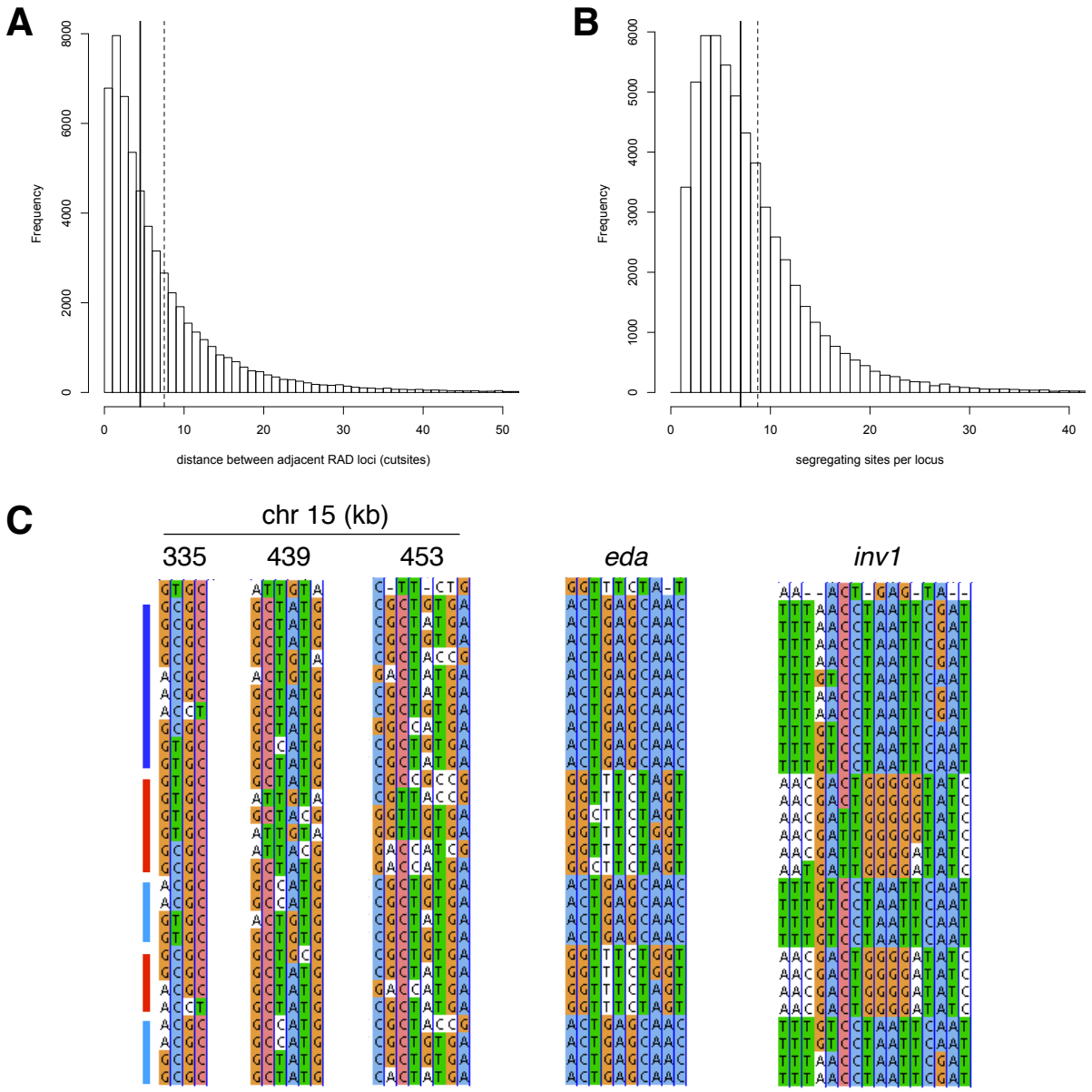
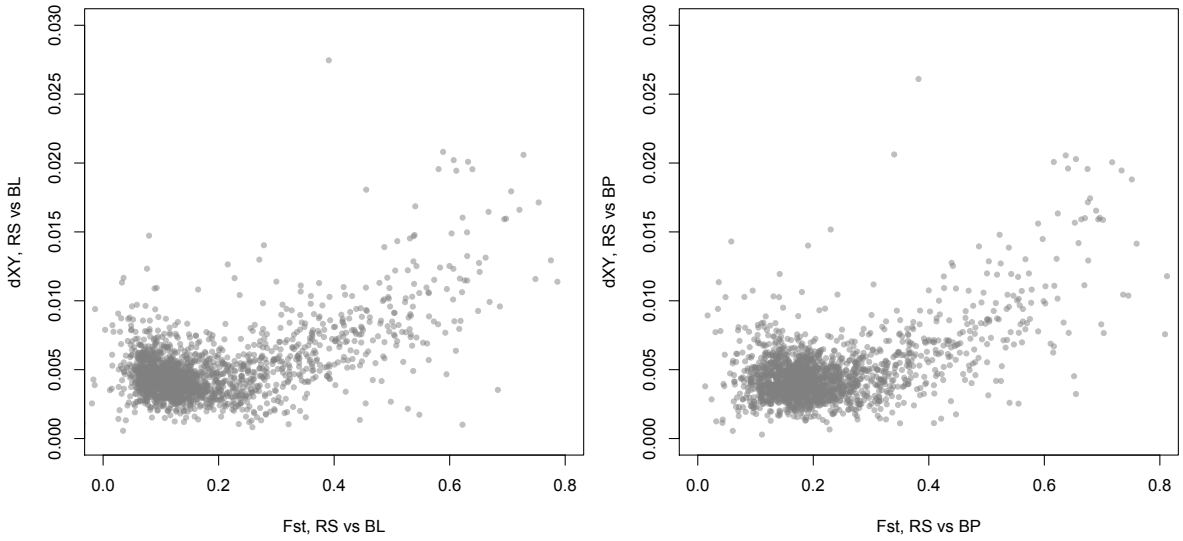


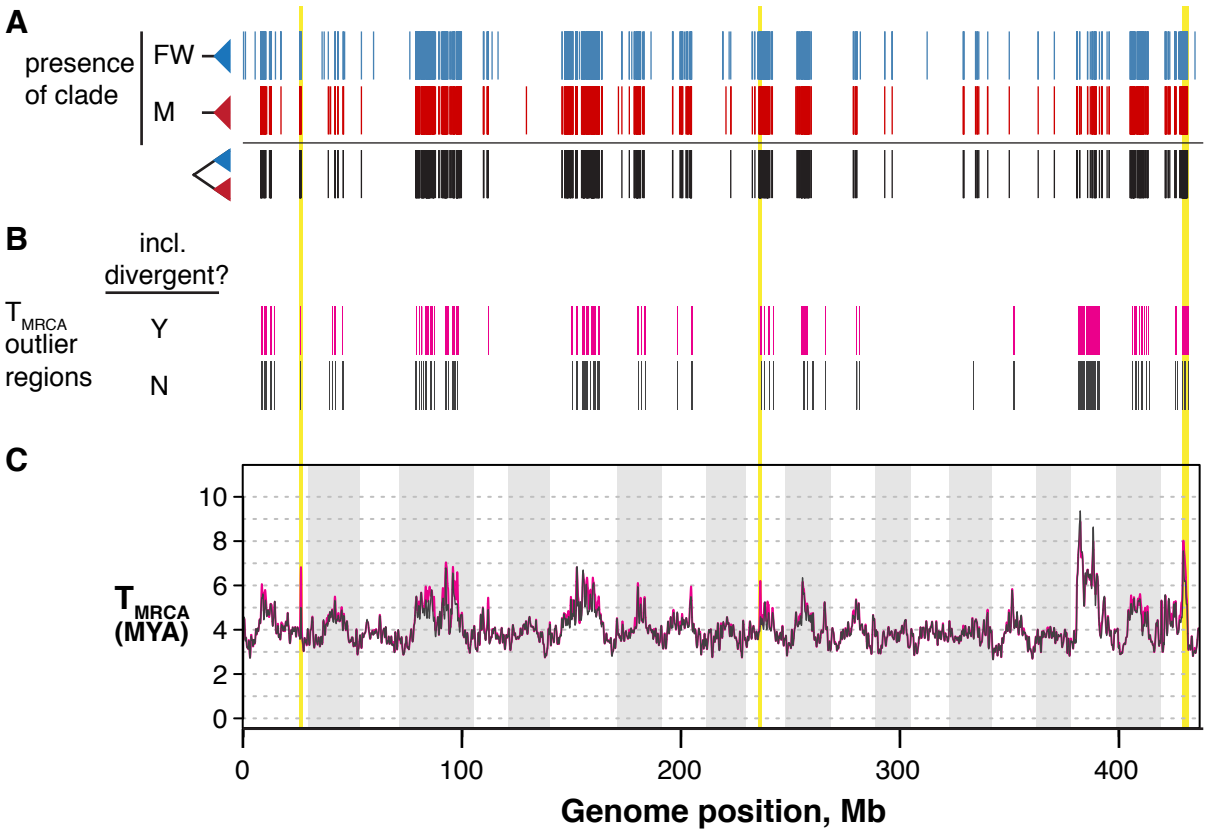

**Figure S2. Accurate phasing of RAD loci even at low population-level sampling.** Neutrally evolving, non-recombining RAD loci were simulated with ms and seq-gen to generate alignments with of 20 to 60 haplotypes (10-30 diploid individuals) and four to 30 segregating sites. Simulated haplotypes were then 'cut' at their midpoints and phased either by inputting all variable sites individually (biallelic 'SNPs', S) or by inputting the haplotype information on either site of the cut as multiallelic loci (M). Boxes represent interquartile range (IQR). Bold lines are medians. Whiskers extend to minimum and maximum values. Even with smaller sample sizes (10-15 individuals), over 75% of phasing attempts resulted in 100% phasing accuracy.




**Supplementary Figure S3. RAD-seq effectively samples genome-wide sequence diversity.** Histograms of (A) the distance between adjacent RAD loci (calculated as the distance between the centers of each restriction site) and (B) the number of variable sites per locus show that most RAD loci were within 4 kb of their nearest neighbor and contained ≥ 7 variable sites. Means for each metric are shown as dashed vertical lines. Medians are solid lines. Each histogram is truncated to highlight the bulk of the distribution. Maximum values: distance = 455 kb; variable sites = 155. C: Example haplotypes from five RAD loci in non-divergent (chromosome 15) and divergent (*eda*, *inv1*) genomic regions. Chromosome 15 loci are labeled by their genomic position. The *eda* RAD locus is within the transcribed region of *eda* and *inv1* is within the breakpoints of the chromosome 1 inversion. Colored bars identify population/ecotype of origin. Red: RS (marine); dark blue: BL (freshwater); light blue: BP (freshwater). Alignments visualized in JalView v1.0 (Waterhouse, *et al*, 2009). Only sites that are variable within threespine stickleback are shown.




**Supplementary Figure S4. Relative ($F_{ST}$) and absolute ($d_{XY}$) sequence divergence are positively correlated genome-wide in two instances of marine-freshwater divergence.** Points are 250 kb non-overlapping genomic windows. Left panel compares the marine Rabbit Slough population (RS) to the freshwater Boot Lake population (BL) (type-II linear model: $r^2 = 0.314$, permuted p-value [reduced major axis] = 0.01). Right panel compares RS to the freshwater Bear Paw Lake population (BL) (type-II linear model: $r^2 = 0.311$, permuted p-value [reduced major axis] = 0.01).




**Supplementary Figure S5. $T_{MRCA}$ outlier regions remain outliers after removing highly differentiated RAD loci.** Panel A is taken from Fig. 2 and shows the genomic distribution of reciprocally monophyletic ("divergent"; black bars) RAD loci. Panel B shows the distributions of $T_{MRCA}$ outlier regions (increased $T_{MRCA}$) including all RAD loci (magenta boxes, "Y"). Below are the $T_{MRCA}$ outlier regions after removing divergent loci and any RAD locus with a marine-freshwater (RS vs. [BL+BP]) $F_{ST} > 0.5$, which is approximately the top 7% of the $F_{ST}$ distribution. Panel C: Genome scans of $T_{MRCA}$ using all RAD loci (magenta) and excluding marine-freshwater outliers (gray).

**Supplementary table 1.** Sequencing summary for threespine stickleback samples

| **Sample** | population | raw reads | filtered reads | merged pairs | mean coverage per locus |
|---|---|---|---|---|---|
| 1827.05 | Rabbit Slough | 10167407 | 10031967 | 7269377 | 12X |
| 1827.06 | Rabbit Slough | 10265078 | 10172621 | 7591801 | 13X |
| 1827.07 | Rabbit Slough | 9175983 | 9040625 | 6771332 | 11X |
| 1827.08 | Rabbit Slough | 7896938 | 7814081 | 5879351 | 10X |
| 1827.09 | Rabbit Slough | 8773502 | 8668261 | 6405777 | 11X |
| 2827.01 | Boot Lake | 8917575 | 8810382 | 6373001 | 11X |
| 2827.07 | Boot Lake | 10064876 | 9917732 | 7255989 | 13X |
| 2827.13 | Boot Lake | 9099831 | 9002717 | 6528704 | 12X |
| 2827.19 | Boot Lake | 11021084 | 10792092 | 7911026 | 14X |
| 2827.25 | Boot Lake | 9920574 | 9814758 | 7287485 | 13X |
| 1902.02 | Bear Paw Lake | 4780489 | 4365505 | 2942926 | 5X |
| 1902.03 | Bear Paw Lake | 5073434 | 4643909 | 3192582 | 5X |
| 1902.04 | Bear Paw Lake | 4902931 | 4600877 | 3138791 | 6X |
| 1902.06 | Bear Paw Lake | 4501906 | 4339253 | 2983345 | 5X |

**Supplementary table 2.** Genome assembly statistics for *Pungitius pungitius*.

| | contig (scaffold) |
|---|---|
| n | 393,037 (391,396) |
| Max length (bp) | 165,088 (182,644) |
| N50 (bp) | 9,202 (9,886) |
| Average length (bp) | 1,314 (1,320) |
| Gaps (%) | 0.03 |
| Total assembly length (bp) | 516,674,741 |