# SUPPORTING INFORMATION

## The genomic signature of parallel adaptation from shared genetic variation

Marius Roesti, Sergey Gavrilets, Andrew P. Hendry, Walter Salzburger, Daniel Berner

(Molecular Ecology, 2014)

Includes:

# Supporting Methods

## Methods S1. Simulation models of parallel adaptation from shared genetic variation

*General model of parallel adaptation from shared genetic variation*

We developed individual-based models in which multiple populations diverge independently from the same source population into a selectively novel environment. This scenario was inspired by threespine stickleback, a species where numerous populations in freshwater environments have been founded from a common marine source population, but is likely relevant to many other biological systems (e.g., Terai et al. 2006; Renaut et al. 2011; Tennessen & Akey 2011; Domingues et al. 2012; Nadeau et al. 2012; Gross & Wilkens 2013; Streisfeld et al. 2013). For consistency with our empirical study (see below), we model eight derived populations. Individuals are monoecious and represented by a single haploid chromosome. A locus with two alleles under divergent selection between the environments is located in the center of that chromosome. The ancestral allele '0' is favored in the environment of the source population whereas the derived allele '1' is favored in the novel environment. The selected locus is flanked on each side by 100 evenly spaced and selectively neutral loci, in analogy to single nucleotide polymorphisms (SNPs) used in genome scans. Among the $n$ colonizers initially founding each of the derived populations, one individual has a haplotype represented by a uniform sequence of 1's. The other colonizers and the source population display the ancestral 0 allele at the selected locus and 0 and 1 alleles drawn at random with equal probability at the neutral loci. We thus explicitly assume that the derived allele at the selected locus is initially embedded in a specific genetic background shared among the derived populations (details on this assumption are discussed in the Discussion S1). Because our interest is in the early stages of population divergence and because the freshwater stickleback populations used for our empirical work are young (postglacial), our models ignore novel mutation.

After initial colonization, each derived population grows according to the Beverton-Holt model in non-overlapping generations (Kot 2001). Specifically, the number of offspring produced by each female is taken from a Poisson distribution with parameter $\frac{b}{1+(b-1)\frac{N}{wK}}$, where $b$ is the expected number of offspring (set to 10 in all simulations), $N$ is the current population size, $K$ is the environment's carrying capacity, and $w$ is the female's fitness. For computational efficiency, we choose $K$ = 1,000, emphasizing that additional exploratory simulations with $K$ = 10,000 produce similar results supporting identical conclusions. Females with the ancestral 0 or the derived 1 allele have a fitness of $w = 1 - s$ and 1 in the novel environment, where $s$ represents the strength of divergent selection between the environments. Males are assigned to females at random. During

reproduction, the female and male chromosomes recombine. The number of recombination events is drawn independently for each offspring from a Poisson distribution with parameter $R$. Recombination occurs with uniform probability across the chromosome. In the beginning of each generation, the derived populations each receive $N_m$ migrants from the source population. After $t$ generations, we calculate the magnitude of population divergence ($F_{ST;}$ Weir & Cockerham 1984) at all neutral loci (we never calculate divergence at the selected locus itself), including all $K$ individuals from all focal populations. ($F_{ST}$ is calculated globally; however, averaging across pairwise population comparisons produced similar results.) The resulting values are averaged across 100 replicate simulations for every parameter combination.

*Parameter space and modeling scenarios*

The default parameterization of our model is tailored to empirical data from the Ectodysplasin (*Eda*) locus in threespine stickleback, the genomic region where the observation of twin peaks flanking a divergence valley (peak-valley-peak) stimulated our hypothesis of a novel signature of adaptation from shared genetic variation (Roesti et al. 2012a). The default settings include $s$ = 0.2 (Barrett et al. 2008), $R$ = 0.05 (Roesti et al. 2013), and $t$ = 5000 (Bell & Foster 1994). With the default recombination rate of 0.05, the simulated chromosome approximates a 10 - 15 megabase (Mb) segment harboring *Eda* on chromosome IV. Phylogenetic evidence from the *Eda* locus justifies our modeling of the derived allele in a single shared genetic background at the onset of the simulations: present-day freshwater stickleback populations still share nearly identical haplotypes at *Eda*, even across continents (Colosimo et al. 2005; Berner et al. 2010b). We further assume $n$ = 100.

Modifications of the default model are used to explore the influence of each parameter on the molecular signatures of adaptation. First, we track population divergence between the *source* population and the derived populations over time ($t$ = 100, 200, 500, 1000, 2500, 5000). These comparisons represent the standard ecological genome scan and hence can serve to validate our general simulation approach. In all subsequent simulations, divergence is calculated among the *derived* populations. Here, we first set $N_m$ = 0 to study how divergence builds up over time in the absence of gene flow. In reality, however, gene flow will often occur between source and derived populations in the early stages of divergence (Wu et al. 2001; Nosil et al. 2009; Feder et al. 2012). Our main modeling effort is therefore devoted to divergence with gene flow, exploring all possible combinations of $N_m$ (1, 5, 10, 15; default = 5), $t$ (100, 200, 500, 1000, 2500, 5000), $n$ (50, 100, 200, 400), $s$ (0.05, 0.1, 0.2, 0.5), and $R$ (0.01, 0.02, 0.05, 0.1).

Finally, we modify the default model to include two selected loci located at equal distances *d/2* from the center of the chromosome, which now harbors 400 total neutral loci. The two derived alleles (one per selected locus) beneficial in the derived environment are initially linked (i.e., within a single neutral background), although they rapidly become dissociated by recombination when their frequency in the derived populations is still low. We perform simulations with different values of $d$ (350, 300, 250, 200, 100, 50, 20, 10) and maintain $N_m = 10$ throughout. To achieve a similar overall selection strength as in the single-locus model, we set $s = 0.1$ for each selected locus. Divergence is calculated at $t = 400, 800, 1200, 1600, 2000$. All other parameter values are the same as in the default single-locus model.

**Methods S2. Stickleback populations for empirical investigation**

Our study uses stickleback samples from two marine ('M') sites and from a lake and stream (freshwater, 'FW') site within each of four independently colonized watersheds (Boot, Joe's Misty, Robert's) on Vancouver Island, British Columbia, Canada (Fig. 1). Sample size was 27 individuals per site. The FW populations are identical to those studied in Roesti et al. 2012a. The M fish were collected with minnow traps from the Cluxewe River estuary (50° 36' 42" N, 127° 11' 02'' W) and the Sayward River estuary (location described in Berner et al. 2010a) on the east coast of Vancouver Island. All our estuarine individuals exhibited full plating along their body and a caudal keel, clearly identifying them as M fish (Bell & Foster 1994). In general, marine stickleback are phenotypically highly stable over space and time, exhibit large population sizes, and show little genetic structure over large geographic distances (Bell & Foster 1994; Walker & Bell 2000; Hohenlohe et al. 2010). Present-day marine stickleback are thus considered good surrogates for the ancestor of recently established FW populations (e.g., Walker & Bell 2000; Berner et al. 2010a). Consistent with this view, the two M samples in the present study did not appreciably differ genetically in any of our analyses: first, haplotype data showed no structure between the two M samples (data not shown). Second, median $F_{ST}$ between the two M samples was zero in the genome-wide analysis (mean and median $F_{ST}$ values for all pairwise population comparisons are presented in Table S3). We therefore pooled the two M samples for the haplotype network analysis (Fig. 3).

**Methods S3. Stickleback candidate genes for parallel M-FW divergence**

To empirically validate the signature of adaptation from shared genetic variation discovered in the simulations, we required loci showing clear signs of parallel divergence in stickleback. We thus focused on three genes suggested to be under strong divergent selection between M and FW

environments. The first candidate gene was *Eda* (Ectodysplasin). M stickleback have a complete plate row along their body, whereas FW populations typically display greatly reduced plating (Bell & Foster 1994). This divergence is thought to primarily reflect differential exposure to predation between the two environments (Reimchen 1992, 1994; Marchinko 2009) and is driven mainly by the repeated fixation of a derived *Eda* allele shared among FW populations (Colosimo et al. 2005, Berner et al. 2010b). Despite selection for the fully plated phenotype (and thus the ancestral M *Eda* allele) in the ocean, individuals heterozygous at *Eda* do still occur at low frequency (Barrett et al. 2008) in the ocean due to recurrent introgression of derived alleles from FW populations (Colosimo et al. 2005; Schluter & Conte 2009). We sequenced a (mainly intronic) 640 bp segment of *Eda* (Table S1).

The second candidate gene was *Atp1a1* (sodium pump subunit alpha-1). This gene is involved in the maintenance of the ion balance and electrolyte homeostasis in different osmoregulatory epithelia (Evans et al. 2005), and has been identified as a physiological key gene in the adaptation to different osmotic environments in many fish species (e.g., stickleback: Hohenlohe et al. 2010; McCairns & Bernatchez 2010; DeFaveri et al. 2011; Shimada et al. 2011; Jones et al. 2012a; killifish: Scott et al. 2004; bull shark: Reilly et al. 2011; brown trout: Larsen et al. 2008; whitefish: Renaut et al. 2011; reviewed in McCormick 2011). We sequenced a (mainly intronic) 380 bp segment of *Atp1a1* (Table S1).

The third candidate gene was *Spg1* (Spiggin). *Spg1* produces a glue-like protein in the kidneys of male stickleback used to stick nesting material together (Wootton 1976; Jakobsson et al. 1999). This glue seems under divergent selection between M and FW environments because of its sensitivity to salinity, pH, and/or temperature (Kawahara & Nishida 2007), and because strong allele frequency shifts between M and FW stickleback have been found at genetic markers in the close neighbourhood of the gene (Hohenlohe et al. 2010; DeFaveri et al. 2011; Shimada et al. 2011). We sequenced a 356 bp segment of *Spg1* (Table S1)**.** This segment was intergenic but directly adjacent to one of the *Spg1* gene copies.

For each of the three candidate genes, we performed Sanger sequencing (see Methods S4), screened these sequences for polymorphisms, and derived haplotype networks (see Methods S5). We then followed the same steps to Sanger sequence an additional 'reference locus' (mainly intergenic, length ranging from 326 – 767 bp) approximately one megabase away from each candidate gene. We predicted that if adaptation to the replicate derived FW environments at each candidate gene occurred through the parallel fixation of a derived variant present at low frequency in a common M source, all lake and stream samples should form a cluster of closely related

haplotypes distinct from the M haplotypes at these loci. Moreover, if M-FW divergence occurred in the face of gene flow, such genealogical structure should not be seen at the three reference loci.

**Methods S4. Targeted Sanger sequencing at candidate genes and reference loci**

PCR amplification primers for the three candidate genes and their associated reference loci (i.e., six total DNA segments) were designed based on the improved assembly (Roesti et al. 2013) of the stickleback reference genome (Jones et al. 2012b), and based on RAD sequences available from previous work (Roesti et al. 2012a). The primer sequences and amplification conditions are provided in Table S1. The resulting sequences were read on an ABI3130*xl* capillary sequencer (Applied Biosystems). Each sequence was run at least twice for each individual, usually with both the forward and reverse primer. This allowed unambiguously identifying the diploid genotype of each individual at each candidate gene and reference locus. On average, each candidate and reference locus was sequenced in 64 FW individuals (128 haplotypes), averaging eight fish per FW sample, and in 23 M individuals (46 haplotypes), including fish from both M samples.

**Methods S5. Haplotype genealogies for candidate genes and reference loci**

To construct haplotype genealogies for the candidate genes and reference loci, we first used CodonCode Aligner v.3.5.6 (CodonCode Corporation) to call diploid consensus sequences and to find SNPs. All polymorphisms were then concatenated (treating indels as a single mutational steps) and phased using PHASE 2.1 (Stephens et al. 2001; Stephens & Donnelly 2003), optimizing the procedure by specifying the polymorphisms' physical positions. Finally, we used jModelTest v0.1.1 (Posada 2008) to identify GTR as the best model of sequence evolution for all polymorphisms, used the maximum-likelihood method implemented in PAUP* v4.0 (Swofford 2003) to determine the most probable genealogical relationship among all individuals at each of the six loci, and visualized these haplotype genealogies following Salzburger et al. (2011).

**Methods S6. Broad-scale analyses around the candidate genes**

To generate broad-scale profiles of divergence and genealogical structure around the three candidate genes, we used consensus sequences from genome-wide RAD (Baird et al. 2008) loci previously generated for all 27 individuals from each of the eight FW samples (details on the wet lab and consensus genotyping protocols are given in Roesti et al. 2012a). We also generated new, comparable RAD data for the M samples based on the same wet lab protocol, with just two modifications: the final library amplification was performed in seven replicate PCRs to reduce amplification variance, and all 54 M individuals were single-end sequenced on a single Illumina

HiSeq lane with 100 cycles. For the M individuals, consensus genotype sequences at the RAD loci were called as in Mateus et al. (2013). After combining the consensus sequences across all M and FW individuals, each RAD locus was screened for SNPs, including a small fraction of micro-indels. All genomic positions in this study refer to the reference genome re-assembly of Roesti et al. (2013).

SNPs in the three 'candidate regions', defined as a 3 - 4 Mb segment around each gene, were used to quantify genetic divergence between M and FW stickleback ($F_{ST}$ based on haplotype diversity; equation 7 in Nei & Tajima 1981). Divergence was calculated for all possible pairwise comparisons between the two M samples and the eight FW samples (16 total comparisons). Robust divergence estimation was ensured by including a SNP only if both populations in a comparison contributed at least 27 nucleotides to the common nucleotide pool, and if the frequency of the minor allele across the nucleotide pool was at least 0.25. The latter criterion eliminated polymorphisms with low information content (Roesti et al. 2012b). In addition, we used only one SNP per RAD locus. Following these same conventions, we then calculated $F_{ST}$ for pairwise comparisons among the derived FW populations. We here considered comparisons among samples from ecologically similar FW environments only (i.e., six lake-lake and six stream-stream comparisons, for 12 comparisons in total). The rationale for excluding lake-stream comparisons was to avoid capturing selective signatures of lake-stream divergence, which is known to be strong (Berner et al. 2008, 2009; Deagle et al. 2012; Roesti et al. 2012a). However, analyses based on *all* possible FW comparisons produced very similar results supporting identical conclusions.

The interaction between selection and heterogeneous recombination rate along stickleback chromosomes can inflate population divergence in chromosome centers relative to their peripheries (Roesti et al. 2012a, 2013). Correcting for this effect by calculating *residual* divergence facilitates the search for signatures of selection (details given in Roesti et al. 2012a). This correction was performed here, although qualitatively similar conclusions emerged either way. Finally, to obtain overall M-FW and FW-FW divergence profiles, we averaged divergence estimates at each RAD locus (residual $F_{ST}$ values) across all pairwise M-FW and all pairwise FW-FW comparisons. This procedure yielded, on average, 6.9 and 6.4 replicate values per RAD site for the overall M-FW and FW-FW contrast. For the *Eda* candidate region (4 Mb in size), the final resolution was 178 and 168 data points for the overall M-FW and FW-FW comparison. The corresponding values for *Atp1a1* (4 Mb) were 193 and 187, and for *Spg1* (3 Mb) 106 and 100. Thus, the median and mean marker spacing in the candidate gene and control regions was 12 and 25 kb respectively (treating markers on sister RAD loci as individual data points).

Parallel divergence between source and derived environments based on shared variation

drives a divergence peak close to the selected locus in source-derived comparisons, but a valley in derived-derived comparisons (see Results). Calculating the *difference* between overall M-FW and FW-FW divergence, hereafter called 'delta divergence', should thus maximize the ability to detect genomic regions underlying parallel divergence (for a proof of principle using simulated data, see Fig. S1). We therefore complemented our standard divergence analyses described above by creating delta divergence profiles for each candidate region. We first averaged overall M-FW and FW-FW divergence separately across non-overlapping 5 kb windows, and then, for each window, we subtracted the resulting FW-FW value from its M-FW counterpart. Working with windows enhanced the power of this analysis because divergence data from *both* the M-FW and FW-FW comparison were not available from all RAD loci.

As a complementary approach to quantifying genetic divergence between M and FW stickleback, we assessed the extent of reciprocal M-FW monophyly captured by phylogenetic trees within the candidate regions. Specifically, we moved a sliding window across the SNPs and, for each window, calculated a distance matrix based on the 'F84' nucleotide substitution model (Felsenstein 1984). We here accepted multiple SNPs on a RAD locus and used a window size of 33 SNPs, which was the smallest number of markers consistently allowing distance matrix calculation across all windows. The genomic position of a window was defined as the RAD locus position of its central SNP. The distance matrices were then translated to midpoint-rooted neighbor joining trees, which in turn allowed calculating the genealogical sorting index (gsi; Cummings et al. 2008). This index ranges from 0 to 1 and quantifies the extent of exclusive ancestry of individuals from defined groups (here M and FW stickleback) in a phylogenetic tree. If multiple gsi values were available for a RAD locus (owing to multiple SNPs at that locus), they were averaged to a single data point. This analysis yielded 167, 178, and 103 gsi values for the *Eda*, *Atp1a1*, and *Spg1* candidate regions, thus resulting in a similar physical resolution as the $F_{ST}$-based divergence analysis. The gsi analysis was performed using the R (R Development Core Team 2013) packages APE (Paradis et al. 2004) and genealogicalSorting (http://www.genealogicalsorting.org).

# Supporting Discussions

**Discussion S1. How are our theoretical models influenced by more complex haplotype structure around the selected variant?**

Our simulations assume that the genetic variant adaptive in the derived populations has a *single* origin and is thus initially embedded in a single genetic background in all derived populations. Indeed, for the *Eda* locus inspiring our theoretical analysis, phylogenetic data have amply demonstrated extensive sharing among multiple FW populations of the *same* haplotype linked to a derived variant (Colosimo et al. 2005; top left haplotype network in Fig. 3A in this study). This strongly suggests a *single* origin of the derived variant (Colosimo et al. 2005). Our phylogenetic data from the two other candidate genes further indicate that this conclusion is not restricted to *Eda* (see Fig. 3A, middle and bottom left haplotype networks). Indeed, whole-genome re-sequencing supports the view that most of the genetic variation used for parallel FW adaptation has a common origin (Jones et al. 2012b). Extensive haplotype sharing among the derived populations is thus an adequate assumption in our models. We also highlight that with the simulation parameters chosen, the 'chromosome' in our models actually corresponds to a relatively narrow segment of a (stickleback) chromosome only.

Nevertheless, it should be kept in mind that the opportunity for a derived FW-adaptive variant to segregate in the M source population prior to selection (and hence to recombine into M genetic backgrounds) will influence the signature of parallel adaptation. Specifically, recombination of the derived variant in the source population will reduce the physical extent of haplotype sharing around the derived variant, eventually causing a *more narrow* divergence valley among the derived populations. This effect is analogous to the erosion of genetic divergence around a selected locus observed in soft sweep models focusing on ancestral versus derived populations (e.g., Hermisson & Pennings 2005; Barrett & Schluter 2008; Messer & Petrov 2013). We can thus make the qualitative predication that loci under strong divergent selection should exhibit a wider divergence valley than weakly selected loci. The reason is that in the former case, a derived variant introduced from a derived population back into the source population by hybridization will be eliminated relatively rapidly from the source population, thus reducing the opportunity for recombination. Similarly, loci situated in low-recombination regions of the genome should exhibit wider divergence valleys (see Discussion S2). We emphasize, however, that the emergence of the flanking *divergence twin peaks* is unaffected by the extent of haplotype sharing around the derived variant (see Fig. 6).

While our models assume a single origin of the derived variant, parallel adaptation among populations can also be based on *multiple* genetic variants produced independently by mutation

(Barrett & Schluter 2008; Messer & Petrov 2013). This scenario was not the focus of our investigation because parallel M-FW divergence from repeated *de novo* mutation is certainly not frequent in stickleback (see above). We can nevertheless make the qualitative predictions that, first, the availability of several independent derived variants in the source population prevents the emergence of a divergence valley among derived populations. This is because the selective sweeps will bring distinct haplotypes to fixation among the derived populations. Second, as above, the emergence of high divergence around the selected locus should still be observed because the barrier to gene flow mechanism operates irrespectively of the initial haplotype structure around the derived variant.

**Discussion S2. Objective and limitations of the genome-wide screen for signatures of parallel adaptation from shared variation.**

The goal of our genome-wide analysis was primarily to illustrate how the signature of parallel adaptation from shared variation can serve as a tool for the genome-wide detection of genes or chromosome regions involved in parallel adaptation – we did not attempt a complete quantitative investigation of the genetic architecture of M-FW divergence in stickleback. A first limitation is that our RAD marker data lack the physical (basepair) resolution to determine whether a selective signature is driven by a single gene, as opposed to multiple genes clustered within a few kb (our median and mean marker spacing is 14 kb, considering both sister tags associated with a restriction site). Nevertheless, our study exhibits an unprecedented *biological* resolution, as we include 8 FW and 2 M population samples, each represented by 27 individuals (Table S3). Overall FW-FW and M-FW divergence estimates at our SNP markers are thus exceptionally robust.

The power of detecting parallel adaptation regions is further complicated by heterogeneous recombination rate. As our models show, the genomic signature of adaptation from shared variation becomes physically more extensive (and hence easier to detect given limited marker resolution) with decreasing recombination rate ('Recombination' in Fig. 2C). Since recombination rate is much higher in the stickleback chromosome peripheries than in chromosome centres (Roesti et al. 2012a, 2013), we certainly overlook small-scale selective signatures in the chromosome peripheries. Although not widely appreciated, this bias potentially also affects other types of genome-wide scans relying on linkage (Roesti et al. 2013).

Furthermore, the selective signature at a locus is weakened when the same FW-beneficial variant is used for adaptation in a *subset* of the replicate FW populations only. This may occur because this allele simply failed to invade some FW watersheds, or because an adaptive phenotypic

change was achieved in some populations through a different genetic pathway. Like previous genomic analyses in the species (Hohenlohe et al. 2010; Jones et al. 2012a,b), our genome-wide screen is thus biased toward discovering signatures of M-FW divergence caused by alleles recycled *with high fidelity* among FW populations.

Finally, if a variant adaptive in FW managed to recombine effectively into diverse M genetic backgrounds prior to selection, we expect a narrow divergence valley only (see Discussion S1). Given relatively coarse marker resolution, this locus might thus escape our screen for the full signature of parallel adaptation from shared variation introduced in this paper (divergence valley *and* twin peaks). For all these reasons, the candidate regions identified in our genome-wide screen certainly represent only a subset of the M-FW adaptation genes in our study populations.

# Supporting Tables

**Table S1. Sanger sequencing of the stickleback candidate genes and their reference loci**

Amplification PCR reaction volume was 12.5 µl, with 1 µl of genomic DNA (concentration: 20 ng/µl) using RedTaq (Sigma-Aldrich) (default) or AmpliTaq (Applied Biosystems) polymerase. The following cycling conditions were used for PCR amplification: 1 x 94 °C for 3 min; followed by 30 x 94 °C for 30 sec, X °C for 45 sec and 72 °C for 45 sec; followed by 1 x 72 °C for 7 min and finally hold at 4 °C. Annealing temperatures (X) for particular primer pairs were (in °C): A/B=52.0, C/D=52.0, E/F=53.0, E/G=55.0, H/I=53.0, J/K=54.0, L/M=51.5, N/O=51.0, P/Q=53.5. Each PCR product was then purified by following the ExoSAP-IT (Affymetrix) standard protocol. For the sequencing PCR, we used the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) and added 0.5 µL primer (forward or reverse) and 1.0 µL BigDye Terminator Reaction Mix to each purified PCR product. The conditions for the subsequent sequencing PCR were: initial denaturation (1 min, 94 °C) followed by 25 cycles of denaturation (10 s, 94 °C), annealing (20 s, 52 °C) and elongation (4 min, 60 °C). Unincorporated BigDye terminators were removed with the BigDye XTerminator™ Purification Kit (Applied Biosystems), by adding 14.5 µL $ddH_2O$, 22.5 µL SAM™ solution and 5.0 µL XTerminator™ beads to the sequencing products. After shaking for 30 min at 2000 rpm, the mix was centrifuged (2 min, 1500 rpm).

| ID | Gene | Region (inferred from FW-FW comparisons using RADseq) | Chromosome | Genomic position (BROAD S1, Feb 2006; Database version: 68.1) | Genomic position (assembly-corrected reference; Roesti et al. 2013) | Direction | Primer sequence | Tm (°C) |
|---|---|---|---|---|---|---|---|---|
| A | *Eda* | candidate | 4 | 12808304 - 12808322 | 12808304 - 12808322 | Forward | 5'-GAC TGG AAG GGA AAG AAG G-3' | 55.3 |
| B | *Eda* | candidate | 4 | 12807800 - 12807820 | 12807800 - 12807820 | Forward | 5'-CTG CGC ACA GAG CGT AAA CAC-3' | 59.5 |
| C | *Eda* | candidate | 4 | 12807758 - 12807775 | 12807758 - 12807775 | Forward | 5'-CAC AAG AGC AGC GAG ACG-3' | 57.1 |
| D | *Eda* | candidate | 4 | 12809097 - 12809122 | 12809097 - 12809122 | Reverse | 5'-GGA ATT CTT TTG TTT TTG TCT TAT CG-3' | 53.4 |
| E | *Eda* | candidate | 4 | 12808548 - 12808569 | 12808548 - 12808569 | Reverse | 5'-CCT GTG AAG AGC GAA AGC AAA G-3' | 57.7 |
| F | *Eda* | reference | 4 | 11721754 - 11721770 | 11721754 - 11721770 | Forward | 5'-CAT CCA GGC CCA CAA TC-3' | 54.0 |
| G | *Eda* | reference | 4 | 11722645 - 11722664 | 11722645 - 11722664 | Reverse | 5'-GGC CTC ATT ACA TAC ATT GC-3' | 53.4 |
| H | *Atp1a1* | candidate | 1 | 21725072 - 21725091 | 25487439 - 25487458 | Forward | 5'-GTG TTT ACT CAA GGG AGA GG-3' | 55.4 |
| I | *Atp1a1* | candidate | 1 | 21724269 - 21724286 | 25488244 - 25488261 | Reverse | 5'-CAG TCC AAC CTG CCC ATC-3' | 57.1 |
| J | *Atp1a1* | reference | 1 | 20601250 - 20601272 | 26611258 - 26611280 | Forward | 5'-GAG CTT TTA TAC GTC TCT GAA GG-3' | 55.8 |
| K | *Atp1a1* | reference | 1 | 20601933 - 20601953 | 26610577 - 26610597 | Reverse | 5'-CAC CTC AGT AGG ACA GAA AGC-3' | 57.4 |
| L | *Spg1* | candidate | 4 | 21189272 - 21189296 | 24986448 - 24986427 | Forward | 5'-GCT GAG TAC AAT GTT TTA TAT AAC C-3' | 53.0 |
| M | *Spg1* | candidate | 4 | 21189764 - 21189788 | 24985956 - 24985980 | Reverse | 5'-CTA CGA ATC TAG AAA TTG TAA GAA G-3' | 53.0 |
| N | *Spg1* | reference | 4 | 20191013 - 20191032 | 25984712 - 25984731 | Forward | 5'-GCT TTA GAT TTC ATC GGG AG-3' | 53.4 |
| O | *Spg1* | reference | 4 | 20191548 - 20191566 | 25984178 - 25984196 | Reverse | 5'-CAT CAG TAT CTG GCT TTG G-3' | 52.9 |
| P | *Spg1* | reference | 4 | 20190803 - 20190821 | 25984923 - 25984941 | Forward | 5'-CGA AGG CCG AAG TTT AAG G-3' | 55.3 |
| Q | *Spg1* | reference | 4 | 20191323 - 20191343 | 25984401 - 25984421 | Reverse | 5'-CTT CTG AAA CGT CCG CTT ATG-3' | 55.8 |

**Table S2. Regions in the stickleback genome identified as candidates for M-FW divergence based on the molecular signature of parallel adaptation from shared variation with gene flow**

A genomic region qualified as M-FW candidate if smoothed delta divergence reached at least 0.2 *and* smoothed gsi was at least 0.6 (see Figure S2). The last column lists strong candidate genes for M-FW divergence contained in these regions, based on evidence from studies in stickleback (references with double asterisk) and/or other (mostly fish) species (references with single asterisk). Some of these candidate regions are visualized in Figure 4.

| Chromosome | Novel candidate region in Mb, corrected for misassembly according to Roesti et al. (2013) (uncorrected in parentheses) | Candidate genes |
|---|---|---|
| 1 | 6.3 - 6.6 (6.3 - 6.6) | *Tyr* (Koga et al. 1995*, Hoegg et al. 2004*; Page-McCaw et al. 2004*; Greenwood et al. 2011**) |
| 1 | 10.2 - 10.6 (10.2 - 10.6) | *Sucnr1* (Deen & Robben 2011*) |
| 4 | 7.3 - 7.7 (7.3 - 7.7) | n.a. |
| 7 | 10.6 - 11.00 (chrUn 7.95 - 8.35) | n.a. |
| 7 | 19.2 - 19.6 (17.45 - 17.85) | n.a. |
| 7 | 19.6 - 20.0 (17.85 - 18.25) | *Ncc* (Inokuchi et al. 2008*; Shimada et al. 2011**) |
| 8 | 8.6 - 9.0 (8.6 - 9.0) | *Adamts10* (Hohenlohe et al. 2010**) |
| 9 | 13.1 - 13.5 (8.7 - 9.1) | n.a. |
| 11 | 5.3 - 5.8 (5.3 - 5.8) | *Atp6v0a1* (Hohenlohe et al. 2010**; Jones et al. 2012b**); *Kcnh4* (Hohenlohe et al. 2010**; Jones et al. 2012b**); *Stat3* (Hohenlohe et al. 2010**); *Fzd2* (Hohenlohe et al. 2010**) |
| 11 | 7.3 - 7.8 (7.3 - 7.8) | *Slc35b1* (Dejima et al. 2009*); *Ndufa4* (van Rooijen et al. 2009*) |
| 11 | 8.4 - 8.9 (8.4 - 8.9) | *Kcnj2* (= *Kir2.1*) (Malinowska et al. 2003*); *Oto2* (Wang et al. 2011*) |
| 12 | 6.8 - 7.2 (12.38 - 12.78) | n.a. |
| 12 | 7.5 - 8.1 (11.52 - 12.12) | n.a. |
| 20 | 6.2 - 6.7 (12.38 - 12.88) | n.a. |
| 20 | 7.2 - 7.6 (11.39 - 11.89) | n.a. |

**Table S3. Genome-wide magnitude of divergence in all focal population comparisons**

Divergence is expressed as median $F_{ST}$ (mean $F_{ST}$ in parentheses) calculated across all SNPs. For details on the stringent SNP filtering conventions applied to maximize the robustness of divergence estimation see main text.

| Ecological contrast | Population 1 | Population 2 | $F_{ST}$ |
|---|---|---|---|
| M-M | Sayward | Cluxewe | 0.0000 (0.0230) |
| M-FW | Sayward | Boot lake | 0.2711 (0.3147) |
| M-FW | Sayward | Robert's lake | 0.0697 (0.1817) |
| M-FW | Sayward | Joe's Lake | 0.3900 (0.4175) |
| M-FW | Sayward | Misty lake | 0.1553 (0.2663) |
| M-FW | Sayward | Boot stream | 0.3690 (0.3718) |
| M-FW | Sayward | Robert's stream | 0.1546 (0.2453) |
| M-FW | Sayward | Joe's Stream | 0.3919 (0.4201) |
| M-FW | Sayward | Misty Stream | 0.1521 (0.2657) |
| M-FW | Cluxewe | Boot lake | 0.2541 (0.2939) |
| M-FW | Cluxewe | Robert's lake | 0.0965 (0.1932) |
| M-FW | Cluxewe | Joe's lake | 0.3637 (0.3848) |
| M-FW | Cluxewe | Misty lake | 0.1715 (0.2449) |
| M-FW | Cluxewe | Boot stream | 0.2826 (0.3438) |
| M-FW | Cluxewe | Robert's stream | 0.1715 (0.2449) |
| M-FW | Cluxewe | Joe's stream | 0.3589 (0.3899) |
| M-FW | Cluxewe | Misty Stream | 0.1765 (0.2588) |
| FW-FW | Boot lake | Robert's lake | 0.2658 (0.3011) |
| FW-FW | Boot lake | Joe's lake | 0.5158 (0.5276) |
| FW-FW | Boot lake | Misty lake | 0.3647 (0.3844) |
| FW-FW | Robert's lake | Joe's lake | 0.3511 (0.3630) |
| FW-FW | Robert's lake | Misty lake | 0.1277 (0.2058) |
| FW-FW | Joe's lake | Misty lake | 0.3706 (0.3902) |
| FW-FW | Boot stream | Robert's stream | 0.3896 (0.3932) |
| FW-FW | Boot stream | Joe's stream | 0.5271 (0.5420) |
| FW-FW | Boot stream | Misty stream | 0.4053 (0.4025) |
| FW-FW | Robert's stream | Joe's stream | 0.4065 (0.4231) |
| FW-FW | Robert's stream | Misty stream | 0.2176 (0.2699) |
| FW-FW | Joe's stream | Misty stream | 0.3816 (0.3957) |

# Supporting Figures

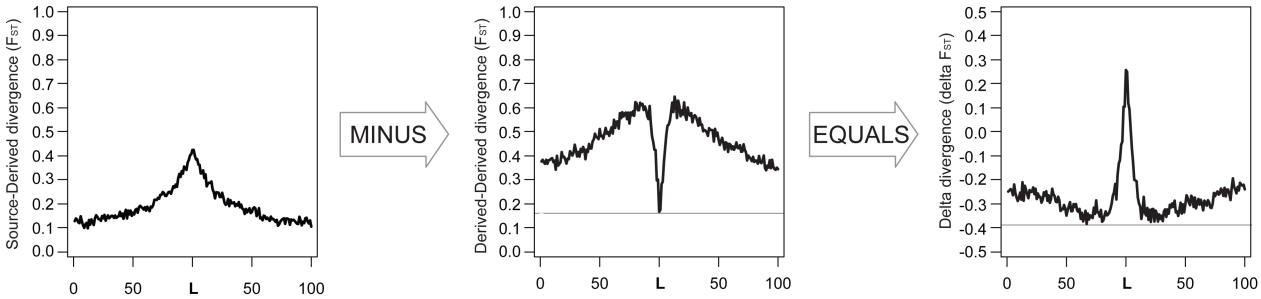


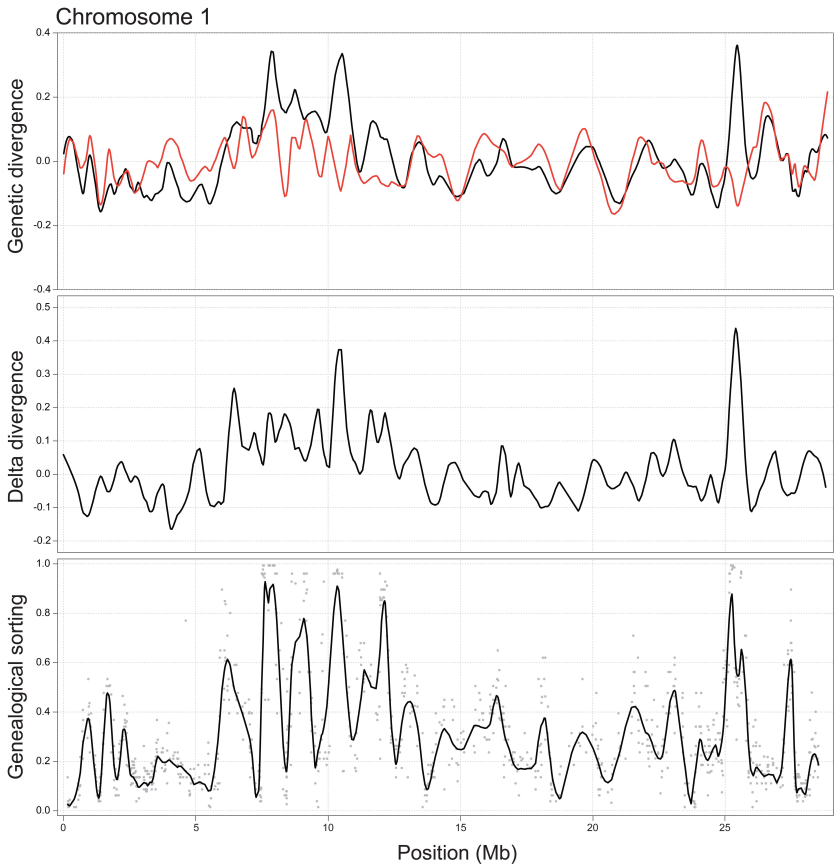

**Figure S1. Delta divergence calculated from simulated data**

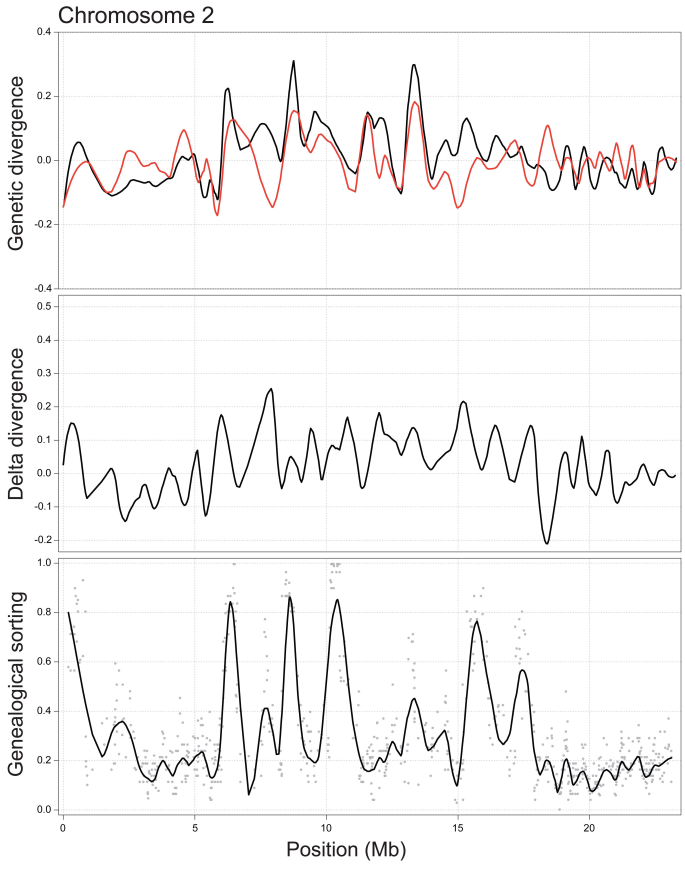
The rationale for using delta divergence to identify genomic regions involved in parallel adaptation from shared variation, illustrated using simulated data generated by the default model (Fig. 2C). Delta divergence is calculated by subtracting the divergence among derived populations (i.e., overall FW-FW divergence in our study) from the divergence between source and derived populations (overall M-FW divergence). The benefit is that the resulting delta divergence peak is higher and sharper than the source-derived peak and the derived-derived valley.

**Figure S2. Divergence and genealogical sorting profiles for all autosomes (presented on the 10 pages that follow)**

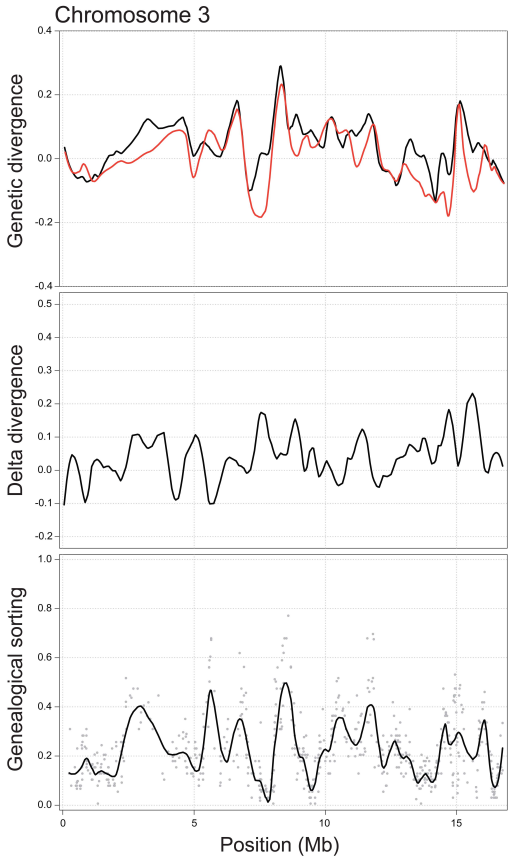
Genetic divergence (based on residual $F_{ST;}$ see Materials and Methods) between M and FW stickleback populations (top panel, black line) and among FW populations (top panel, red line), resulting delta divergence (middle panel), and M-FW genealogical sorting (bottom panel) plotted for all autosomes. Plotting conventions are as in Fig. 3 (B - D) and Fig. 4.

## Figure S2 (chromosomes 1 and 2)



## Figure S2 (chromosomes 3 and 4)



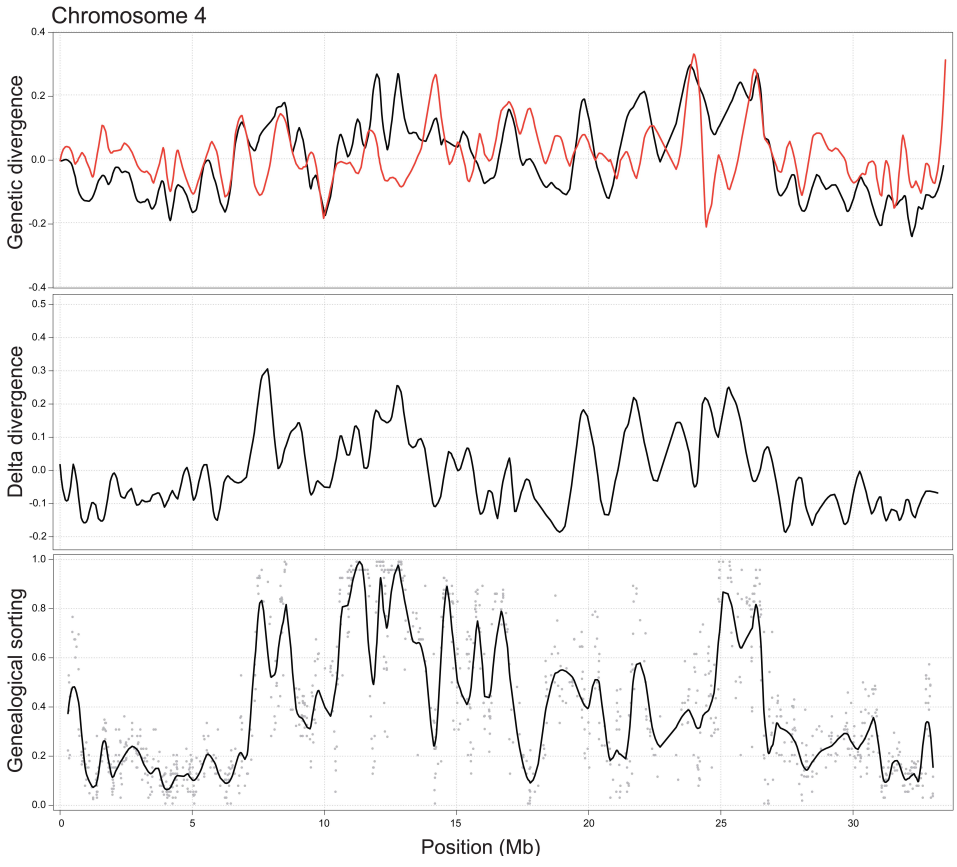
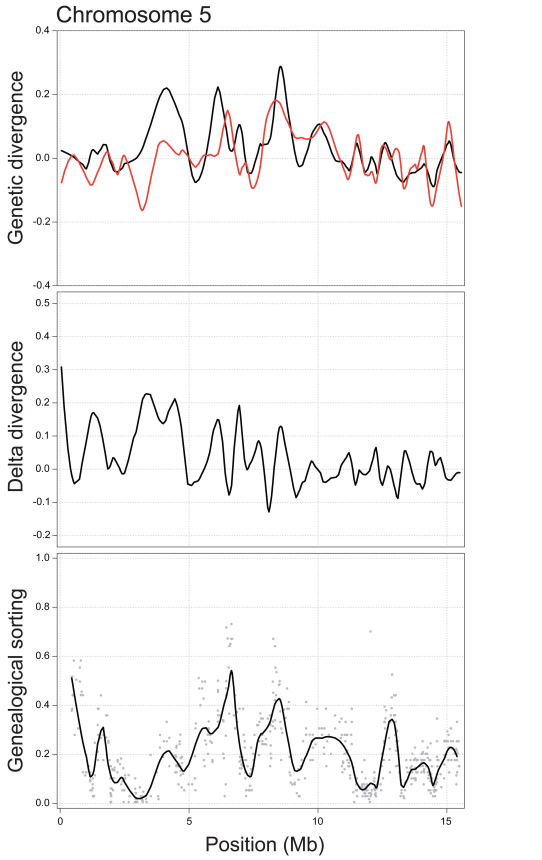
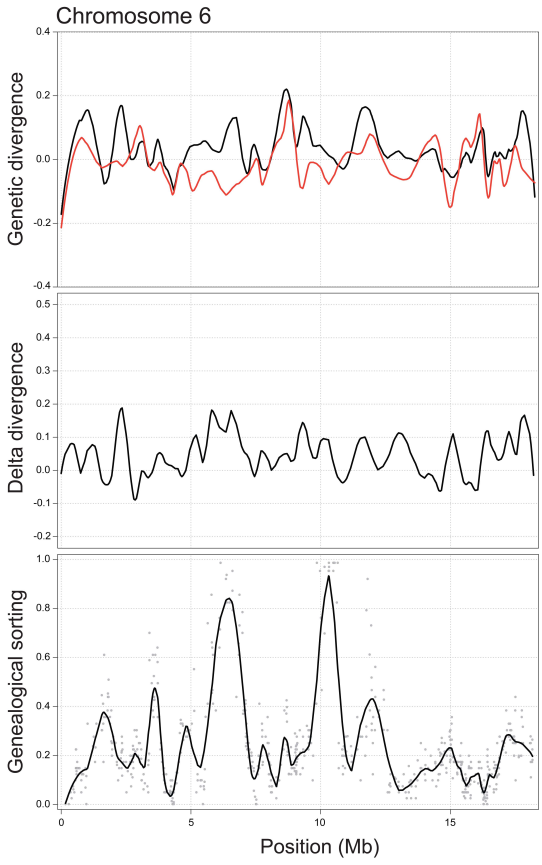
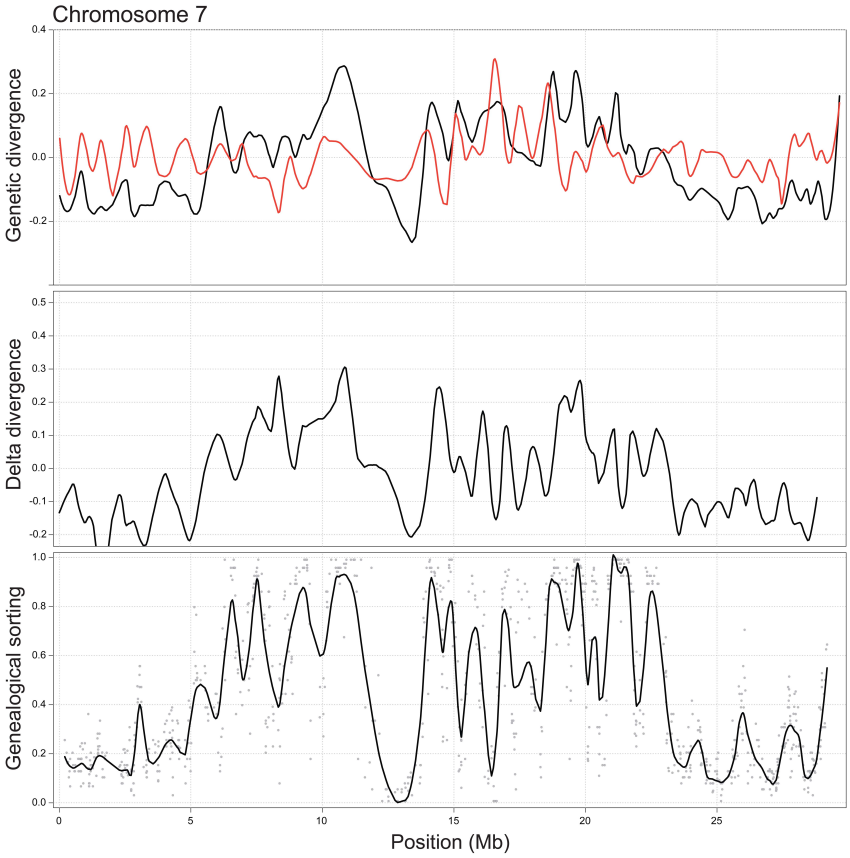



**Figure S2 (chromosomes 5 and 6)**



## Figure S2 (chromosomes 7 and 8)



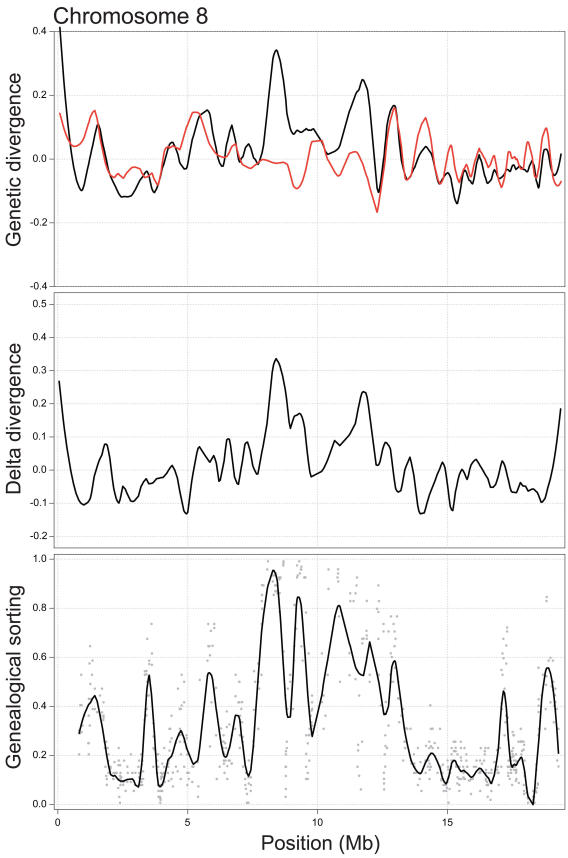
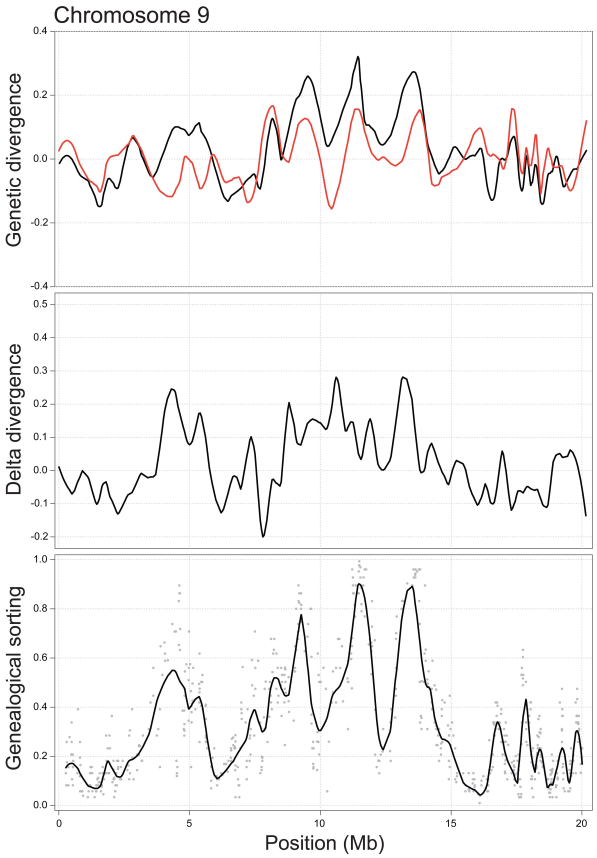



**Figure S2 (chromosomes 9 and 10)**



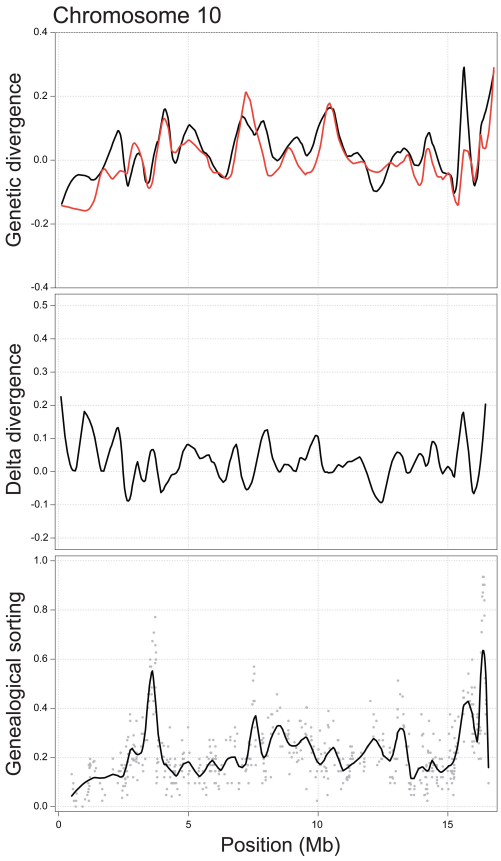
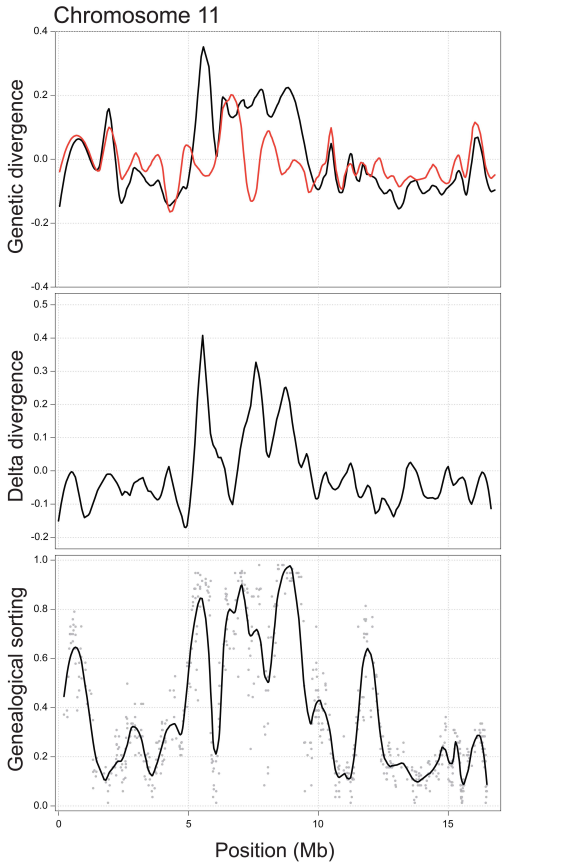



**Figure S2 (chromosomes 11 and 12)**



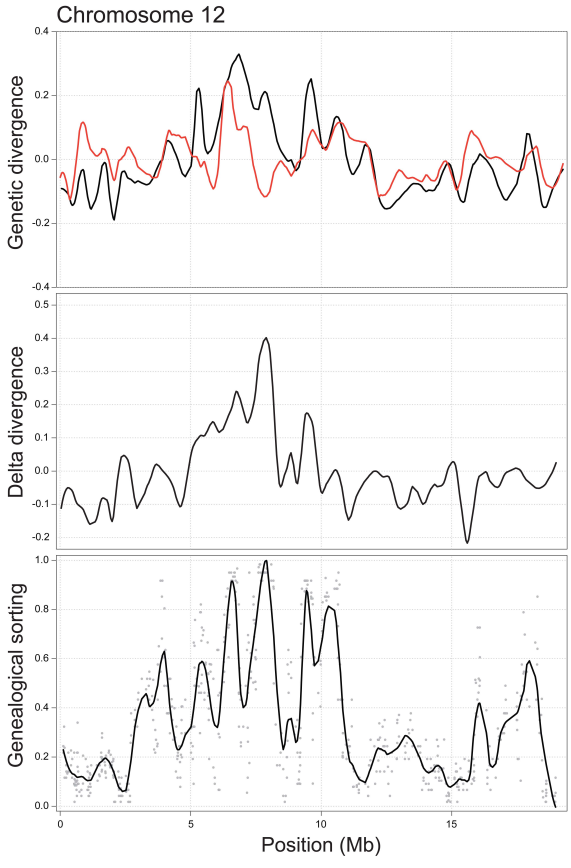
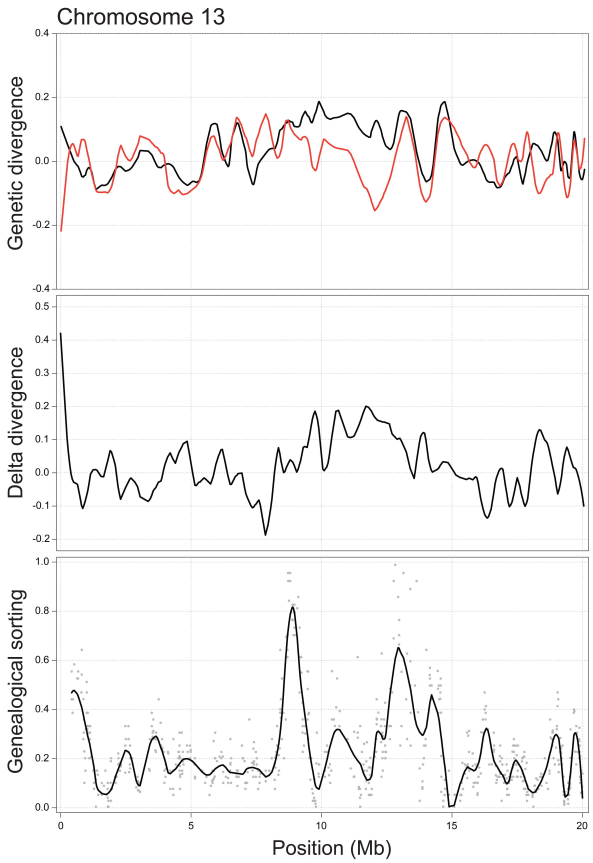



**Figure S2 (chromosomes 13 and 14)**



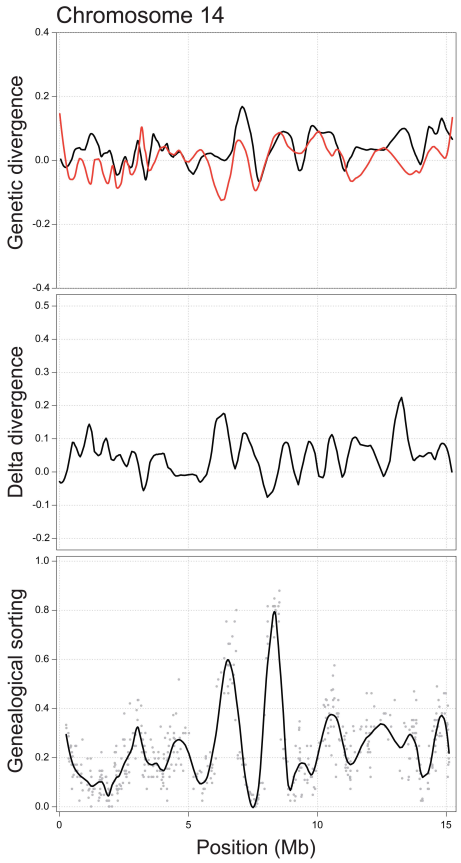
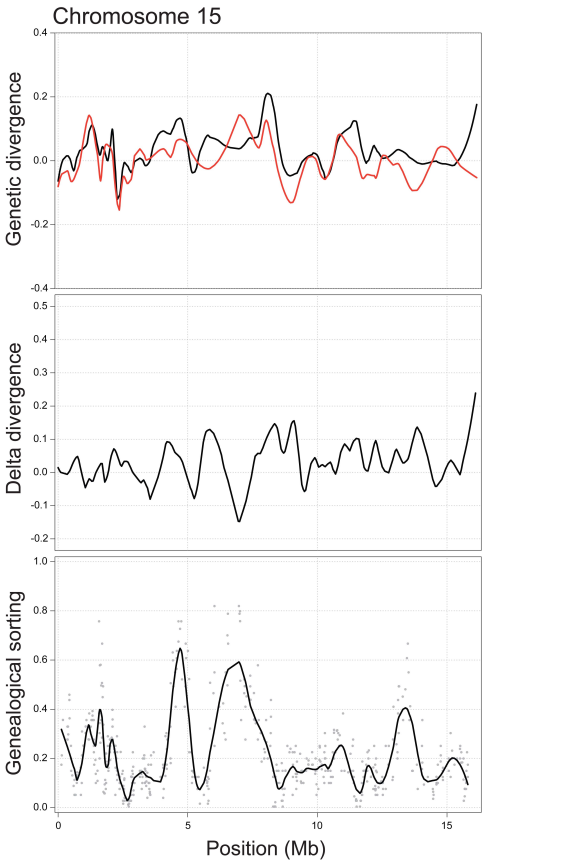



**Figure S2 (chromosomes 15 and 16)**



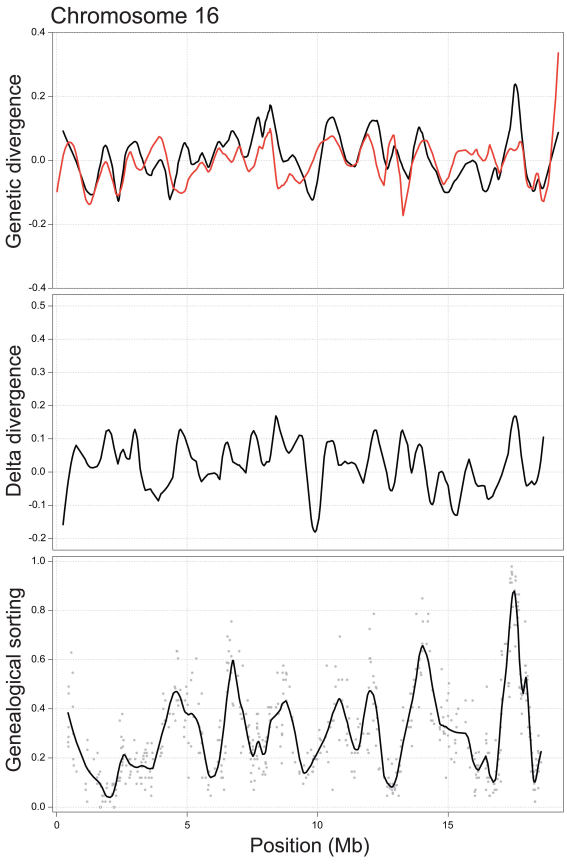
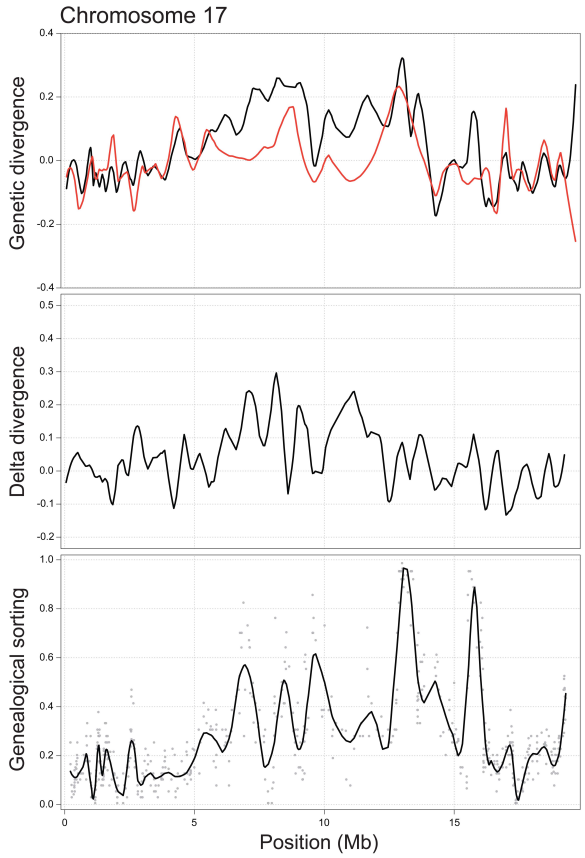



**Figure S2 (chromosomes 17 and 18)**



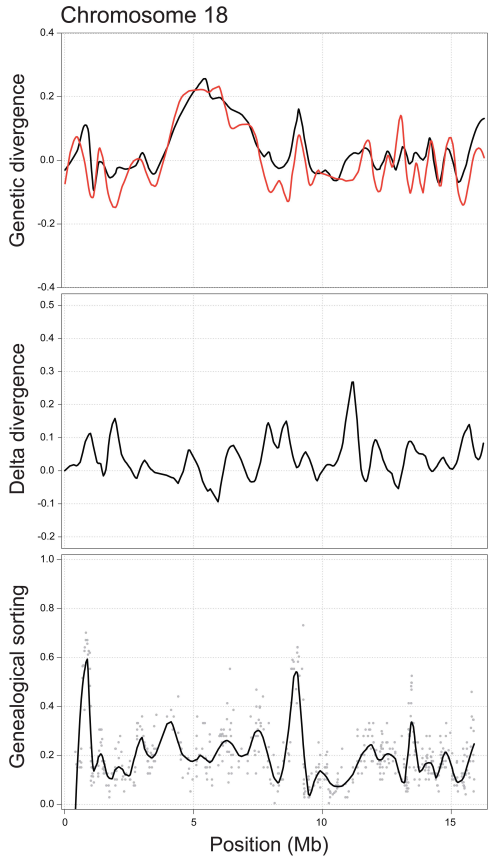
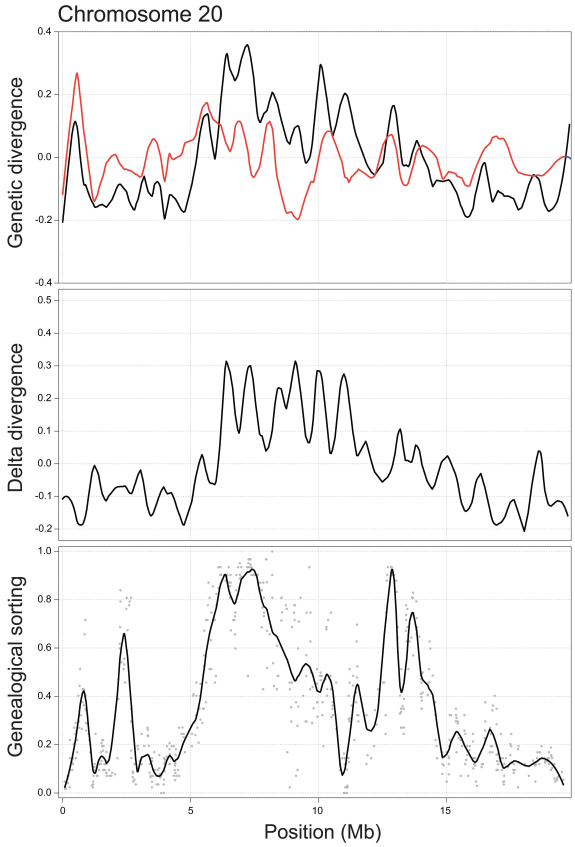
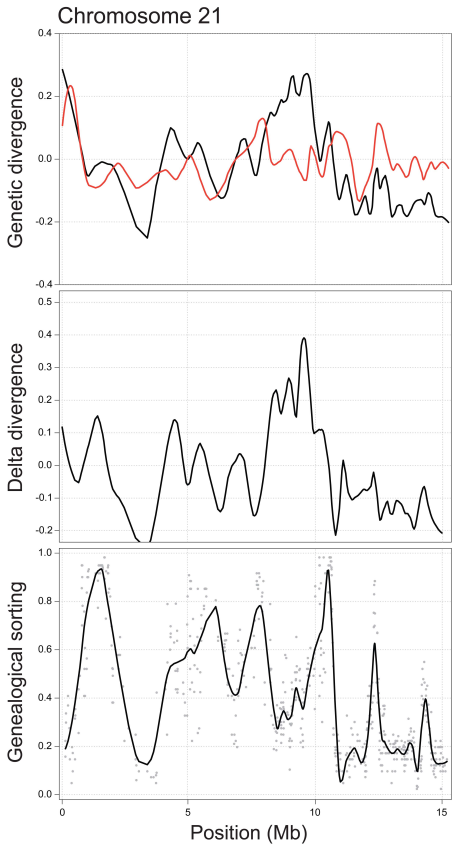



**Figure S2 (chromosomes 20 and 21)**



## Supporting References


Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE,* **3**, e3376.

Barrett SCH, Rogers SM, & Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. *Science,* **322**, 255-257.

Barrett RDH & Schluter D (2008) Adaptation from standing genetic variation. *Trends in Ecology and Evolution,* **23**, 38-44.

Bell MA & Foster SA (1994) *The evolutionary biology of the threespine stickleback* (Oxford University, Oxford).

Berner D, Adams DC, Grandchamp AC, & Hendry AP (2008) Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *Journal of Evolutionary Biology,* **21**, 1653-1665.

Berner D, Grandchamp A-C, & Hendry AP (2009) Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution,* **63**, 1740-1753.

Berner D, Stutz WE, & Bolnick DI (2010a) Foraging trait (co)variances in stickleback evolve deterministically and do not predict trajectories of adaptive diversification. *Evolution,* **64**, 2265-2277.

Berner D, Roesti M, Hendry AP, & Salzburger W (2010b) Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Molecular Ecology,* **19**, 4963-4978.

Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science,* **307**, 1928-1933.

Cummings MP, Neel MC, & Shaw KL (2008) A genealogical approach to quantifying lineage divergence. *Evolution,* **62**, 2411-2422.

Deagle BE, Jones FC, Chan YF *et al.* (2012) Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences,* **279**, 1277-1286.

Deen PM & Robben JH (2011) Succinate receptors in the kidney. *Journal of the American Society of Nephrology,* **22**, 1416-1422.

DeFaveri J, Shikano T, Shimada Y, Goto A, & Merilä J (2011) Global analysis of genes involved in

freshwater adaptation in threespine sticklebacks (*Gasterosteus aculeatus*). *Evolution,* **65**, 1800-1807.

Dejima K, Murata D, Mizuguchi S *et al.* (2009) The ortholog of human solute carrier family 35 member B1 (UDP-galactose transporter-related protein 1) is involved in maintenance of ER homeostasis and essential for larval development in *Caenorhabditis elegans*. *The FASEB Journal,* **23**, 2215-2225.

Domingues VS, Poh Y-P, Peterson BK *et al.* (2012) Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution,* **66**, 3209-3223.

Evans DH, Piermarini PM, & Choe KP (2005) The multifunctional fish gill: dominant site of gas exchange, osmoregulation, acid-base regulation, and excretion of nitrogenous waste. *Physiological Reviews,* **85,** 97-177.

Feder JL, Egan SP, & Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics,* **28**, 342-350.

Felsenstein J (1984) Distance methods for inferring phylogenies: A justification. *Evolution,* **38**, 16-24.

Greenwood AK, Jones FC, Chan YF *et al.* (2011) The genetic basis of divergent pigment patterns in juvenile threespine sticklebacks. *Heredity,* **107**, 155-166.

Gross JB & Wilkens H (2013) Albinism in phylogenetically and geographically distinct populations of *Astyanax* cavefish arises through the same loss-of-function *Oca2* allele. *Heredity,* **111**, 122-130.

Hermisson J & Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics,* **169**, 2335-2352.

Hoegg S, Brinkmann H, Taylor JS, & Meyer A (2004) Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *Journal of Molecular Evolution,* **59**, 190-203.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics,* **6**, e1000862.

Inokuchi M, Hiroi J, Watanabe S, Lee KM, & Kaneko T (2008) Gene expression and morphological localization of NHE3, NCC and NKCC1a in branchial mitochondria-rich cells of Mozambique tilapia (*Oreochromis mossambicus*) acclimated to a wide range of salinities. *Comparative Biochemistry and Physiology,* **151,** 151–158.

Jakobsson S, Borg B, Haux C, & Hyllner SJ (1999) An 11-ketotestosterone induced kidney-secreted protein: the nest building glue from male three-spined stickeback, *Gasterosteus*

*aculeatus*. *Fish Physiology and Biochemistry,* **20,** 79-85.

Kawahara R & Nishida M (2007) Extensive lineage-specific gene duplication and evolution of the spiggin multi-gene family in stickleback. *BMC Evolutionary Biology,* **7**, 209.

Koga A, Inagaki H, Bessho Y, & Hori H (1995) Insertion of a novel transposable element in the thyrosinase gene is responsible for an albino mutation in the medaka fish *Oryzias lapites*. *Molecular Biology and Evolution,* **240**, 400-405.

Jones FC*,* Chan YF, Schmutz J *et al.* (2012a) A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Current Biology,* **22**, 83-90.

Jones FC*,* Grabherr MG, Chan YF *et al.* (2012b) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature,* **484**, 55-61.

Kot M (2001) *Elements of Mathematical Ecology* (Cambrige University Press, Cambridge, UK).

Larsen PF, Nielsen EE, Koed A *et al.* (2008) Interpopulation differences in expression of candidate genes for salinity tolerance in winter migrating anadromous brown trout (*Salmo trutta* L.). *BMC Genetics*, **9,** 12.

Malinowska DH, Sherry AM, Tewari KP, & Cuppoletti J (2003) Gastric parietal cell secretory membrane contains PKA- and acid-activated Kir2.1 K+ channels. *American Journal of Physiology - Cell Physiology,* **286,** 495-506.

Marchinko KB (2009) Predation's role in repeated phenotypic and genetic divergence of armor in threespine stickleback. *Evolution,* **63,** 127-138.

Mateus CS*,* Stange M, Berner D *et al.* (2013) Strong genome-wide divergence between sympatric European river and brook lampreys. *Current Biology,* **23**, 649-650.

McCairns RJS & Bernatchez L (2010) Adaptive divergence between freshwater and marine sticklebacks: insights into the role of phenotypic plasticity from an intergrated analysis of candidate gene expression. *Evolution,* **64**, 1029-1047.

McCormick SD (2001) Endocrine control of osmoregulation in teleost fish. *American Zoologist,* **41**, 781-794.

Messer PW & Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution,* **28**, 659-669.

Nadeau NJ*,* Whibley A, Jones RT *et al.* (2012) Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences,* **367**, 343-353.

Nei M & Tajima F (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics,* **97**, 145-163.

Nosil P, Funk DJ, & Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology,* **18**, 375-402.

Page-McCaw PS, Chung SC, Muto A *et al.* (2004) Retinal network adaptation to bright light requires tyrosinase. *Nature Neuroscience,* **7,** 1329-1336.

Paradis E, Claude J, & Strimmer K (2004) Analyses of Phylogenetics and Evolution in R language. *Bioinformatics,* **20**, 289-290.

Posada D (2008) jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution,* **25**, 1253-1256.

Reilly BD, Cramp RL, Wilson JM, Campbell HA, & Franklin CE (2011) Branchial osmoregulation in the euryhaline bull shark, *Carcharhinus leucas*: a molecular analysis of ion transporters. *Journal of Experimental Biology,* **214,** 2883-2895.

Reimchen TE (1992) Injuries on stickleback from attacks by a toothed predator (*Oncorhynchs*) and implication for the evolution of lateral plates. *Evolution,* **46,** 1224-1230.

Reimchen TE (1994) Predators and morphological evolution in threespine stickleback. *The evolutionary biology of the threespine stickleback*, eds Bell MA & Foster SA (Oxford University, Oxford), pp 240-273.

Renaut S, Nolte AW, Rogers SM, Derome N, & Bernatchez L (2011) SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (*Coregonus* spp.). *Molecular Ecology,* **20**, 545-559.

Roesti M, Hendry AP, Salzburger W, & Berner D (2012a) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology,* **21**, 2852-2862.

Roesti M, Salzburger W, & Berner D (2012b) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology,* **12**, 94.

Roesti M, Moser D, & Berner D (2013) Recombination in the threespine stickleback genome – patterns and consequences. *Molecular Ecology,* **22**, 3014-3027.

Salzburger W, Ewing GB, & von Haeseler A (2011) The performance of phylogenetic algorithms in estimating haplotype genealogies with migration. *Molecular Ecology,* 20, 1952-1963.

Schluter D & Conte GL (2009) Genetics and ecological speciation. *Proceedings of the National Academy of Sciences, USA,* **106,** 9955-9962.

Scott GR, Rogers JT, Richards JG, Wood MC, & Schulte PM (2004) Intraspecific divergence of ionoregulatory physiology in the euryhaline teleost *Fundulus heteroclitus*: possible mechanisms of freshwater adaptation. *Journal of Experimental Biology,* **207,** 3399-3410.

Shimada Y, Shikano T, & Merilä J (2011) A high incidence of selection on physiologically important genes in the three-spined stickleback, *Gasterosteus* aculeatus. *Molecular Biology and Evolution,* **28**, 181-193.

Stephens M, Smith NJ, & Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics,* **68**, 978-989.

Stephens M & Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics,* **73**, 1162-1169.

Streisfeld MA, Young WN, & Sobel JM (2013) Divergent selection drives genetic differentiation in an R2R3-MYB transcription factor that contributes to incipient speciation in *Mimulus aurantiacus*. *PLoS Genetics,* **9**, e1003385.

Swofford DL (2003) *PAUP*: Phylogenetic analysis using parsimony (*and other methods)* (Sinauer Associates, Sunderland).

Team RDC (2013) *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria).

Tennessen JA, Akey JM (2011) Parallel adaptive divergence among geographically diverse human populations. *PLoS Genetics,* **7**, e1002127.

Terai Y, Seehausen O, Sasaki T *et al.* (2006) Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. *PLoS Biology,* **4**, e433.

van Rooijen E, Voest EE, Logister I *et al.* (2009) Zebrafish mutants in the von Hippel-Lindau tumor suppressor display a hypoxic response and recapitulate key aspects of Chuvash polycythemia. *Blood,* **113**, 6449-6460.

Walker JA & Bell MA (2000) Net evolutionary trajectories of body shape evolution within a microgeographic radiation of threespine sticklebacks (*Gasterosteus aculeatus*). *Journal of Zoology,* **252**, 293-302.

Wang X, Tan Y, Sievers Q *et al.* (2011) Thyroid hormone-response genes mediate otolith growth and development during flatfish metamorphosis. *Comparative Biochemistry and Physiology –*

*Part A: Comparative Physiology,* **158,** 163-168.

Weir BS & Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure. *Evolution,* **38**, 1358-1370.

Wootton RJ (1976) *The biology of the sticklebacks* (Academic, London).