MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 3944–3956 doi: 10.1111/mec.12720

FROM THE COVER

The genomic signature of parallel adaptation from shared genetic variation

MARIUS ROESTI,* SERGEY GAVRILETS,† ANDREW P. HENDRY,‡ WALTER SALZBURGER* and DANIEL BERNER*

*Zoological Institute, University of Basel, Vesalgasse 1, 4051 Basel, Switzerland, †Department of Ecology and Evolutionary Biology and Department of Mathematics, National Institute for Mathematical and Biological Synthesis (NIMBioS), University of Tennessee, Knoxville, TN 37996, USA, ‡Department of Biology and Redpath Museum, McGill University, 859 Sherbrooke St W., Montreal, Quebec, Canada

Abstract

Parallel adaptation is common and may often occur from shared genetic variation, but the genomic consequences of this process remain poorly understood. We first use individual-based simulations to demonstrate that comparisons between populations adapted in parallel to similar environments from shared variation reveal a characteristic genomic signature around a selected locus: a low-divergence valley centred at the locus and flanked by twin peaks of high divergence. This signature is initiated by the hitchhiking of haplotype tracts differing between derived populations in the broader neighbourhood of the selected locus (driving the high-divergence twin peaks) and shared haplotype tracts in the tight neighbourhood of the locus (driving the low-divergence valley). This initial hitchhiking signature is reinforced over time because the selected locus acts as a barrier to gene flow from the source to the derived populations, thus promoting divergence by drift in its close neighbourhood. We next empirically confirm the peak-valley-peak signature by combining targeted and RAD sequence data at three candidate adaptation genes in multiple marine (source) and freshwater (derived) populations of threespine stickleback. Finally, we use a genome-wide screen for the peak-valley-peak signature to discover additional genome regions involved in parallel marine-freshwater divergence. Our findings offer a new explanation for heterogeneous genomic divergence and thus challenge the standard view that peaks in population divergence harbour divergently selected loci and that low-divergence regions result from balancing selection or localized introgression. We anticipate that genome scans for peak-valley-peak divergence signatures will promote the discovery of adaptation genes in other organisms.

Keywords: barrier to gene flow, evolutionary genomics, *Gasterosteus aculeatus*, genetic hitchhiking, genome scan, population divergence

Received 21 January 2014; revision received 12 March 2014; accepted 12 March 2014

Introduction

Understanding how selection shapes the genome and identifying the loci underlying adaptive divergence are major goals of biology (Wu 2001; Nielsen 2005; Stinchcombe & Hoekstra 2008; Nosil & Schluter 2011; Feder *et al.* 2012). Recent studies have indicated that genomic differentiation between diverging populations can be highly heterogeneous and can involve selection on numerous loci throughout the genome, with some of these loci now having been identified (e.g. Hohenlohe *et al.* 2010; Lawniczak *et al.* 2010; Fournier-Level *et al.* 2011; Jones *et al.* 2012b; Nadeau *et al.* 2012; Roesti *et al.* 2012a; Renaut *et al.* 2013; Streisfeld *et al.* 2013). Nevertheless, understanding how evolutionary processes cause heterogeneous genomic divergence remains challenging (e.g. Slatkin & Wiehe 1998; Barton 2000;

Correspondence: Daniel Berner, Fax: +41 (0) 61 267 0301; E-mail: daniel.berner@unibas.ch

Hermisson & Pennings 2005; Excoffier & Ray 2008; Bierne 2010; Feder & Nosil 2010; Bierne *et al.* 2011; Roesti *et al.* 2012a, 2013; reviewed in Wu 2001; Nosil *et al.* 2009). Traditional population genetic theory has primarily focused on a scenario in which a new genetic variant arises by mutation in a population colonizing a new environment (hereafter called a 'derived' population) where the variant is beneficial (Orr 1998; Barrett & Schluter 2008; Messer & Petrov 2013). The new genetic variant is then expected to fix in the derived population, whereas the initial genetic variant remains favoured and is thus retained in the 'source' population inhabiting the ancestral environment. Consequently, the source and derived populations are differentiated at the locus under divergent selection and, due to genetic hitchhiking, also in the selectively neutral genetic neighbourhood of that locus (Maynard Smith & Haigh 1974; Kaplan *et al.* 1989). Genomic regions of high population divergence, as identified in marker-based genome scans, are thus generally assumed to harbour genes involved in adaptive divergence (Nielsen 2005; Storz 2005).

This traditional theoretical scenario might not be adequate when adaptation occurs from standing (pre-existing) genetic variation rather than from novel mutations (Hermisson & Pennings 2005; Barrett & Schluter 2008; Pritchard *et al.* 2010). This realization has stimulated theory focusing on 'soft' selective sweeps, where a novel genetic variant is segregating in an ancestral source population before becoming selected in a derived population. In this case, the divergence signature driven by the selective sweep will be weakened relative to the classical 'hard' sweep expected from a novel mutation (Hermisson & Pennings 2005; Barrett & Schluter 2008; Messer & Petrov 2013). The reason is that the derived variant can become associated through recombination with diverse genetic backgrounds in the ancestral population before the derived population becomes established in the new environment. This diversity reduces the selective sweep in the locus' genomic neighbourhood when the variant eventually becomes selected. Although hard and soft sweep models differ in the age (or origin) of the selected variant and in the expected strength of the associated selective signature, they share the focus on comparing populations inhabiting *selectively different* environments (i.e. source vs. derived).

In this study, we consider the patterns of genomic divergence that might be expected among multiple *derived* populations adapting in parallel to *selectively similar* environments. We scrutinize these genomic patterns through theoretical modelling and through targeted and genome-wide sequencing in multiple natural populations of threespine stickleback fish (*Gasterosteus aculeatus*) that have adapted in parallel to freshwater environments from a common marine source population. We find that a locus involved in parallel adaptation from shared genetic variation generates a novel and characteristic pattern of genomic divergence, which provides a new perspective on how to interpret high- and low-divergence outliers detected in genome scans.

Materials and methods

Models of parallel adaptation from shared genetic variation

We developed individual-based simulation models in which multiple derived populations diverge independently into a selectively novel environment from a shared source population inhabiting a selectively different, ancestral environment. A detailed description of these simulations is provided in the Methods S1 (Supporting information), so that we here give a brief overview only. Individuals are represented by a single haploid chromosome. The centre of the chromosome holds a locus under divergent selection between the ancestral and novel environment, with the allele favourable in the novel environment occurring at low frequency in the source population. The selected locus is flanked on each side by 100 evenly spaced, selectively neutral polymorphic loci, in analogy to single nucleotide polymorphisms (SNPs) used in genome scans. After the colonization of the novel environments, the derived populations evolve, with each generation including migration from the source population, followed by reproduction with fertility selection and recombination.

Our simulations started with a default parameterization 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 genomic signature of adaptation from shared genetic variation (Roesti *et al.* 2012a). The default model was then expanded to explore the influence of migration rate, time, the number of founder individuals, the strength of divergent selection and recombination rate on the molecular signatures of adaptation, as captured by the magnitude of population divergence (F_{ST}; Weir & Cockerham 1984) across the neutral loci. We also modified the default model to include *two* selected loci located at equal distances from the centre of the chromosome, which now harboured 400 total neutral loci. Our models first considered divergence between the *source* and the derived populations. These comparisons represent the standard ecological genome scan, as described earlier, and hence served to validate our general simulation approach. In

all subsequent simulations, we focused on divergence among *derived* populations.

Stickleback populations for empirical investigation

Our empirical analyses used marine and freshwater (hereafter M and FW) populations of threespine stickleback. These populations provide an excellent natural system for studying the genomics of parallel adaptation from shared variation because numerous FW populations have been derived independently and recently (<10 000 years ago) from a common M source population (Bell & Foster 1994). Moreover, FW stickleback display relatively consistent phenotypic shifts from their M ancestors as a response to shared selective conditions among FW habitats – that is, parallel adaptation (e.g. Taylor & McPhail 1986; Walker 1997; Walker & Bell 2000; Schluter *et al.* 2004; Berner *et al.* 2010a). Our study considered M stickleback sampled from two estuarine sites on the east coast of Vancouver Island (British Columbia, Canada), and FW samples from a lake and a stream population within each of four independently colonized drainages on Vancouver Island (Berner *et al.* 2008, 2009; Roesti *et al.* 2012a) (Fig. 1; Methods S2, Supporting information). Each of the ten total samples comprised 27 individuals.

Fig. 1 Origin of the stickleback samples used for the empirical analysis. The map shows Vancouver Island (British Columbia, Canada), with the lake and stream populations from four independently colonized freshwater (FW) drainages shown as light and dark coloured circles. The two sites where the marine (M) source population was sampled are shown as black circles.

Targeted sequencing and haplotype networks at M-FW candidate genes and reference loci

We Sanger-sequenced DNA segments at three stickleback candidate genes thought to be under strong divergent selection between M and FW environments. These genes included (i) *Eda*, the major gene underlying the reduction in lateral plate number typically observed when M stickleback colonize FW (Colosimo *et al.* 2005); (ii) *Atp1a1*, a key player in physiological adaptation to osmotically different environments in many organisms (McCormick 2001); and (iii) *Spg1*, which encodes a presumably pH and salinity sensitive glue-like protein used by stickleback males to build their nests (Kawahara & Nishida 2007) (further details on these genes is given in Methods S3, Supporting information). We additionally sequenced a 'reference locus' approximately one megabase away from each of the above three genes. Details on primers and Sanger sequencing are provided in Table S1 and Methods S4 (Supporting information). SNPs derived from these sequences were used to construct haplotype genealogies for each candidate gene and reference locus (Methods S5, Supporting information). 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.

Broad-scale genetic divergence in the candidate regions

To explore divergence at a broader scale around each candidate gene (i.e. across 3–4 Mb 'candidate regions' centred at the genes), we derived SNPs from consensus sequences at genome-wide RAD loci (Baird *et al.* 2008) generated for all 27 individuals from the M and FW samples. (For details on the wet laboratory and consensus genotyping protocols, see Roesti *et al.* 2012a and Methods S6, Supporting information.) These SNPs were used to quantify genetic divergence (F_{ST} based on haplotype diversity; Nei & Tajima 1981 eq. 7) for all possible pairwise comparisons between the two M samples and the eight FW samples (16 total comparisons). We here used only one SNP per RAD locus. We further ensured robust divergence estimation by including a SNP only if both populations in a comparison contributed at least 27 nucleotides to the common nucleotide pool, and if the minor allele frequency across this pool was ≥0.25, thereby eliminating polymorphisms with low information content (Roesti *et al.* 2012b). Moreover,

we corrected for inflated population divergence in chromosome centres relative to their peripheries due to heterogeneous recombination rate along the stickleback chromosomes (Roesti *et al.* 2012a, 2013) by calculating *residual* divergence (details in Roesti *et al.* 2012a), although qualitatively similar conclusions emerged without this correction. Following these same conventions, we also calculated F_{ST} for pairwise comparisons between the derived FW populations. We here considered comparisons between samples from ecologically similar FW environments only (i.e. six lake–lake and six stream–stream comparisons, 12 in total). The rationale was to avoid capturing selective signatures of lake–stream divergence, which is known to be strong (Berner *et al.* 2008, 2009, 2010b; Deagle *et al.* 2012; Roesti *et al.* 2012a). However, analyses based on *all* possible FW comparisons produced very similar results.

Finally, we generated *overall* M-FW and FW-FW divergence values by averaging residual F_{ST} at each RAD locus across all pairwise M-FW and all pairwise FW-FW comparisons. On average, 6.9 and 6.4 replicate estimates were available per RAD locus for the overall M-FW and FW-FW comparisons, and we achieved a median and mean marker spacing of 12 and 25 kb across the candidate regions (Methods S6, Supporting information). For visualization, we produced sliding window divergence profiles for each candidate region by using the R (R Development Core Team 2013) implementation of LOESS (robust locally weighted scatterplot smoothing; Cleveland 1979) (R was also used for all other operations unless stated otherwise). We chose a polynomial degree of two and adjusted the smoothing span to achieve equal smoothing resolution across all chromosomes. All genomic positions in this study refer to the improved assembly of the initial (Jones *et al.* 2012b) stickleback reference genome (Roesti *et al.* 2013; http://datadryad.org/resource/doi:10.5061/dryad.846nj.2).

Delta divergence and genealogical sorting in the candidate regions

Parallel divergence between source and derived environments based on shared variation drives opposed patterns of genomic divergence in source–derived versus derived–derived population 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, Supporting information). We therefore complemented our standard F_{ST}-based divergence analyses described above by generating delta divergence profiles for each of the three candidate regions (Methods S6, Supporting information).

As an alternative to quantifying genetic divergence between M and FW stickleback based on F_{ST}, we additionally 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, calculated a distance matrix for each window, translated each distance matrix to a neighbour joining tree and finally extracted the genealogical sorting index (gsi; Cummings *et al.* 2008) from each tree (details in Methods S6, Supporting information). This index ranges from 0 to 1 and quantifies the extent of exclusive ancestry of individuals from defined groups (here M and FW stickleback). The gsi data showed a similar physical resolution as the F_{ST} data, and smoothed profiles were generated as described above.

Genome-wide search for signatures of parallel adaptation

To discover additional genomic regions potentially involved in parallel M-FW divergence from shared variation, we performed genome-wide screens of population divergence and genealogical sorting using the RAD-based SNP data and analytical approaches described above. The genome-wide M-FW and FW-FW divergence analyses based on F_{ST} used 16 687 and 16 269 data points (each representing the average of multiple pairwise population comparisons), while the gsi-based analysis used 14 890 data points integrating 29 787 phylogenetic trees across the 21 chromosomes. Both types of analyses achieved an approximate genome-wide median and mean marker spacing of 14 kb. We considered a genome region a new candidate if smoothed delta divergence was >0.2 and smoothed gsi exceeded 0.6. For each region meeting these criteria, we retrieved all genes located within a window of 400–600 kb centred at the delta divergence peak (generally coinciding exactly with the gsi peak) from the Ensembl Genome Browser and assessed whether these genes were known from other (mostly fish) species to be important to saltwater versus freshwater adaptation.

Heterogeneity among chromosomes in M-FW divergence

The presence versus absence of barriers to gene flow (i.e. genes under divergent selection) on specific chromosomes could lead to heterogeneity among chromosomes in the magnitude of population divergence. We considered this possibility by testing for a difference in overall divergence between autosomes under strong

versus weak M-FW selection. The difference between these two chromosome types was defined as those displaying ≥2 versus no candidate regions for parallel adaptation from shared variation, as defined in the previous paragraph. For each chromosome, we calculated median raw (not residual, as above) F_{ST} and gsi for each M-FW comparison and averaged these replicate values. We then tested whether the magnitude of divergence differed between the two chromosome categories by permuting F_{ST} and gsi over the chromosomes 9999 times and using the absolute difference between the chromosome categories as test statistic.

Results

Models of parallel adaptation from shared genetic variation

Our models of multiple derived populations diverging from a shared source population into selectively similar environments produced a single peak of high divergence around the selected locus when comparing the source to the derived populations (Fig. 2A). This contrast is the type typically considered in divergence mapping studies. However, our main interest was in *derived–derived* population comparisons, where we found that the parallel fixation of a shared variant leads to a valley of reduced divergence (hereafter 'divergence valley') around the locus under selection (Fig. 2B). The divergence valley was initially flanked by regions of slightly elevated divergence that then declined towards the chromosome peripheries. In the absence of migration (hence no gene flow), this decline became less striking over time as overall baseline divergence increased owing to drift (Fig. 2B). By contrast, source–derived migration caused the divergence valley to be flanked on either side by striking peaks of high divergence ('Migration' in Fig. 2C). Although these 'divergence twin peaks' emerged consistently across our simulations when comparing derived populations, their height and width were influenced by several factors. First, the peaks grew higher and sharper with increasing time ('Time' in Fig. 2C) and with a decreasing number of founder individuals ('Founders' in Fig. 2C). Second, the physical extent of the divergence twin peaks and of the divergence valley was greater – and could be quite extensive (kilobases to megabases) – when divergent selection was strong and/or recombination was low ('Selection' and 'Recombination' in Fig. 2C).

In our simulations with two loci under divergent selection, separate peak-valley-peak signatures emerged when the selected loci were far apart ('Distant' in Fig. 2D). When the loci were closer together, the entire chromosome segment between them reached high

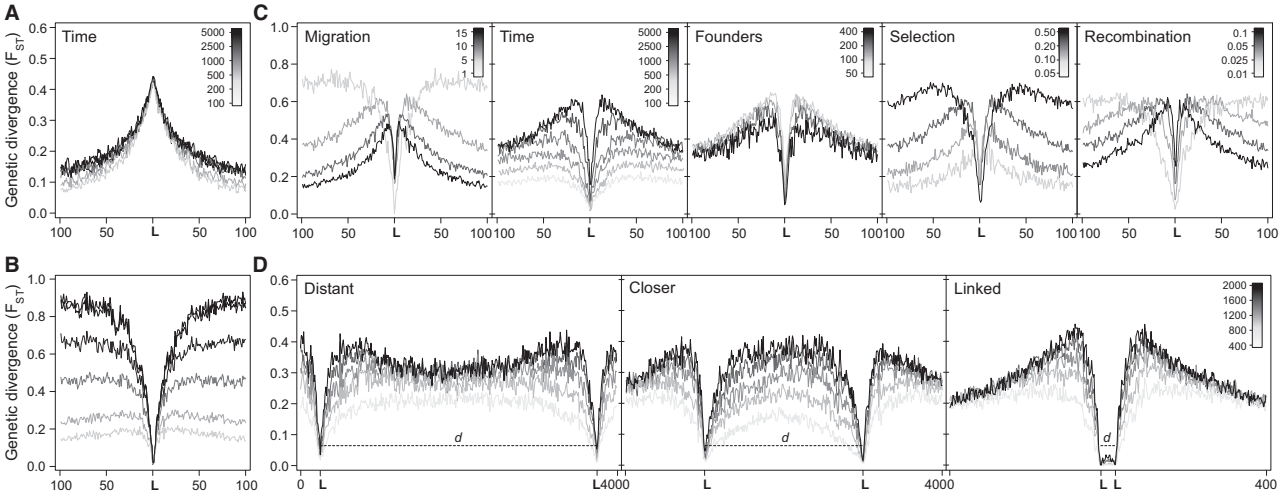

Fig. 2 Molecular signatures of parallel adaptation from shared genetic variation. Shown is the magnitude of population divergence (F_{ST}) at neutral markers along a chromosome segment holding a locus (or loci; L) under divergent selection between a source and multiple derived populations. (A) Traditional genome scan comparing the source to the derived populations, exhibiting a divergence peak at L. (B) By contrast, comparing multiple derived populations adapting in parallel produces a divergence valley around L. In the absence of migration, the rest of the chromosome diverges over time [timescale as in (A)]. (C) Allowing for migration from the source to the derived populations generates a characteristic peak-valley-peak signature of selection ('Migration'). The other panels in (C) show how this signature is influenced by variation in divergence time ('Time', in generations), the number of initial colonizers ('Founders'), the strength of divergent selection on L ('Selection') and the recombination rate across the chromosome segment ('Recombination'). (D) Simulations with two loci occurring at different distances *d* on the chromosome (the scale indicates divergence time).

divergence ('Closer' in Fig. 2D). Finally, when the selected loci occurred in very close proximity to each other, the two divergence valleys collapsed to a single large valley flanked by particularly pronounced divergence twin peaks ('Linked' in Fig. 2D).

Signatures of parallel adaptation from shared variation at stickleback candidate genes

Haplotype genealogies generated from targeted sequence data at three candidate genes for M-FW adaptation *(Eda, Atp1a1, Spg1)* consistently revealed the pattern that our simulations suggested should characterize parallel adaptation from shared genetic variation. That is, lake and stream FW samples shared closely related haplotypes that were clearly distinct from the haplotypes predominant in M stickleback ('Candidate gene' in Fig. 3A). In marked contrast, the reference loci approximately one megabase away from the candidate genes showed little or no habitat-related haplotype structure ('Reference locus' in Fig. 3A). This result indicates the parallel fixation of shared alleles at the candidate genes in FW, despite high M-FW gene flow in other parts of the genome.

We next used SNPs generated through RAD sequencing to assess broad-scale divergence (F_{ST} and genealogical sorting index, gsi) around the three candidate genes for the overall M-FW (source vs. derived) and FW-FW (derived vs. derived) comparison. As expected from the above small-scale targeted Sanger sequencing, M-FW divergence was exceptionally strong close to the three candidate genes (black lines in Fig. 3B), and gsi indicated striking phylogenetic separation between M and FW stickleback in these regions

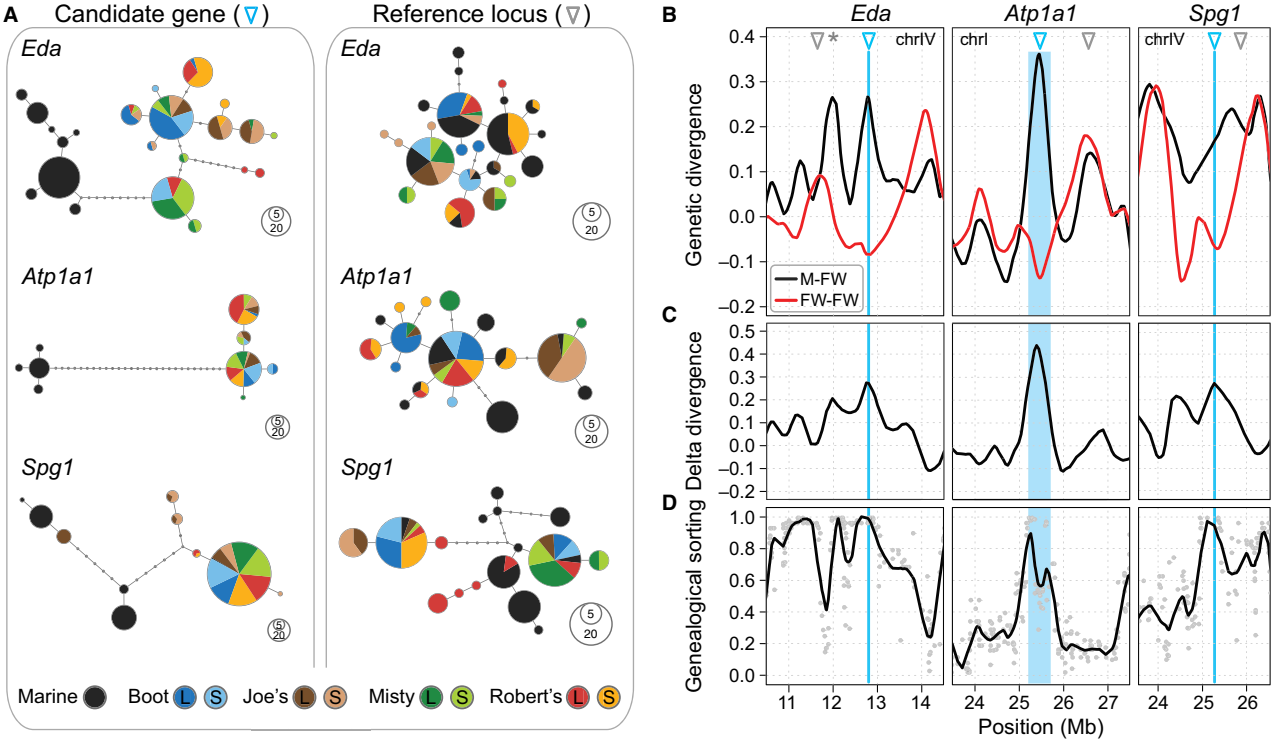

Fig. 3 Genetic structure and divergence at three candidate genes for M-FW divergence in stickleback. (A) Haplotype networks based on targeted sequencing at each candidate gene (left column), and at a corresponding reference locus (right column). The pies represent unique haplotypes, pie sizes reflect their frequency and the internodes are mutational steps. At the candidate genes, lake (L) and stream (S) populations from four independently colonized watersheds form a haplotype cluster separated from the marine (M) haplotype cluster, a genealogical structure not seen (or that is much weaker) at the reference loci. (B) Genetic divergence profiles reveal that the candidate genes (position given by the blue triangle and vertical line; the reference loci are indicated by grey triangles) generally coincide with peaks in M-FW divergence (black) and valleys in FW-FW divergence (red), as expected for parallel adaptation from shared variation. (C) Consequently, delta divergence, calculated as the difference between the two profiles in (B) (i.e. M-FW divergence minus FW-FW divergence; see also Fig. S1) peaks right at the candidate genes. (D) In the same regions, genealogical sorting (gsi) profiles reveal that M and FW stickleback exhibit completely separate ancestry at a broad physical scale (the gsi values underlying the smoothed profile are drawn as grey dots). Note that the neighbourhood of *Eda* contains further regions displaying complete genealogical separation, for instance, at the *Abcb7* gene indicated by a grey asterisk in (B). The location of *Atp1a1* is given as a wide blue vertical line because it is associated with an inversion (Jones *et al.* 2012b).

(Fig. 3D). Importantly, however, the comparison between the *derived* FW populations revealed a valley of low divergence around each candidate gene, as predicted by our simulations (red lines in Fig. 3B). Moreover, these divergence valleys in the FW-FW comparisons were often flanked by striking divergence peaks – some of which were absent in the M-FW comparison, a pattern specifically predicted by our simulations with gene flow. We also found that these signatures of adaptation from shared variation were particularly obvious when M-FW and FW-FW divergence was combined into delta divergence profiles, yielding peaks exactly at the candidate genes (Fig. 3C).

Genome-wide signatures of parallel adaptation from shared variation

We used (delta) divergence and gsi profiles to search the stickleback genome for additional regions likely involved in parallel adaptation from shared variation. This screen discovered 15 such regions on eight chromosomes. Details on these regions, including strong candidates genes for M-FW adaptation (some of which have been suggested previously for stickleback; Hohenlohe *et al.* 2010; Jones *et al.* 2012a,b), are provided in Table S2 (Supporting information), and (delta) divergence and gsi profiles for seven representative new candidate regions are presented in Fig. 4. Full genome-wide divergence and genealogical sorting profiles are provided in Fig. S2 (Supporting information).

Chromosome-level relationship between candidate regions and divergence

The six autosomes displaying multiple genomic signatures of parallel adaptation from shared variation (i.e. the chromosomes 1, 4, 7, 11, 12, 20) also exhibited exaggerated overall divergence (45% higher F_{ST} on average, $P = 0.0023$; 35% higher gsi, $P = 0.0186$) between M and FW populations compared with the 12 chromosomes lacking such signatures (Fig. 5).

Discussion

We combined simulations and empirical data to shed light on the genomic patterns that arise when multiple populations diverge into selectively similar environments by using shared genetic variation from the ancestral source population. Our main finding is that the immediate neighbourhood of the selected genetic locus underlying parallel adaptation will remain undifferentiated among the derived populations, whereas the broader neighbourhood around the locus will be driven to high divergence. In combination, this produces a characteristic peak-valley-peak signature of genomic divergence among derived populations.

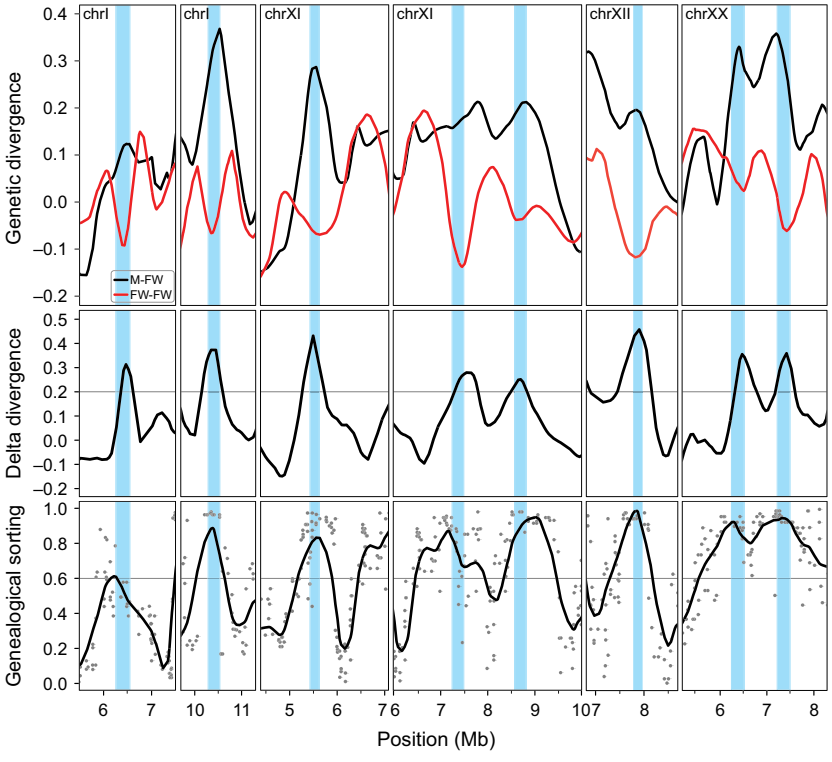

Fig. 4 Exemplary novel regions in the stickleback genome potentially harbouring genes involved in parallel FW adaptation from shared genetic variation. These regions were identified as candidate adaptation hotspots because they displayed high delta divergence (>0.2; threshold shown as grey horizontal line) as a consequence of opposed divergence profiles in M-FW versus FW-FW comparisons, and strong M-FW genealogical sorting (>0.6; grey horizontal line). The plotting conventions follow the ones in Fig. 3B–D. Strong candidate genes located within the blue regions are listed in Table S2.

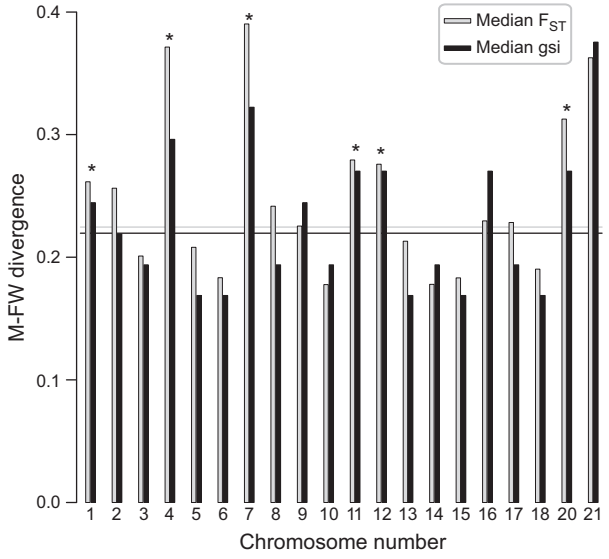

Fig. 5 Heterogeneity among chromosomes in overall divergence (F_{ST} and gsi) between M and FW stickleback. The values represent averages across the 16 replicate population comparisons. Genome-wide median F_{ST} and gsi are given as horizontal grey and black lines. Chromosomes exhibiting two or more signatures of parallel adaptation from shared variation are indicated by an asterisk. Note that the strongly divergent chromosomes (4, 7) are among the three largest ones in the stickleback genome and exhibit particularly low average recombination rates (Roesti *et al.* 2013). The sex chromosome (19) was excluded from this analysis.

Distinct mechanisms drive the peak-valley-peak divergence signature

The peak-valley-peak signature of divergence among derived populations arises from an interaction between two distinct evolutionary mechanisms operating at different timescales. The first mechanism is hitchhiking (Maynard Smith & Haigh 1974; Kaplan *et al.* 1989) of different genomic regions surrounding the genetic variant that is selected in parallel within the derived populations (Fig. 6A). This process halts as soon as the adaptive variant reaches fixation (or some other migration-selection balance) within the derived populations. During this phase, the opportunity for the derived variant to become associated with new genetic backgrounds through recombination is limited. Therefore, in the *close* neighbourhood of the selected locus, the derived populations become fixed for the same haplotype tract linked to the shared adaptive variant. Comparing populations adapted in parallel will thus reveal a low-divergence valley surrounding the locus under common selection (Fig. 6B). In the *broader* neighbourhood of the selected locus, however, recombination during the hitchhiking phase will occur sufficiently frequently to associate the adaptive variant with genetic backgrounds *specific* to each derived population. These population-specific haplotypes increase in frequency along with the adaptive variant, causing elevated divergence among derived populations on either side of the divergence valley. Even further away from the selected locus, divergence declines again because recombination increasingly associates the derived variant with random haplotypes from the source population. This first hitchhiking phase thus establishes a divergence valley surrounded by a broad region of elevated divergence among the derived populations (Fig. 6B; also see Fig. 2B).

The second mechanism shaping the peak-valley-peak signature is a *long-term* barrier to gene flow caused by the locus under selection (Barton & Bengtsson 1986; Gavrilets & Cruzan 1998; Bierne 2010; Feder & Nosil 2010). Specifically, the selected locus blocks introgression from the source to the derived populations in its chromosomal neighbourhood, while recombination makes introgression increasingly easy with increasing distance from the locus (Fig. 6C). In other words, effective population size is reduced around the selected locus, promoting localized divergence by genetic drift. The divergence valley persists despite elevated drift, however, because the derived populations continue to share the same haplotype tract around the adaptive variant (Fig. 6C). This second mechanism – the barrier to gene flow – thus reinforces and sharpens the broad peak-valley-peak divergence signature that is initiated by hitchhiking.

Determinants of the peak-valley-peak signature

The physical extent of the peak-valley-peak divergence signature is influenced by several factors (Fig. 2C). First, decreasing gene flow between the source and the derived populations causes the peak-valley-peak to become more extensive – because the overall opportunity for introgression decreases. In the extreme case of no gene flow at all, relatively elevated divergence around the divergence valley produced initially by hitchhiking is rapidly eroded because drift causes divergence among the derived populations across the entire chromosome (Fig. 2B). Second, in the presence of gene flow, the peak-valley-peak signature becomes narrower over time as genetic homogenization through introgression moves closer to the selected locus. (Note that a narrower divergence valley is also expected when the derived variant recombines extensively while standing in the source population prior to parallel adaptation; see Discussion S1, Supporting information.) Third, the divergence twin peaks become higher with a decreasing number of individuals founding the derived populations, which increases stochasticity in the haplotypes linked to the derived variant, hence promoting drift. Finally, the peak-valley-peak signature becomes more

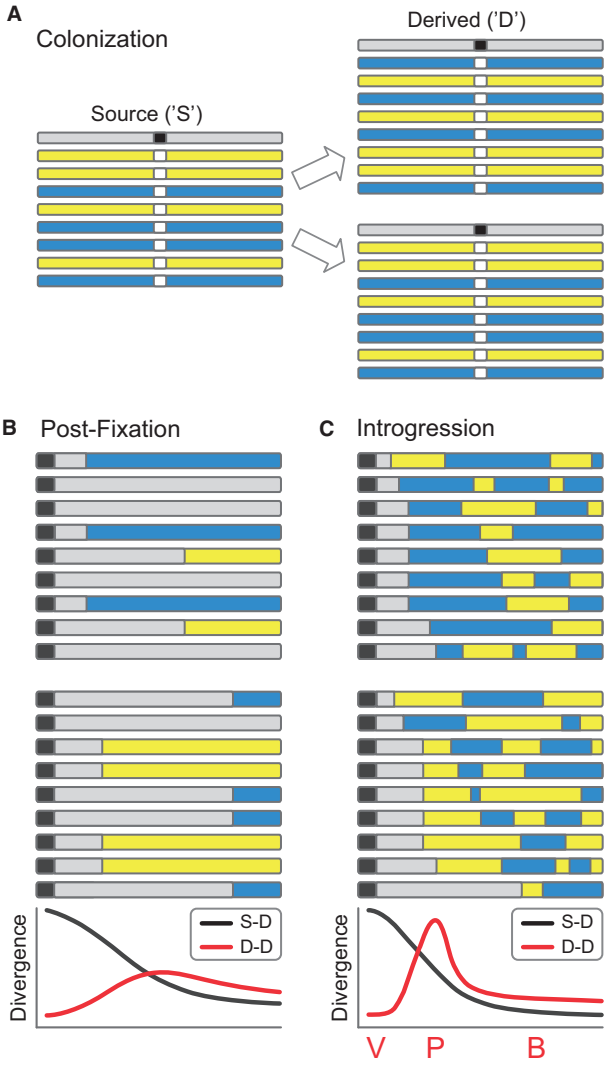

Fig. 6 Mechanisms generating the peak-valley-peak signature of parallel adaption from shared genetic variation. (A) Multiple novel, selectively similar environments are colonized by a source population occupying a selectively different environment. Individuals are represented by a single haploid chromosome, with different colours indicating different genetic backgrounds. The centre of the chromosome holds a locus under divergent selection, with the white variant favoured in the source population, and the black variant favoured in the derived populations but standing at low frequency in the source population as well. (B) Immediately after the parallel fixation of the selected variant, the derived populations share identical haplotype tracts (grey) near the selected locus, whereas population-specific haplotypes (blue, yellow) have hitchhiked further away from the locus [(B) and (C) show the locus and one side of the chromosome only]. As a result, comparisons between derived populations (red line in the bottom panel) show minimal divergence around the selected locus, flanked by a region of elevated divergence. By contrast, comparisons between the source and derived populations reveal the classical signature of a selective sweep (black line). (C) Continuous migration from the source population causes introgressive hybridization in the derived populations. Introgression is impeded in the neighbourhood of the locus, however, where divergent selection produces a barrier to gene flow that locally promotes population divergence by drift. Consequently, comparisons between derived populations reveal a characteristic genomic signature including a divergence valley ('V' in the bottom panel) caused by haplotype sharing flanked by a divergence peak ('P') reflecting elevated drift. Further away from the locus, population divergence decays to the genome-wide migration-drift baseline level ('B').

extensive with an increasing strength of divergent selection between the source and the derived populations, and with decreasing recombination rate. The reason is that both factors render the barrier to gene flow associated with the selected locus more effective (Barton & Bengtsson 1986; Feder & Nosil 2010).

In our simulations with a single selected locus, the physical extent of the peak-valley-peak divergence signature can be quite substantial – many kilobases to a few megabases. Our two-locus models, however, indicate that even more extensive signatures can emerge when multiple loci are simultaneously under selection. Interestingly, the divergence patterns driven in this latter situation vary qualitatively as a function of the recombination distance between the two loci under selection. When these loci are relatively close to each other, a large region of *high* divergence can arise between them ('Closer' in Fig. 2D), although this region does not hold either of the selected loci. This pattern arises because the barriers to gene flow associated with the two loci overlap in this region, making introgression particularly difficult. When the selected loci occur in even closer proximity to each other, however, they together bring to fixation a larger genomic segment shared among the derived populations, resulting in a remarkably wide region of *low* divergence ('Linked' in Fig. 2D). Also, the divergence twin peaks flanking this divergence valley are higher than the peaks driven by each locus alone ('Distant' in Fig. 2D), because the two tightly linked loci together drive a single, stronger barrier to gene flow.

Empirical insights from stickleback

Our empirical system provides an appropriate natural analogue for the conditions specified in our simulations. First, no appreciable genetic divergence was present between our two M samples taken 100 km apart (Fig. 1) (median and mean F_{ST} for all pairwise population comparisons are provided in Table S3, Supporting information). This result is consistent with previous reports of very weak genetic structure within M stickleback

(Hohenlohe *et al.* 2010; Jones *et al.* 2012a; Catchen *et al.* 2013), and it generally supports the established idea that present-day M stickleback provide an appropriate surrogate for the ancestors of derived FW populations. Second, haplotype genealogies confirmed that our FW stickleback populations adapted in parallel at three candidate genes involved in M-FW adaptive divergence, specifically by recycling shared variants from a common M source population (see also Colosimo *et al.* 2005; Jones *et al.* 2012b; Deagle *et al.* 2013). Given these results, we scrutinized patterns of genetic divergence around the three candidate genes to empirically test for the signatures of parallel adaptation from shared variation that were suggested by the simulations.

All three candidate regions exhibited the expected genomic signature of parallel adaptation from shared variation: in comparisons between the derived (FW) populations, the selected loci showed low-divergence valleys that were flanked by high-divergence twin peaks. At the same time, classical source–derived (M-FW) comparisons revealed the expected strong divergence at the candidate genes. Combining these opposed FW-FW and M-FW profiles into 'delta divergence' proved a particularly effective way to reveal parallel adaptation from shared variation. One reason is that these profiles reduce heterogeneity in genomic divergence unrelated to a focal ecological factor (here M vs. FW), such as selective sweeps driven by genetic variants favoured in *all* types of habitats (Bierne 2010). An excellent complementary method was to use genealogical sorting in phylogenetic trees (gsi; Cummings *et al.* 2008) to confirm shared ancestry among the FW populations but exclusive ancestry between M and FW populations. Generally, our ability to detect signatures of parallel adaptation from shared variation was greatly enhanced by high replication at the population level. That is, F_{ST} profiles from *single* pairwise population comparisons (M-FW, FW-FW) exhibited substantial noise (details not presented), which would have made interpretations difficult in the absence of multiple such pairs.

The physical scales of the signatures of parallel adaptation from shared variation were extensive – and similar to those suggested by the simulations. For instance, almost full genealogical sorting occurred over several hundred kilobases around each candidate gene. Moreover, the divergence valley around *Eda* was remarkably wide and displayed *two* divergence minima separated by a small rebound in divergence (at ca. 12.5 Mb in Fig. 3B), as well as massive divergence peaks on either side. This pattern strikingly resembles our simulations with two closely linked loci under selection ('Linked' in Fig. 2D). We therefore propose that the broad neighbourhood of *Eda* is influenced by selection on two genes (or gene clusters) that together produce a particularly effective barrier to gene flow from the M source population. Consistent with this idea, the second divergence minimum near *Eda* coincides with the ATP-binding cassette *Abcb7* (at 12.0 Mb in Fig. 3B), a gene recently suggested to be under divergent selection between M and FW stickleback (Jones *et al.* 2012b). Similarly, a second M-FW adaptation gene near *Spg1* likely influences divergence profiles in that region (Fig. 3B).

A screen of the whole stickleback genome for the joint occurrence of peak-valley-peak signatures of divergence and strong M-FW genealogical sorting identified additional regions on multiple chromosomes likely involved in parallel adaptation from shared variation. As was the case with our initial three candidate genes, these new regions were often flanked by striking divergence twin peaks in the FW-FW comparison, but not in the M-FW comparison (Fig. 4), as predicted by our simulations with gene flow. (Gene flow is known to occur between M and FW populations; Hagen 1967; Jones *et al.* 2006.) Our study thus provides further molecular evidence for divergence in the face of gene flow between contemporary M and FW populations (Catchen *et al.* 2013; Deagle *et al.* 2013). Furthermore, our genome-wide analysis makes a strong case for the notion that adaptation involves numerous loci (e.g. Hohenlohe *et al.* 2010; Lawniczak *et al.* 2010; Fournier-Level *et al.* 2011; Jones *et al.* 2012b; Roesti *et al.* 2012a; Renaut *et al.* 2013), although our methods certainly underestimate the number of loci involved in adaptive divergence between M and FW stickleback (Discussion S2, Supporting information). Finally, our empirical analysis indicated that loci under divergent selection may hinder introgression and drive heterogeneous genomic divergence at the scale of entire chromosomes (Fig. 5).

Implications for ecological genomics

Our results add complexity to the interpretation of regions of low and high divergence discovered in genome scans for signatures of selection. On the one hand, we demonstrate that the common interpretation of regions exhibiting exceptionally low population divergence – that is, localized introgression and balancing selection (Nielsen 2005; Storz 2005) – is potentially problematic; the same pattern can arise when populations use shared genetic variation for parallel adaptation. On the other hand, we also demonstrate that peaks of high population divergence do not necessarily indicate divergent selection. They might instead reflect selectively neutral regions under the influence of neighbouring loci involved in parallel adaptation from shared variation to *similar* environments (for related caveats see Excoffier & Ray 2008; Bierne 2010; Bierne *et al.* 2011). Inference in ecological genomics thus benefits strongly from the integration of multiple complementary analytical

approaches (e.g. source–derived vs. derived–derived comparisons, delta divergence, genealogical sorting; see also Grossman *et al.* 2010), requiring extensive population-level replication within a clear-cut ecological context. On the bright side, genome scans specifically looking for the signature described in our study might help discover adaptation genes in empirical systems where ecological divergence is likely to have occurred repeatedly by recycling genetic variation (e.g. Terai *et al.* 2006; Renaut *et al.* 2011; Tennessen & Akey 2011; Domingues *et al.* 2012; Nadeau *et al.* 2012; Gross & Wilkens 2013).

Acknowledgements

R. Taylor, J.-S. Moore and K. Oke provided marine stickleback samples. A.-C. Grandchamp, J.-S. Moore and K. Hudson aided freshwater stickleback sampling. B. Aeschbach and N. Boileau facilitated wet laboratory work. I. Nissen and C. Beisel performed Illumina sequencing at the Quantitative Genomics Facility, D-BSSE, ETH Zürich. M. Matschiner stimulated the gsi analysis and constructed the haplotype networks. N. Bierne and three anonymous reviewers provided valuable comments on the manuscript. M. Hansen, T. Vines and J. Gow speedily processed the manuscript. Financial support came from the National Institute for Mathematical and Biological Synthesis (NSF Award EF-0830858) and the National Institutes of Health (Grant GM56693) (SG); the Natural Sciences and Engineering Research Council of Canada (APH); the European Research Council (Starting Grant INTERGENADAPT, WS); the Swiss National Science Foundation (Sinergia Grant CRSII3_136293, WS; Ambizione PZ00P3_126391/1, DB); the University of Basel (WS, DB), and the Freiwillige Akademische Gesellschaft Basel (DB). We are grateful to all these people and institutions.

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

Barton NH (2000) Genetic hitchhiking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **355**, 1553–1562.

Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridizing populations. *Heredity*, **57**, 357–376.

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.

Bierne N (2010) The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution*, **64**, 3254–3272.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Catchen J, Bassham S, Wilson T *et al.* (2013) The population structure and recent colonization history of Oregon threespine stickleback determined using RAD-seq. *Molecular Ecology*, **22**, 2864–2883.

Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

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.

Deagle BE, Jones FC, Absher DM, Kingsley DM, Reimchen TE (2013) Phylogeography and adaptation genetics of stickleback from the Haida Gwaii archipelago revealed using genome-wide single nucleotide polymorphism genotyping. *Molecular Ecology*, **22**, 1917–1932.

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.

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, **23**, 347–351.

Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, **64**, 1729–1747.

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

Fournier-Level A, Korte A, Cooper MD *et al.* (2011) A map of local adaptation in *Arabidopsis thaliana*. *Science*, **334**, 86–89.

Gavrilets S, Cruzan MB (1998) Neutral gene flow across single locus clines. *Evolution*, **52**, 1277–1284.

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.

Grossman SR, Shylakhter I, Karlsson EK *et al.* (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, **327**, 883–886.

Hagen DW (1967) Isolating mechanisms in threespine sticklebacks (*Gasterosteus*). *Journal of the Fisheries Research Board of Canada*, **24**, 1637–1692.

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

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.

Jones F, Brown C, Pemberton J, Braithwaite V (2006) Reproductive isolation in a threespine stickleback hybrid zone. *Journal of Evolutionary Biology*, **19**, 1531–1544.

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.

Kaplan NL, Hudson RR, Langley CH (1989) The 'hitchhiking effect' revisited. *Genetics*, **123**, 887–899.

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.

Lawniczak MKN, Emrich SJ, Holloway AK *et al.* (2010) Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science*, **330**, 512–514.

Maynard Smith J, Haigh J (1974) Hitch-hiking effect of a favorable gene. *Genetics Research*, **23**, 23–35.

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 & 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.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Nosil P, Schluter D (2011) The genes underlying the process of speciation. *Trends in Ecology & Evolution*, **26**, 160–167.

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

Orr HA (1998) The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution*, **52**, 935–949.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, 208–215.

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.

Renaut S, Grassa CJ, Yeaman S *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, **4**, 1827.

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.

Schluter D, Clifford EA, Nemethy M, McKinnon JS (2004) Parallel evolution and inheritance of quantitative traits. *American Naturalist*, **163**, 809–822.

Slatkin M, Wiehe T (1998) Genetic hitch-hiking in a subdivided population. *Genetical Research*, **71**, 155–160.

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

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.

Taylor EB, McPhail JD (1986) Prolonged and burst swimming in anadromous and freshwater threespine stickleback, *Gasterosteus aculeatus*. *Canadian Journal of Zoology*, **64**, 416–420.

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.

Walker JA (1997) Ecological morphology of lacustrine threespine stickleback *Gasterosteus aculeatus* L. (Gasterosteidae) body shape. *Biological Journal of the Linnean Society*, **61**, 3–50.

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.

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

Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

M.R. designed the study, carried out the wet laboratory work, analysed and interpreted data and drafted the initial manuscript; S.G. performed the simulations; A.P.H. contributed stickleback samples; W.S. provided infrastructure and financial resources; D.B. designed and directed the study, analysed and interpreted data and wrote the final article, with input from all co-authors.

Data accessibility

Marine and freshwater stickleback RAD sequences: NCBI SRA accession numbers SRP036088 and SRP007695. MATLAB code for the simulations: Supplemental material (Appendix S2). Phased Sanger sequences: Supplemental material (Appendix S3).

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Containing Methods S1–S6, Discussions S1 & S2, Tables S1–S3, Figures S1 & S2.

Appendix S2 MATLAB code for the simulations.

Appendix S3 Phased Sanger sequences for all candidate genes and their reference loci.