PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2012) **279**, 1277–1286
doi:10.1098/rspb.2011.1552
Published online 5 October 2011

Population genomics of parallel phenotypic evolution in stickleback across stream–lake ecological transitions

Bruce E. Deagle[1,*], Felicity C. Jones[2], Yingguang F. Chan[2], Devin M. Absher[4], David M. Kingsley[2,3] and Thomas E. Reimchen[1]

[1]*Department of Biology, University of Victoria, Victoria, British Columbia, Canada V8W 3NS*
[2]*Department of Developmental Biology, and* [3]*Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305-5329, USA*
[4]*HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA*

Understanding the genetics of adaptation is a central focus in evolutionary biology. Here, we use a population genomics approach to examine striking parallel morphological divergences of parapatric stream–lake ecotypes of threespine stickleback fish in three watersheds on the Haida Gwaii archipelago, western Canada. Genome-wide variation at greater than 1000 single nucleotide polymorphism loci indicate separate origin of giant lake and small-bodied stream fish within each watershed (mean F_{ST} between watersheds = 0.244 and within = 0.114). Genome scans within watersheds identified a total of 21 genomic regions that are highly differentiated between ecotypes and are probably subject to directional selection. Most outliers were watershed-specific, but genomic regions undergoing parallel genetic changes in multiple watersheds were also identified. Interestingly, several of the stream–lake outlier regions match those previously identified in marine–freshwater and benthic–limnetic genome scans, indicating reuse of the same genetic loci in different adaptive scenarios. We also identified multiple new outlier loci, which may contribute to unique aspects of differentiation in stream–lake environments. Overall, our data emphasize the important role of ecological boundaries in driving both local and broadly occurring parallel genetic changes during adaptation.

Keywords: genome scan; F_{ST} outlier; ecological speciation; *Gasterosteus*; single nucleotide polymorphism

1. INTRODUCTION

Uncovering the genetic basis of local adaptation in natural populations will refine our understanding of natural selection [1], allow insight into how species respond to environmental change [2] and shed light on the process of speciation [3]. Many studies investigating the genetics of adaptation have taken a population genomics approach in which genome-wide patterns of genetic variation are documented in many individuals within a species [4–6]. Using this approach, regions of the genome that are under divergent selection between local populations (outlier loci) can be identified by their high level of differentiation compared with the background levels [7–9]. Putatively, neutral non-outlier loci also provide insight by providing a clearer window into population history [10]. In cases where subpopulations have diverged enough to become reproductively isolated, the genetics of speciation can be examined [3,11]. To date, the majority of population genomics studies on wild species have used anonymous genetic markers [4–6]. However, with advances in high-throughput genetics and mounting numbers of completed genome projects, markers with known locations within the genome are increasingly being examined, narrowing the search for underlying genes [12].

The threespine stickleback (*Gasterosteus aculeatus*) has become a powerful ecological model species with both well-studied natural history and extensive genetic resources. This small fish inhabits marine environments throughout the temperate Northern Hemisphere and has colonized countless freshwater rivers, streams, ponds and lakes [13,14]. Several studies have uncovered the genetic basis of phenotypic traits that have evolved during repeated colonizations of the freshwater environment by marine ancestors [15,16]. These studies have highlighted cases of parallel phenotypic evolution occurring via selection on the same genomic loci. More broadly, parallel evolution has produced genome-wide patterns, with many of the same genomic regions being under selection in independently derived freshwater populations [12].

Within freshwater habitats formed since the end of the last ice age (approx. 12 000 years ago), there has been a considerable radiation in stickleback morphology, with numerous adaptations documented among populations [17,18]. This recent radiation presents further superb opportunities to examine the genetics of adaptation. These freshwater populations include divergent parapatric and sympatric populations. Parapatric populations

* Author for correspondence (bdeagle@uvic.ca).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2011.1552 or via http://rspb.royalsocietypublishing.org.

Received 28 July 2011
Accepted 14 September 2011

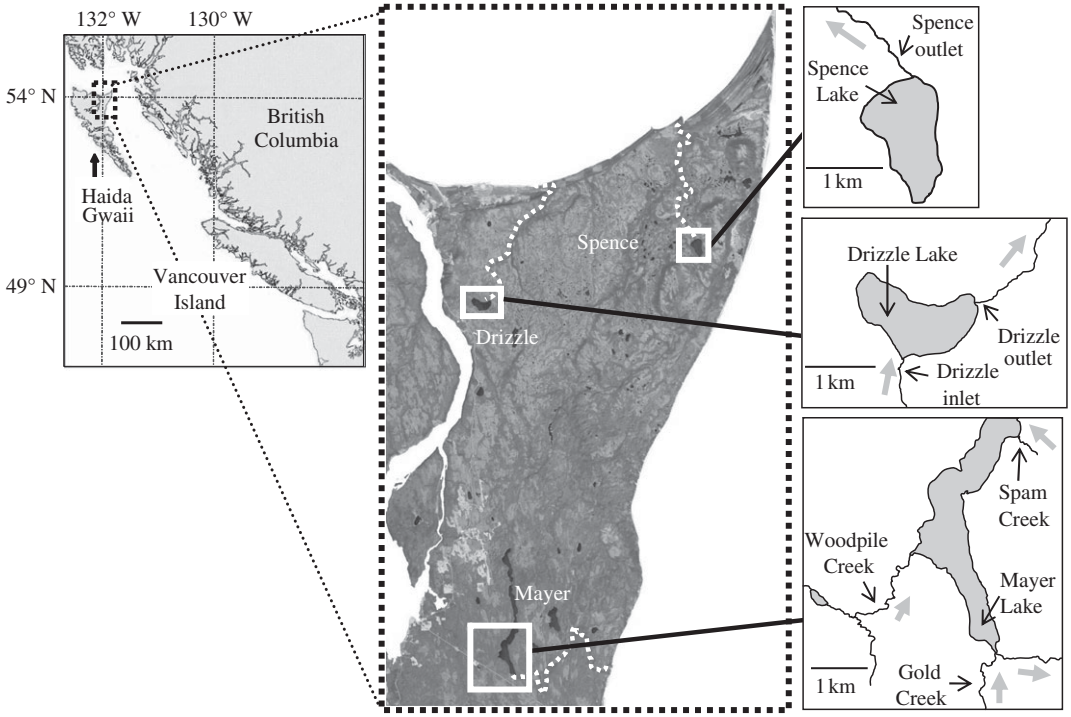

Figure 1. Map showing location of study populations. Insets show location of lake and stream sampling sites within watersheds. White dotted lines trace path of outlet streams.

of stickleback inhabiting adjacent streams and lakes are particularly well studied and show varying levels of habitat-specific morphological adaptations, with some distinct pairs providing convincing examples of ecological speciation [19–24]. Morphological differences between parapatric stream–lake sticklebacks were originally described in Mayer Lake [19] and Drizzle Lake [20] on the Haida Gwaii Archipelago. In these two systems, the divergence between lake and stream fish is remarkably parallel; lake fish have higher gill raker counts, a larger more streamlined body, longer spines and an unusual melanistic coloration [19,20,25]. There is evidence that the differences are adaptations to both divergent predation and trophic selective landscapes [18,20,26]. The most exceptional feature of stickleback in these lakes is their large body size. Among more than 100 Haida Gwaii lakes and streams surveyed, only eight include 'giant' stickleback (length greater than 75 mm) [25,27]. These giants include stickleback from Drizzle, Mayer and nearby Spence lakes—populations located on a post-glacial outwash plain dotted with dozens of lakes, ponds and streams with stickleback of typical body size [25].

Here, we evaluate genome-wide patterns of single nucleotide polymorphism (SNP) variation at greater than 1000 loci in stickleback from stream and lake habitats in Drizzle, Mayer and Spence drainage systems. While independent colonization and evolution seem likely [20], the flat topology, close proximity of populations and striking similarity of the fish make it difficult to rule out historical connections. SNP data were used to address two sets of questions: (i) are these stream–lake sticklebacks (and giant stickleback) independently derived? and (ii) what are the adaptive genetic differences between the stream–lake fish, and have parallel genomic changes occurred between systems?

2. MATERIAL AND METHODS

(a) *Study area and stickleback collection*

The three study systems are located on northeast Haida Gwaii, off the Pacific coast of Canada (figure 1). In the Drizzle system, sticklebacks were collected from the lake, and the only significant inlet and outlet streams [20,28]. Mayer system sticklebacks were collected from the lake and three inlet streams [19,25], and sticklebacks in the Spence system were collected from the lake and outlet. Collections were made in May–June of 2009 and 2010 using minnow traps and fish preserved in 95 per cent ethanol. When numbers permitted, 20 fish were arbitrarily selected (avoiding juveniles less than 40 mm) for genotyping and morphological analysis (table 1). In streams flowing into Mayer Lake, sticklebacks morphologically similar to those from Mayer Lake were captured alongside typical stream-form fish (differentiated based on colour, shape and size; photo in the electronic supplementary material, figure S1). In these inlet streams, 20 of the stream-form fish were studied along with additional lake-form fish (table 1).

(b) *Morphological analysis*

We measured six metric and three meristic traits: standard length, body depth, first dorsal spine length, left pelvic spine length, gape length, eye diameter, gill raker number on first left branchial arch (upper and lower arms) and number of lateral plates. All metric traits were size standardized to allow size-independent comparisons. This was accomplished by fitting a general linear model (GLM) for each trait with standard length as a covariate and population as a factor. Standard length by population interaction was non-significant for all traits ($p > 0.12$), we therefore used population coefficients from GLMs fit without interaction to calculate population-specific expected values for each trait corresponding to a length of 60 mm [29]. Standardized

Table 1. List of study populations, sample sizes and pairwise F_{ST} between populations within watersheds.

location	number of fish with SNP data	H_{obs}[a]	ID	pairwise F_{ST}[a]		
				DrizOut	DrizIn	
Drizzle Lake	19	0.234	DrizLk	0.15	0.192	
Drizzle outlet	20	0.224	DrizOut		0.082	
Drizzle inlet	9	0.215	DrizIn			
				Gold	Wood	Spam
Mayer Lake	18	0.259	MayLk	0.079	0.091	0.092
Gold Creek	16 + 8 LF[b]	0.229	Gold		0.049	0.093
Woodpile Creek	16 + 3 LF[b]	0.232	Wood			0.074
Spam Creek	19 + 4 LF[b]	0.231	Spam			
				SpOut		
Spence Lake	18	0.234	SpLk	0.082		
Spence outlet	17	0.283	SpOut			
Total	167					

[a]H_{obs} (observed heterozygosity) and F_{ST} based on all evenly distributed SNPs ($n = 760$).
[b]LF (lake-form) fish were captured in streams along with stream-form fish. They are morphologically and genetically like Mayer Lake fish ($F_{ST} = 0$).

trait values for each fish were calculated by adding a size-scaled residual (residual × (60/length)) to expected values.

Multi-variate morphological differentiation of the study populations was assessed using principal components analysis (PCA). Data from all variables (size standardized if appropriate) were scaled to have unit variance before calculation by a singular value decomposition of the matrix in R (v. 2.9.0) statistical software [30].

(c) *DNA extraction and SNP genotyping*

Stickleback genomic DNA was genotyped at 1536 biallelic SNP loci using Illumina's BeadArray Technology and GoldenGate assay (Illumina, San Diego, CA, USA) following the methodology described by Jones *et al.* [31]. SNPs were ascertained from two marine and three freshwater stickleback populations, geographically distant (greater than 800 km) from those in the current study [31]. GenomeStudio software (v. 2010.2; Illumina) was used to visualize and manually adjust all intensity clusters. SNPs with poorly separated clusters or low signals ($n = 342$) were excluded. In the exported data, SNPs missing greater than 10 per cent of genotypes calls were removed ($n = 24$) as were any stickleback with greater than 5 per cent missing data. Repeatability of genotype calls was greater than 99 per cent in two samples run in triplicate. The final dataset comprised 1170 SNPs from 167 sticklebacks (table 1).

The SNPs cover all 21 stickleback linkage groups, mtDNA and unassembled scaffolds (electronic supplementary material, table S1). They fall into three groups [31]: (i) SNPs chosen to be evenly distributed across the genome based on local recombination rate ($n = 773$); (ii) assembly SNPs, chosen to tag unoriented or unassembled parts of the genome ($n = 117$); and (iii) candidate SNPs chosen to target specific genomic regions of interest (primarily regions differentiated between marine and freshwater populations or potentially linked to traits of interest; $n = 280$).

(d) *Population differentiation based on SNP data*

We used PCA and tree-based clustering methods to evaluate structure in the genetic data, using all evenly distributed SNPs except sex-linked loci (760 loci). We also re-ran analyses with outlier loci removed ($n = 27$, defined with a low stringency Bayesian prior of 1, see below). PCA has been used extensively in analysis of SNP data as an unsupervised clustering method to identify population structure [32]. Since PCA requires a dataset without missing values, we filled in the less than 1 per cent missing entries in our final SNP dataset by randomly sampling genotype data for the particular locus (across all localities). This conservative approach homogenizes genotype frequencies across populations and re-sampling had little effect on the PCAs. For tree-based analysis, we calculated F_{ST} (with sample size correction) and used the program PopTree2 [33] to produce neighbour-joining (NJ) trees based on population allele frequencies. Alternate genetic distance measures (Nei's DA and Nei's standard genetic distance DST [33]) produced congruent results (data not shown). Finally, we constructed individual-based distance trees in MEGA [34]. In this case, we created an artificial nucleotide sequence by concatenating all diploid SNP data (missing data coded as N) for each individual and calculated a pairwise uncorrected P distance matrix (equivalent to allele sharing distance), then produced a NJ tree.

(e) *Outlier detection*

We performed a genome scan for F_{ST} outliers using the Bayesian approach implemented in BayeScan v. 2.01 [6,8,9]. BayeScan estimates the probability that a given SNP is under selection by calculating the posterior odds (POdds), which is the ratio of the posterior probabilities of two models (selection/neutral) for each locus, given the allele frequency data [8,9]. Analyses were carried out separately for each of the physically isolated watersheds, rather than a global analysis [4,9]. Initially, the prior probability of the model with selection was set at 1/10 (assumes *a priori* that the neutral model is 10 times more likely than the model including selection). We also ran analyses with a prior probability of 1 allowing identification of less-stringently defined outliers. Outliers identified with a prior of 1 were excluded in some analyses (see §2*d*) and only reported when detected in multiple independent genome scans (see §3). Default parameters were used in all BayeScan runs. To define outliers, the expected false discovery rate was kept constant (less than 0.05) and the POdds threshold defining outliers varied correspondingly [6].

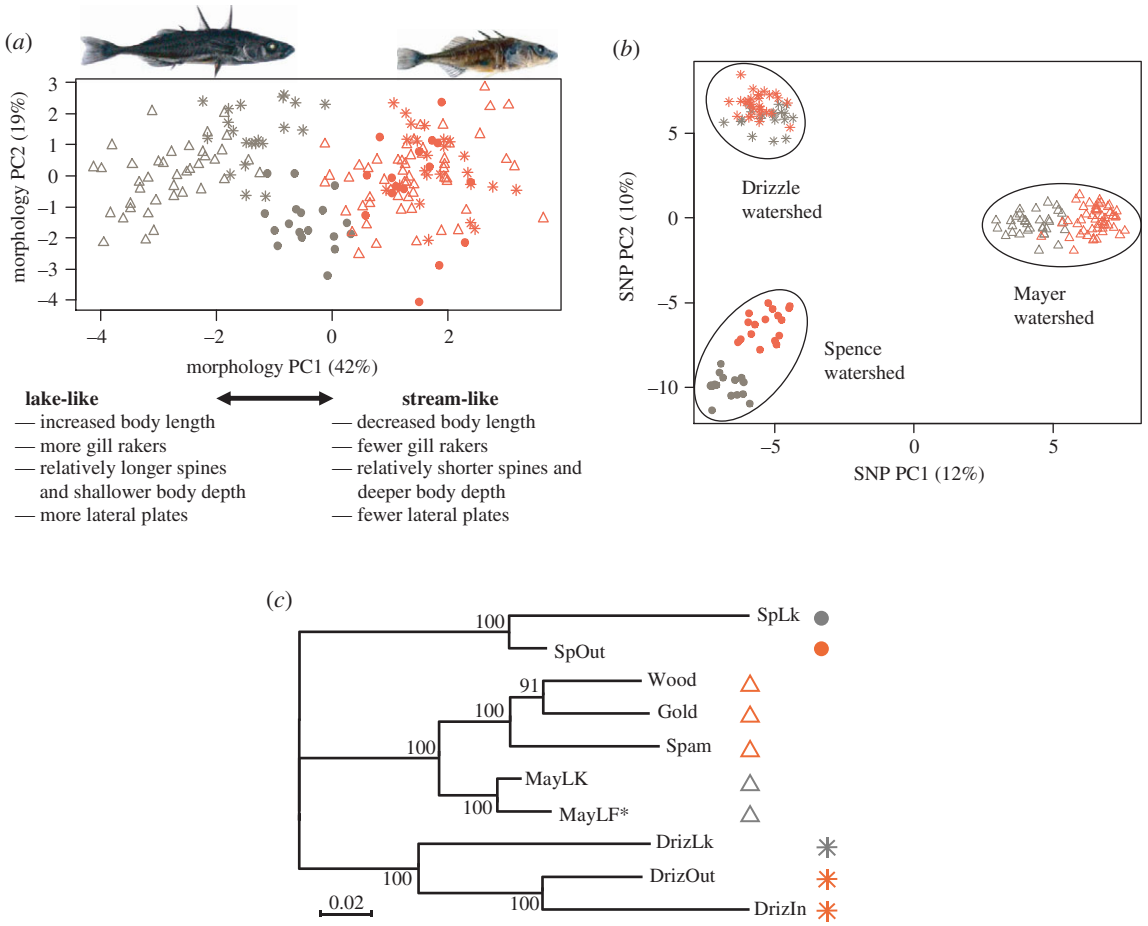

Figure 2. Differentiation of stream–lake stickleback based on morphological versus genetic data. Colour distinguishes stream (orange) and lake (grey) fish, symbols identify watershed. (*a*) First two principal components (PCs) from nine morphological variables (typical Mayer Lake and Gold Creek fish are shown) (*b*) first two PCs from 760 SNPs (evenly distributed, non-sex-linked loci). (*c*) Population-level neighbour-joining tree based on F_{ST} across the 760 loci. Per cent bootstrap support (1000 replicates) shown at nodes. Removal of stream–lake outlier loci has little effect on PCA or tree (electronic supplementary material, figures S3 and S4).

3. RESULTS

(a) *Morphology*

In all three watersheds, lake fish have greater standard length, more gill rakers and longer size-adjusted pelvic spines compared with adjoining stream fish. Lake fish also have shallower body depth and longer dorsal spines in the Mayer and Drizzle systems, but not in the Spence system. The lateral plate number varies among watersheds, and was consistently lower in streams (not significantly in Drizzle). These results are consistent with larger morphological datasets collected from some sites previously [19,20,25,28] and confirm strong parallels in phenotypic divergence between stream and lake fish, especially in Mayer and Drizzle systems [20]. The morphology of lake-form stickleback found in Mayer watershed streams matched Mayer Lake stickleback except they had shallower body depth. Morphology data are summarized in electronic supplementary material, figure S2.

PCA of morphological variables differentiated fish from stream and lake habitats on the first axis (figure 2*a*). This axis accounted for 42 per cent of the variance and had large positive loadings for body depth combined with large negative loadings for standard length, number of lower gill rakers and length of spines. On PC1, there is no overlap between stream and lake fish collected from the same watershed. There is considerable overlap between the stream fish from separate watersheds, reflecting their similar overall morphology. PC2 (19%) has large loadings for gape length and eye diameter; these traits did not consistently differ between habitats or watersheds.

(b) *Population differentiation based on SNP data*

In these analyses, we used all evenly distributed, non-sex-linked SNPs (760 loci) and also with the sub-set of these identified as outlier loci excluded ($n = 27$; defined with a low stringency prior of 1). By definition, removal of divergent outliers generates lower genetic distances within watersheds, but this did not substantially change clustering results. We present data with outliers removed in electronic supplementary material, figures S3 and S4.

(i) *PC analysis*

PCA of the genetic data clearly separates stickleback into collection locations in a hierarchical fashion. The first two PCs separated stickleback into three clusters corresponding to watershed of origin (figure 2*b*). PC1 (12.7%) separates Mayer from Drizzle and Spence watersheds, which are in turn separated on PC2 (9.7%). In subsequent PCs, separation is seen between stream and lake stickleback in Drizzle (PC3), Mayer (PC4) and Spence (PC6). PC5 separates Spam Creek from other

Mayer creeks (see the electronic supplementary material, figure S3).

(ii) *Tree-based analysis*

Our population-level trees separated the three watersheds at the basal node with 100 per cent bootstrap support (figure 2*c*), reflecting the higher between watersheds F_{ST} (mean = 0.244) compared with within watersheds (mean = 0.098; table 1). The next nodes separate lake populations from stream populations within each of the watersheds (figure 2*c*; mean F_{ST} = 0.114). For pairwise F_{ST} with and without outliers, see the electronic supplementary material, table S2. Individual-based trees produce congruent results (electronic supplementary material, figure S4).

(c) *Stream–lake outlier loci*

All variable SNPs (including assembly and candidate SNPs) were used in outlier loci analyses. Genome scans were performed separately for each watershed with differing number of variable loci (Drizzle: $n = 864$, Mayer: $n = 917$ and Spence: $n = 966$). Initially, we used a prior probability of 10, and in the three comparisons identified 34 SNPs showing a pattern of differentiation indicative of divergent selection ('Prior10' outliers; table 2 and figure 3). The Prior10 outlier SNPs include 21 genomic regions (when SNPs less than 20 kb apart are grouped), three of the outlier regions were identified in multiple watersheds (two between Drizzle and Mayer, one between Drizzle and Spence; table 2). One of the regions (Chr4–19.8 Mb) contained a high density of candidate SNP markers with 14 individual outlier SNPs identified over a 90 kb block (table 2 and figure 3). The remainder of the Prior10 outliers were defined by a single SNP.

To identify additional loci under divergent selection in multiple watersheds, we set a less-stringent outlier threshold by lowering the BayeScan prior to 1. Under this criterion, the SNPs identified within each watershed may include some false positives; however, since each watershed is independent, the chance of an SNP being incorrectly picked as an outlier multiple times is minimal (no data were used in multiple genome scans, unlike many previous studies [5,6]). With the lower prior, the number of genomic regions identified as outliers in at least one watershed increased from 21 to 73. Of the additional 52 genomic regions, six were identified in multiple watersheds (table 2). All of the new shared outliers have Bayes factors of greater than 7 (corresponding to substantial evidence), most are greater than 10 (strong). One new outlier region (Chr19–14.8 Mb) contains two SNPs identified as outliers in all three watersheds. Several additional SNP loci were identified within the Chr4–19.8 Mb Prior10 outlier region, one of these is an outlier in all three watersheds (table 2).

Despite clear morphological separation of the stream–lake stickleback in some traits, fixed differences in allele frequencies between habitats were not observed (figure 3; and for all outlier loci allele frequencies see the electronic supplementary material, figure S5). For shared outliers, allele frequencies mostly (but not universally) diverged in the same direction between habitats. Within watersheds where multiple streams were sampled, divergence was between all streams versus lake fish for 12 of 15 outliers (i.e. in three cases one stream had the lake allele at a high frequency).

4. DISCUSSION

(a) *Origin of the stream–lake stickleback*

Our genome-wide data indicate that the stream–lake stickleback we examined originated in separate divergences within each of the three watersheds. This implies three origins of the 'giant' stickleback in this small area on Haida Gwaii. These events presumably represent independent selection on standing genetic variation present in marine ancestors since few new genetic mutations would have been expected since post-glacial colonization. It is possible that similar watershed-dominated genome-wide genetic structure could be produced by secondary gene flow within watersheds after a single origin and dispersal by a giant-like ancestor. However, this would require maintenance of habitat-specific morphological distinctiveness in the face of extensive and long-term homogenizing gene flow within each watershed. In either case, it is clear that habitat, rather than history, has played a deterministic role in shaping the current morphological diversity. The separate evolution of these stream–lake stickleback contrasts with those in Germany, where primary genetic divergence was among habitat type [35]; our results are consistent with genetic data from stream–lake stickleback on Vancouver Island (400 km to the south of Haida Gwaii) [24].

Given that the typical freshwater form of stickleback has evolved from the marine ancestor countless times throughout the stickleback distribution [13], the limited geographical distribution of giant lake stickleback that are highly distinctive from adjoining streams remains a conundrum. It may be that differences in predation regimes [18,26], or other biological and physical factors beyond the benthic–limnetic ecological contrast typically invoked [22,23], are required to drive divergence to the level observed in these Haida Gwaii populations. Or perhaps, the particular haplotypes under selection are restricted to this geographical area. While SNPs on the array are present elsewhere (since they were ascertained from other populations), they may be tagging haplotype variants that are restricted to Haida Gwaii. Gene flow via marine stickleback [11,15] could have facilitated movement of some genomic components between these closely located watersheds. In this scenario, while the forms are assembled independently, some of the same allelic variants may be used in each process (see §4*b* below).

There are no current physical barriers to dispersal between sampling sites within each system; this is highlighted by the presence of lake-form fish in the streams entering Mayer Lake. These 'lake' fish are genetically indistinguishable from lake-collected fish, suggesting they are not permanent stream residents. Lake-form fish were not identified in these creeks during previous sampling [19]. Despite this, they made up a substantial proportion (approx. 25%) of fish trapped in Mayer streams during the current study and presumably can have extensive ecological interactions with the stream-form. Sympatric coexistence of these two ecotypes is reminiscent of benthic–limnetic species pairs that coexist in some lakes [36]. Stream-form fish were not detected in

Table 2. Outlier loci identified in each watershed and corresponding posterior odds (POdds) from BayeScan [9]. Shared outliers were identified in at least two watersheds. Loci shown in bold were outliers in all three watersheds.

	outlier region	SNP group[a]	Drizzle POdds[b]	Mayer POdds[b]	Spence POdds[b]	genome region	benthic–limnetic[d]	marine–freshwater[d]
shared								
chrIV:12005099[c]	1	candidate	24[c]		24[c]	Chr4–12.0 Mb	—	no
chrIV:18425274[c]	2	even dist.	16[c]	23[c]		Chr4–18.4 Mb	no	no
chrIV:19812956	3-1	candidate	∞	∞		Chr4–19.8 Mb	—	yes
chrIV:19814842[c]	3-2	candidate	39[c]	61[c]		Chr4–19.8 Mb	—	yes
chrIV:19819889	3-3	candidate	∞	∞		Chr4–19.8 Mb	—	yes
chrIV:19826019	3-4	candidate	∞	∞		Chr4–19.8 Mb	—	yes
chrIV:19827176	3-5	candidate	∞	∞		Chr4–19.8 Mb	—	yes
chrIV:19856347	3-6	candidate	2498	9[c]		Chr4–19.8 Mb	—	yes
chrIV:19863404	3-7	candidate	332	178		Chr4–19.8 Mb	—	yes
chrIV:19872201	3-8	candidate	∞	80[c]		Chr4–19.8 Mb	—	yes
chrIV:19872520	3-9	candidate	∞	87[c]		Chr4–19.8 Mb	—	yes
chrIV:19881291	3-10	candidate	∞	66[c]		Chr4–19.8 Mb	—	yes
chrIV:19881370	3-11	candidate	∞	142		Chr4–19.8 Mb	—	yes
chrIV:19881515	3-12	candidate	∞	65[c]		Chr4–19.8 Mb	—	yes
chrIV:19890632	3-13	candidate	∞	69[c]		Chr4–19.8 Mb	—	yes
chrIV:19896811	**3-14**	candidate	**31**[c]	**262**	**18**[c]	Chr4–19.8 Mb	—	yes
chrIV:19906553	3-15	candidate	∞		24[c]	Chr4–19.8 Mb	—	yes
chrIV:26063824	4	candidate	39[c]	∞		Chr4–26.0 Mb	—	yes
chrVII:13205977[c]	5	candidate	48[c]	21[c]		Chr7–13.2 Mb	yes	yes
chrXIX:14796728[c]	6-1	candidate		54[c]	142[c]	Chr19–14.8 Mb	—	yes
chrXIX:14798132[c]	**6-2**	candidate	7[c]	**37**[c]	**84**[c]	Chr19–14.8 Mb	—	yes
chrXIX:14799088[c]	**6-3**	candidate	7[c]	**41**[c]	**118**[c]	Chr19–14.8 Mb	—	yes
chrXX:9279241	7	even dist.	184	1665		Chr20–9.3 Mb	no	no
chrXX:12622695[c]	8	assembly		8[c]	17[c]	Chr20–12.6 Mb	no	yes
chrUn:1279794	9	assembly	∞		2499	ChrUn-1	no	—
watershed-specific								
Drizzle								
chrIX:10468143	10	even dist.	95			Chr9–10.5 Mb	no	no
chrXI:7635920	11	candidate	40			Chr11–7.6 Mb	—	yes
chrXIX:3309372	12	even dist.	90			Chr19–3.3 Mb	yes	no
chrXX:12810044	13	assembly	124			Chr20–12.8 Mb	yes	yes
chrXX:13893619	14	even dist.	554			Chr20–13.9 Mb	—	no
chrUn:2154566	15	assembly	∞			ChrUn-3	no	—
chrUn:2632376	16	assembly	160			ChrUn-4	yes	—
Mayer								
chrII:14991358	17	even dist.		1665		Chr2–15.0 Mb	yes	no
chrIV:9220132	18	assembly		2499		Chr4–9.2 Mb	no	no
chrVIII:4503012	19	even dist.		178		Chr8–4.5 Mb	no	no
chrVIII:9768150	20	even dist.		832		Chr8–9.8 Mb	—	no
chrXX:232763	21	candidate		713		Chr20–0.2 Mb	—	no
chrXX:16912820	22	candidate		108		Chr20–16.9Mb	—	no
chrXXI:1893294	23	candidate		160		Chr21–1.9 Mb	no	no
chrUn:7381868	24	assembly		216		ChrUn-2	yes	—
Spence								
chrIV:23965307	25	candidate			624	Chr4–24.0 Mb	yes[e]	yes
chrVII:5936068	26	even dist.			171	Chr7–5.9 Mb	no	no

[a] *a priori* classification of SNPs (see §2 for details).
[b] Posterior odds for SNP being under divergent selection with a prior of 1; this is equivalent to Bayes factor.
[c] Only outliers when prior probability is set at 1 to identify shared outliers (see §2 for other parameters and justification).
[d] Outlier region also identified in genomes scans of benthic–limnetic [31] or marine–freshwater species pairs [12]. Dash indicates data not comparable owing to differences in markers. Marine–freshwater outliers [12] defined by overall comparison with elevated F_{ST} differentiation ($p \leq 10^{-5}$).
[e] chrIV:23937349 is an outlier in benthic–limnetic.

any lakes during the current study, and despite extensive sampling in Mayer and Drizzle lakes, they have only been reported near stream mouths in Mayer Lake [26]. In the Drizzle and Mayer systems (where we sampled multiple streams), stream-form fish are generally more genetically similar to each other than to lake fish, and this is a genome-wide effect that persists when outlier loci are removed. This indicates continued gene flow through the lake, perhaps facilitated by phenotype-dependent habitat preferences [37]. Why the 'benthic' stream-forms do not become established in these lakes is an intriguing question and remains a topic for future ecological investigations.

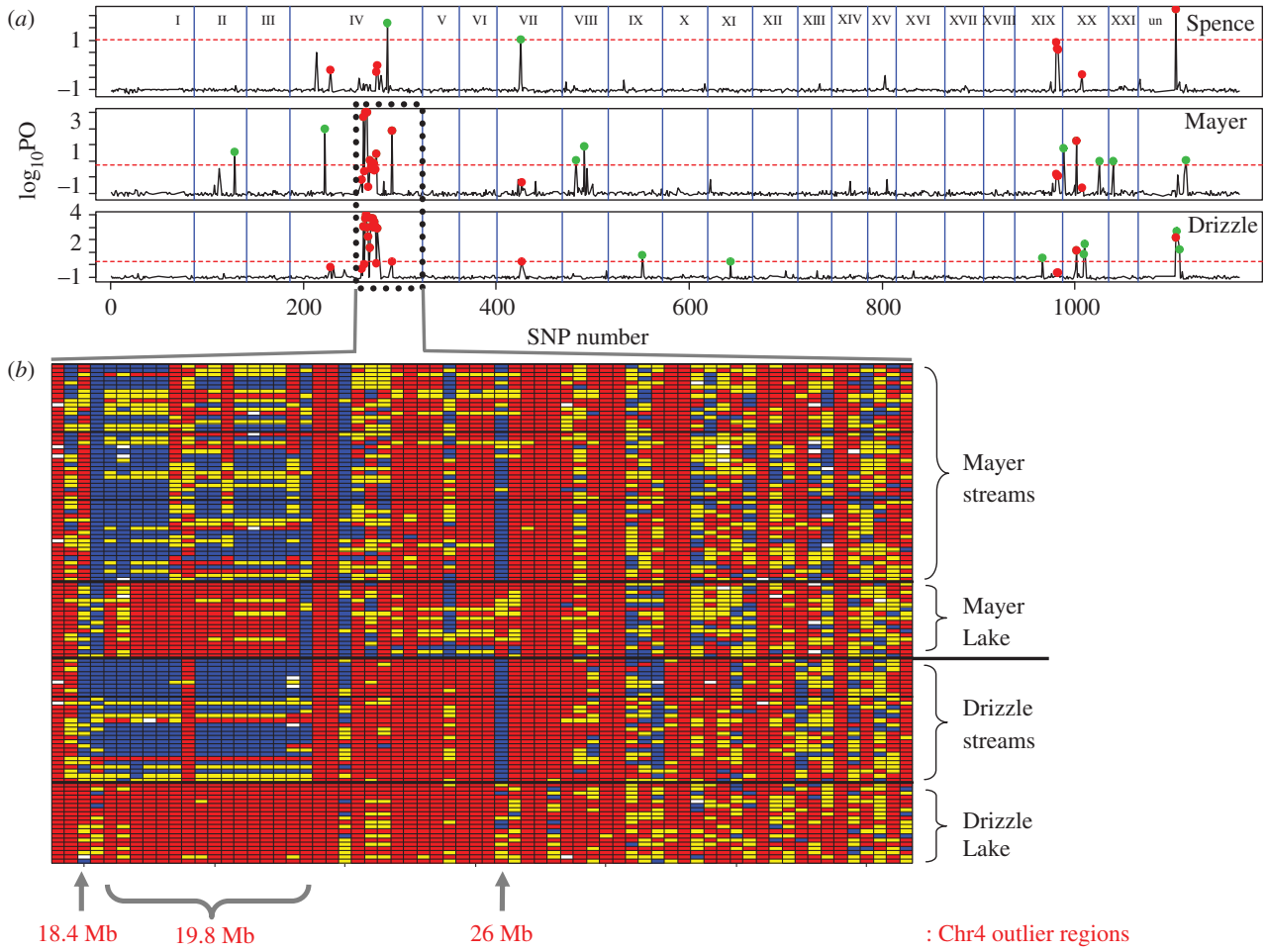

Figure 3. Genome scans in stream–lake sticklebacks. (*a*) Plots show posterior odds of each SNP marker for each of the three watersheds (BayeScan [9]; prior of 10). Vertical lines separate 21 chromosomes and unassembled scaffolds. Horizontal dotted lines show outlier thresholds corresponding to a false discovery rate of less than 0.05. SNPs with a red dot are outliers in multiple watersheds (including reduced stringency outliers) and green dots indicate watershed-specific outliers. (*b*) Image plot showing genotypes of SNPs located along part of chromosome 4 (15.7–32.6 Mb) for Mayer and Drizzle stickleback (red, AA; yellow, Aa; blue, aa). Genotypes were colour-coded, so markers homozygous for the most common allele in Drizzle Lake are red. Three shared outlier regions are labelled, including Chr4–19.8 Mb (a candidate region with a high density of SNPs).

(b) *Stream–lake outlier loci*

Genome-wide scans identified several genomic regions subject to divergent selection in each watershed. Given the density of markers used, the outlier SNPs are probably linked to selected haplotypes, rather than representing causative adaptive mutations. The proportion of adaptive genetic variation detected is dependent on both the number of markers and the level of linkage disequilibrium across the genome. Overall, our genomic coverage is roughly one SNP every 600 kb (450 Mb/760 evenly spaced SNPs) with some areas of higher coverage owing to the candidate SNPs on the array ($n = 280$). So, while our coverage is substantially higher than many previous scans of sticklebacks [38,39], ours is almost certainly not an exhaustive catalogue of the genomic loci under divergent selection. Overall, we identified 21 stringently defined outlier regions; eight of these were from the evenly distributed SNP group (i.e. approx. 1% of these non-candidate SNPs were outliers). This is low compared with 2–10% of loci generally found to be under diversifying selection [7]; however, comparisons between studies are difficult owing to differences in markers, outlier detection methodologies and number of samples.

Given the parallels in phenotypic differentiation between the stream and lake stickleback and the numerous examples of parallel evolution at the genetic level in recent stickleback literature [12,15,16], one might expect that a substantial proportion of outliers would be shared among watersheds. This is not the case; only three out of 21 stringent outlier regions (14%) were identified in more than one system (using a less-stringent outlier threshold, it is 9 out of 73; 12% shared). Including the less-stringently defined outliers, only two genomic regions are differentiated in all three watersheds. While not prevalent, these repeated genetic contrasts provide strong evidence for habitat-specific selection driving adaptive evolution at these loci.

The Chr4–19.8 Mb outlier region was covered by a relatively high density of SNP markers on the array. A closer look at this area highlights issues that can complicate interpretation of data from lower coverage genome scans. First, within this outlier region, there is often an imperfect association between SNPs and the presumed adaptive variant. For example, within the region in the Mayer system stickleback, four of the 15 SNPs show weak differentiation between stream–lake habitats

(figure 3*b*). This can happen when mutations accumulate within the outlier region, when recombination breaks down an ancestral region, or when selection acts on multiple ancestral alleles within the pool of standing genetic variation (i.e. a soft-sweep [40]). In the Spence system, we see evidence that recombination has disconnected the link between most SNPs in the Chr4–19.8 Mb region and the adaptive variant (only two SNPs are outliers). If we extrapolate these observations to the other SNPs in our dataset, it is apparent that some SNPs we have characterized may be within, or close to, a genome region under divergent selection, but are no longer diagnostically linked to the adaptive change.

Population-specific outliers are often discounted as potentially erroneous owing to non-repeatability [6,9]. However, in the current study, many of these are strongly supported and this divergence is being maintained despite the potential for gene flow. The occurrence of population-specific outliers makes sense for several reasons. First, ecological pressures are unlikely to be fully parallel, as demonstrated by slightly different morphological divergences seen between habitats in each stream–lake pair (see also [24]). Second, it is possible that for polygenic traits, different loci may be recruited. Third, particular adaptive alleles may be absent in some of the independently colonized watersheds. Finally, as discussed above, linkage between the adaptive variant and a specific outlier SNPs may have broken down in some cases. Given these possibilities, strongly supported population-specific outlier loci should not be discounted in subsequent studies characterizing adaptive genetic variation.

(c) *Comparison with outlier loci in other stickleback divergences*

In his initial description of the giant Mayer Lake stickleback, Moodie [19] pointed out that they are closer morphologically to the marine stickleback than to typical freshwater populations, but clarified that it was the 'character complex' that set this population apart rather than each particular character. This indicates that some characters (and associated genetic loci) distinguishing giant stickleback may be shared with ancestral marine populations and others with derived freshwater populations. We can explore this by comparing the stream–lake outliers we have identified with regions defined as outliers found in previous marine–freshwater [12] and benthic–limnetic comparisons [31].

Hohenlohe *et al.* [12] used a high-density genome scan to compare two marine and three freshwater Alaskan stickleback populations. They identified several candidate regions differentiating marine–freshwater stickleback and suggested this was a result of co-selection on multiple functionally related genomic regions. Many of these marine–freshwater outlier regions were also tagged by SNPs on our array, and eight out of 22 chromosomal regions we identified as stream–lake outliers are within candidate regions from marine–freshwater study [12] (table 2). This clearly shows that some of the genomic regions under divergent selection between marine and freshwater populations can be broken down and retained in some freshwater populations. These findings will focus the search for functionality of these particular genomic regions to traits that not only differ between marine and freshwater species pairs, but also between these stream–lake sticklebacks (e.g. osmoregulatory genes are unlikely to diverge between adjoining freshwater populations).

Since the phenotypic divergence seen in sympatric benthic–limnetic lake stickleback morphology mirrors that seen in stream versus lake in many ways [36], again we might expect common genomic regions to be involved. Based on a comparison with a genome scan on benthic–limnetic stickleback in three lakes using a similar set of SNP as used in the current study [31], we find seven genome regions that are outliers in at least one stream–lake pair and benthic–limnetic pair (table 2). None of the outlier regions that differentiated all stream–lake populations, or all benthic–limnetic pairs, is shared between studies. In fact, the maximum number of study systems that shared a common outlier region is three (out of the possible six). It may be possible to look for phenotypic commonalities between the stream–lake and benthic–limnetic pairs that do share outliers to suggest the possible underlying causes; however, with relatively low-density genome scans, much uncertainty remains.

(d) *Candidate genes and QTLs in stream–lake genomic outlier regions*

While having a reference genome sequence does allow outlier loci from genome scans to be connected with particular genomic regions, the identified outlier loci are based on an individual SNP (or small number of linked SNPs) potentially representing a region containing many genes. For example, within the Chr4–19.8 Mb outlier region, there are 15 outliers SNPs covering approximately 100 kb (approx. 19.81–19.90 Mb) in the Mayer and Drizzle systems, and the flanking SNPs are at 19.3 and 21.2 Mb. Therefore, this is potentially a 2 Mb region in which the adaptive variant could be found. In the Spence system, only two outlier SNPs were identified at the end of the region (at 19.89 and 19.90 Mb) narrowing the window to just over 1 Mb. Hohenlohe *et al.* [12] also identified marine–freshwater differentiation in this part of the genome (Chr4/LG IV Peak 2); a 1.1 Mb area centred at 20 Mb containing 31 protein coding genes (from which they listed two candidate genes possibly related to morphology: Wnt7B and FBLN1). These comparisons illustrate how data from multiple populations can help reduce the size of the candidate regions. However, further dedicated studies with full sequence will be required to examine fine-scale divergence in this region, and around other outliers, before strong candidate genes can be proposed (169 genes within 50 kb of all outlier loci identified are listed in the electronic supplementary material, table S3).

An alternative way to focus the search for genes under divergent selection is through comparison with quantitative trait loci (QTL) studies for known phenotypic differences. This has the advantage of allowing specific phenotypic traits to be linked to the genomic regions under divergent selection. Limitations are that mapping resolution is low in experimental crosses, the same loci may not always be used in different environments, and many interesting phenotypic differences have not yet been studied by QTL mapping. Nonetheless, several studies have been carried out in stickleback, which link markers to morphological traits that differ between the stream–lake sticklebacks. For example, in a cross between

benthic and limnetic sticklebacks, Peichel *et al.* [41] mapped the location of QTLs influencing length of the pelvic spines (marker: Chr8 17.7 Mb) and first dorsal spine (marker: Chr1 18.9 Mb; Chr2 20.2 Mb). Although stream–lake fish in the current study differ in dorsal and pelvic spine lengths, none of the outlier regions we identified map near the previously described QTLs. By contrast, variation in body depth has been linked to a QTL (marker: Stn321) in another cross between marine and freshwater sticklebacks [42], and this marker has been shown to be highly differentiated between a stream–lake stickleback pair using a candidate marker approach [37]. The body depth Stn321 QTL marker is located on Chr7 at approximately 13.66 Mb [42], approximately 450 kb from an SNP, we identified as an outlier in the Mayer and Drizzle systems. It is plausible that these two markers are detecting signals from a common gene controlling body depth, providing a strong candidate region for further characterization.

5. CONCLUSIONS

The population genomics approach we used to examine parallel morphological adaptation of stream–lake stickleback allowed us to establish the independent origin of stream and giant lake stickleback, in three geographically proximate watersheds. The majority of genomic outlier regions identified through genome scans were watershed-specific. However, several were shared between watersheds, and interestingly, several of the stream–lake outliers match those previously identified in marine–freshwater and benthic–limnetic comparisons. Further characterization of the shared regions we have identified will clarify whether the same genetic variants are found in all systems, or if the same loci have been altered in different ways. Our results are a first step to delve further into the search for genomic regions involved in morphological differentiation and reproductive isolation between stream and lake sticklebacks in this model system of ecological speciation. The large amount of standing genetic variation present in stickleback may distinguish this species from others that have undergone adaptive radiations, especially those originating from a small number of founders. However, the patterns of genetic changes observed in stickleback are likely to be mirrored in many other species undergoing rapid adaptation to new environments [43]. The divergences we have observed also emphasize the importance of ecological boundaries to differentiation in a broader context, especially since sharp gradients are probably just as widespread in terrestrial ecosystems.

Stickleback sampling followed guidelines for scientific fish collection in British Columbia, Canada (Ministry of Environment permits: SM09-51584 and SM10-62059). Sampling in Naikoon Provincial Park and Drizzle Lake Ecological Reserve were carried out under park use permits: 103171, 103172, 104795 and 104796.

Research supported by NSERC Discovery grant NRC 2354 (TER), grant P50 HG002568 from US National Institutes of Health (D.M.A. and D.M.K.), a NSERC Postdoctoral Fellowship (B.E.D.) and a Stanford Affymetrix Bio-X Graduate Fellowship (Y.F.C.). D.M.K. is an investigator of the Howard Hughes Medical Institute. Shannon Brady and HudsonAlpha staff facilitated laboratory analysis. John Taylor provided laboratory space and useful discussion.

REFERENCES

1 Schluter, D., Marchinko, K. B., Barrett, R. D. H. & Rogers, S. M. 2010 Natural selection and the genetics of adaptation in threespine stickleback. *Phil. Trans. R. Soc. B* **365**, 2479–2486. (doi:10.1098/rstb.2010.0036)
2 Gienapp, P., Teplitsky, C., Alho, J. S., Mills, J. A. & Merila, J. 2008 Climate change and evolution: disentangling environmental and genetic responses. *Mol. Ecol.* **17**, 167–178. (doi:10.1111/j.1365-294X.2007.03413.x)
3 Butlin, R. K. 2010 Population genomics and speciation. *Genetica* **138**, 409–418. (doi:10.1007/s10709-008-9321-3)
4 Wilding, C. S., Butlin, R. K. & Grahame, J. 2001 Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *J. Evol. Biol.* **14**, 611–619. (doi:10.1046/j.1420-9101.2001.00304.x)
5 Nosil, P., Egan, S. P. & Funk, D. J. 2008 Heterogeneous genomic differentiation between walking-stick ecotypes: 'isolation by adaptation' and multiple roles for divergent selection. *Evolution* **62**, 316–336. (doi:10.1111/j.1558-5646.2007.00299.x)
6 Fischer, M. C., Foll, M., Excoffier, L. & Heckel, G. 2011 Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (*Microtus arvalis*). *Mol. Ecol.* **2**, 1450–1462. (doi:10.1111/j.1365-294X.2011.05015.x)
7 Nosil, P., Funk, D. J. & Ortiz-Barrientos, D. 2009 Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* **18**, 375–402. (doi:10.1111/j.1365-294X.2008.03946.x)
8 Beaumont, M. A. & Balding, D. J. 2004 Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* **13**, 969–980. (doi:10.1111/j.1365-294X.2004.02125.x)
9 Foll, M. & Gaggiotti, O. 2008 A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977–993. (doi:10.1534/genetics.108.092221)
10 Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. 2003 The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* **4**, 981–994. (doi:10.1038/nrg1226)
11 Schluter, D. & Conte, G. L. 2009 Genetics and ecological speciation. *Proc. Natl Acad. Sci. USA* **106**, 9955–9962. (doi:10.1073/pnas.0901264106)
12 Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A. & Cresko, W. A. 2010 Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* **6**, e1000862. (doi:10.1371/journal.pgen.1000862)
13 Bell, M. A. & Foster, S. A. 1994 *The evolutionary biology of the threespine stickleback*. Oxford, UK: Oxford University Press.
14 Wooton, R. 1976 *The biology of the stickleback*. London, UK: Academic Press.
15 Colosimo, P. F. *et al.* 2005 Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* **307**, 1928–1933. (doi:10.1126/science.1107239)
16 Chan, Y. F. *et al.* 2010 Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx*1 enhancer. *Science* **327**, 302–305. (doi:10.1126/science.1182213)
17 Hagen, D. W. & Gilbertson, L. G. 1972 Geographic variation and environmental selection in *Gasterosteus aculeatus* L in Pacific Northwest, America. *Evolution* **26**, 32–51. (doi:10.2307/2406981)
18 Reimchen, T. E. 1994 Predators and morphological evolution in threespine stickleback. In *The evolutionary biology of the threespine stickleback* (eds M. A. Bell & S. A. Foster), pp. 240–276. Oxford, UK: Oxford University Press.
19 Moodie, G. E. E. 1972 Morphology, life-history, and ecology of an unusual stickleback (*Gasterosteus aculeatus*)

in Queen Charlotte Islands, Canada. *Can. J. Zool.* **50**, 721–732. (doi:10.1139/z72-099)
20 Reimchen, T. E., Stinson, E. M. & Nelson, J. S. 1985 Multivariate differentiation of parapatric and allopatric populations of threespine stickleback in the Sangan River watershed, Queen Charlotte Islands. *Can. J. Zool.* **63**, 2944–2951. (doi:10.1139/z85-441)
21 Lavin, P. A. & McPhail, J. D. 1993 Parapatric lake and stream sticklebacks on northern Vancouver Island: disjunct distribution or parallel evolution? *Can. J. Zool.* **71**, 11–17. (doi:10.1139/z93-003)
22 Hendry, A. P., Taylor, E. B. & McPhail, J. D. 2002 Adaptive divergence and the balance between selection and gene flow: lake and stream stickleback in the Misty system. *Evolution* **56**, 1199–1216.
23 Berner, D., Adams, D. C., Grandchamp, A. C. & Hendry, A. P. 2008 Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *J. Evol. Biol.* **21**, 1653–1665. (doi:10.1111/j.1420-9101.2008.01583.x)
24 Berner, D., Grandchamp, A. C. & Hendry, A. P. 2009 Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution* **63**, 1740–1753. (doi:10.1111/j.1558-5646.2009.00665.x)
25 Moodie, G. E. E. & Reimchen, T. E. 1976 Phenetic variation and habitat differences in *Gasterosteus* populations of Queen Charlotte Islands. *Syst. Zool.* **25**, 49–61. (doi:10.2307/2412778)
26 Moodie, G. E. E. 1972 Predation, natural selection and adaptation in an unusual threepine stickleback. *Heredity* **28**, 155–167. (doi:10.1038/hdy.1972.21)
27 Reimchen, T. E. 1988 Inefficient predators and prey injuries in a population of giant stickleback. *Can. J. Zool.* **66**, 2036–2044. (doi:10.1139/z88-299)
28 Stinson, E. M. 1983. Threespine sticklebacks (*Gasterosteus aculeatus*) in Dizzle Lake and its inlet, Queen Charlotte Islands. MSc thesis, University of Alberta, Edmonton, Canada.
29 Reist, J. D. 1986 An empirical-evaluation of coefficients used in residual and allometric adjustment of size covariation. *Can. J. Zool.* **64**, 1363–1368. (doi:10.1139/z86-203)
30 R Development Core Team 2009 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
31 Jones, F. C. *et al.* Submitted. A Genome-wide SNP genotyping array reveals patterns of global and species pair divergence in sticklebacks.
32 Patterson, N., Price, A. L. & Reich, D. 2006 Population structure and eigenanalysis. *PLoS Genet.* **2**, 2074–2093. (doi:10.1371/journal.pgen.0020190)
33 Takezaki, N., Nei, M. & Tamura, K. 2010 PopTree2: Software for constructing population trees from allele frequency data and computing other population statistics with Windows interface. *Mol. Biol. Evol.* **27**, 747–752. (doi:10.1093/molbev/msp312)
34 Tamura, K., Dudley, J., Nei, M. & Kumar, S. 2007 MEGA4: molecular evolutionary genetics analysis (MEGA) v. 4.0. *Mol. Biol. Evol.* **24**, 1596–1599. (doi:10.1093/molbev/msm092)
35 Reusch, T. B. H., Wegner, K. M. & Kalbe, M. 2001 Rapid genetic divergence in postglacial populations of threespine stickleback (*Gasterosteus aculeatus*): the role of habitat type, drainage and geographical proximity. *Mol. Ecol.* **10**, 2435–2445. (doi:10.1046/j.0962-1083.2001.01366.x)
36 McPhail, J. D. 1994 Speciation and the evolution of reproductive isolation in the sticklebacks (*Gasterosteus*) of southwestern British Columbia. In *The evolutionary biology of the threespine stickleback* (eds M. A. Bell & S. A. Foster), pp. 399–437. Oxford, UK: Oxford University Press.
37 Bolnick, D. I., Snowberg, L. K., Patenia, C., Stutz, W. E., Ingram, T. & Lau, O. L. 2009 Phenotype-dependent native habitat preference facilitates divergence between parapatric lake and stream stickleback. *Evolution* **63**, 2004–2016. (doi:10.1111/j.1558-5646.2009.00699.x)
38 DeFaveri, J., Shikano, T., Shimada, Y., Goto, A. & Merila, J. 2011 Global analysis of genes involved in freshwater adaptation in threespine sticklebacks (*Gasterosteus aculeatus*). *Evolution* **65**, 1800–1807. (doi:10.1111/j.1558-5646.2011.01247.x)
39 Makinen, H. S., Cano, M. & Merila, J. 2008 Identifying footprints of directional and balancing selection in marine and freshwater three-spined stickleback (*Gasterosteus aculeatus*) populations. *Mol. Ecol.* **17**, 3565–3582. (doi:10.1111/j.1365-294X.2008.03714.x)
40 Barrett, R. D. H. & Schluter, D. 2008 Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44. (doi:10.1016/j.tree.2007.09.008)
41 Peichel, C. L., Nereng, K. S., Ohgi, K. A., Cole, B. L. E., Colosimo, P. F., Buerkle, C. A., Schluter, D. & Kingsley, D. M. 2001 The genetic architecture of divergence between threespine stickleback species. *Nature* **414**, 901–905. (doi:10.1038/414901a)
42 Albert, A. Y. K., Sawaya, S., Vines, T. H., Knecht, A. K., Miller, C. T., Summers, B. R., Balabhadra, S., Kingsley, D. M. & Schluter, D. 2008 The genetics of adaptive shape shift in stickleback: pleiotropy and effect size. *Evolution* **62**, 76–85.
43 Elmer, K. R. & Meyer, A. 2011 Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends Ecol. Evol.* **26**, 298–306. (doi:10.1016/j.tree.2011.02.008)