nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-020-1222-6

On the causes of geographically heterogeneous parallel evolution in sticklebacks

Bohao Fang[1,2]✉, Petri Kemppainen[1,2]✉, Paolo Momigliano[1], Xueyun Feng[1] and Juha Merilä[1]

The three-spined stickleback (*Gasterosteus aculeatus*) is an important model system for the study of parallel evolution in the wild, having repeatedly colonized and adapted to freshwater from the sea throughout the northern hemisphere. Previous studies identified numerous genomic regions showing consistent genetic differentiation between freshwater and marine ecotypes but these had typically limited geographic sampling and mostly focused on the Eastern Pacific region. We analysed population genomic data from global samples of the three-spined stickleback marine and freshwater ecotypes to detect loci involved in parallel evolution at different geographic scales. Most signatures of parallel evolution were unique to the Eastern Pacific and trans-oceanic marine–freshwater differentiation was restricted to a limited number of shared genomic regions, including three chromosomal inversions. On the basis of simulations and empirical data, we demonstrate that this could result from the stochastic loss of freshwater-adapted alleles during the invasion of the Atlantic basin and selection against freshwater-adapted variants in the sea, both of which can reduce standing genetic variation available for freshwater adaptation outside the Eastern Pacific region. Moreover, the elevated linkage disequilibrium associated with marine–freshwater differentiation in the Eastern Pacific is consistent with secondary contact between marine and freshwater populations that evolved in isolation from each other during past glacial periods. Thus, contrary to what earlier studies from the Eastern Pacific region have led us to believe, parallel marine–freshwater differentiation in sticklebacks is far less prevalent and pronounced in all other parts of the species global distribution range.

The extent to which the evolution of similar phenotypes arises by selection acting on shared ancestral polymorphism (parallel evolution[1]) or via distinct molecular evolutionary pathways (convergent evolution[2–4]) is a major question in evolutionary biology. A powerful approach to disentangle these processes is to study the genomic architecture underlying the repeated evolution of similar phenotypes in similar environments. After the retreat of Pleistocene glaciers, marine three-spined sticklebacks (*Gasterosteus aculeatus*) colonized many newly formed freshwater habitats and adopted similar changes in morphological, physiological, life history and behavioural traits[5–9]. Thus, this species has become one of the most widely used model systems to study the molecular basis of adaptive evolution in vertebrates in the wild[10].

Previous studies of the three-spined stickleback model system have quantified the extent of parallel evolution by identifying genomic regions that are consistently differentiated between marine and freshwater ecotypes sampled across different geographic areas[3,11–18]. The focus has historically been on the Eastern Pacific region[11,13,14,19–21] but several recent studies have focused on Atlantic populations[12,15–18]. However, only two studies have so far included samples from a larger (global) geographic range[3,11]. On the basis of whole-genome sequence data from a few individuals from Eastern Pacific and Atlantic populations ($n=21$), Jones et al.[11] identified ~200 genomic regions that consistently separated marine and freshwater individuals globally, representing ~0.5% of the dataset. They also found that 2.83% of the genome showed signatures of parallel selection in Eastern Pacific freshwater locations—approximately six times more than that at the global scale (tree i, supplementary fig. 2 and supplementary table 7 in Jones et al.[11])—suggesting that more loci contribute to parallel evolution at smaller geographic (regional) scales. However, since this pattern was unique to the Eastern Pacific (and the focus was on global parallelism) its implications for a holistic understanding of marine–freshwater differentiation at both regional and global scales was never discussed.

Such global heterogeneous ecotype divergence is consistent with the results of several other studies as well. Focusing on 26 candidate genes in six pairs of marine–freshwater populations across the globe, DeFaveri et al.[3] found that only ~50% of the genes under divergent selection were shared across more than three population pairs and none was shared among all populations. This suggested a limited reuse of ancestral polymorphism at the global scale, implicating either an important role of convergent evolution at larger geographic scales[3] or geographic heterogeneity in selective pressure among different[3] freshwater ecosystems[3,4]. Furthermore, studies focusing on parallel evolution within oceans, and even smaller geographic regions, show striking differences in the proportion of loci involved in parallel freshwater adaptation between Pacific and Atlantic regions[11–13,15–18,21]. For instance, Terekhanova et al.[17,18] recovered only 21 highly localized genomic regions involved in parallel freshwater differentiation in the White Sea, in contrast to Hohenlohe et al.[13] and Nelson and Cresko[21] who found large genomic regions involved in parallel freshwater differentiation across almost all chromosomes in the Eastern Pacific populations.

Therefore, the potential mechanisms underlying this apparent large-scale geographic heterogeneity in genome-wide patterns of parallel evolution in three-spined sticklebacks remain unexplored. To this end, we analysed population genomic data from a comprehensive sampling of all major geographic areas inhabited by the three-spined stickleback, and used unsupervised and supervised methods to detect loci involved in parallel marine–freshwater

[1]Ecological Genetics Research Unit, Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland. [2]These authors contributed equally: Bohao Fang, Petri Kemppainen. ✉e-mail: bohao.fang@helsinki.fi; petri.kemppainen@helsinki.fi

Fig. 1 | Linkage disequilibrium network analysis. a, LD-cluster reflects geographical structure between the Pacific and Atlantic oceans. **b**–**d**, LD-clusters associated with marine–freshwater genetic parallelism in the Eastern Pacific. **e**–**h**, Examples of LD-clusters involved in trans-oceanic genetic parallelism. In each panel (LD-cluster), the top and middle plots present the marine–freshwater differentiation (F_{ST}) between Atlantic and Eastern Pacific samples, respectively. The bottom plot shows the PCA based on the LD-cluster loci. The seven different colours represent the geographic origin of populations. Solid and open circles refer to freshwater and marine ecotypes, respectively. All identified LD-clusters (29 in total) and corresponding information are presented in Extended Data Fig. 1 and Extended Data Fig. 7. **i**, Map of the sampled populations; colours match those in the PCA results. A Mercator projection of the sampling map is shown in Extended Data Fig. 6.

differentiation at different geographical scales. On the basis of earlier observations[3,12,15–18,22], we suggest that the genetic parallelism in response to freshwater colonization by marine sticklebacks is heterogeneous at the global scale, and that the degree of genetic parallelism is much stronger in the Eastern Pacific region than elsewhere.

We further seek to understand and discuss the ultimate causes of the marked regional differences in genome-wide signatures of parallel genetic differentiation among ecotypes. To explain the mechanism behind the repeated use of the same alleles in independent freshwater populations of sticklebacks, Schluter and Conte[1] proposed the 'transporter hypothesis'. This postulates that three-spined sticklebacks have repeatedly colonized and adapted to freshwater environments via selection on standing genetic variation in large marine populations. These freshwater-adapted alleles are in turn maintained in the marine populations by recurrent gene flow with previously colonized freshwater populations. Three-spined stickleback populations have persisted in the Eastern Pacific since ~26 Ma (million years ago[23–26]) and from there recolonized the Western Pacific and Atlantic Ocean basins following local extinctions much more recently, during the late Pleistocene (36.9–346.5 ka [thousand years ago][27–29]). During bottlenecks and founder events, rare alleles are lost at a higher rate than common alleles[30,31]. Since freshwater-adapted alleles exist in the marine populations only at low frequencies[1], it is likely that they were lost to a higher degree than neutral variation during geographic range expansions from the Eastern Pacific (via the sea), thereby reducing the standing genetic variation available for freshwater adaptation outside of the Eastern Pacific. To test this hypothesis, we used individual-based forward simulations designed to mimic the transporter hypothesis and the general global population demographic history of three-spined sticklebacks outlined above. We conclude with a discussion on other potential biological and demographic explanations for the high degree of geographic heterogeneity in patterns of parallel genomic differentiation, and reflect upon the representativeness of the Eastern Pacific three-spined stickleback populations as a general model for the study of parallel evolution.

Results

Marine–freshwater divergence determined by unsupervised and supervised approaches. The Linkage Disequilibrium Network Analysis (LDna) was applied on a dataset including 2,511,922 single nucleotide polymorphisms (SNPs) derived from 166 individuals worldwide to identify and extract clusters of highly correlated loci; that is, sets of loci affected by the same evolutionary processes (LD-clusters). The first step of LDna identified 214,326 loci that were in high LD with at least one other locus within windows of 100 SNPs (Supplementary Information 1). The next step of performing LDna on each chromosome separately (only using one locus from each LD-cluster from step one; Supplementary Information 1) resulted in 81 distinct LD-clusters. From these, a final 29 LD-clusters were obtained (pooling within chromosome LD-clusters whenever they were grouped by LDna in the final step; Supplementary Information 1), containing a total of 71,064 loci (namely clusters 1–29). Eight of these LD-clusters associated with geographic structure and genetic parallelism are highlighted in Fig. 2a–h. Details of all 29 clusters can be found in Extended Data Fig. 1 and Extended Data Fig. 7.

LDna successfully recovered most of the previously identified regions from Jones et al. [11] that differentiated marine from freshwater

Fig. 2 | Genetic parallelism identified by the unsupervised and supervised methods. a, Comparison of marine–freshwater differentiation (F_{ST}) in the Atlantic (*x* axis) and Eastern Pacific (*y* axis) datasets for the three LD-clusters (LD-clusters 2, 21 and 29) associated with strong marine–freshwater parallelism in the Eastern Pacific. **b**, Genome-wide F_{ST} of the Eastern Pacific samples for loci of the LD-clusters coloured as in **a**. **c**, The same as **a** but for the 12 LD-clusters (5, 6, 10, 11, 12, 13, 16, 18, 20, 22, 25 and 27) that are involved in global marine–freshwater genetic parallelism. **d**,**e**, Genome-wide F_{ST} of the Atlantic and Eastern Pacific samples, respectively, with colours corresponding to LD-cluster loci in **c**. The position of the EDA locus is indicated in **d** and **e**.

ecotypes and failed to recover only small regions that had low coverage and relatively low levels of marine–freshwater differentiation (Supplementary Information 2 and Extended Data Fig. 2).

Trans-oceanic marine–freshwater parallelism. LD-clusters 5, 6, 10, 11, 12, 13, 16, 18, 20, 22, 25 and 27 (a total of 2,502 loci, 0.100% of the dataset; see four representatives in Fig. 1e–h and all in Extended Data Figs. 1 and 2) grouped multiple freshwater individuals from different geographic regions across the Pacific and Atlantic Oceans (for all $P < 0.05$, permutation test for ecotype differentiation), reflecting genetic marine–freshwater parallelism on a global (trans-oceanic) scale. Within those, LD-clusters 6, 11, 12, 16 and 27 (a total of 1,639 loci, 0.065% of the dataset) similarly showed high marine–freshwater allelic differentiation (F_{ST}; Fig. 1e–h and Fig. 2c–e) in both the Eastern Pacific and Atlantic, further suggesting global parallelism. Particularly, loci from LD-cluster 11 mapped to four distinct regions on chromosome V of which one (Fig. 2c–e) marks the position of the ectodysplasin A (EDA) gene that is known to be responsible for marine–freshwater differences in lateral armour plate development worldwide[20]. In contrast, although the remaining clusters (5, 10, 13, 18, 20, 22, 25, a total of 863 loci, 0.034% of the dataset) also grouped freshwater individuals from both the Pacific and Atlantic Oceans (similar to LD-cluster 29 above; Fig. 2c–e and Extended Data Fig. 1), they showed much less marine–freshwater differentiation in the Atlantic Ocean than in the Eastern Pacific (Fig. 2c–e). Among the LD-clusters associated with marine–freshwater differentiation, LD-clusters 6 and 22 covered previously known chromosomal inversions on chromosomes I and XI, respectively[11] (Fig. 1e, Extended Data Fig. 1 and Extended Data Fig. 7). In addition, we also found a putative novel inversion on chromosome V (LD-cluster 19, 241 loci) that was not associated with marine–freshwater differentiation (Extended Data Fig. 1 and Extended Data Fig. 7). LDna and F_{ST} analyses did not detect any regions showing marine–freshwater differentiation in the Western Pacific (Extended Data Fig. 3c) and thus, this ocean is not considered further here.

Eastern Pacific marine–freshwater parallelism. LD-clusters 2 (53,785 loci, 2.141% of the dataset; Fig. 1b) and 21 (183 loci, 0.007% of the dataset; Fig. 1c) separated Eastern Pacific freshwater individuals exclusively from the remaining samples, a pattern that is not expected by chance alone (permutation test $P < 0.001$; Extended Data Fig. 1 and Extended Data Fig. 7). Rather, this reflects a shared adaptive response among Eastern Pacific freshwater populations. The exception was two freshwater individuals from the Eastern Pacific (Alaska) that did not group with the other freshwater individuals from the Eastern Pacific but instead with Atlantic (marine and freshwater), Western Pacific (marine and freshwater) and the marine individuals from Eastern Pacific (Fig. 1c and Extended Data Fig. 4). LD-cluster 29 (2,728 loci, 0.109% of the dataset) covering a known inversion on chromosome XXI (Fig. 1d) grouped the Eastern Pacific freshwater individuals (except the two Alaskan individuals above) together with six Atlantic freshwater individuals and one Eastern Pacific marine individual. Because this LD-cluster maps to an inversion, the groups also represent putative inversion karyotypes[22].Thus, this inversion shows strong ecotype differentiation not only in the Eastern Pacific but also in a small proportion of individuals outside of the Eastern Pacific that putatively carry the freshwater-adapted karyotype (the karyotype with the highest frequency among Eastern Pacific freshwater individuals). Notably, no cluster of similar magnitude to LD-cluster 2—which separates freshwater individuals from one specific region from all remaining samples in the data—could be detected outside of the Eastern Pacific, demonstrating that parallel marine–freshwater differentiation in the Eastern Pacific is much more prevalent than anywhere else in the world.

A small proportion of the loci from LD-cluster 2 (28 SNPs) mapped to regions that have showed global parallelism[11]. (Extended Data Fig. 2). In addition, <1% of all loci in LD-cluster 2 (243 SNPs) showed $F_{ST} > 0.2$ also in the Atlantic (as is evident for example from Fig. 2a and Extended Data Fig. 3a). These loci appear to be non-randomly distributed in the genome (Extended Data Fig. 3a), indicating that indeed they are likely to be linked to genomic regions involved in marine–freshwater differentiation in both the Atlantic and the Eastern Pacific. Due to small sample sizes, the F_{ST} Manhattan plots display a considerable amount of noise, particularly in the datasets from the Eastern and Western Pacific Oceans (Fig. 2b, e and Extended Data Fig. 3).

Geographic structure and regional local adaptation. LD-cluster 1 (10,184 loci, 0.405% of the dataset) separated all Pacific individuals (Eastern and Western) from the Atlantic individuals (Fig. 1a), thus mainly reflecting trans-oceanic geographic structure. LD-clusters 4, 8, 9, 14 and 24 (a total of 526 loci, 0.021% of the dataset; Extended Data Fig. 1) separated freshwater individuals from only one geographic region; this probably reflects geographic clustering but could also contain some loci involved in non-parallel freshwater adaptation. We therefore could not determine the underlying evolutionary phenomena that produced these clusters (Extended Data Fig. 2). Accordingly, loci from these LD-clusters showed little marine–freshwater differentiation in both the Eastern Pacific and Atlantic (Extended Data Fig. 1) and only two loci (from LD-cluster 14) mapped to the marine–freshwater divergent regions identified by Jones et al.[11] regions (Extended Data Fig. 2).

Proof of concept simulations. Several potential explanations for geographic heterogeneity in parallel patterns of marine–freshwater differentiation in three-spined sticklebacks have been suggested[3]. One such explanation that has not received much attention in the context of three-spined sticklebacks is the stochastic loss of freshwater-adapted alleles due to founder events when three-spined sticklebacks colonized the rest of the world from the Eastern Pacific in the late Pleistocene (see earlier). Thus, as a proof of concept, we used forward-in-time simulations to investigate the conditions under which parallel islands of differentiation between marine and freshwater ecotypes can arise under such a scenario.

In the simulated data, before Atl (the simulated Atlantic Ocean) was colonized from Pac (the simulated Pacific Ocean) before the closing of the Bering Strait (38–40 ka), all five freshwater populations in Pac were fixed or nearly fixed for the freshwater-adapted alleles of all locally adapted quantitative trait loci (QTL; Fig. 3f). Following the colonization of Atl, the increased frequency of the freshwater allele in the Atlantic freshwater populations depended on both QTL density and the level of gene flow between Pac and Atl (Fig. 3f). The highest increase in freshwater-adapted alleles in Atl was observed when QTL density was low (three QTL per chromosome) and trans-oceanic gene flow was high (five migrants/generation, Fig. 3f). During post-glacial colonization of new freshwater habitats from the sea (10 ka to present), freshwater-adapted alleles (in both Pac and Atl) gradually increased in the newly formed freshwater populations (Fig. 3f), reflecting local adaptation. This increase was similarly dependent on the QTL density (both in Pac and Atl) and trans-oceanic gene flow (only affecting Atl, Fig. 3f). These patterns probably reflect the underlying levels of ancestral variation in the sea available for subsequent freshwater adaptation (Supplementary Fig. 1a). The lowest frequencies of freshwater-adapted alleles in the sea were always observed when QTL density was the highest (in both Pac and Atl) and trans-oceanic gene flow was the lowest (only affecting the Atl; Supplementary Fig. 1a). Furthermore, the frequency of freshwater-adapted alleles in both the sea (ancestral variation) and in the post-glacial freshwater populations (local adaptation) depended on whether the

Fig. 3 | Ecological genetics in simulated data. a–e, A schematic of the demographic scenario used for the simulations that is consistent with the 'transporter hypothesis'. **a**, Initial local adaption of the freshwater populations in the Pacific. **b**, The colonization of stickleback populations from the Pacific to the Atlantic. **c**, Geographic isolation between the two oceans. **d**, Extinction of lakes during the LGP with the survival of refuge populations. **e**, The post-glacial colonization of the new freshwater populations. **f**, Frequency of selected (freshwater-adapted) alleles in the newly established freshwater populations through generations at high and low levels of trans-oceanic gene flow and different QTL densities. **g**, PCA of the empirical data (LD-cluster 2; left) and the simulated data (right), with ecotypes and geographical regions as shown in the figure legend. **h**, CSS of the empirical and simulated data in the Pacific and Atlantic oceans, respectively. **i**, Boxplots of observed heterozygosity in different geographical regions in the empirical and simulated data (empirical data: generalized linear model [GLM], $F_{2,64} = 43.05$, $P < 0.001$; simulated data: GLM, $F_{1,238} = 509.7$, $P < 0.001$; Supplementary Information 3). Only trends, rather than absolute values, of heterozygosity should be compared between empirical and simulated data (refer to Extended Data Fig. 5 for further information). Chr., chromosome.

QTL were located in low- or high-recombination regions; the lowest frequencies of freshwater-adapted alleles were always observed in low-recombination regions (Supplementary Fig. 1b,c). The freshwater-adapted alleles in both Pac and Atl freshwater populations never reached similar frequencies during the post-glacial colonization (10 ka; Fig. 3f) to those found before post-glacial colonization (>10 ka; Fig. 3f), showing that ancestral variation in the sea was not sufficient to allow complete local adaptation (that is, fixation of all original freshwater-adapted alleles) in our simulations. Note that with the rapid fixation of all the freshwater-adapted alleles (that started at frequency 0.5 in the freshwater populations in Pac) and the low mutation rate used (1×10^{-8} per site and generation), the contributions of de novo mutations (at the QTL) to freshwater adaptation in these simulations are negligible.

In the simulations, present-day marine–freshwater differentiation (mean neutral F_{ST}) was always low for the two chromosomes without QTL as well as in high-recombination regions of chromosomes that contained QTL (Fig. 4 and Supplementary Fig. 1d). In contrast, F_{ST} for low-recombination regions of QTL-containing chromosomes was high for Pac (for all parameter settings), indicating strong islands of parallel marine–freshwater differentiation. This was also true for Atl when QTL density was low (three or six QTL per chromosome) and when trans-oceanic gene flow was high but not when QTL density was high (nine QTL per chromosome; Fig. 4 and Supplementary Fig. 1d) and trans-oceanic gene flow was low.

In the LD-clusters with the strongest Pacific freshwater (PF) versus non-PF genetic differentiation in the simulated data, the cluster separation score (CSS; the scaled centroid distance based on the two first principal components [PCs], with range [0,1]) were always high (>0.75) between PF and the Atlantic populations (Atlantic marine and freshwater, AF and AM, respectively), similar to LD-cluster 2 (Fig. 3g, h and Supplementary Fig. 1e). However, the CSS between PF and Pacific marine (PM) for LD-cluster 2 was also high (0.62), in contrast to the simulated data where this score was low (starting from <0.2) but increased with QTL density and more so when migration rate during colonization of Atl from Pac was high (five migrants/generation; Fig. 3g,h and Supplementary Fig. 1e). This is probably due to QTL density increasing the distance (in the principal component analyses, PCA) between PF and PM and migration rate decreasing the distance between Pac and Atl individuals, as CSS here is scaled by the maximum Euclidean distance between any two points in the data. Furthermore, when migration rate was low, CSS between AF and AM was small for all LD-clusters. However, when migration rate was high and with increasing QTL density, LD-clusters similar to LD-cluster 2 were readily observed also in Atl. This shows that in simulations, low migration rates and high QTL densities are required to produce patterns similar to the observed data.

Discussion

Using genome-wide SNP data from a comprehensive global sampling of marine and freshwater stickleback ecotypes, we demonstrate that a much smaller proportion of the genome (0.208% of the dataset) is involved in global parallel marine–freshwater differentiation than exclusively in the Eastern Pacific (2.149% of the dataset). This shows

Fig. 4 | Genomic differentiation in simulated data. a,**b**, Genome-wide marine–freshwater differentiation (F_{ST}) from simulated data (data from the last generation representing present-day sampling): Eastern Pacific (**a**) and Atlantic (**b**). For each parameter combination, loci from all 20 replicates were pooled. Red dots indicate QTL and blue dots indicate loci from LD-clusters that were the most similar to LD-cluster 2 (empirical data) showing the strongest marine–freshwater differentiation in the Eastern Pacific (grey represents non-LD-cluster loci). **c**, F_{ST} distribution of QTL in the simulations (all replicates pooled), indicating the proportion of loci that were either fixed (F_{ST} ~ 1), lost (F_{ST} ~ 0) or were fixed to different degrees in only one, two or three of the four freshwater populations ($0.1 \lesssim F_{ST} \lesssim 0.9$) since post-glacial colonization. A small amount of noise (along the *x* axis) has been added to the QTL positions to improve their visibility. Chr., chromosome.

that parallel evolution in the three-spined stickleback is much more pervasive in the Eastern Pacific than anywhere else in the world. Indeed, the LD signal from marine–freshwater differentiation in the Eastern Pacific is even stronger than that from geographic structuring between the Pacific and Atlantic Oceans—LD-clusters separating freshwater individuals from the Eastern Pacific comprised five times as many loci as the LD-cluster reflecting geographic structuring between Pacific and Atlantic Oceans. With simulations, we demonstrate that this pattern could partly be explained by the stochastic loss of low-frequency freshwater-adapted alleles in the sea during range expansion from the Eastern Pacific. As predicted, the discrepancy between the simulated Pacific and Atlantic populations in both F_{ST} and CSS analyses was the highest when trans-oceanic gene flow was low (stronger founder event) but this also required the QTL density of locally adapted loci to be high, as this reduced the levels of standing variation of freshwater-adapted alleles in the Atlantic. However, the loss of ancestral variation due to founder effects and the transporter hypothesis is probably not the only explanation for the large discrepancy in patterns of marine–freshwater differentiation between the Eastern Pacific and Atlantic Oceans. In the following, we discuss other alternative biological processes that could potentially contribute to this discrepancy.

Geographic heterogeneity in standing genetic variation. The 'transporter hypothesis'[1] postulates that a low frequency of freshwater-adapted alleles is maintained in the sea via recurrent gene flow between ancestral marine and previously colonized freshwater populations. This standing genetic variation is what selection acts on during the subsequent colonization of freshwater habitats. This implicitly assumes that a large pool of locally adapted alleles has accumulated over a long period of time, as gene flow is expected to spread potentially beneficial mutations across demographically independent populations[32,33]. In support of this hypothesis, it has been shown that haplotypes repeatedly used in freshwater adaptation are identical by descent[20,34] and old—on average, 6 million years (Myr) but some are reported to be as old as 15 Myr (ref. [21]). Notably, these studies analysed populations from the Eastern Pacific region, which represents the oldest and most ancestral marine population[27,28] where three-spined sticklebacks are thought to have persisted since the split from their close relative, the nine-spined stickleback (*Pungitius pungitius;* ~26 Ma; refs. [23–26,35]). However, populations in the Western Pacific and the Atlantic are much younger, as they were colonized from the Eastern Pacific during the late Pleistocene (36.9–346.5 ka; refs. [27,28]). There is no evidence for trans-oceanic admixture[27,28] following the split of Pacific and

Atlantic clades, and there are no extant populations of three-spined sticklebacks in arctic Russia between the Kara Sea and the Eastern Siberian Sea. Thus, the spread of freshwater-adapted alleles from the Eastern Pacific to elsewhere via migration through the Bering Strait is unlikely and has probably not occurred in recent times. Our simulations show that following colonization of freshwater populations from the sea, the accessibility of freshwater-adapted alleles—which is a function of colonization history, QTL density and recombination rate—largely determines the number of loci that show high marine–freshwater differentiation. Thus, consistent with previous simulations[34,36], genomic islands of differentiation in linked neutral loci require several QTL to cluster in low-recombination regions (Fig. 4 and Supplementary Fig. 1d). Furthermore, when trans-Atlantic gene flow was low and QTL density was high, we readily observed LD-clusters that showed high marine–freshwater differentiation only in the Eastern Pacific, not in the Atlantic.

Our simulations and empirical data suggest that both stochastic loss of genetic diversity and selection against freshwater-adapted variants played a role in reducing the pool of standing genetic variation of freshwater-adapted alleles in the Atlantic region. During range expansions, genetic diversity is expected to decrease with distance from the source population from which the expansion started[37]. This pattern was clear in our simulations as well as in the empirical data, both of which show a statistically significant reduction in heterozygosity in the Atlantic region as compared to the Eastern Pacific region (Fig. 3i and Supplementary Information 3). These results are consistent with a moderate founder effect following colonization of the Atlantic basin from the Eastern Pacific (Supplementary Information 3), which could account for the random loss of standing genetic variation of freshwater-adapted alleles. Furthermore, since freshwater-adapted alleles are selected against in the sea and thus occur at low frequencies in marine environments, they are even less likely to spread to new geographic regions than are neutral alleles[30,31]. Consistent with this hypothesis, the mean individual heterozygosity for LD-cluster 2 loci was 29 times higher in the Pacific compared to the Atlantic Ocean, a very pronounced difference compared to that in the rest of the genome (Extended Data Fig. 5b). However, the between-ocean differences in marine–freshwater differentiation in the simulated data (Fig. 4) were much less pronounced compared to the empirical data (Fig. 2d). Thus, founder effects and selection are probably not the only factors affecting patterns of marine–freshwater differentiation in the three-spined sticklebacks at the global level.

Alternative explanations for the observed discrepancy in patterns of marine–freshwater differentiation between the Eastern Pacific and Atlantic include: (1) stronger spatial genetic structure in marine populations outside of the Eastern Pacific causing heterogeneity in standing genetic variation available for freshwater adaptation and (2) heterogeneity in selective regimes among freshwater habitats, both between Atlantic and Eastern Pacific Oceans and between different geographic areas in the Atlantic. We have further tested these hypotheses but found little, or inconclusive, support in our data and in other studies (Supplementary Information 3).

Secondary contact in the Eastern Pacific. All of the above hypotheses assume that the original source of the ancestral variation in the Eastern Pacific and elsewhere is the same. That is, ancient Eastern Pacific marine populations carried most ancestral variation of freshwater-adapted alleles at low frequencies, sourcing the Atlantic region with freshwater-adapted alleles which were partially lost either stochastically or due to selection. However, an alternative hypothesis is that modern Eastern Pacific freshwater variants were not present in the marine ancestors but rather in land-locked, ice-lake freshwater populations[38]. As glaciers melted, those populations could have followed meltwater downstream, establishing freshwater populations with a different stock of alleles. Secondary contact with marine sticklebacks during this time might have eroded genetic differentiation across most of the genome, with the exception of those regions involved in freshwater adaptation. In this scenario, the standing genetic variation responsible for Eastern Pacific freshwater adaptation may not have entered the Eastern Pacific marine population until after the end of the last glaciation. Since there is no evidence of gene flow between Atlantic and Pacific populations after the closing of the Bering Strait (60–30 ka), this ancestral variation could have remained unique to the Pacific Ocean.

There is ample evidence for large ice-lakes during the last glacial period (LGP) in North America, with little (if any) connection to the sea[39–42]. Thus, a large part of the genetic variation underlying marine and freshwater adaptation in the Eastern Pacific could in principle have evolved in allopatry; that is, separately among the freshwater ice-lake populations and in the sea (and any other potential freshwater water bodies the sea is in contact with). Consistent with this is the strong pattern of long-range LD observed among Eastern Pacific marine individuals[43], as well as our LDna results which revealed one large cluster separating the Pacific and Atlantic Ocean individuals, and one that specifically separates all Eastern Pacific freshwater individuals from all other individuals. The secondary contact hypothesis is also consistent with close to zero heterozygosity in Atlantic marine individuals observed for LD-cluster 2 loci (Extended Data Fig. 5a), as well as the mismatches in marine–freshwater differentiation in the simulated compared to the empirical data. Curiously, we found significant isolation-by-distance in the Atlantic but not in the Eastern Pacific where overall population structuring was nevertheless higher than in the Atlantic (Extended Data Fig. 5d). This could be consistent with the secondary contact hypothesis, if introgression was stronger in some regions of the Eastern Pacific compared to others. However, further empirical and simulation studies are needed to test the extent to which this secondary contact hypothesis provides a better explanation for the observed data than the transporter hypothesis alone.

Conditions that allow global parallelism. Genomic islands of parallel ecotype divergence were more likely to arise in the simulations when several QTL clustered in the same low-recombination region. Surprisingly these were also the QTL where the frequency of the freshwater-adapted allele was lowest in the sea and, thus, were least likely to spread to Atlantic during colonization from Pacific. Since QTL in low-recombination regions are less likely to be separated by recombination when freshwater-adapted individuals migrate to the sea, it is reasonable to assume that the selection pressure against these 'haplotypes' in the sea is stronger[43]. However, this is not consistent with the empirical data showing that the genomic regions most likely to show global parallel ecotype divergence are inversions, where recombination in heterokaryotypes is particularly restricted. Our simulations assume that freshwater-adapted alleles are selected against in the sea (and the strength of this selection is equal for all QTL) while in reality, selection against some of the 'freshwater haplotypes/karyotypes' in the sea may be weak or even absent, allowing them to easily spread during range expansions. Consistent with this reasoning, in PCAs based on loci from LD-clusters corresponding to inversions (LD-clusters 6, 22 and 29) several marine individuals also cluster with the freshwater individuals (Fig. 1d,e and Extended Data Fig. 1), indicating frequent occurrence of the 'freshwater karyotypes' in the sea. Indeed, Terekhanova et al.[17] found that the genomic regions most commonly involved in local adaptation in many independent freshwater populations were also those with the highest frequencies in the sea. In other words, the most geographically widespread genomic regions involved in freshwater adaptation (sensu the transporter hypothesis) are likely to experience the weakest selection against them in the sea, allowing them to remain at higher frequencies in the sea as standing genetic variation[17].

Are three-spined sticklebacks a representative model to study parallel evolution?. Since the pattern of parallel genetic differentiation between marine and freshwater stickleback ecotypes in the Eastern Pacific is in stark contrast to what is seen across other parts of the species distribution range, it is reasonable to question the generality of the findings from the Eastern Pacific stickleback studies with respect to parallel evolution at broader geographic scales. A recent review of parallel evolution suggests that even dramatic phenotypic parallelism can be generated by a continuum of parallelism at the genetic level[44]. For instance, the coastal ecotypes of *Senecio lautus* show only partial reuse of particular QTL among replicate populations[45] and genetic redundancy frequently underlies polygenic adaptation in *Drosophila*[46]. Similarly, using F_{ST} outliers to detect putative genomic targets of selection, Kautt et al.[47] (cichlid fish), Le Moan et al.[48] (anchovy) and Westram et al.[49] (periwinkles) showed that phenotypically very similar populations often share only a small proportion of their F_{ST} outliers.

One exception that seems more general across taxa is the repeated involvement of chromosomal inversions in parallel evolution. Chromosomal inversions could store standing variation as a balanced polymorphism and distribute it to fuel parallel adaptation[50]. For instance, the same chromosome I inversion is involved in global marine–freshwater differentiation in three-spined sticklebacks[11,15,17,18], as in this study, also differentiates stream and lake ecotypes in the Lake Constance basin in Central Europe[51]. Two other clear examples where most genetic differentiation between ecotypes at larger geographic scales is partitioned into inversions come from monkey flowers (*Mimulus guttatus*[52]) and marine periwinkles (*Littorina saxatilis*[53,54]).

While our study focuses exclusively on marine–freshwater ecotype pairs of three-spined sticklebacks, other ecotype pairs within freshwater habitats, such as stream versus lake and benthic versus limnetic, also exist. A recent study focusing on stream–lake populations found that putative selected loci showed greater parallelism in the Eastern Pacific (Vancouver Island) than at the global scale (North America and Europe[55]), that is a similar pattern as reported by our study. Furthermore, Conte et al.[56] studied the extent of QTL reuse in parallel phenotypic divergence of limnetic and benthic three-spined sticklebacks within Paxton and Priest Lakes (British Columbia) and found that although 76% of 42 phenotypic traits diverged in the same direction, only 49% of the underlying QTL evolved in parallel in both lakes. For highly parallel traits in two other pairs of benthic–limnetic sticklebacks, only 32% of the underlying QTL were reported to be shared[57]. Thus, these studies are also in stark contrast to the original conclusions of widespread genetic parallelism in three-spined sticklebacks. Notably, the two freshwater individuals from the Eastern Pacific that did not cluster with the remaining freshwater individuals from the Eastern Pacific (and were subsequently removed from the datasets used for F_{ST} genome scans) were from Alaska. These two individuals are also phylogenetically distinct from other freshwater individuals from the Eastern Pacific[28]. One explanation for this could be that the highly divergent freshwater populations in the Eastern Pacific have a different colonization history than the Alaskan lakes. More specifically, the former could have been colonized from some divergent ice-lake refugia (see above), whereas the latter could have independently been colonized from the sea.

Conclusions

Our results demonstrate that genetic parallelism in the marine–freshwater three-spined stickleback model system is not as pervasive as some earlier studies focusing on Eastern Pacific populations have led us to believe. Our analysis of geographically more comprehensive data, with similar and less assumption-burdened methods as used in earlier studies, shows that the extraordinary genetic parallelism observed in the Eastern Pacific Ocean is not detectable elsewhere in the world (for example, Atlantic Ocean and Western Pacific Ocean). Hence, the focus on the Eastern Pacific has generated a perception bias—the patterns detected there do not actually apply to the rest of the world. Furthermore, our simulations show that the spread of freshwater-adapted alleles from the Pacific to the Atlantic could have been hampered by founder events, particularly for QTL clustered in low-recombination regions (those most likely to result in parallel islands of ecotype differentiation). Therefore, geographic differences in the incidence and pervasiveness of parallel evolution in three-spined sticklebacks probably stem from geographic heterogeneity in access to, and amount of, standing genetic variation, which in turn has been influenced by selection as well as historical population demography. Such historical demographic factors include founder events as well as the potential accumulation of genetic ecotypic differences in allopatry during the last glacial maximum, followed by a secondary contact only after the Atlantic Ocean was colonized via the sea from the Eastern Pacific. Hence, while striking genome-wide patterns of genetic parallelism exist (for example, in Eastern Pacific sticklebacks), the conditions under which such patterns can occur may be far from common, perhaps even exceptional.

Methods

Sample collection. We obtained population genomic data from 166 individuals representing both marine and freshwater ecotypes from the Eastern and Western Pacific, as well as from the Eastern and Western Atlantic Oceans (Fig. 1i, Extended Data Fig. 6 and Supplementary Table 1). Additional data from previously published studies were retrieved from GenBank. Fish collected for this study were sampled with seine nets, minnow traps and electrofishing. Specimens were preserved in ethanol after being euthanized with an overdose of tricaine mesylate (MS222). The samples were collected under appropriate national fishing or ethical licences granted to collectors of the respective countries listed in acknowledgement in Fang et al.[28]. In Finland, the fishing is permitted by land owner according to the Finnish Fishing Law (5§ 27.5.2011/600). The research does not involve animal experiments according to the Act of Animal Experimentation (FINLEX 497/2013).

To study the extent of genetic parallelism among freshwater sticklebacks with different phylogeographic histories, we classified global samples into seven biogeographic regions based on their phylogenetic affinities: (1) Eastern Pacific, (2) Western Pacific, (3) Western Atlantic, (4) White Sea and Barents Sea, (5) North Sea and British Isles, (6) Baltic Sea and (7) Norwegian Sea[28] (Fig. 1i). A summary of coordinates, ecotype and population information on the sampled individuals and reacquired samples is given in Supplementary Table 1.

Sequencing and genotype likelihood estimation. Restriction site associated DNA sequencing (RAD-seq) using the enzyme PstI was performed for the 62 individuals sampled in this study, using the same protocol as in Fang et al.[28], where DNA library preparation and sequencing method are described in detail. The raw RAD-seq data have been uploaded to GenBank. Previously published RAD-seq and whole-genome sequencing (WGS) data for an additional 104 individuals from 62 populations were retrieved from GenBank (Supplementary Table 1). All RAD-seq and WGS datasets were mapped to the three-spined stickleback reference genome (release-92), retrieved from Ensembl[58] using BWA mem v.0.7.17 (ref. [59]). PCR duplicates were removed using the program Stacks v.2.5 (ref. [60]) for pair-end RAD data and SAMtools v.1.9 (function markdup; ref. [61]) for whole-genome data. Given the heterogeneity in sequencing depth among different datasets, and particularly the very low coverage of the data retrieved from Jones et al.[11], most of the analyses were performed directly using genotype likelihoods, avoiding variant calling whenever possible. Genotype likelihoods were estimated from the mapped reads using the model of SAMtools[61] as implemented in the program suite ANGSD v.0.929 (ref. [62]). Full scripts for the genotype likelihood estimation and filtering parameters are publically available through DRYAD. Bases with a *q*-score below 20 (-minQ 20) and reads with mapping quality below 25 (-minMapQ 25) were removed and variants were only retained if they had a *P* value smaller than 1×10^{-6} (-SNP_pval 1×10^{-6} flag in ANGSD). We retained sites with a minimum read depth of two (-minIndDepth 2) in at least 80% of the sampled individuals (-minInd 133). The sex chromosome (chromosome XIX; refs. [63,64]) was excluded from downstream analyses due to sex-specific genomic heterogeneity[65,66]. The raw output of genotype likelihoods from all 166 individuals comprised 2,511,922 genome-wide loci.

Unsupervised approach to determine marine–freshwater differentiation. LDna uses a pairwise matrix of LD values, estimated by r^2, to produce a single linkage clustering tree. The hierarchical clustering algorithm uses the LD matrix to combine two clusters connected to each other by at least one edge. In the resulting tree, the nodes represent clusters of loci connected by LD values above thresholds, where the threshold value is proportional to the distance from the root[22]. As the LD

threshold is sequentially lowered, an increasing number of loci will be connected to each other in a fashion that reflects their similarity in phylogenetic signals. For each cluster merger (with decreasing LD threshold), the change in median LD between all pairwise loci in a cluster before and after the merger is estimated as λ. When two highly interconnected clusters merge, λ will be large (unlike when only a single locus is added to an existing cluster), signifying that these two clusters bear distinct phylogenetic signals. LDna is currently limited to ~20,000 SNPs at a time due to its dependence on LD estimates for all pairwise comparisons between loci in the dataset. To analyse the whole dataset, we applied a novel three-step LDna approach to reduce the complexity of the data in a nested fashion. First, we started with non-overlapping windows within each chromosome[67], then performed the analysis on each chromosome individually[22], and finally on the whole dataset (Supplementary Information 1). In all steps of LDna, we estimated LD between loci from genotype likelihoods using the program ngsLD[68], setting the minimum SNP minor allele frequency at 0.05. Full scripts for the LDna analyses are provided in DRYAD. In the first step, we only kept loci that were in high LD with at least one locus ($r^2 > 0.8$) within a window of 100 SNPs, as most SNPs in the data were not correlated with any other adjacent loci (so-called singleton clusters[67]) and thus, are unlikely to be informative in the LDna analyses (Supplementary Information 1).

The main evolutionary phenomena that cause elevated LD between large sets of loci in population genomic datasets are polymorphic inversions, population structure and local adaptation, all of which are expected to be present in our dataset[22]. There are specific and distinct predictions about the population genetic signal and the distribution of loci in the genome that arise from these evolutionary phenomena[22]. First, clusters with LD signals caused by polymorphic inversions are expected to predominantly map to the specific genomic region where the inversions are situated. In addition, PCAs of these loci are expected to separate individuals on the basis of karyotypes. In general, the heterokaryotype is expected to be intermediate to the two alternative homokaryotypes (provided that all karyotypes exist in the dataset), and the heterokaryotypic individuals are expected to show higher observed heterozygosity than do the homokaryotypes. However, this is not always so clear, for instance when the inversion is new and mutational differences have not yet accumulated. Second, a PCA based on loci whose frequencies are shaped by genetic drift is expected to separate individuals on the basis of geographic location, with no (or very little) separation between marine and freshwater ecotypes. Third, an LD signal caused by local adaptation (globally) is expected to cluster individuals on the basis of ecotype, regardless of geographic location, with both the locus distribution and LD patterns to some extent being negatively correlated with local recombination rate[34,69]. The reason for this correlation is that gene flow between ecotypes erodes genetic differentiation in sites linked to locally adapted loci with the exception of regions where recombination is restricted (for instance in inversions, or close to centromeres or telomeres). No such pattern is expected for LD caused by population structuring, as the main source of this LD is the random genetic drift that, in the absence of gene flow, generates LD in a fashion that is independent of genome position[22] and background selection is also not expected to result in strong patterns of within-species genetic differentiation, particularly when there is at least some level of gene flow[70,71].

If a set of loci contributes to local adaptation exclusively in a particular geographic area, a PCA based on these loci will only separate individuals based on ecotype in that region. We considered loci to be involved in parallel evolution only if they grouped individuals of the same ecotype from more than one independent location. Otherwise, it is not possible to discern drift from local adaptation, particularly if the effective population size is small (that is, genetic drift is strong).

To determine if an LD-cluster was associated with parallel freshwater differentiation, we first used expectation maximization and hierarchical clustering methods to identify clusters of individuals in PCAs that contained a minimum of seven individuals, of which at least 85% are freshwater ecotypes (the 'in-group'; dotted line; Fig. 1a–h and Extended Data Fig. 1). With less than seven in-group individuals, there was no power to detect significant associations, even if all individuals were freshwater ecotypes. Second, if such in-groups were detected, we used permutations to further test whether this cluster contained more freshwater individuals than expected by chance (Supplementary Information 4). We benchmarked LDna by quantifying the proportion of regions previously identified by Jones et al.[11] as involved in marine–freshwater ecotype differentiation (globally and within the Eastern Pacific) that were correctly recovered by LD-cluster loci (Supplementary Information 2). Note that freshwater individuals from locations without marine individuals were also important for the analyses, as they can inform us about the geographic scale of parallel marine–freshwater differentiation (for the LD-clusters where marine–freshwater differentiation also involved geographic regions where marine samples were available).

Supervised approaches to determine marine–freshwater differentiation. Genome-wide allelic differentiation (F_{ST} estimated from genotype likelihoods in ANGSD) between marine and freshwater ecotypes was estimated separately for the three major oceans in our study: Eastern Pacific, Western Pacific and Atlantic Oceans. All available samples were always used but, due to the small number of available Pacific marine individuals, marine individuals from the Eastern ($n = 4$) and Western Pacific ($n = 13$) were pooled and treated as a combined Pacific marine group in Eastern and Western Pacific ecotype comparisons (Extended Data Fig. 8) to reduce the noise of non-window-based (single SNP) analyses. To determine whether pooling of Eastern and Western Pacific marine individuals could bias F_{ST} estimates, we first estimated F_{ST} in 100 kilobase (kb) windows for Eastern Pacific marine versus Eastern Pacific freshwater and Western Pacific marine versus Eastern Pacific freshwater individuals, as using large windows allowed us to obtain precise estimates even when the marine group comprised of only four marine individuals from the Eastern Pacific. The two sets of window-based pairwise F_{ST} estimates were highly correlated ($r = 0.904$, $P < 0.001$; Extended Data Fig. 3d–g), suggesting that pooling marine individuals from Eastern and Western Pacific should not strongly affect SNP-based estimates. Note that from the results of the unsupervised LDna, two Eastern Pacific freshwater individuals from Kodiak Island, Alaska (ALA population) never grouped with the other Eastern Pacific freshwater individuals. Therefore, in agreement with earlier phylogenetic analyses[28], these two individuals were excluded from the supervised analyses. The squared correlation coefficient of F_{ST} before and after this exclusion was 0.88, indicating that this did not affect the results.

In each comparison, the sites were firstly filtered from raw mapped reads, retaining sites with <25% missing data with quality control (-minIndDepth 1, -uniqueOnly 1, -remove_bads 1, -minMapQ 20, -minQ 20). We retained only variable sites (-SNP_pval 1×10^{-6}) in each region, resulting in 1,218,858 SNPs in the Eastern Pacific, 1,072,257 SNPs in the Western Pacific and 1,681,923 SNPs in the Atlantic. We then obtained genotype likelihoods and site allele frequency likelihoods of the variants (-GL 1, -doSaf 1). On the basis of these likelihoods, we estimated the two-dimensional site-frequency spectrum (SFS) for each pair of ecotypes (realSFS) and calculated the pairwise weighted F_{ST} (realSFS fst).

Proof of concept using simulated data. The simulations were performed with quantiNemo[72] and aimed to recreate the transporter hypothesis model in the Eastern Pacific (referred to as Pac in the context of simulations), to simulate the colonization of the Atlantic (referred to as Atl in the context of simulations) from Pac 60–30 ka during the last known opening of the Bering Strait[27,73,74] and the subsequent post-glacial (10 ka) colonization of newly formed freshwater habitats in both oceans (simulation details can be found in Supplementary Information 4). In short, simulations begin with one marine population in Pac connected to five independent freshwater populations by symmetrical gene flow (that is, no gene flow exists between any of the freshwater populations; Fig. 3a) for 10^4 generations (40–50 ka). This is followed by colonization of Atl from Pac (with Atl having identical population structure to Pac) by allowing one or five migrants per generation between the oceans for 2,000 years (38–40 ka), after which no further gene flow is possible. The retreat of the Pleistocene continental ice sheets (at 10 ka; Fig. 3d) and the colonization of newly formed freshwater habitats is simulated by removing four of the freshwater populations, immediately followed by the emergence of four new (post-glacial) freshwater populations (in both Pac and Atl; Fig. 3e). The fifth freshwater population remains as a 'glacial refugia' that continues to feed freshwater-adapted alleles to the sea as standing genetic variation. Post-glacial local adaptation is thus only possible due to the spread of freshwater-adapted alleles from the sea in accordance with the transporter hypothesis[1] (Fig. 3a–e).

Marine–freshwater differentiation was based on bi-allelic QTL with allelic effects of either zero or ten, with the selection optima in the marine habitat being zero and the selection optima in all freshwater populations being 20. Thus, a freshwater individual homozygous for allele 2 for a given QTL meant that the individual was at its optimal phenotype and vice versa for marine individuals. Selection intensities were such that a sufficient amount of standing genetic variation was allowed in the sea and rapid local adaptation in freshwater was possible (see Supplementary Information 4 for details). In simulations, all allele frequencies started from 0.5 in all populations (including the QTL in the freshwater habitats). The simulated genome was comprised of ten equally sized chromosomes, with a total genome size of 1,000 basepairs. Regions of both low (centromeric regions) and high (chromosome arms) recombination were represented (Supplementary Information 4). Three, six or nine QTL per chromosome were randomly placed in eight of the chromosomes, after which the positions were fixed, leaving the last two chromosomes without any QTL. Twenty replicate simulations were run for each of the six different parameter settings (two levels of trans-oceanic gene flow rates and three different QTL densities). The frequency of freshwater-adapted alleles was recorded at 50-generation intervals throughout the simulations. Population genomic data were saved at the end of the simulations (representing present-day sampling).

Linking empirical data to simulated data. LDna identified one major cluster (LD-cluster 2; see Results) that separated all Eastern Pacific freshwater individuals from the remaining individuals (Atlantic, Western Pacific and marine samples from the Eastern Pacific, pooled). From the simulated data, we first subsampled individuals from Pac and Atl to match the samples size of the empirical data (excluding the Western Pacific samples, as this ocean was not included in the simulations) and used LDna to detect clusters similar to LD-cluster 2 using CSS (custom R-scripts available from DRYAD). CSS were calculated as the Euclidean centroid distance in a PCA (based on coordinates from the two first PC scaled by their eigenvalues) between two groups of individuals, standardized by the longest

distance between any two individuals in the PCA (CSS range [0,1]). PCA of the simulated datasets were performed by the function snpgdsPCA from the R-package SNPRelate[75]. CSS scores are known to correlate with F_{ST} but give higher resolution when genetic differentiation is high and are less sensitive to small sample sizes[11]. In LDna, we are only interested in clusters with high λ values (see earlier and Supplementary Information 1). Therefore, from the ten LD-clusters with the highest λ values (from each simulated dataset), we considered the cluster with the highest CSS between PF (Eastern Pacific freshwater) and non-PF individuals to be the strongest candidate for showing high ecotype differentiation specifically in Pac, and thus, the most similar to LD-cluster 2 in the empirical data. The non-PF individuals were comprised of PM (Eastern Pacific marine), AF (Atlantic freshwater) and AM (Atlantic marine) individuals pooled. To further assess how similar the patterns of population differentiation (in PCAs) were in the above LD-clusters (simulated data) and empirically obtained LD-cluster 2, we compared the CSS for all pairwise comparisons between PF and the other three groups of individuals (PF versus PM, PF versus AF and PF versus AM) in the simulated and empirical datasets. To further assess the extent to which clusters similar to LD-cluster 2 could be produced in the Atl, we used the same procedure as above to look for the LD-clusters with the highest CSS scores between AF and non-AF individuals (AM, PF and PM) and calculated CSS scores between AF individuals and the three non-AF groups.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The RAD-seq data have been uploaded to the GenBank under accession numbers SAMN14078677 to SAMN14078738 (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA605695). Previously published sequencing data are retrieved from studies specified in Supplementary Table 1.

Code availability

The scripts used for analysing empirical data (genotype likelihood estimation, filtering, LDna) and simulated data are available in DRYAD repository: https://doi.org/10.5061/dryad.b2rbnzsb1.

Received: 1 April 2019; Accepted: 14 May 2020;
Published online: 22 June 2020

References

1. Schluter, D. & Conte, G. L. Genetics and ecological speciation. *Proc. Natl Acad. Sci. USA* **106**, 9955–9962 (2009).
2. Arendt, J. & Reznick, D. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends Ecol. Evol.* **23**, 26–32 (2008).
3. DeFaveri, J., Shikano, T., Shimada, Y., Goto, A. & Merila, J. Global analysis of genes involved in freshwater adaptation in threespine sticklebacks (*Gasterosteus aculeatus*). *Evolution* **65**, 1800–1807 (2011).
4. Stern, D. L. The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).
5. Bell, M. A. & Foster, S. A. *The Evolutionary Biology of the Threespine Stickleback* (Oxford Univ. Press, 1994).
6. Gibson, G. The synthesis and evolution of a supermodel. *Science* **307**, 1890–1891 (2005).
7. Hendry, A. P., Peichel, C. L., Matthews, B., Boughman, J. W. & Nosil, P. Stickleback research: the now and the next. *Evol. Ecol. Res.* **15**, 111–141 (2013).
8. Lescak, E. A. et al. Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proc. Natl Acad. Sci. USA* **112**, E7204–E7212 (2015).
9. Östlund-Nilsson, S., Mayer, I. & Huntingford, F. A. *Biology of the Three-spined Stickleback* (CRC Press, 2006).
10. McKinnon, J. S. & Rundle, H. D. Speciation in nature: the threespine stickleback model systems. *Trends Ecol. Evol.* **17**, 480–488 (2002).
11. Jones, F. C. et al. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
12. Ferchaud, A. L. & Hansen, M. M. The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: three-spine sticklebacks in divergent environments. *Mol. Ecol.* **25**, 238–259 (2016).
13. Hohenlohe, P. A. et al. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* **6**, e1000862 (2010).
14. Hohenlohe, P. A. & Magalhaes, I. S. in *Population Genomics* (eds Oleksiak, M. F. & Rajora, O.P.) 249–276 (Springer, 2020).
15. Liu, S., Ferchaud, A. L., Gronkjaer, P., Nygaard, R. & Hansen, M. M. Genomic parallelism and lack thereof in contrasting systems of three-spined sticklebacks. *Mol. Ecol.* **27**, 4725–4743 (2018).
16. Pujolar, J. M., Ferchaud, A. L., Bekkevold, D. & Hansen, M. M. Non-parallel divergence across freshwater and marine three-spined stickleback *Gasterosteus aculeatus* populations. *J. Fish Biol.* **91**, 175–194 (2017).
17. Terekhanova, N. V., Barmintseva, A. E., Kondrashov, A. S., Bazykin, G. A. & Mugue, N. S. Architecture of parallel adaptation in ten lacustrine threespine stickleback populations from the White Sea area. *Genome Biol. Evol.* **11**, 2605–2618 (2019).
18. Terekhanova, N. V. et al. Fast evolution from precast bricks: genomics of young freshwater populations of threespine stickleback *Gasterosteus aculeatus*. *PLoS Genet.* **10**, e1004696 (2014).
19. Chan, Y. F. et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**, 302–305 (2010).
20. Colosimo, P. F. et al. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* **307**, 1928–1933 (2005).
21. Nelson, T. C. & Cresko, W. A. Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. *Evol. Lett.* **2**, 9–21 (2018).
22. Kemppainen, P. et al. Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Mol. Ecol. Resour.* **15**, 1031–1045 (2015).
23. Betancur, R. R., Orti, G. & Pyron, R. A. Fossil-based comparative analyses reveal ancient marine ancestry erased by extinction in ray-finned fishes. *Ecol. Lett.* **18**, 441–450 (2015).
24. Matschiner, M., Hanel, R. & Salzburger, W. On the origin and trigger of the notothenioid adaptive radiation. *PLoS ONE* **6**, e18911 (2011).
25. Meynard, C. N., Mouillot, D., Mouquet, N. & Douzery, E. J. A phylogenetic perspective on the evolution of Mediterranean teleost fishes. *PLoS ONE* **7**, e36443 (2012).
26. Sanciangco, M. D., Carpenter, K. E. & Betancur, R. R. Phylogenetic placement of enigmatic percomorph families (Teleostei: Percomorphaceae). *Mol. Phylogenet. Evol.* **94**, 565–576 (2016).
27. Fang, B., Merila, J., Matschiner, M. & Momigliano, P. Estimating uncertainty in divergence times among three-spined stickleback clades using the multispecies coalescent. *Mol. Phylogenet. Evol.* **142**, 106646 (2020).
28. Fang, B., Merila, J., Ribeiro, F., Alexandre, C. M. & Momigliano, P. Worldwide phylogeny of three-spined sticklebacks. *Mol. Phylogenet. Evol.* **127**, 613–625 (2018).
29. Orti, G., Bell, M. A., Reimchen, T. E. & Meyer, A. Global survey of mitochondrial DNA sequences in the threespine stickleback: evidence for recent migrations. *Evolution* **48**, 608–622 (1994).
30. Halliburton, R. & Halliburton, R. *Introduction to Population Genetics* (Pearson/Prentice Hall, 2004).
31. Hyten, D. L. et al. Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl Acad. Sci. USA* **103**, 16666–16671 (2006).
32. Johannesson, K. et al. Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. *Philos. Trans. R. Soc. Lond. B* **365**, 1735–1747 (2010).
33. Kemppainen, P., Lindskog, T., Butlin, R. & Johannesson, K. Intron sequences of arginine kinase in an intertidal snail suggest an ecotype-specific selective sweep and a gene duplication. *Heredity* **106**, 808–816 (2011).
34. Roesti, M., Gavrilets, S., Hendry, A. P., Salzburger, W. & Berner, D. The genomic signature of parallel adaptation from shared genetic variation. *Mol. Ecol.* **23**, 3944–3956 (2014).
35. Varadharajan, S. et al. A high-quality assembly of the nine-spined stickleback (*Pungitius pungitius*) genome. *Genome Biol. Evol.* **11**, 3291–3308 (2019).
36. Feder, J. L. & Nosil, P. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution* **64**, 1729–1747 (2010).
37. Ramachandran, S. et al. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl Acad. Sci. USA* **102**, 15942–15947 (2005).
38. Bierne, N., Gagnaire, P. A. & David, P. The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. *Curr. Zool.* **59**, 72–86 (2013).
39. Baker, V. R. & Bunker, R. C. Cataclysmic Late Pleistocene flooding from glacial Lake Missoula—a review. *Quat. Sci. Rev.* **4**, 1–41 (1985).
40. Bretz, J. H. The Lake Missoula floods and the channeled scabland. *J. Geol.* **77**, 505–543 (1969).
41. Oviatt, C. G. Chronology of Lake Bonneville, 30,000 to 10,000 yr BP. *Quat. Sci. Rev.* **110**, 166–171 (2015).
42. Upham, W. *The Glacial Lake Agassiz* Vol. 25 (US Government Printing Office, 1896).
43. Hohenlohe, P. A., Bassham, S., Currey, M. & Cresko, W. A. Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philos. Trans. R. Soc. Lond. B* **367**, 395–408 (2012).
44. Bolnick, D. I., Barrett, R. D. H., Oke, K. B., Rennison, D. J. & Stuart, Y. E. (Non)parallel evolution. *Annu. Rev. Ecol. Evol. Syst.* **49**, 303–330 (2018).
45. Roda, F., Walter, G. M., Nipper, R. & Ortiz-Barrientos, D. Genomic clustering of adaptive loci during parallel evolution of an Australian wildflower. *Mol. Ecol.* **26**, 3687–3699 (2017).
46. Barghi, N. et al. Genetic redundancy fuels polygenic adaptation in *Drosophila*. *PLoS Biol.* **17**, e3000128 (2019).

47. Kautt, A. F., Elmer, K. R. & Meyer, A. Genomic signatures of divergent selection and speciation patterns in a 'natural experiment', the young parallel radiations of Nicaraguan crater lake cichlid fishes. *Mol. Ecol.* **21**, 4770–4786 (2012).
48. Le Moan, A., Gagnaire, P. A. & Bonhomme, F. Parallel genetic divergence among coastal–marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Mol. Ecol.* **25**, 3187–3202 (2016).
49. Westram, A. et al. Do the same genes underlie parallel phenotypic divergence in different *Littorina saxatilis* populations? *Mol. Ecol.* **23**, 4603–4616 (2014).
50. Morales, H. E. et al. Genomic architecture of parallel ecological divergence: beyond a single environmental contrast. *Sci. Adv.* **5**, eaav9963 (2019).
51. Roesti, M., Kueng, B., Moser, D. & Berner, D. The genomics of ecological vicariance in threespine stickleback fish. *Nat. Commun.* **6**, 8767 (2015).
52. Twyford, A. D. & Friedman, J. Adaptive divergence in the monkey flower *Mimulus guttatus* is maintained by a chromosomal inversion. *Evolution* **69**, 1476–1486 (2015).
53. Faria, R. et al. Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. *Mol. Ecol.* **28**, 1375–1393 (2018).
54. Westram, A. M. et al. Clines on the seashore: the genomic architecture underlying rapid divergence in the face of gene flow. *Evol. Lett.* **2**, 297–309 (2018).
55. Paccard, A. et al. Repeatability of adaptive radiation depends on spatial scale: regional versus global replicates of stickleback in lake versus stream habitats. *J. Hered.* **111**, 43–56 (2020).
56. Conte, G. L. et al. Extent of QTL reuse during repeated phenotypic divergence of sympatric threespine stickleback. *Genetics* **201**, 1189–1200 (2015).
57. Conte, G. L., Arnegard, M. E., Peichel, C. L. & Schluter, D. The probability of genetic parallelism and convergence in natural populations. *Proc. Biol. Sci.* **279**, 5039–5047 (2012).
58. Hubbard, T. et al. Ensembl 2005. *Nucleic Acids Res.* **33**, D447–D453 (2005).
59. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
60. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. & Cresko, W. A. Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–3140 (2013).
61. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
62. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinform.* **15**, 356 (2014).
63. Kitano, J. et al. A role for a neo-sex chromosome in stickleback speciation. *Nature* **461**, 1079–1083 (2009).
64. Natri, H. M., Shikano, T. & Merilä, J. Progressive recombination suppression and differentiation in recently evolved neo-sex chromosomes. *Mol. Biol. Evol.* **30**, 1131–1144 (2013).
65. Hedrick, P. W. Sex: differences in mutation, recombination, selection, gene flow, and genetic drift. *Evolution* **61**, 2750–2771 (2007).
66. Schaffner, S. F. The X chromosome in population genetics. *Nat. Rev. Genet.* **5**, 43–51 (2004).
67. Li, Z., Kemppainen, P., Rastas, P. & Merila, J. Linkage disequilibrium clustering-based approach for association mapping with tightly linked genomewide data. *Mol. Ecol. Resour.* **18**, 809–824 (2018).
68. Fox, E. A., Wright, A. E., Fumagalli, M. & Vieira, F. G. ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics* **35**, 3855–3856 (2019).
69. Roesti, M., Moser, D. & Berner, D. Recombination in the threespine stickleback genome—patterns and consequences. *Mol. Ecol.* **22**, 3014–3027 (2013).
70. Matthey-Doret, R. & Whitlock, M. C. Background selection and FST: consequences for detecting local adaptation. *Mol. Ecol.* **28**, 3902–3914 (2019).
71. Stankowski, S. et al. Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *PLoS Biol.* **17**, e3000391 (2019).
72. Neuenschwander, S., Hospital, F., Guillaume, F. & Goudet, J. quantiNemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. *Bioinformatics* **24**, 1552–1553 (2008).
73. Hu, A. et al. Influence of Bering Strait flow and North Atlantic circulation on glacial sea-level changes. *Nat. Geosci.* **3**, 118–121 (2010).
74. Meiri, M. et al. Faunal record identifies Bering isthmus conditions as constraint to end-Pleistocene migration to the New World. *Proc. Biol. Sci.* **281**, 20132167 (2014).
75. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

Acknowledgements

We are grateful to the following people who helped in obtaining the samples used in this study: J. DeFaveri, A. Adill, W. Aguirre, T. Bakker, A. Bell, M. Bell, B. Borg, F. Franzén, A. Goto, A. Hendry, G. Herczeg, F. von Hippel, A. Hirvonen, J. Hämäläinen, M. Kaukoranta, A. Kijewska, D. Kingsley, Y. Kosaka, L. Kvarnemo, D. Lajus, T. Leinonen, A. Levsen, S. McCairns, A. Millet, J. Morozinska, C. Munk, H. Mäkinen, A. Nolte, K. Østbye, W. Pekkola, J. Pokela, M. Ravinet, K. Räsänen, D. Schluter, M. Seymor, T. Shikano, P. Sjöstrand, G. Staines, B. Stelbrink, I. Syvänperä, A. Vasemägi, M. Webster, J. Willacker, H. Winkler and L. Zaveik. Our research was supported by Academy of Finland grant nos. 250435, 263722, 265211 and 1307943 to J.M. and grant no. 316294 to P.M., the Finnish Cultural Foundation grant no. 00190489 to P.K. and the Chinese Scholarship Council grant no. 201606270188 to B.F. We thank J. DeFaveri for feedback and linguistic corrections.

Author contributions

P.K. and J.M. conceived the concept of the study, with contributions from P.M. and B.F. B.F. and P.K. carried out analyses with significant contributions from P.M. P.K. and B.F. led the writing, with significant contributions from P.M. and J.M. X.F. contributed to LiftOver analysis. B.F. visualized the data. All authors accepted the final version of this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-020-1222-6.

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-020-1222-6.

Correspondence and requests for materials should be addressed to B.F. or P.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Extended Data Fig. 1 | Visualization of all LD-clusters identified by LDna. In each panel, I) the top and II) middle plots represent the marine–freshwater differentiation (F_{ST}) of the clustered loci of the individuals in the Atlantic and Eastern Pacific, respectively. III) The bottom left plot shows population differentiation based on loci in each LD-cluster (principal component analysis; PCA). Only one chromosome is presented on the x axis when the clustered loci were located on a single chromosome. IV) The bottom right plot depicts the number of in-group samples (as positive value) and the remaining samples (as negative value). Global samples from various regions are shown in different colours; freshwater ecotypes are indicated by light-colour and marine ecotype by dark-colour. The same colour scheme was used in the PCA. The p-values were obtained from permutation tests of cluster separation (Supplementary Information 1).

Extended Data Fig. 2 | Ability of LDna to recover marine–freshwater differentiated regions from Jones et al.[11]. Jones et al.[11] identified 812 regions showing parallel marine–freshwater differentiation in the Eastern Pacific ("*i*-regions") and 81 regions showing global parallelism ("*m-f* regions"). (**a**) The proportions of *m-f* and *i*-regions that were correctly recovered by LDna (red; at least one SNP from 29 LD-clusters mapped to these regions), the proportion or regions for which we had data but LDna analyses failed to recover (cyan), and regions for which we had no genetic data (blue). (**b**) Number of high LD-SNPs (produced by the first LDna-filtering step) and raw SNPs (bottom row) in regions that were and were not recovered by LDna and (**c**) size of the regions that were and were not recovered by LDna (on log_{10} scale). (**d**) F_{ST} from raw SNPs located within regions that were and were not recovered by LDna. Overall, *m-f* regions and *i*- regions that were not recovered by LDna were generally smaller, contained fewer SNPs (that is had lower sequencing coverage) and exhibited lower F_{ST} than the regions correctly recovered by LDna.

Extended Data Fig. 3 | Genome-wide marine–freshwater differentiation (F_{ST}) in the Atlantic, Eastern Pacific and Western Pacific Oceans. (**a**–**c**) SNP-based F_{ST} of the individuals in the Atlantic (ATL), Eastern Pacific (EP) and Western Pacific (WP), respectively. Ecotype pairs follow the main analyses (Extended Data Table 2). (**d**) Window-based F_{ST} (win-size=100 kb) between EP freshwater samples (n = 13) and EP marine samples (n = 4). (**e**) Window-based F_{ST} between EP freshwater samples (n = 13) and all Pacific marine samples (n = 13). (**d**, **e**) are significantly correlated ($r = 0.904$, $p < 0.0001$). (**f**, **g**) SNP-based EP genetic parallelism (LD-clusters 2, 21, 29) for the same ecotype comparison as (**d**, **e**), respectively. Loci from LD-clusters involved in genetic parallelism are colour-coded for all panels (refer to main Fig. 2).

Extended Data Fig. 4 | PCA plot of LDna clusters with population identification. See Supplementary Table 1 for population identifiers.

Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Population diversity and Isolation-By-Distance (IBD) in marine three-spined stickleback populations. (**a**) Boxplots of individual heterozygosity (proportion heterozygous positions per individual) of marine individuals in different geographical regions (EP = Eastern Pacific, WP = Western Pacific and ATL = Atlantic; GLM, $F_{2,64} = 43.05$, $P < 0.001$). (**b**) Boxplots of individual heterozygosity of LD-cluster 2 in different geographical regions (GLM, $F_{2,64} = 91.9$, $P < 0.001$). (**c**) IBD between marine populations. Note that the different scales of empirical and simulated heterozygosity in (**a**, **b**) are not relevant. This is because in the simulations of all allele frequencies started from 0.5 and while a burn in of 10k generations was appropriate for loci linked to QTL, neutral loci would have required four times more generations to reach equilibrium (see Supplementary Information 3). However, the trends in terms of loss of heterozygosity away from the ancestral Eastern Pacific marine populations is still informative and consistent with the empirical data.

Extended Data Fig. 6 | Mercator projection of global three-spined stickleback populations used in the study. 166 three-spined stickleback individuals from 63 localities were used, including 119 freshwater individuals and 47 marine individuals. For a complete list of samples, see Supplementary Table 1.

Cluster ID	N_{loci}	LD_{mean}	LD_{median}	$P_{randomisation\text{-}test}$	Located Chromosome	$N_{overlap_ecotype_divergent_regions}$ (regions/SNPs) *	$N_{overlap_tree_i}$ (regions/SNPs) **	Inferred cause
Cluster 1	10184	0.536	0.537	NA	All chromosomes	1/2	50/153	Geographic structure
Cluster 2	53790	0.385	0.391	0.001	All chromosomes	12/28	539/8482	Local adaptation (genetic parallelism of **Eastern Pacific** freshwater populations)
Cluster 3	595	0.395	0.398	0.252	Chr. I, IV, VI, VIII, IX, XI, XII, XIV, XVI	0/0	0/0	Inconclusive
Cluster 4	367	0.358	0.356	0.244	Chr. I, II, III, VII, XVI, XX, XXI	0/0	0/0	Inconclusive
Cluster 5	47	0.463	0.494	0.027	Chr. I	0/0	0/0	Local adaptation (genetic parallelism of **global** freshwater populations)
Cluster 6	992	0.654	0.701	0.001	Chr. I (chromosomal inversion)	32/486	4/110	Local adaptation (genetic parallelism of **global** freshwater populations)
Cluster 7	32	0.575	0.597	0.094	Chr. II	0/0	0/0	Inconclusive
Cluster 8	34	0.331	0.354	NA	Chr. II	0/0	0/0	Inconclusive
Cluster 9	39	0.301	0.290	0.803	Chr. III	0/0	0/0	Inconclusive
Cluster 10	70	0.467	0.497	0.002	Chr. IV	0/0	0/0	Local adaptation (genetic parallelism of **global** freshwater populations)
Cluster 11	331	0.416	0.452	0.001	Chr. IV	11/166	1/13	Local adaptation (genetic parallelism of **global** freshwater populations)
Cluster 12	60	0.535	0.537	0.001	Chr. IV	3/38	0/0	Local adaptation (genetic parallelism of **global** freshwater populations)
Cluster 13	115	0.702	0.715	0.027	Chr. IV	0/0	0/0	Local adaptation (genetic parallelism of **global** freshwater populations)
Cluster 14	39	0.480	0.485	1.000	Chr. IV	0/0	2/4	Inconclusive
Cluster 15	54	0.498	0.527	0.393	Chr. VIII	0/0	0/0	Inconclusive
Cluster 16	34	0.626	0.660	0.017	Chr. IX	0/0	0/0	Local adaptation (genetic parallelism of **global** freshwater populations)
Cluster 17	22	0.696	0.719	0.777	Chr. V	0/0	0/0	Inconclusive
Cluster 18	30	0.597	0.599	0.016	Chr. V	0/0	0/0	Local adaptation (genetic parallelism of **global** freshwater populations)
Cluster 19	241	0.467	0.474	NA	Chr. V (putative chromosomal inversion)	0/0	0/0	Putative chromosomal inversion
Cluster 20	45	0.554	0.557	0.001	Chr. V	0/0	0/0	Local adaptation (genetic parallelism of **global** freshwater populations)
Cluster 21	183	0.616	0.625	0.009	Chr. X	0/0	0/0	Local adaptation (genetic parallelism of **Eastern Pacific** freshwater populations)
Cluster 22	520	0.745	0.754	0.001	Chr. XI (chromosomal inversion)	0/0	2/4	Local adaptation (genetic parallelism of **global** freshwater populations)
Cluster 23	65	0.567	0.568	NA	Chr. XIII	0/0	0/0	Inconclusive
Cluster 24	47	0.581	0.586	0.006	Chr. XIV	0/0	0/0	Local adaptation of two Western Pacific freshwater populations from restricted region
Cluster 25	36	0.455	0.475	0.035	Chr. XVI	0/0	0/0	Local adaptation (genetic parallelism of **global** freshwater populations)
Cluster 26	71	0.587	0.601	NA	Chr. XVI	0/0	0/0	Inconclusive
Cluster 27	222	0.316	0.312	0.001	Chr. XX	0/0	1/87	Local adaptation (genetic parallelism of **global** freshwater populations)
Cluster 28	69	0.699	0.718	0.052	Chr. XX	0/0	0/0	Inconclusive
Cluster 29	2730	0.691	0.704	0.001	Chr. XXI (including a chromosomal inversion)	0/0	30/2551	Local adaptation (genetic parallelism of **global** freshwater populations but rare outside of EP populations)

* number of regions/SNPs covered in the top freshwater-marine genomic regions (81 regions) in Jones et al. 2012.

** number of regions/SNPs covered in the 812 regions (after liftover between ref. versions) showing parallel selection in Eastern Pacific freshwater samples identified by SOM/HMM in Jones et al. 2012.

Extended Data Fig. 7 | Summary of all LD-clusters. Shaded rows (LD-clusters) contribute to genetic parallelism of regional or trans-oceanic freshwater populations.

No.	Figrue ID	Region	Freshwater samples	Marine samples
1	Fig. 2b,e; Extended Data Fig. 3b	Eastern Pacific (EP)	13 EP	4 EP + 9 WP
2	Fig. 2d; Extended Data Fig. 3a	Atlantic (ATL)	92 ATL	34 ATL
3	Extended Data Fig. 3c	Western Pacific (WP)	12 WP	4 EP + 9 WP

Extended Data Fig. 8 | Sampling schemes for F_{ST} analyses. The table specifies sampling schemes used for F_{ST} analyses and figures.

nature research

Corresponding author(s): petri.kemppainen@helsinki.fi
bohao.fang@helsinki.fi

Last updated by author(s): Apr 19, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We did not use any code to collect data
Data analysis	We used published code, programs, R packages to analyze our data. The bioinformatic analyses were conducted using standard bioinformatic programs, e.g. ANGSD program. And the LDna analysis was conducted by R package LDna v.65.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The authors declare that all the data supporting the findings of this study are available within the paper and its supplementary information files. And accession codes is available in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The primary focus was to identify patterns of parallel allele frequency changes between two ecotypes (marine and freshwater) from global samples of three-spined sticklebacks (Gasterosteus aculeatus) using both unsupervised and supervised methods. The study uses population genomic analyses of samples collected from the wild, either by us or publicly available from previous studies. The empirical data set comprised 166 samples from covering almost the whole distribution range of the focal species. Simulations mimicking biology of the focal species were used to deepen our understanding of the differences between our results and previous studies.
Research sample	Samples of three-spined stickleback (Gasterosteus aculeatus) collected from (East and West Coasts of) both the Pacific and Atlantic Oceans were included in this study, encompassing the main distribution range of this species. Sampling was done opportunistically to include as many individuals as possible. Mostly previously available data were used, but we also complemented these data with samples collected by our lab for previous (but unpublished) projects.
Sampling strategy	We aimed to collect sequence data from as many individuals/populations as possible from the main distribution range of the focal species.
Data collection	The specimens used for DNA extraction were collected primarily for the purposes of earlier studies. The following persons helped in obtaining the samples used in this study: Anders Adill, Windsor Aguirre, Theo Bakker, Alison Bell, Mike Bell, Bertil Borg, Jacquelin DeFaveri, Fredrik Franzén, Akira Goto, Andrew Hendry, Gabor Herczeg, Frank von Hippel, Aki Hirvonen, Jenni Hämäläinen, Markku Kaukoranta, Agnieszka Kijewska, David Kingsley, Yoshinobu Kosaka, Lotta Kvarnemo, Dmitry Lajus, Tuomas Leinonen, Arne Levsen, Scott McCairns, Antoine Millet, Jola Morozinska, Corey Munk, Hannu Mäkinen, Arne Nolte, Kjartan Østbye, Wäinö Pekkola, Jouko Pokela, Mark Ravinet, Katja Räsänen, Dolph Schluter, Mat Seymor, Takahito Shikano, Per Sjöstrand, Garrett Staines, Björn Stelbrink, Ilkka Syvänperä, Anti Vasemägi, Mike Webster, James Willacker, Helmut Winkler, and Linda Zaveik.
Timing and spatial scale	The data for this study was collected in between 2003-2015. The rationale behind sampling design was to have geographically as comprehensive sampling as possible, whereas the timing of sampling was opportunistic and dictated by availability of samples. Nevertheless, most of the used samples were collected during the species breeding season.
Data exclusions	We did not excluded any of the collected data from our analyses. Regarding the inclusion of whole genome sequences from previous studies, (Jones et al., 2012; Feulner et al., 2013; Yoshida et al., 2014), we included all the sequence data available in the public database.
Reproducibility	Our study did not include any laboratory or field experiments. Results of the analyses based on empirical data and simulations were reported. Scripts for the analyses is provided via DRYAD, and can thus also be replicated by others.
Randomization	Prior grouping was based in sampling location (marine/freshwater) and geographic location. However emphasis of this study was on unsupervised methods where prior grouping is not relevant. No randomization was necessary as we aimed to sample as many individuals as possible.
Blinding	Blinding was not considered to relevant for this study: we used mainly unsupervised methods in which no a priori grouping of samples is done. Even when using supervised methods, informed interpretation of population genetic analyses requires information about geographic origin of the samples, and blinding would make inference quite problematic.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Not applicable
Wild animals	The study utilized mature breeding threespined sticklebacks (Gasterosteus aculeatus) collected from the wild. A random sample of fish caught with minnow traps or hand seines were collected.
Field-collected samples	The caught fish were immediately euthainsed with MS-222
Ethics oversight	Collection of samples for this study was covered with appropriate licences from each of the country and locality concerned. In many cases, the the samples provided to us by colleges utilizing their own licenses. or the data was retrieved from published sources. The Finnish samples collected by us were fished under national (and personal) fishing licences.

Note that full information on the approval of the study protocol must also be provided in the manuscript.