Ancient genomic variation underlies recent and repeated ecological adaptation

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

[1]Institute of Ecology and Evolution, University of Oregon, Eugene, OR
*To whom correspondence should be addressed: wcresko@uoregon.edu

Adaptation in the wild often involves the use of standing genetic variation (SGV), allowing rapid responses to selection on ecological timescales. Despite increasing documentation of evolutionarily important SGV in natural populations, we still know little about how the genetic and genomic structure and molecular evolutionary history of SGV relate to adaptation. Here, we address this knowledge gap using the threespine stickleback fish (*Gasterosteus aculeatus*) as a model. We demonstrate that adaptive genetic variation is structured genome-wide into distinct marine and freshwater haplogroups. This divergent variation averages six million years old, nearly twice the genome-wide average, but has been evolving over the 15-million-year history of the species. Divergent marine and freshwater genomes maintain regions of ancient ancestry that include multiple chromosomal inversions and extensive linked variation. These discoveries about ancient SGV demonstrate the intertwined nature of selection on ecological timescales and genome evolution over geological timescales.

The mode and tempo of adaptive evolution depend on the sources of genetic variation affecting fitness[1,2]. While new mutation is ultimately the source of all genetic variation, recent studies of adaptation in the wild document adaptive genetic variation that was either segregating in the ancestral population as standing genetic variation (SGV)[3-5] or introgressed from a separate population or species[6,7]. The use of SGV appears particularly important when dramatic responses to selection occur on ecological timescales, in dozens of generations or fewer[3]. When environments change rapidly, SGV can propel rapid evolution in ecologically relevant traits even in populations of long-lived organisms like Darwin's finches[8], monkeyflowers[9], and threespine stickleback fish[10].

Existing genetic variants have evolutionary histories that are often unknown but that may have significant impacts on subsequent adaptation[9,11]. The abundance, genomic distribution, and fitness effects[10-14] of SGV are themselves 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 distribution? And how does this evolutionary history of SGV potentially channel future evolutionary change?

Answers to these questions are critical for our understanding of the importance of SGV in nature and our ability to predict the paths available to adaptation on ecological timescales[9]. Biologists are beginning to probe evolutionary histories of SGV using genome-wide sequence variation across multiple individuals in numerous populations[15], but this level of inference has been unavailable for most natural systems because of methodological limitations that remove phase information (e.g. pool-seq[16]) or produce very short reads (e.g. RAD-seq[17]). Here, we investigate the structure and evolutionary history of divergent SGV by implementing a novel haplotyping method based on restriction site-associated DNA sequencing (RAD-seq). This approach creates nearly 1kb haplotypes at thousands of densely sampled loci, allowing us to accurately measure sequence variation and estimate divergence times across the genome.

SGV is likely critical to adaptation in this species. Marine stickleback have repeatedly colonized freshwater lakes and streams[18,19], and adaptive divergence in isolated freshwater habitats is highly parallel at the phenotypic[20,21] and genomic levels[19,22] (but see Stuart et al.[23]). In addition, analyses of haplotype variation at the genes *eda*[10,24] and *atp1a1*[24] 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*[10]. 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. To address this fundamental question we utilize the new haplotype RAD-seq approach to assay genome-wide variation associated with adaptive divergence in two young freshwater ponds, which formed during the end- the end-Pleistocene glacial retreat (c. 12,000 years ago[21,25]; Fig. 1). Our results demonstrate a suite of adaptive variation structured into distinct marine and freshwater haplotypes that evolved over millions of years.

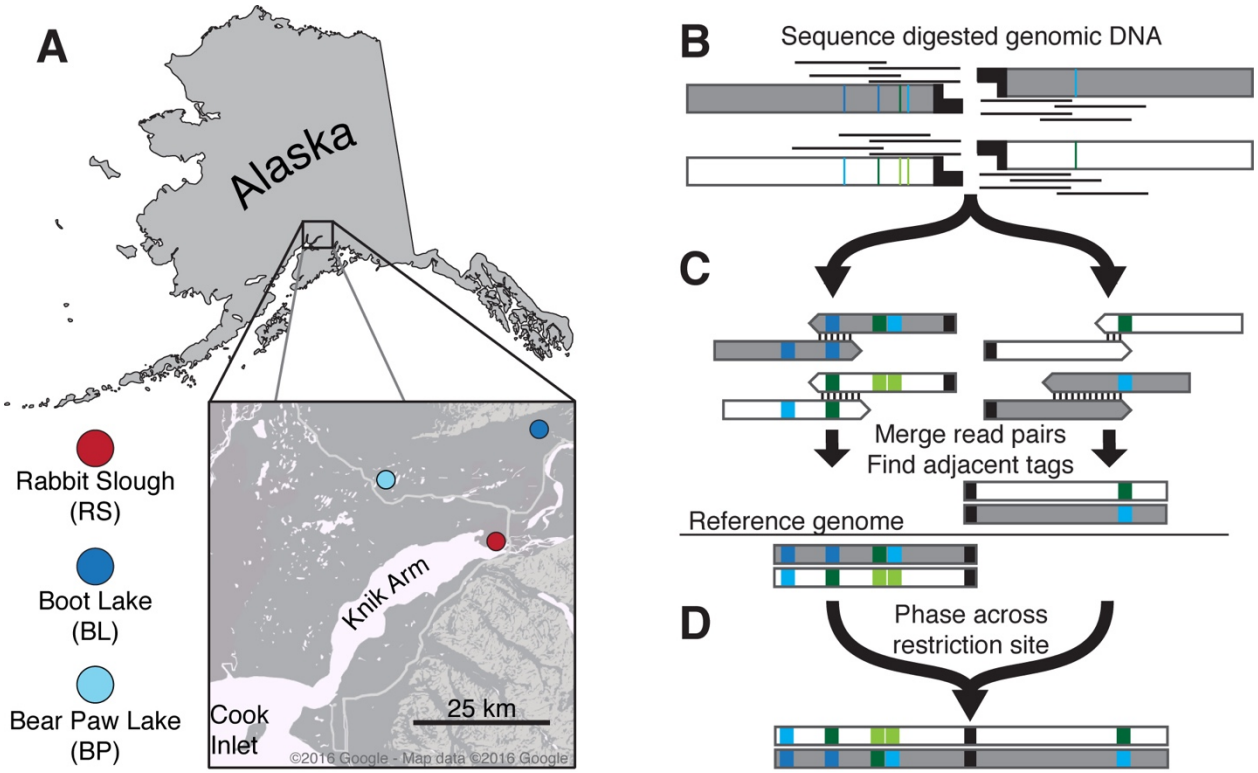


Figure 1. Stickleback sampling and RAD sequencing to measure haplotype variation. A) Threespine stickleback sampling locations in this study. Colors represent habitat type: red: marine; blue: freshwater. B-D: We modified the original RAD-seq protocol to generate local haplotypes. Colored bars represent 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.

RESULTS AND DISCUSSION

Parallel adaptation to freshwater environments has been a major theme of stickleback evolutionary history[26]. Stereotypical morphological changes (e.g. bony armor[20] and craniofacial structures[27]) presumably reflect adaptation to similar selective regimes[28,29] and are accompanied by parallel genomic divergence[19,22], which involves large regions spanning many megabases[24,30], including multiple chromosomal inversions[19]. The leading hypothesis for the genetics of parallel divergence in stickleback posits that distinct freshwater-adaptive haplotypes that are identical-by-descent (IBD) are shared among fresh water populations due to historical gene flow between marine and freshwater populations[30]. To test for the presence of these haplotypes directly, we characterized the genomic architecture and evolutionary history

79 of SGV by modifying the RAD-seq protocol[31] to generate phased haplotypes similar in
80 length to Sanger sequencing reads, each anchored to tens of thousands of *PstI*
81 restriction sites spread across the genome (Fig. 1.B-D). We sampled five fish (10
82 haploid genomes) each from Boot Lake (BL) and Rabbit Slough (RS), and four fish (8
83 genomes) from Bear Paw Lake (BP). After stringent data filtering (see Methods), this
84 resulted in a dataset of 57,992 RAD loci (locus = two tags representing one cut site)
85 with 694 potential variable sites per locus and a median of seven segregating sites per
86 locus (range: 2-155, Suppl. Fig. 1, Suppl. Table 1). We then used these phased
87 haplotypes to estimate genealogies at each RAD locus. By including haplotypes from all
88 three populations in these genealogical analyses, we were able to jointly calculate
89 population genetic statistics (F_{ST}, π, d_{XY}), estimate the degree of lineage sorting within
90 populations, and identify patterns of IBD among populations.
91 We find that indeed, parallel population genomic divergence in each freshwater
92 site consistently involved haplotypes that were IBD among both freshwater populations
93 (Fig. 2). Background F_{ST} between populations ranged from 0.139-0.226, with
94 divergence between the freshwater populations BL and BP being highest ($F_{ST(RS-BL)}$ =
95 0.139, $F_{ST(RS-BP)}$ = 0.194, $F_{ST(BL-BP)}$ = 0.226; two-sided Mann-Whitney test for all pairwise
96 comparisons: $p \leq 1x10^{-10}$). The degree and genomic distribution of pairwise F_{ST} between
97 the BL, BP, and RS populations were similar to those previously reported[22], including
98 marine-freshwater F_{ST} outlier regions on chromosome 4 over a broad span in which the
99 *eda* gene is embedded (orange triangle in Fig. 2A) and three regions now known to be
100 associated with chromosomal inversions on chromosomes 1, 11, and 21 (yellow bars in
101 Fig. 2; hereafter referred to as *inv1*, *inv11*, and *inv21*). The gene *atp1a1* (green triangle
102 in Fig. 2A) is contained within *inv1*. As expected, we found distinct haplogroups
103 associated with marine and freshwater habitats at both *eda* and *atp1a1* (Fig. 3, insets).
104 Strikingly, this finding of habitat specific haplogroups was not at all unique to
105 these well studied genes or chromosomal inversions. The two isolated freshwater
106 populations shared IBD haplotypes within all common marine-freshwater F_{ST} peaks
107 even though IBD was rare elsewhere (Fig. 2B). Furthermore, we observed a separate
108 clade of haplotypes representing the marine RS population at the majority (1129 of

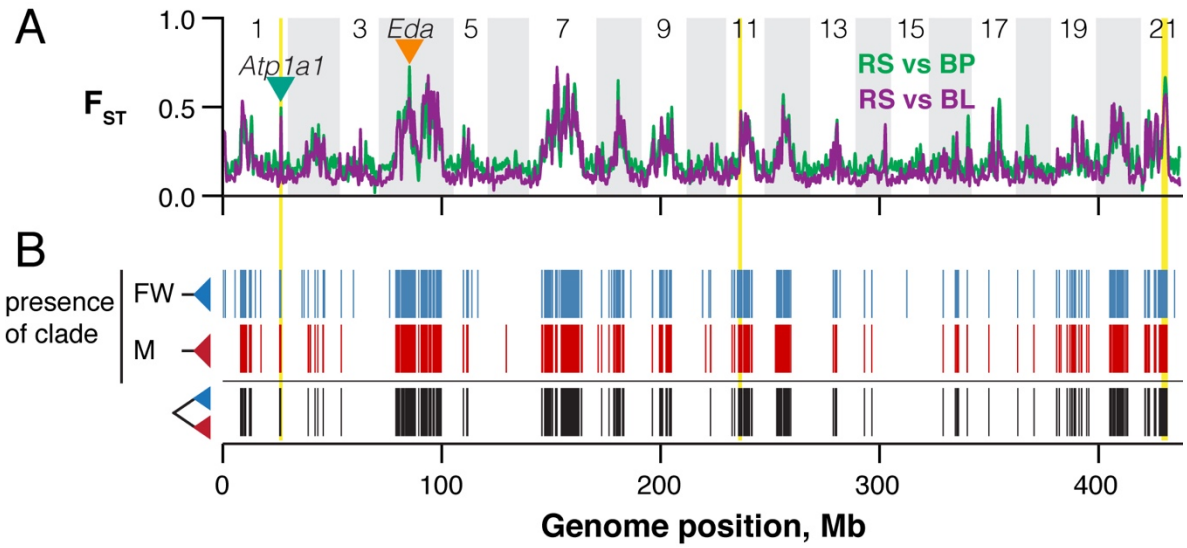


Figure 2. The genealogical structure of parallel genomic divergence. A) Genome-wide F_{ST} for both marine-freshwater comparisons was kernel-smoothed using a normally distributed kernel with a window size of 500 kb. Inverted triangles indicate the locations of two genes known to show extensive marine-freshwater 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.

2172, 52%) of RAD loci showing freshwater IBD. The result was a genome-wide pattern of reciprocal monophyly between marine and freshwater haplotypes. Notably, this is the same genealogical structure previously reported at *eda* [10,24] and *atp1a1* [24], demonstrating that these loci are but a small part of a genome-wide suite of genetic variation sharing similar habitat-specific evolutionary histories, and the previous documentation of their genealogies was a harbinger of a much more extensive pattern across the genome revealed here. Hereafter, we refer collectively to this class of RAD loci as ‘divergent loci’.

Because the genealogical structure of divergence across the genome mirrors that at *eda* and *atp1a1*, we asked whether levels of sequence variation and divergence also showed consistent genomic patterns. At all RAD loci we therefore calculated π within each population, as well as in the combined freshwater populations, and d_{XY} between marine and freshwater habitat types. Genome-wide diversity was similar across populations and habitat types (m ean $\pi_{RS} = 0.0032$, $\pi_{BL} = 0.0034$, $\pi_{BP} = 0.0026$, $\pi_{FW} = 0.0038$) and comparable to previous estimates[22]. Likewise, genome-wide d_{XY}

among habitat types was modest (0.0049) when compared to π across all populations (π = 0.0042, two-sided Mann-Whitney test: $p \leq 1\text{x}10^{-10}$). Among divergent loci, however, we observed reductions in diversity in both habitats (mean $\pi_{\text{RS-divergent}}$ = 0.0012, $\pi_{\text{RS-divergent}}$ = 0.0016, two-sided permutation test: $p \leq 1\text{x}10^{-4}$, Fig. 3, Suppl. Fig. 2), indicating natural selection in both habitats. Sequence divergence associated with reciprocal monophyly was striking, however, averaging nearly three times the genome-wide mean (mean $d_{XY\text{-divergent}}$ = 0.0124). This divergence ranged more than an order of magnitude (0.0013–0.0442), from substantially lower than the genome-wide average to ten times greater than the average. These findings indicate that much of the genetic variation underlying adaptive divergence was not just standing and structured by habitat, but has been segregating and accumulating for millennia.

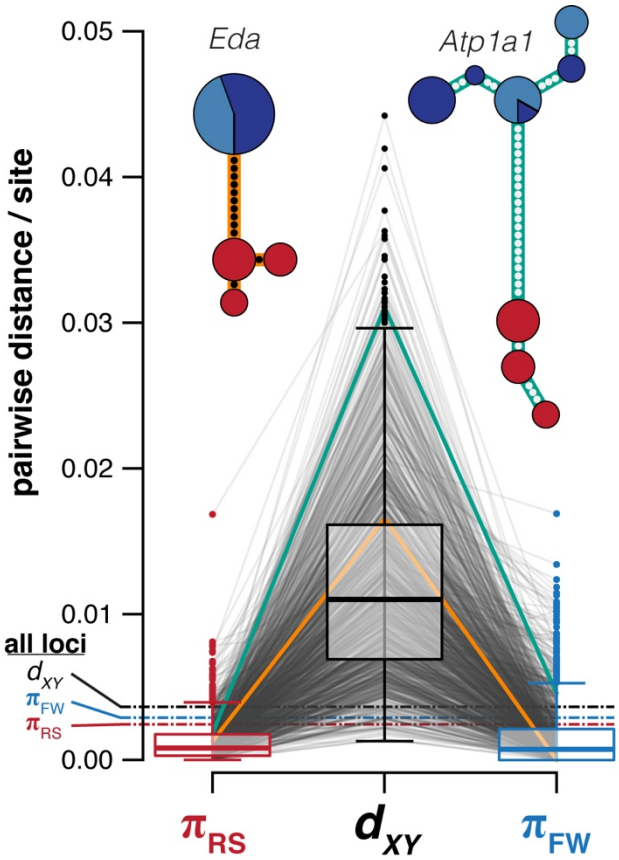


Figure 3. Extensive sequence divergence between marine and freshwater haplogroups accompanies reciprocal monophyly. For each reciprocally monophyletic RAD locus, we calculated sequence variation (π) within and sequence divergence between habitat types (d_{XY}). Each RAD locus is shown as a pair of lines 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.

These data clearly support the hypothesis of Schluter and Conte[30] of ancient haplotypes 'transported' among freshwater populations. Much of the divergence we observed was ancient in origin, with levels of sequence divergence at some RAD loci exceeding that observed at *eda* (Fig. 3, gold line) and suggestive of divergence times of at least two million years ago[10]. Our

154 observation that sequence variation was consistently reduced in both habitat types
155 emphasizes that alternative haplotypes at these loci are likely selected for in the marine
156 population as well as the freshwater. These alternative fitness optima — driven by
157 different ecologies — provide a favorable landscape for the maintenance of
158 variation[32,33], but also lead to a more potent barrier to gene flow among freshwater
159 populations if there are fitness consequences in the marine habitat for stickleback
160 carrying freshwater-adaptive variation. Conditional fitness effects through genetic
161 interactions (e.g., dominance[34] or epistasis[35]) and genotype-by-habitat interactions[36]
162 could potentially extend the residence time of freshwater haplotypes in the marine
163 habitat. Future work should consider the phenotypic effects of divergently adaptive
164 variation in different external environments[36,37].

165 A steady accumulation of divergently adaptive variation between marine and
166 freshwater stickleback genomes may also have been critical to the rapid divergence in
167 the young pond populations we study here. We found reciprocal monophyly associated
168 with a spectrum of sequence divergence, including a substantial fraction of divergent
169 loci (11.0%, 124/1129) with d_{XY} below the genome-wide average. Thus, ongoing
170 marine-freshwater ecological divergence has yielded continuing marine-freshwater
171 genomic divergence. Moreover, while this younger variation is shared between the
172 freshwater populations in this study, and localizes to genomic regions of divergence
173 shared globally[19], some adaptive variants may be distributed only locally (e.g. to
174 southern Alaska or the eastern Pacific basin). In addition to the globally distributed suite
175 of variation, there may also exist a substantial amount of regional variation contributing
176 to stickleback genomic and phenotypic diversity.

177 Sequence divergence provides an important relative evolutionary timescale.
178 However, to more directly compare the timescales of ecological adaptation and genomic
179 evolution, we translated patterns of sequence variation into the time to the most recent
180 common ancestor (T_{MRCA}) of allelic variation, in years. To do so, we performed a *de*
181 *novo* genome assembly of the ninespine stickleback (*Pungitius pungitius*), a member of
182 the Gasterosteidae that diverged from the threespine stickleback lineage approximately
183 15 million years ago[38] (Fig. 4A, Suppl. Table 2). We then aligned our RAD dataset to

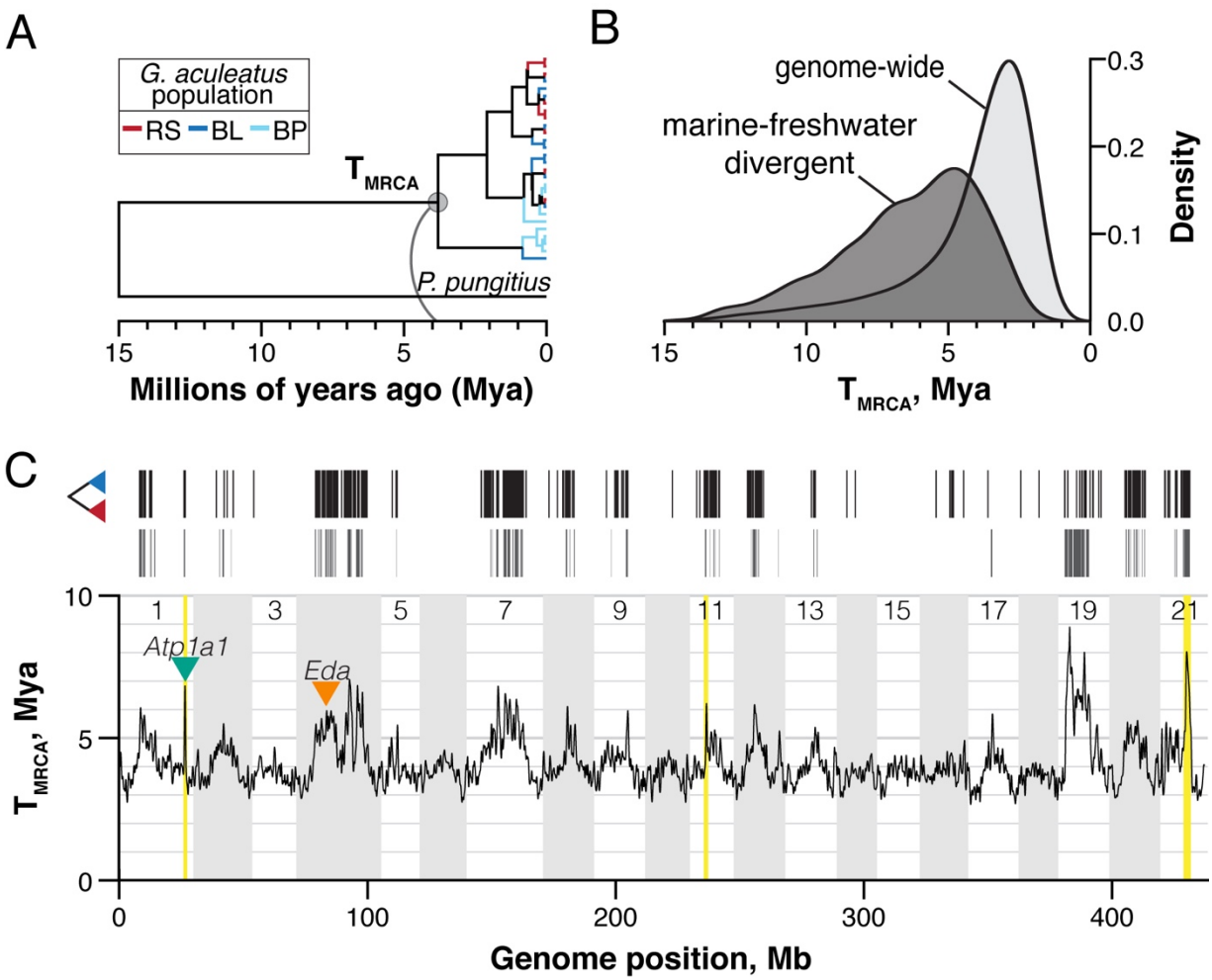


Figure 4. Marine-freshwater divergence has evolved over millions of years, affecting large genomic regions. We performed Bayesian estimation of the time to the most recent common ancestor (T_{MRCA}) of alleles at threespine stickleback RAD loci. We calibrated coalescence times within threespine stickleback 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.

184 this assembly and estimated gene trees for each alignment with BEAST[39], setting
185 divergence to the ninespine stickleback at 15 MYA (see Methods).
186 We find that the divergence of key marine and freshwater haplotypes has been
187 ongoing for millions of years and extends back to the split with the ninespine stickleback

188 lineage (Fig. 4B). Genome-wide variation averaged 4.1 MY old, and T_{MRCA} for the vast
189 majority of RAD loci was under 5 MY old. In contrast, divergent loci averaged 6.4 MY
190 old and, amazingly, the most ancient 10% (118 of 1129 loci) are estimated at over 10
191 MY old. This deep genomic divergence not only underscores that the marine-freshwater
192 transition has been occurring throughout the history of the threespine stickleback
193 lineage, for which there is evidence in the fossil record going back 10 million years[40],
194 but it also demonstrates that at least some of the variation fueling those ancient events
195 has persisted until the present day. In some genomic regions, then, marine and
196 freshwater threespine stickleback are as divergent as threespine and ninespine
197 stickleback, which are classified into separate genera.
198 Adaptive divergence has impacted the history of the stickleback genome as a
199 whole (Fig. 4C). We identified 32.6 Mb, or 7.5%, of the genome as having elevated
200 T_{MRCA} (gray boxes in Fig. 4C; two-sided permutation test, $p \leq 0.001$). Outside of the
201 non-recombining portion of the sex chromosome (chr. 19), the oldest regions of the
202 stickleback genome were those enriched for divergent loci. Patterns of ancient ancestry
203 closely mirrored recent divergence in allele frequencies (Fig. 2A) and it appears that
204 historical and contemporary marine-freshwater divergence has impacted ancestry
205 across much of the length of some chromosomes. Chromosome 4, for example,
206 contains at least three broad peaks in T_{MRCA} and a total of 5.9 Mb identified as genome-
207 wide outliers (two-sided permutation test, $p \leq 0.001$). This chromosome has been of
208 particular interest because of its association with a number of phenotypes [20,41], including
209 fitness[42]. We found the major-effect armor plate locus *eda* comprised a local peak
210 (mean T_{MRCA} = 6.4 MYA) nested within a large region of deep ancestry spanning 8.1
211 Mb. Moreover, at least two other peaks distal to *eda,* centered at 21.4 Mb and 26.6 Mb,
212 were also several million years older than the genomic average at 6.8 MYA and 7.0
213 MYA, respectively.
214 Intriguingly, genomic regions of elevated T_{MRCA} remained outliers even after
215 removing marine-freshwater relative divergence outliers as measured by F_{ST} (Suppl.
216 Fig. 3). We estimated that 7.5% of the genome had increased T_{MRCA} even though only
217 1.9% of RAD loci (1129 of 57,992) were classified as divergent. When we removed

218 these loci along with loci with extreme values of marine-freshwater F_{ST} ($F_{ST} > 0.5$),
219 many of the regions in which they resided were still T_{MRCA} outliers. It is possible that the
220 remainder of this old variation is neutral with respect to fitness. However, we identified
221 divergence outliers based on only a single axis of divergence: the marine-freshwater
222 axis. Throughout the entire species range, populations are locally experiencing multiple
223 axes of divergence, including lake-stream and benthic-limnetic axes[43], that often shares
224 a common genomic architecture[44,45]. Our data may indicate underlying similarities in
225 selection regimes. Alternatively, this co-localized ancient variation may represent the
226 accumulation of adaptive divergence along multiple axes in the same genomic regions,
227 whether or not the underlying adaptive variants are the same. Aspects of the genomic
228 architecture, such as gene density or local recombination rates, may in part govern
229 where in the genome adaptive divergence can occur[46-48]. Multiple axes of divergence
230 may therefore act synergistically to maintain genomic variation across the stickleback
231 metapopulation.

232 Nevertheless, much of the ancient variation we observe may in fact itself be
233 neutral, having been maintained by close linkage to loci under divergent selection
234 between the marine and freshwater habitats[32]. Indeed, the broadest peaks of T_{MRCA} we
235 observe occur in genomic regions with low rates of recombination[47,49] in other
236 stickleback populations, which would extend the size of the linked region affected by
237 divergent selection. On ecological timescales, low recombination rates in stickleback
238 are thought to promote divergence by making locally adapted genomic regions resistant
239 to gene flow[47]. Our results potentially extend the inferred impact of recombination rate
240 variation on genomic variation to timescales that are 1000-fold longer, maintaining both
241 multimillion-year-old adaptive variation and large stores of linked genetic variation.
242 Future modeling efforts will be needed to explore the range of population genetic
243 parameter values (e.g. selection coefficients, migration rates, and recombination rates)
244 required to produce the extent of divergence we see here.

245 Lastly, our findings demonstrate that known chromosomal inversions maintain
246 globally distributed, multilocus haplotypes. The three chromosomal inversions (*inv1*,
247 *inv11*, and *inv21*; yellow bars in Fig. 4C) all showed sharp spikes in T_{MRCA}. Genomic

248 signatures of these inversions are distributed throughout the species range, including
249 coastal marine-freshwater population pairs in the Pacific and Atlantic basins[19] and
250 inland lake-stream pairs in Switzerland. [44]. Despite our limited geographic sampling, our
251 finding that all three of these inversions are over six million years old is further evidence
252 of single, ancient origins of each, followed by their spread across the species range.
253 Each inversion contained a high density of divergent RAD loci (*inv1*: 64% of loci
254 divergent; *inv11*: 60%; *inv21*: 71%) but we also identified regions within these inversions
255 in which haplotypes from marine or freshwater habitats, or both, were not monophyletic.
256 *inv1* and *inv11* both contained two regions separated by loci in which neither habitat
257 type was monophyletic; *inv21*, the largest of the three, contained ten such regions.
258 Additionally, T_{MRCA} and F_{ST} decreased sharply to background levels outside of the
259 inversions, demonstrating the potential for gene flow and recombination to homogenize
260 variation in these regions. We interpret this as evidence that these inversions help
261 maintain linkage disequilibrium among multiple divergently adaptive variants in regions
262 susceptible to homogenization[11,50]. The presence of these inversions, therefore, further
263 supports the hypothesis that the recombinational landscape can influence where in the
264 genome adaptive divergence can occur and emphasizes the degree to which gene flow
265 among divergently adapted stickleback populations has impacted global genomic
266 diversity.
267
268 CONCLUSIONS
269 Selection operating on two very different timescales — the ecological and the
270 geological — has shaped genomic patterns of SGV in the threespine stickleback.
271 Selection on ecological timescales drives phenotypic divergence in decades or millennia
272 by sorting SGV across geography and throughout the genome[22,44,51,52]. Our findings
273 show that the persistence of this ecological diversity and local adaptation of stickleback
274 has set the stage for long-term divergent selection and for the continual accumulation
275 and maintenance of adaptive variation over millions of years. A number of genetic
276 variants fueling contemporary, rapid adaptation may even have been present - and
277 under selection - since before the threespine-ninespine stickleback lineages split. The

extent to which ecological adaptation in a single population drew on haplotypes that have evolved over millions of years and persisted in multiple populations, many of which are now extinct, underscores the need to understand macroevolutionary patterns when studying microevolutionary processes, and vice versa.

METHODS

Sample collection and library preparation

Wild threespine stickleback were collected from Rabbit Slough (N 61.5595, W 149.2583), Boot Lake (N 61.7167, W 149.1167), and Bear Paw Lake (N 61.6139, W 149.7539). Rabbit Slough is an offshoot of the Knik Arm of Cook Inlet and is known to be populated by anadromous populations of stickleback that are stereotypically oceanic in phenotype and genotype [22,53]. Boot Lake and Bear Paw Lake are both shallow lakes formed during the end-Pleistocene glacial retreat. Fish were collected in the summers of 2009 (Rabbit Slough), 2010 (Bear Paw Lake), and 2014 (Boot Lake) using wire minnow traps and euthanized *in situ* with Tricaine solution. Euthanized fish were immediately fixed in 95% ethanol and shipped to the Cresko Laboratory at the University of Oregon (Eugene, OR, USA). DNA was extracted from fin clips preserved in 95% ethanol using either Qiagen DNeasy spin column extraction kits or Ampure magnetic beads (Beckman Coulter, Inc) following manufacturer's instructions. Yields averaged 1-2 µg DNA per extraction (~30 mg tissue). Treatment of animals followed protocols approved the University of Oregon Institutional Animal Care and Use Committee (IACUC).

We designed our library preparation strategy to identify sufficient sequence variation for gene tree reconstruction and to simplify downstream sequence processing and analysis by taking advantage of the phase information captured by paired-end sequencing. We generated RAD libraries from these samples using the single-digest sheared RAD protocol from Baird et al. with the following specifications and adjustments: 1 µg of genomic DNA per fish was digested with the restriction enzyme *PstI-HF* (New England Biolabs), followed by ligation to P1 Illumina adaptors with 6 bp inline barcodes. Ligated samples were multiplexed and sheared by sonication in a Bioruptor (Diagenode). To ensure that most of our paired-end reads would overlap

308 unambiguously and produce longer contiguous sequences, we selected a narrow
309 fragment size range of 425-475 bp. The remainder of the protocol was per Baird et al.
310 [31]. All fish were sequenced on an Illumina HiSeq 2500 using paired-end 250 bp
311 sequencing reads at the University of Oregon's Genomics and Cell Characterization
312 Core Facility (GC3F).
313

314 *Sequence preparation*

315 Raw Illumina sequence reads were demultiplexed, cleaned, and processed
316 primarily using the Stacks pipeline[54]. Paired-end reads were demultiplexed with
317 **`process_shortreads`** and cleaned using **`process_radtags`** using default criteria
318 (throughout this document, names of scripts, programs, functions, and command-line
319 arguments will appear in **`fixed-width font`**). Overlapping read pairs were then
320 merged with **`fastq-join`** [55] (Fig. S1). Pairs that failed to merge were removed from
321 further analysis. In order to retain the majority of the sequence data for analysis in
322 Stacks and still maintain adequate contig lengths, merged contigs were trimmed to 350
323 bp and all contigs shorter than 350 bp were discarded. We aligned these contigs to the
324 stickleback reference genome[19,49] using **`bbmap`** with the most sensitive alignment
325 settings ('**`vslow=t`**'; http://jgi.doe.gov/data-and-tools/bbtools/) and used the **`pstacks`**,
326 **`cstacks`**, and **`sstacks`** components of the Stacks pipeline to create stacks and call
327 SNPs and haplotypes, create a catalog of RAD tags across individuals, and match tags
328 across individuals. All data were then passed through the Stacks error correction
329 module **`rxstacks`** to prune unlikely haplotypes. We ran the Stacks component program
330 populations on the final dataset to filter loci genotyped in fewer than four individuals in
331 each population and to create output files for sequence analysis. We use the naming
332 conventions of Baird et al. [56]: A "RAD tag" refers to sequence generated from a single
333 end of a restriction site and the pair of RAD tags sequenced at a restriction site
334 comprises a "RAD locus" (Figure 2.1).

335 We used the program **`phase`** [57] to phase pairs of RAD tags originating from the
336 same restriction site. We coded haplotypes present at each RAD tag, which often

337 contain multiple SNPs, into multiallelic genotypes. This both simplified and reduced
338 computing time for the phasing process. Custom Python scripts automated this process
339 and are included as supplementary files. We required that each individual had at least
340 one sequenced haplotype at each tag for phasing to be attempted. If a sample had
341 called genotypes at only one tag in the pair, the sample was removed from further
342 analysis of that locus. The resultant phased haplotypes were used to generate
343 sequence alignments for import into BEAST.
344 We recovered a total of 236,787 RAD tags after filtering, mapping to 151,813 *PstI*
345 restriction sites. At 84,974 restriction sites, we recovered and successfully phased
346 adjacent RAD tags (169,948 RAD tags) into single RAD loci. We retained these 84,974
347 RAD loci for our analysis.
348
349 *Ninespine stickleback genome assembly*
350 In order to estimate the T_{MRCA} of threespine stickleback RAD alleles, we used the
351 ninespine stickleback (*Pungitius pungitius*) as an outgroup (Figure 3.1, see Figure 1.2).
352 RAD sequence analysis, however, relies on the presence of homologous restriction
353 sites among sampled individuals and results in null alleles when mutations occur within
354 a restriction site[58]. Because this probability increases with greater evolutionary distance
355 among sampled sequences, we elected to use RAD-seq to only estimate sequence
356 variation within the threespine stickleback. We then generated a contig-level *de novo*
357 ninespine stickleback genome assembly from a single ninespine stickleback individual
358 from St. Lawrence Island, Alaska (collected by J. Postlethwait) using DISCOVAR *de*
359 *novo* (https://software.broadinstitute.org/software/discovar). We used this single
360 ninespine stickleback haplotype to estimate threespine-ninespine sequence divergence
361 and time calibrate coalescence times within the threespine stickleback. DISCOVAR de
362 novo requires a single shotgun library of paired-end 250-bp sequence reads from short-
363 insert-length DNA fragments. High molecular weight genomic DNA was extracted from
364 an ethanol-preserved fin clip by proteinase K digestion followed by DNA extraction with
365 Ampure magnetic beads. Purified genomic DNA was mechanically sheared by
366 sonication and size selected to a range of 200-800 bp by gel electrophoresis and

367 extraction. We selected this fragment range to agree with the recommendations for *de*
368 *novo* assembly using the DISCOVAR *de novo*
369 (https://software.broadinstitute.org/software/discovar/blog). This library was sequenced
370 on a single lane of an Illumina HiSeq2500 at the University of Oregon's Genomics and
371 Cell Characterization Core Facility (GC3F: https://gc3f.uoregon.edu/). We assembled
372 the draft ninespine stickleback genome using DISCOVAR de novo. Raw sequence read
373 pairs were first quality filtered and adaptor sequence contamination removed using the
374 program **process_shortreads**, which is included in the Stacks analysis pipeline [59]. We
375 ran the genome assembly on the University of Oregon's Applied Computational
376 Instrument for Scientific Synthesis (ACISS: http://aciss-computing.uoregon.edu).
377
378 *Alignment of RAD tags to the ninespine assembly*
379 We included the single ninespine stickleback haplotype into our sequence
380 analyses by aligning a single phased threespine stickleback RAD haplotype from each
381 locus to the ninespine genome assembly. For those that aligned uniquely (59,254 RAD
382 loci), we used a custom Python script to parse the output BAM file [60] and reconstruct the
383 ninespine haplotype from the query sequence and alignment fields. The final dataset
384 consists of 57,992 RAD loci that mapped to the 21 threespine stickleback chromosomes
385 and aligned uniquely to the ninespine assembly.
386
387 *Lineage sorting and time to the most recent common ancestor*
388 Allelic divergence can occur by multiple modes of lineage sorting during
389 adaptation. To identify patterns of lineage sorting associated with freshwater
390 colonization, we analyzed gene tree topologies at all RAD loci using BEAST v. 1.7[39,61].
391 We used blanket parameters and priors for BEAST analyses across all RAD loci.
392 Markov chain Monte Carlo (MCMC) runs of 1,000,000 states were specified, and trees
393 logged every 100 states. We used a coalescent tree prior and the GTR+Γ substitution
394 model with four rate categories and uniform priors for all substitution rates. We identified
395 evidence of lineage sorting by using the program **`treeannotator`** to select the
396 maximum clade credibility (MCC) tree for each RAD locus and the

397 **`is.monophyletic()`** function included in the R package 'ape' [62]. We determined for
398 each MCC tree whether tips originating from marine (RS) or freshwater (BL+BP) formed
399 monophyletic clades.
400 To convert node ages estimated in BEAST into divergence times, in years, we
401 assumed a 15 million-year divergence time between threespine and ninespine
402 stickleback at each RAD locus [38]. The T_{MRCA} of all alleles in each gene tree was set at
403 15 Mya at each node age of interest was converted into years relative to the total height
404 of the tree. Additionally, to use the ninespine stickleback as an outgroup, we required
405 that threespine stickleback haplotypes at a RAD locus were monophyletic to the
406 exclusion of the ninespine haplotype. Doing so reduced our analysis to 49,672 RAD loci
407 for analyses included in Fig. 4 of the main text. RAD loci not showing this pattern of
408 lineage sorting did not show evidence of a genome-wide correlation with marine-
409 freshwater divergence and thus do not impact the assertions in the main text. We used
410 medians of the posterior distributions as point estimates of T_{MRCA} for each RAD locus.
411 Because of the somewhat limited information from any single RAD locus, and because
412 the facts of the genealogical process mean that the true T_{MRCA} at any locus likely differs
413 from the 15 My estimate [63-65], we do not rely heavily on T_{MRCA} estimates at individual
414 RAD loci. Rather, we use these estimates to understand patterns of broad patterns of
415 ancestry throughout the threespine stickleback genome — spatially along chromosomes
416 and genome-wide patterns.
417 We determined T_{MRCA} outlier genomic regions by permuting and kernel
418 smoothing the genomic distribution of T_{MRCA} estimates using the same window sizes as
419 we present in the main text. Windows where the actual T_{MRCA} value exceeded 99.9% of
420 permuted windows were considered outliers. This method controls for the local density
421 of RAD loci (poorly sampled regions will have larger confidence bands) and the size of
422 the windows used.
423
424 *Sequence diversity and haplotype networks*
425 We quantified sequence diversity within and among populations and sequence
426 divergence between populations using R (R Core Team [66]). We used the R package

427 'ape' [62] to compute pairwise distance matrices for all alleles at each RAD locus and
428 used these matrices to calculate the average pairwise nucleotide distances, π, within
429 and among populations along with d_{XY}, the average pairwise distance between two
430 sequences using only across-population comparisons [67]. We also calculated the
431 haplotype-based F_{ST} from Hudson et al. [68] implemented in the R package 'PopGenome'
432 [69]. We used permutation tests written in R to identify differences in variation within- and
433 between-habitat type at divergent RAD loci versus the genome-wide distributions.
434 Mann-Whitney-Wilcoxon tests implemented in R were used to identify variation in
435 genome-wide diversity among populations and habitat types.
436 We constructed haplotype networks of the RAD loci at *eda* and *atp1a1* using the
437 infinite sites model with the function **`haploNet()`** in the R package 'pegas'[70]. The
438 *atp1a1* network was constructed from from a RAD locus spanning exon 15 of *atp1a1*
439 and including portions of introns 14 and 15 at (chr1:21,726,729-21,727,381 [BROAD
440 S1, v89]; chr1: 26,258,117-26,257,465 [re-scaffolding from Glazer, *et al* [49]]). The *eda*
441 network spans exon 2 and portions of introns 1 and 3 of *eda* (chr4: 12,808,396-
442 12,809,030).
443
444 *Code availability*
445 Scripts used to phase RAD-tags, summarize gene trees, calculate population genetic
446 statistics, and produce figures and statistics presented in paper are available at
447 https://github.com/thomnelson/ancient-divergence. Scripts for processing raw sequence
448 data are available from the authors upon request.
449
450 DATA AVAILABILITY
451 Raw sequence data supporting these findings are available on the Sequence Read
452 Archive at PRJNAXXXXXX. The final datasets needed to reproduce the figures and
453 statistics presented in the paper are available at https://github.com/thomnelson/ancient-
454 divergence.
455
456 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, 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.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

LITERATURE CITED

1 Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress on Genetics* **1**, 356-366 (1932).

2 Orr, H. A. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics* **6**, 119-127, doi:10.1038/nrg1523 (2005).

3 Barrett, R. D. H. & Schluter, D. Adaptation from standing genetic variation. *Trends in Ecology and Evolution* **23**, 38-44, doi:10.1016/j.tree.2007.09.008 (2008).

4 Domingues, V. S. *et al.* Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution* **66**, 3209-3223, doi:10.1111/j.1558-5646.2012.01669.x (2012).

5 Schrider, D. R. & Kern, A. D. Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular Biology and Evolution*, doi:10.1093/molbev/msx154 (2017).

6 Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194-197, doi:10.1038/nature13408 (2014).

7 Fontaine, M. C. *et al.* Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* **347**, 1258524, doi:10.1126/science.1258524 (2015).

8 Grant, P. R. & Grant, B. R. Unpredictable evolution in a 30-year study of Darwin's finches. *Science* **296**, 707-711, doi:DOI 10.1126/science.1070315 (2002).

9 Wright, K. M., *et al*. Indirect Evolution of Hybrid Lethality Due to Linkage with Selected Locus in Mimulus guttatus. *PLoS Biology* **11**, doi:10.1371/journal.pbio.1001497 (2013).

10 Colosimo, P. F. *et al.* Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* **307**, 1928-1933, doi:10.1126/science.1107239 (2005).

11 Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419-434, doi:10.1534/genetics.105.047985 (2006).

12 Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289-1303 (1993).

13 Linnen, C. R., *et al*. On the origin and spread of an adaptive allele in deer mice. *Science* **325**, 1095-1098, doi:10.1126/science.1175826 (2009).

14 Stankowski, S. & Streisfeld, M. A. Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers. *Proceedings of the Royal Society B: Biological Sciences* **282**, 20151666, doi:10.1098/rspb.2015.1666 (2015).

15 Pease, J. B., *et al*. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology* **14**, e1002379, doi:10.1371/journal.pbio.1002379 (2016).

16 Schlotterer, C., *et al*. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics* **15**, 749-763, doi:10.1038/nrg3803 (2014).

17 Davey, J. W. *et al.* Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510, doi:10.1038/nrg3012 (2011).

18 Bell, M. A. & Foster, S. A. in *The Evolutionary Biology of the Threespine Stickleback* (eds M. A. Bell & S. A. Foster) Ch. 1, 1-27 (Oxford University Press, 1994).

19 Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55-61, doi:10.1038/nature10944 (2012).

20 Colosimo, P. F. *et al.* The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS Biology* **2**, 635-641, doi:10.1371/journal.pbio.0020109 (2004).

21 Cresko, W. A. *et al.* Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proc Natl Acad Sci U S A* **101**, 6050-6055, doi:10.1073/pnas.0308479101 (2004).

22 Hohenlohe, P. A. *et al.* Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *Plos Genet* **6**, e1000862, doi:10.1371/journal.pgen.1000862 (2010).

23 Stuart, Y. E. *et al.* Contrasting effects of environment and genetics generate a continuum of parallel evolution. *Nature Ecology & Evolution* **1**, 0158, doi:10.1038/s41559-017-0158 (2017).

24 Roesti, M., *et al*. The genomic signature of parallel adaptation from shared genetic variation. *Mol Ecol* **23**, 3944-3956, doi:10.1111/mec.12720 (2014).

25 Francis, R. C., *et al*. Historical and ecological sources of variation among lake populations of threespine sticklebacks, *Gasterosteus aculeatus*, near Cook Inlet, Alaska. *Canadian Journal of Zoology* **64**, 2257-2265 (1986).

26 Bell, M. A. & Foster, S. A. in *The Evolutionary Biology of the Threespine Stickleback* (eds M. A. Bell & S. A. Foster) Ch. 16, 472-486 (Oxford University Press, 1994).

27 Kimmel, C. B. *et al.* Evolution and development of facial bone morphology in threespine sticklebacks. *Proc Natl Acad Sci U S A* **102**, 5791-5796, doi:10.1073/pnas.0408533102 (2005).

28 Reimchen, T. E. in *The Evolutionary Biology of the Threespine Stickleback* (eds M. A. Bell & S. A. Foster) Ch. 9, 240-276 (Oxford University Press, 1994).

29 Arnegard, M. E. *et al.* Genetics of ecological divergence during speciation. *Nature* **511**, 307-311, doi:10.1038/nature13301 (2014).

30 Schluter, D. & Conte, G. L. Genetics and ecological speciation. *Proc Natl Acad Sci U S A* **106 Suppl 1**, 9955-9962, doi:10.1073/pnas.0901264106 (2009).

31 Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *Plos One* **3**, e3376, doi:10.1371/journal.pone.0003376 (2008).

32 Charlesworth, B., Nordborg, M. & Charlesworth, D. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetics Research* **70**, 155-174, doi:Doi 10.1017/S0016672397002954 (1997).

33 Lenormand, T. Gene flow and the limits to natural selection. *Trends Ecol Evol* **17**, 183-189, doi:Doi 10.1016/S0169-5347(02)02497-7 (2002).

34 Otto, S. P. & Bourguet, D. Balanced polymorphisms and the evolution of dominance. *American Naturalist* **153**, 561-574, doi:Doi 10.1086/303204 (1999).

35 Phillips, P. C. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics* **9**, 855-867, doi:10.1038/nrg2452 (2008).

36 McGuigan, K., *et al*. Cryptic genetic variation and body size evolution in threespine stickleback. *Evolution* **65**, 1203-1211, doi:10.1111/j.1558-5646.2010.01195.x (2011).

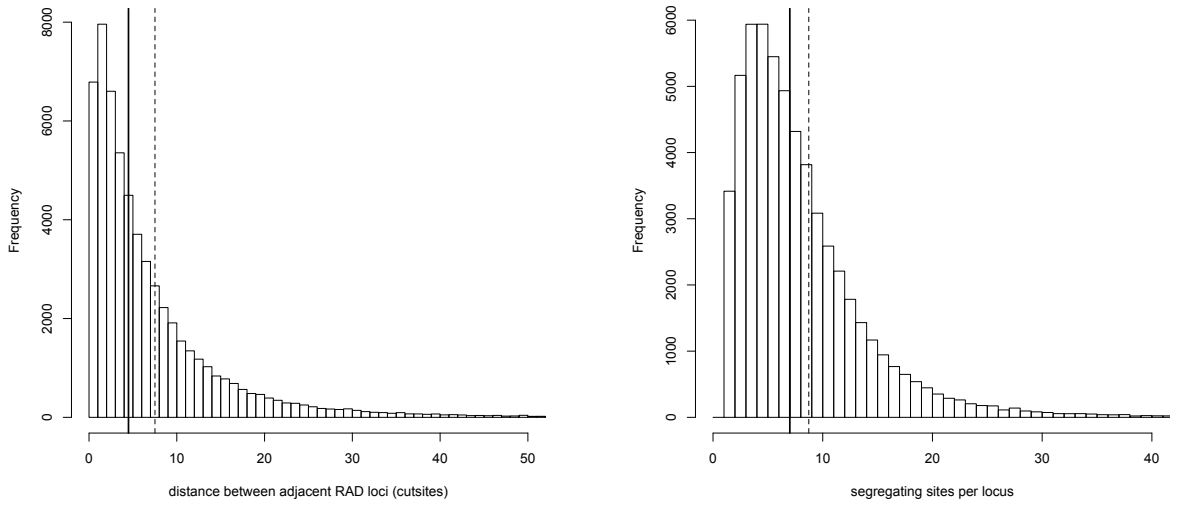
37 McCairns, R. J. S. & Bernatchez, L. Plasticity and heritability of morphological variation within and between parapatric stickleback demes. *Journal of*

575 *Evolutionary Biology* **25**, 1097-1112, doi:10.1111/j.1420-9101.2012.02496.x
576 (2012).
577 38 Aldenhoven, J. T., *et al*. Phylogeography of ninespine sticklebacks (*Pungitius*
578 *pungitius*) in North America: glacial refugia and the origins of adaptive traits. *Mol*
579 *Ecol* **19**, 4061-4076, doi:10.1111/j.1365-294X.2010.04801.x (2010).
580 39 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics
581 with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**, 1969-1973,
582 doi:10.1093/molbev/mss075 (2012).
583 40 Bell, M. A., Baumgartner, J. V. & Olson, E. C. Patterns of temporal change in
584 single morphological characters of a Miocene stickleback fish. *Paleobiology* **11**,
585 258-271 (1985).
586 41 Miller, C. T. *et al.* Modular skeletal evolution in sticklebacks is controlled by
587 additive and clustered quantitative trait loci. *Genetics* **197**, 405-420,
588 doi:10.1534/genetics.114.162420 (2014).
589 42 Barrett, R. D., Rogers, S. M. & Schluter, D. Natural selection on a major armor
590 gene in threespine stickleback. *Science* **322**, 255-257,
591 doi:10.1126/science.1159978 (2008).
592 43 McKinnon, J. S. & Rundle, H. D. Speciation in nature:the threespine stickleback
593 model systems. *Trends in Ecology and Evolution* **17**, 480-488 (2002).
594 44 Roesti, M., Kueng, B., Moser, D. & Berner, D. The genomics of ecological
595 vicariance in threespine stickleback fish. *Nature Communications* **6**, 8767,
596 doi:10.1038/ncomms9767 (2015).
597 45 Deagle, B. E. *et al.* Population genomics of parallel phenotypic evolution in
598 stickleback across stream-lake ecological transitions. *Proceedings of the Royal*
599 *Society B: Biological Sciences* **279**, 1277-1286, doi:10.1098/rspb.2011.1552
600 (2012).
601 46 Samuk, K. *et al.* Gene flow and selection interact to promote adaptive divergence
602 in regions of low recombination. *Molecular Ecology*, doi:10.1111/mec.14226
603 (2017).
604 47 Roesti, M., Moser, D. & Berner, D. Recombination in the threespine stickleback
605 genome - Patterns and consequences. *Molecular Ecology* **22**, 3014-3027,
606 doi:10.1111/mec.12322 (2013).
607 48 Aeschbacher, S., *et al*. Population-genomic inference of the strength and timing
608 of selection against gene flow. *Proceedings of the National Academy of Sciences*
609 *USA* **114**, 7061-7066, doi:10.1073/pnas.1616755114 (2017).
610 49 Glazer, A. M., *et al*. Genome assembly improvement and mapping of
611 convergently evolved skeletal traits in sticklebacks with genotyping-by-
612 sequencing. *G3-Genes Genomes Genetics* **5**, 1463-1472,
613 doi:10.1534/g3.115.017905 (2015).
614 50 Guerrero, R. F., Rousset, F. & Kirkpatrick, M. Coalescent patterns for
615 chromosomal inversions in divergent populations. *Philosophical Transactions of*
616 *the Royal Society B: Biological Sciences* **367**, 430-438,
617 doi:10.1098/rstb.2011.0246 (2012).

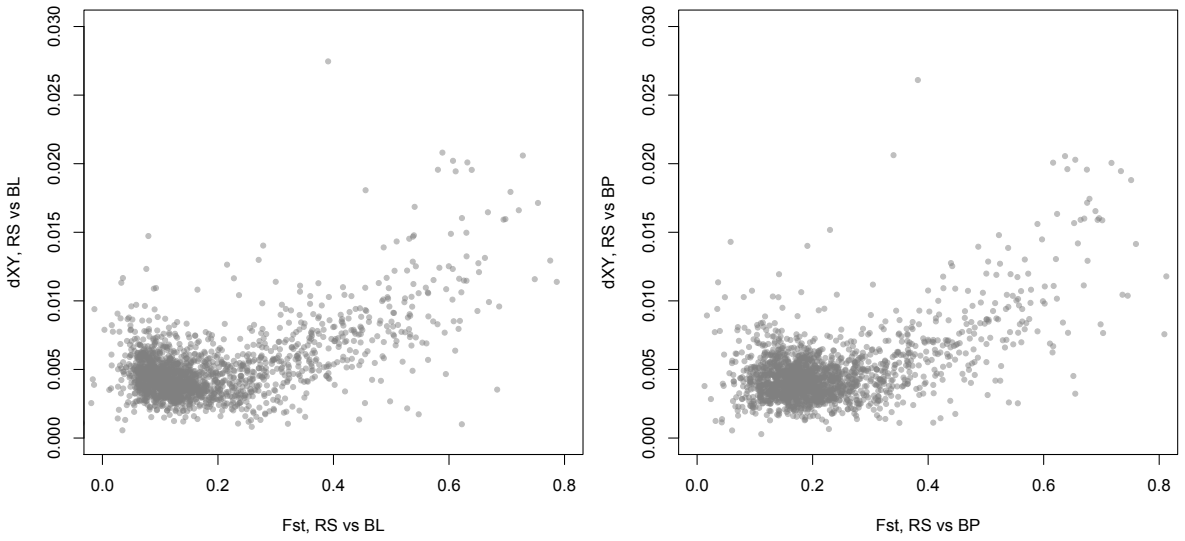
51 Hendry, A. P., Taylor, E. B. & McPhail, J. D. Adaptive divergence and the balance between selection and gene flow: lake and stream stickleback in the Misty system. *Evolution* **56**, 1199-1216, doi:10.1554/0014-3820(2002)056[1199:ADATBB]2.0.CO;2 (2002).
52 Lescak, E. A. *et al.* Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proceedings of the National Academy of Sciences USA* **112**, E7204-7212, doi:10.1073/pnas.1512020112 (2015).
53 Cresko, W. A. *et al.* Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proceedings of the National Academy of Sciences USA* **101**, 6050-6055, doi:10.1073/pnas.0308479101 (2004).
54 Catchen, J. M., *et al*. Stacks: building and genotyping Loci de novo from short-read sequences. *G3 - Genes Genomes Genetics* **1**, 171-182, doi:10.1534/g3.111.000240 (2011).
55 Aronesty, E. ea-utils: Command-line tools for processing biological sequencing data. (2011).
56 Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *Plos One* **3**, 1-7, doi:10.1371/journal.pone.0003376 (2008).
57 Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978-989, doi:Doi 10.1086/319501 (2001).
58 Arnold, B., *et al*. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol Ecol* **22**, 3179-3190, doi:10.1111/mec.12276 (2013).
59 Catchen, J., *et al*. Stacks: An analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140, doi:10.1111/mec.12354 (2013).
60 Li, H. *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
61 Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *Bmc Evol Biol* **7**, 214, doi:10.1186/1471-2148-7-214 (2007).
62 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289-290, doi:10.1093/bioinformatics/btg412 (2004).
63 Kingman, J. F. C. The coalescent. *Stochastic Processes and their Applications* **13**, 235-248, doi:10.1016/0304-4149(82)90011-4 (1982).
64 Kingman, J. F. C. On the genealogy of large populations. *Journal of Applied Probability* **19**, 27-43 (1982).
65 Tajima, F. Evolutionary relationship of DNA-sequences in finite populations. *Genetics* **105**, 437-460 (1983).
66 R Core Team. R Foundation for Statistical Computing, Vienna, Austria (2016).
67 Nei, M. *Molecular Evolutionary Genetics*. (Columbia university press, 1987).
68 Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA-sequence data. *Genetics* **132**, 583-589 (1992).

69 Pfeifer, B., *et al*. PopGenome: An efficient swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution* **31**, 1929-1936, doi:10.1093/molbev/msu136 (2014).

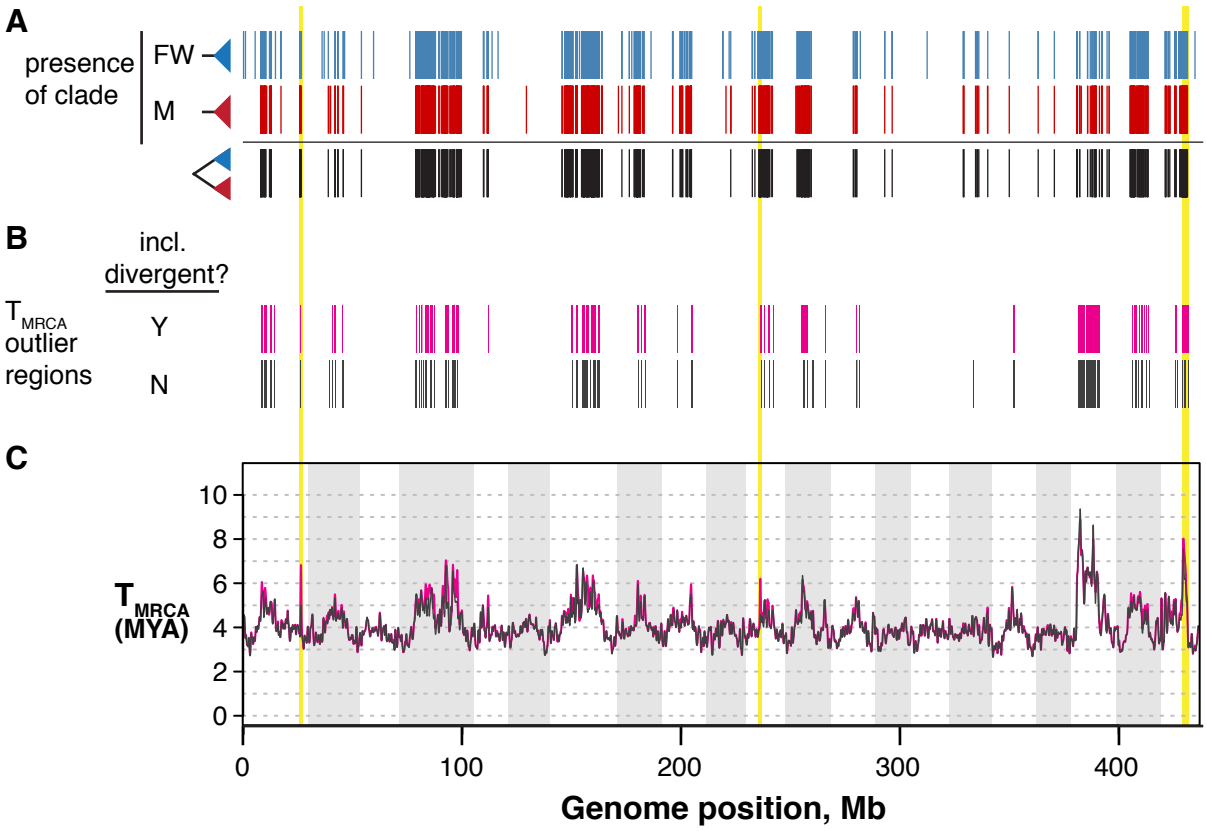
70 Paradis, E. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419-420, doi:10.1093/bioinformatics/btp696 (2010).



Supplementary Figure S1. RAD-seq effectively samples genome-wide sequence diversity. Histograms of the distance between adjacent RAD loci (left: calculated as the distance between the centers of each restriction site) and 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.



Supplementary Figure S2. 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 S3. 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