Oceans apart: Heterogeneous patterns of parallel evolution in sticklebacks

Bohao Fang*, Petri Kemppainen*, Paolo Momigliano and Juha Merilä

Ecological Genetics Research Unit, Organismal and Evolutionary Biology Research Program, Faculty of Biological and Environmental Sciences, University of Helsinki, FI-00014 Helsinki, Finland

*These authors contributed equally to this work

Correspondence to: PK and BF

E-mail: petri.kemppainen@helsinki.fi; bohao.fang@helsinki.fi

Abstract

Reuse of standing genetic variation is thought to be the main mechanism behind the repeated evolution of the same phenotypes in similar environments. An important model system for the study of genomic mechanisms underlying parallel ecological adaptation in the wild is the three-spined stickleback (*Gasterosteus aculeatus*), which has repeatedly colonized and adapted to freshwater from the sea. Previous studies have identified numerous genomic regions showing consistent genetic divergence between freshwater and marine ecotypes, but most of these studies have been based on limited geographic sampling, and with strong biases towards studies conducted in the Eastern Pacific area. We analysed population genomic data from three-spined sticklebacks with comprehensive global sampling using unsupervised methods to detect loci involved in parallel evolution at different geographical areas. In line with previous studies, we find several genomic regions, including two chromosomal inversions, contributing to global differentiation of marine and freshwater ecotypes. However, signatures of parallel evolution were far stronger in the Eastern Pacific region than anywhere else in the world. With simulations, we demonstrate that since freshwater-adapted alleles exist in the marine populations only at low frequencies, they are

28 easily lost during founder events, thereby reducing the amount of standing genetic variation
29 available for freshwater adaptation outside of the ancestral Eastern Pacific region. Hence,
30 geographic heterogeneity in access to standing genetic variation due to historical
31 demographic factors appears to provide an explanation to marked geographic differences in
32 the pervasiveness of parallel evolution in the three-spined stickleback. Therefore, the degree
33 of genetic parallelism in the three-spined stickleback model system appears not be as
34 pervasive as earlier studies focused on Eastern Pacific populations have led us to believe.

35

36 **Keywords:** *Gasterosteus aculeatus*; genetic differentiation; linkage disequilibrium; local
37 adaptation; parallel evolution

Introduction

The extent to which the evolution of similar phenotypes arises by selection acting on shared ancestral polymorphism (i.e. parallel evolution, Conte and Schluter 2009) or via distinct molecular evolutionary pathways (i.e. convergent evolution; Arendt and Reznick 2008; DeFaveri et al., 2011; Stern 2013) 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 akin phenotypes in similar environments. After the retreat of Pleistocene glaciers, marine three-spined sticklebacks (*Gasterosteus aculeatus*) colonised and adapted to many newly formed freshwater habitats, evolving repeated changes in a number of morphological, physiological, life history and behavioural traits (Bell and Foster, 1994; Gibson, 2005; Östlund-Nilsson et al., 2006; Hendry et al., 2013; Lescak et al., 2015). 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 (McKinnon and Rundle, 2002).

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 (Hohenlohe et al., 2010, 2019; DeFaveri et al., 2011; Jones et al., 2012; Terekhanova et al., 2014, 2019; Ferchaud and Hansen, 2016; Pujolar et al., 2017; Liu et al., 2018). Studies of parallel genomic patterns of ecotype divergence have historically focused on the Eastern Pacific region (Colosimo et al., 2005; Chan et al., 2010; Hohenlohe et al., 2010, 2019; Jones et al., 2012; Nelson and Cresko, 2018), but with several recent studies emerging also from Atlantic populations (Terekhanova et al., 2014, 2019; Ferchaud and Hansen, 2016; Pujolar et al., 2017; Liu et al., 2018). However, only two studies have thus far included samples from a larger (global) geographic range (DeFaveri et al. 2011; Jones et al. 2012). Based on whole genome sequence data from a limited number of individuals (21) that included samples from Eastern Pacific and Atlantic populations, Jones et al. (2012) identified ~200 genomic regions consistently separating marine and freshwater individuals globally, representing roughly 0.5% of the dataset. They also found that 2.83% of the dataset showed signatures of parallel selection in Eastern Pacific freshwater locations, approximately six times more than at the global scale (Supplementary Fig. 3 and Supplementary Table 7 in Jones et al. [2012]),

suggesting that more loci contribute to parallel evolution at smaller geographic (regional) scales. 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. (2011) found that only ~50% of genes under divergent selection were shared across more than three population pairs and none across all. This suggested a limited re-reuse of ancestral polymorphism at the global geographic scale, implicating either an important role of convergent evolution at larger geographic scales (DeFaveri et al., 2011) or geographic heterogeneity in selective pressure among different freshwater ecosystems (DeFaveri et al. 2011; Stern 2013). Based on a subset of the data from Jones et al. (2012) and using a novel method based on linkage disequilibrium for identifying sets of highly correlated loci, Kemppainen et al. (2015) found that approximately 10 times more loci i showed consistent marine-freshwater differentiation in the Eastern Pacific than at the global scale. 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 (Terekhanova et al., 2014, 2019; Hohenlohe et al., 2010; Jones et al., 2012; Ferchaud and Hansen 2016; Pujolar et al., 2017; Nelson and Cresko, 2018; Liu et al., 2018). For instance, Terekhanova et al. (2014) recovered only 19 highly localised genomic regions involved in parallel freshwater divergence in the White Sea in contrast to Hohenlohe et al. (2010) and Nelson and Cresko (2018) who found large genomic regions involved in parallel freshwater divergence across almost all chromosomes in the Eastern Pacific. However, no study has yet addressed these conflicting results, and therefore, the potential mechanisms underlying this apparent large-scale geographic heterogeneity in genome-wide patterns of parallel evolution in three-spined sticklebacks remain unaddressed. 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 supervised and unsupervised methods to detect loci involved in parallel marine-freshwater divergence at different geographical scales. Based on earlier observations (DeFaveri et al., 2011; Terekhanova et al., 2014, 2019; Kemppainen et al., 2015; Ferchaud and Hansen 2016; Pujolar et al., 2017; Liu et al., 2018), we hypothesize that the genetic parallelism in response to colonisation of freshwater environments by marine

sticklebacks is heterogeneous at the global scale, and 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 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 (2009) proposed the “transporter hypothesis”. This hypothesis postulates that three-spined sticklebacks have repeatedly colonized and adapted to freshwater environments via selection on standing genetic variation in large marine populations, where recurrent gene flow from previously colonized freshwater habitats maintains freshwater-adapted alleles. Three-spined stickleback populations have persisted in the Eastern Pacific for approximately 26 Mya (Matschiner et al., 2011; Meynard et al., 2012; Betancu-R et al., 2015; Sanciangco et al., 2016), from which the Western Pacific and the Atlantic were colonized much more recently, during late Pleistocene (36.9-346.5 kya, Orti et al., 1994; Fang et al., 2018, 2019). During bottlenecks and founder events, rare alleles are lost at a higher rate than common alleles (Halliburton and Halliburton, 2004; Hyten et al., 2006). Since freshwater-adapted alleles exist in the marine populations only at low frequencies (Schluter and Conte, 2009), it is likely that they were lost to a higher degree than neutral variation during founder events, thereby reducing the amount of standing genetic variation available for freshwater adaptation outside of the Eastern Pacific. However, this hypothesis has not received much attention. To this end, as proof of concept, we used individual-based forward simulations designed to mimic the transporter hypothesis, and the general global ppulation demographic history of three-spined sticklebacks outlined above, to test his hypothesis. We conclude with a discussion on potential biological and demographic explanations for the high degree of geographic heterogeneity in patterns of parallel genomic divergence, and reflect upon the representativeness of the eastern Pacific three-spined stickleback populations as a general model for the study of parallel evolution.

Material and Methods

Sample collection

To study parallel evolution at the global scale, 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. 2j, Supplementary Table 1 and Supplementary Fig. 1). Of these, 62 individuals from 17 populations were sampled and sequenced specifically for this study while data from 66 individuals from 38 populations from Fang et al. (2018) and additional 38 individuals (from 24 populations) from other published studies (Jones et al., 2012; Feulner et al., 2013; Yoshida et al., 2014; Supplementary Table 1) 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).

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: Eastern Pacific, Western Pacific, Western Atlantic, White and Barents Seas, North Sea and British Isles, Baltic Sea and Norwegian Sea (Fang et al., 2018; Fig. 2j). A summary of coordinates, ecotype and population information concerning sampled individuals and re-acquired samples are given in the Supplementary Table 1.

Sequencing and genotype likelihood estimation

Restriction site associated DNA sequencing (RADseq) using the enzyme *PstI* was performed for the 62 individuals sampled in this study using the same protocol as in Fang et al. (2018), where DNA library preparation and sequencing method are described in detail. RADseq data for 66 individuals from Fang et al. (2018) were retrieved from GenBank (Accessions SRR7067148 - SRR7067275), as were the whole genome sequencing (WGS) data of 38 additional samples from published studies (Supplementary Table 1). All RADseq and WGS datasets were mapped using BWA mem v0.7.17 (Li and Durbin, 2010) to the three-spined stickleback reference genome (release-92) retrieved from Ensembl database (Hubbard et al., 2005). Given the heterogeneity in sequencing depth among different datasets, and particularly the very low coverage of the data retrieved from Jones et al. (2012), most of the

157 analyses were performed directly using genotype likelihoods avoiding variant calling
158 whenever possible. Genotype likelihoods where estimated from the mapped reads using the
159 model of SAMtools (Li, 2011) as implemented in the program suite ANGSD v0.929
160 (Korneliussen et al., 2014). Full scripts of genotyping and filtering parameters are publically
161 available through DRYAD (doi: XX). Bases with a q-score below 20 (-minQ 20) and reads with
162 mapping quality below 25 (-minMapQ 25), were removed and variants where only retained if
163 they had a p-value smaller than 1e-6 (-SNP_pval $1e^{-6}$ flag in ANGSD). We retained sites with
164 at least 80% effective sample size (-minInd 133) sequenced at a minimum read depth of 2 (-
165 minIndDepth 2). The sex chromosome (Chr. IX; Kitano et al., 2009, Natri et al., 2013) was
166 excluded from downstream analyses due to sex-specific genomic heterogeneity (Schaffner,
167 2004; Hedrick, 2007). The raw output of genotype likelihoods from 166 individuals comprised
168 2,511,922 genome-wide loci.

169

170 **Supervised approaches to determine marine-freshwater divergence**

171 To test to what extent *a priori* population grouping can bias inferences of the pervasiveness of
172 parallel divergence patterns, we estimated allelic differentiation (as estimated by the fixation
173 index, F_{ST}; Weir and Cockerham [1984], as implemented in VCFtools) between the two
174 ecotypes similarly to previously published studies (see Introduction), but using different
175 geographical subsamples from the complete dataset. Using ANGSD, we called the genotype
176 from the raw dataset, keeping only genotypes with a posterior probability over 0.95 and minor
177 allele frequency above 0.1 (-postCutoff 0.95 in ANGSD; --maf 0.1 in VCFtools). Four datasets
178 with different geographic sampling schemes were analysed (Fig. 1 a-d and Supplementary
179 Fig. 2).

180 In dataset 1 (Fig. 1a), for each SNP we estimated F_{ST} was estimated between all available
181 freshwater individuals from the Eastern Pacific (n=13 and Supplementary Table 3) and a
182 group of marine individuals (n=12) comprising four marine samples from the Eastern Pacific
183 (same samples as in Jones et al., 2012), two marine samples from the Western Pacific and
184 the remaining individuals were randomly sampled from non-Eastern Pacific locations (n=6).
185 This dataset was the one most biased towards freshwater individuals from the Eastern
186 Pacific. The inclusion of two Western Pacific marine individuals was due to the small number

of available Eastern Pacific marine individuals. Dataset 2 (Fig. 1b and Supplementary Table 3) was constructed to simulate the sampling scheme of Jones et al. (2012): it contrasted freshwater individuals (n=26) against marine individuals (n=26), with half of the individuals in each group always coming from the Eastern Pacific (with the exception of the two Western Pacific individuals as above) and the other half from elsewhere. In dataset 3 (Fig. 1c and Supplementary Table 3), a random sample of all available freshwater individuals (n=26) were contrasted against a random sample (n=26) of all available marine individuals (regardless of geographic location). The dataset 4 (Fig. 1d and Supplementary Table 3) was similar to the second dataset, except with the Pacific individuals coming from the Western Pacific instead of from the Eastern Pacific. Thus, the dataset 4 represents a form of the negative control. Random sampling was repeated 100 times under each sampling scheme, F_{ST} was calculated for each locus in each replicate (VCFtools), and the mean F_{ST} values of all replicates were used.

To summarize the pattern of differentiation (genome-wide F_{ST}) for the four datasets, we calculated moving averages of 95% quantiles and medians in windows of 100 SNPs with a step size of 50 SNPs. Since background F_{ST} was likely to be highly dependent on sample sizes and mixtures of individuals from different populations, we focused on the difference between the median and the 95% quantile of genome divergence. A large difference between the two (median and 95% quantile value) indicating a skewed distribution due to an excess of "outlier" loci, i.e. genomic islands of parallel marine-freshwater divergence.

Unsupervised approach to determine marine-freshwater divergence

We conducted Linkage Disequilibrium Network Analysis (LDna) on the whole dataset (2,511,922 SNPs) to identify and extract clusters of highly correlated loci, i.e. sets of loci affected by the same evolutionary processes. LDna starts by producing a single linkage clustering tree (a hierarchical clustering algorithm which combines two clusters connected to each other by at least one edge) based on a pairwise matrix of LD values (as estimated by r^2; Hill 1968) where nodes represent loci and edges represent LD values above given thresholds (Kemppainen et al., 2015). As the LD threshold is sequentially lowered, more and more loci will be connected to each other in a fashion that reflects how similar phylogenetic signals they

contain. For each cluster merger (with decreasing LD threshold) the change in median LD between all pairwise loci in a cluster before and after merger is estimated as λ. When two highly interconnected clusters merge, λ will be large (as compared to when e.g. 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 three-step LDna approach to reduce complexity of population genomic data in a nested fashion starting from non-overlapping windows within chromosomes (Li et al. 2018), separately for each chromosome and finally for the whole dataset (Supplementary Information 1). In all steps of LDna, we estimated LD between loci from genotype likelihoods using the program ngsLD (Fox et al., 2019), setting the minimum SNP minor allele frequency at 0.05. Full scripts the LDna analyses are provided in the DRYAD (doi: xxx). In the first step, we only kept loci that were in high LD with at least one more 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, Li et al. 2018), and thus are unlikely to be informative in our analyses.

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 (Kemppainen et al., 2015). 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 (Kemppainen et al., 2015). First, clusters with LD signals caused by inversions are expected to predominantly map to the specific genomic position where the inversion is situated. In addition, a Principal Component Analysis (PCA) of these loci is expected to separate individuals based on karyotype, with the heterokaryotype being intermediate to the two alternative homokaryotypes (provided that all karyotypes exist in the dataset) and the heterokaryotypic individuals showing higher observed heterozygosities than the homokaryotypes (although this is not always so clear, for instance when the inversion is new and enough 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, clusters with LD signal caused by local adaptation (globally) are expected to cluster individuals based on ecotype, regardless of geographic location, with both the locus distribution and LD patterns being, to some extent, negatively correlated with local recombination rate (Roesti et al., 2013, 2014). The reason for this correlation is that gene flow between ecotypes erodes genetic differentiation on 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 random genetic drift that, in the absence of gene flow, generates LD in fashion that is independent of genome position (Kemppainen et al. 2015). If a set of loci contributes to local adaptation exclusively in a particular geographic area, 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, as otherwise, it is not possible to discern drift from local adaptation, particularly if N_e is small (i.e. genetic drift is strong). To determine if an LD-cluster was likely associated with parallel freshwater divergence we first used expectation maximisation (EM) and hierarchical clustering methods to identify clusters of individuals in PCAs that contained a minimum of seven individuals of which at least 90% of the individuals are freshwater ecotypes (the “in-group”; dotted line; Fig 2a-i, Supplementary Fig. 3). Second, if such in-groups were detected, we further tested whether this cluster contained more freshwater individuals than expected by chance by permutation (Supplementary Information 2). With less than seven in-group individuals, there was no power to detect significant associations, even if all of them were freshwater individuals.

Proof of concept using simulated data

Several potential explanations for geographic heterogeneity in parallel patterns of marine-freshwater divergence in three-spines sticklebacks have been suggested (DeFaveri, et al. 2011). 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 colonised the rest of the world from the Eastern Pacific in late Pleistocene (see Introduction). Thus, as proof of concept, we here use forward-in-time

simulations to investigate the conditions for such a scenario under which parallel islands of differentiation (i.e. parallel genomic divergence between marine and freshwater ecotypes) can arise.

Our simulations were aimed at recreating the transporter hypothesis model in the Eastern Pacific, and to simulate the colonization of the Atlantic from the Pacific 30-60 Kya during the last known opening of the Bering strait (Hu et al., 2010; Meiri et al., 2014; Fang et al., 2019) and the subsequent colonization of newly available freshwater habitats. Our model starts by allowing local adaptation in the Pacific until mutation/drift/selection equilibrium (Fig. 4a) for 20 Ky after which the Atlantic is colonised from the Pacific between 36 to 40 Kya with no further trans-oceanic gene flow possible beyond 36 Kya. Both oceans are connected to five independent freshwater populations (gene flow between them is only possible via standing genetic variation in the sea) by symmetrical gene flow of N_em=1, where N_em is the number of migrants per generation. While the freshwater populations in the Pacific are populated already from the simulation start, the freshwater populations in the Atlantic can naturally only be colonised following colonisation from the Pacific. To simulate the retreat of the Pleistocene continental ice sheets, and the colonisation of newly formed freshwater habitats, we firstly removed four of the freshwater populations (at 10 Kya) in both oceans. Immediately following this, four new freshwater locations were allowed to be colonised (from the sea) while keeping one of the original freshwater populations as “glacial refugia”, that could continue to feed freshwater adapted alleles to the sea as standing genetic variation. While this process in reality most likely was gradual and not instant, as in our simulations, the end result is the same; most Pleistocene freshwater populations are no longer present today (bar potential glacial refugia), and most present day populations were colonised from the sea (following the retreated of the Pleistocene ice sheets). Thus, post glacial local adaptation is only possible due to the spread of freshwater adapted alleles from the sea in accordance with the transporter hypothesis (Schluter & Conte, 2009; Fig. 4a-e). The carrying capacity (K), that under mutation/drift equilibrium equates to N_e, was kept at 10,000 individuals in the sea and 1,000 individuals in the freshwater populations. This historical global population demographic scenario and the population sizes are in line with findings from Liu, et al. (2016) and Fang et al. (2018, 2019). Generation time was assumed to be two years. We allowed two different levels of trans-oceanic gene flow during the colonisation of the Atlantic (N_em=1 or N_em=5).

From previous studies (Roesti et al. 2013, 2014), we know that recombination rate variation plays an important role in the formation of differentiation islands around locally adaptive loci. Thus, we configured a genetic map where the recombination distance (as measured in centiMorgans, cM; Haldane [1919]) was shorter at the chromosome centres to simulate the presence of centromeres (recombination distance versus physical distance between loci can be seen in Supplementary Fig. 4). We simulated ten equally sized chromosomes of a total length of 100 cM. This relatively short total map length was chosen such that fewer numbers of neutral markers (n=1000; in the interest of computational speed) would be needed to detect differentiation islands. To demonstrate the effect of the number of differentially selected freshwater-adapted traits (each coded by a single quantitative trait locus, QTL), we simulated datasets with 24, 48 and 72 QTL equally distributed among the eight first chromosomes (3, 6 or 9 QTL per chromosome), leaving the remaining two without any QTL. The positions of these QTL within chromosomes were randomly selected (Fig. 5), and then fixed for all simulation replicates (n=20). We also ensured that there was always some recombination distance between a QTL and the nearest neutral locus (but sometimes by necessity very small, e.g. around the simulated centromeres and telomeres, i.e. no QTL position was exactly the same as any of the positions for the neutral markers). The allelic effects of the QTL were either zero (allele 1) or 10 (allele 2), 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 individual was at its optimal phenotype, and *vice versa* for marine individuals. The selection intensity was set to 100 for the freshwater habitat and 200 for the marine habitat, lower values translating to stronger selection intensity (see quantiNemo manual for details). Smaller selection intensity in the sea allowed higher frequencies of the freshwater allele (allele 2) in the sea as standing genetic variation to facilitate rapid local adaption in the newly formed freshwater populations from the sea. Per-site germ-line mutation rate was set to 1.0×10^{-8} per generation for all loci. Full simulation details can be obtained from DRYAD (doi: xx). Population genomic datasets for both neutral loci and QTL were saved every 50 generations. All allele frequencies started at 0.5 (hence the initial 10,000 generations to allow mutation-drift-selection equilibrium), thus also initially allowing all QTL to be fixed in the Pacific freshwater populations.

The combination of two trans-oceanic gene flow rates (N_em=1 or N_em=5) and three different QTL settings (ca. 3, 6 or 9 QTL per chromosome) resulted in six different parameter settings. Custom R scripts were used to retrieve and parse outputs from the simulations. We monitored the allele frequency of freshwater-adapted alleles for the QTL to assess levels of standing genetic variation at 50-generation intervals in both marine and freshwater habitats. Genome divergence was estimated for the last generation in the simulations; in each replicate simulations, F_{ST} was estimated for neutral loci between the marine and the four freshwater populations that were colonised after the Pleistocene glaciations (pooled) separately for each ocean. The F_{ST} calculations were based only on the neutral loci, thus assuming any genetic signal of selection is due to LD with a QTL. To determine how repeatable evolution was in the simulated data, we obtained the squared correlation coefficient between all pairs of simulation replicates for F_{ST} averaged over 20 bps windows with a step size of 10 bps.

Results

Supervised approaches to determine marine-freshwater divergence

Genome-wide allelic differentiation of four different datasets suggested that parallel marine-freshwater divergence in Eastern Pacific freshwater populations was much stronger and consistent across the genome than in the other geographic regions (Atlantic and Western Pacific). The strongest pattern of differentiation was found when F_{ST} was estimated between Eastern Pacific freshwater individuals and marine individuals randomly sampled from all geographic regions in the dataset (dataset 1, Fig. 1a), with 95%-quantile reaching above 0.5 on most of the 21 chromosomes but with the median F_{ST} remaining close to zero (Fig. 1a, Supplementary Fig. 2). The second dataset (dataset 2, Fig. 1b) mirroring the sampling scheme of Jones et al. (2012) also exhibited strong ecotype differentiation. The overall F_{ST} was lower than in the dataset 1 because of the inclusion of freshwater individuals also from Atlantic, but these two datasets showed similar patterns of marine-freshwater divergence (Fig. 1a-b) as previous studies from the Eastern Pacific (Hohenlohe et al., 2010; Jones et al., 2012; Nelson and Cresko, 2018). When marine and freshwater individuals were randomly sampled

369 across the data (dataset 3), much less consistent marine-freshwater divergence was found
370 (Fig. 1c, Supplementary Fig. 2) with Chr. I and XX standing out as clear outlier regions. Since
371 our random global sampling is biased towards Atlantic samples (126 individuals *vs*. 40
372 individuals from the Pacific, Supplementary Table. 1) we also analysed a negative control
373 (dataset 4), focusing on the Western Pacific freshwater ecotypes. In this dataset, no regions
374 showed consistent marine-freshwater divergence with very little variation in the 95%-quantile
375 across the genome (Fig. 1d, Supplementary Fig. 2).

376 Since F_{ST} is negatively correlated to within-group diversity, the neutral F_{ST} is expected to
377 decrease when pooling many divergent populations (datasets 2,3 and 4; Fig. 1 b-c,
378 Supplementary Fig. 2). However, if allelic differentiation is caused by selection to freshwater-
379 marine environments, F_{ST} is still expected to be high if divergent population are pooled
380 because allele fixation is not random and concordant with how populations were pooled,
381 leading to a strong heterogeneity in the difference between the median and 95% quantile F_{ST}
382 across the genome. Furthermore, the dataset 4 (biased towards sampling of freshwater
383 individuals from the Western Pacific) provided a control test to show that biased geographical
384 sampling alone does not lead to high heterogeneity in 95% quantiles across the genome
385 (Supplementary Fig. 2).

386

387 **Unsupervised approach to determine marine-freshwater divergence**

388 The first step of LDna on the empirical dataset (2,511,922 SNPs derived from 166 individuals
389 worldwide) identified 214,326 loci that were in high LD with at least one other locus within
390 windows of 100 SNPs (Supplementary Information 1). Performing LDna on each chromosome
391 separately (only using one locus from each LD-cluster from step one; Supplementary
392 information 1) resulted in 81 distinct LD-clusters and from these a final 29 LD-clusters were
393 obtained (pooling within chromosome LD-clusters whenever they were grouped by LDna in
394 the final step; Supplementary information 1), containing in total 71,064 loci (*viz*. Cluster 1-29).
395 Nine of these LD-clusters associated with local adaptation, inversions and divergence
396 between Pacific and Atlantic populations are highlighted in Fig. 2a-i. All 29 clusters are
397 detailed in Supplementary Fig. 3 and Supplementary Table 2.

Cluster 1 (10,184 loci, 0.405% of the dataset) separated all Pacific individuals (East and West) from the Atlantic individuals (Fig. 2a), thus reflecting trans-oceanic geographic structure. Clusters 2 (53,785 loci, 2.141% of the dataset, Fig. 2b) and 21 (2,728 loci, 0.007% of the dataset; correspond to an inversion in Chr. XXI, Fig. 2c) separated most of the Eastern Pacific freshwater individuals from the remaining samples, a pattern that is not expected by chance alone (randomisation test $P < 0.001$, Supplementary Fig. 3, Table 2), reflecting a shared adaptive response amid Eastern Pacific freshwater populations. Two Eastern Pacific freshwater individuals from Kodiak Island, Alaska (ALA population) did not group with the other Eastern Pacific individuals. Therefore, in agreement with earlier phylogenetic analyses (Fang et al., 2018), these two individuals were considered separate from the other Eastern Pacific freshwater individuals for the remaining analyses. Notably, no cluster of similar magnitude separating freshwater individuals from any other large-scale geographic region could be detected, demonstrating that parallel marine-freshwater divergence in the Eastern Pacific is much more prevalent than anywhere else in the world.

Clusters 5, 6, 10, 11, 12, 13, 16, 18, 20, 22, 25, 27 and 29 (in total of 5,232 loci, 0.208% of the dataset; see six representatives in Fig. 2 d-i, and all in Supplementary Fig. 3, Table 2; $P < 0.05$) grouped multiple freshwater individuals from different geographic regions across the Pacific and Atlantic Oceans, reflecting genetic marine-freshwater parallelism on a global (trans-oceanic) scale. LD-clusters 6, 22 and 29 correspond to previously known chromosomal inversions (on chromosomes I, XI and XXI, respectively) associated with marine-freshwater divergence (Jones et al. 2012Fig. 2d, g, i; Supplementary Fig. 3, Table 2). LD-clusters 6, 11, 12 and 27 showed similar marine-freshwater divergence (F_{ST}: Fig. 3a) in datasets biased towards Eastern Pacific freshwater samples (dataset 2) and datasets biased towards Atlantic individuals (dataset 3), suggesting extensive global parallelism. While clusters 5, 10, 13, 22, 25 and 29 also separate freshwater individuals from both Pacific and Atlantic populations from all other individuals, the divergent Atlantic freshwater individuals (in the in-group) were few (Supplementary Fig. 3) and thus these LD-clusters were still biased towards marine-freshwater divergence in the Eastern Pacific (Fig 3 a).

Four LD-clusters (Clusters 4, 8, 9 and 14, in total of 479 loci, 0.019% of the dataset, Supplementary Fig. 3) separated freshwater individuals from only one geographic region

likely reflecting geographic clustering but could also contain some loci involved in non-parallel freshwater adaptation. They are therefore interpreted as inconclusive with respect to their underlying evolutionary phenomena (Supplementary Table 2). Accordingly, loci from these LD-clusters showed little marine-freshwater divergence in the global dataset (Supplementary Fig. 5). Since LD-cluster 24 (47 loci) separated most of the freshwater individuals from the Western Pacific (bar one) from the remaining individuals (Supplementary Fig. 3) it could reflect marine-freshwater divergence exclusively in the Western Pacific. However, the marine and freshwater individuals in this region were sampled far from each other (all freshwater individuals came from Japan and all marine individuals came from south North Eastern Russia; Supplementary Fig. 6 and Supplementary Fig. 1), thus this marine-freshwater divergence could also have been driven by geographic structuring and it was thus also classified as inconclusive. Furthermore, the LD-cluster 19 (241 loci) was identified as a putative chromosomal inversion but it was not associated with marine-freshwater differentiation (Supplementary Fig. 3 and Supplementary Table 2).

Proof of concept using simulated data

In the simulated data, before the Atlantic was colonised from the Pacific, all five freshwater populations in the Pacific were fixed, or nearly fixed for the freshwater-adapted alleles of all locally adapted QTL (Fig. 4f). Following the invasion of the Atlantic (36-40 Kya; Fig. 4f), the increase of freshwater allele frequency in the Atlantic freshwater populations depended on both the number of QTL and level of gene flow from Pacific to Atlantic (Fig. 4f). The highest increase in freshwater-adapted alleles was observed when the number of QTL was small (3 QTL per chromosome) and trans-oceanic gene flow was high (N_em=5, g. 11,000-25,000, Fig. 4f). During post-glacial colonisation of new freshwater habitats from the sea (g. 25,002-30,000), freshwater-adapted alleles (in both Pacific and Atlantic) gradually increased in the newly formed freshwater populations (Fig. 4f), thus reflecting local adaptation. Likewise, this increase was similarly dependent on the number of QTL under selection (both in Pacific and Atlantic) and trans-oceanic gene flow (only affecting the Atlantic, Fig. 4f). These patterns likely reflect the underlying levels of ancestral variation in the sea available for subsequent freshwater adaptation (Fig. 4g). The lowest frequencies of freshwater-adapted alleles in the

458 sea was always observed in conjunction of the highest number of locally adapted QTL (in
459 both Pacific and Atlantic) and lowest trans-oceanic gene flow (only affecting the Atlantic, Fig.
460 4g). Furthermore, both the frequency of freshwater adapted alleles in the sea (ancestral
461 variation) and in the post-glacial freshwater populations (local adaptation) depended on
462 whether the QTL were located in low or high recombination regions, with lowest frequencies
463 of freshwater-adapted alleles always observed in low recombination regions (Supplementary
464 Fig. 7a,b). Frequencies of the freshwater-adapted alleles in both Pacific and Atlantic
465 freshwater populations never reached similar frequencies during the post-glacial colonisation
466 (10 Kya; Fig. 4f) as during the last glacial period (10,000-36,000 Kya; Fig. 4f) showing that
467 ancestral variation in the sea in our simulations was not sufficient to allow complete local
468 adaptation (i.e. fixation of all freshwater adapted alleles).

469 In the simulations, present-day marine-freshwater divergence (mean neutral F_{ST}) was always
470 low for the two chromosomes without QTL and in high recombination regions (Fig. 5). In
471 contrast, F_{ST} for low recombination regions was high for Pacific (for all parameter settings),
472 indicating strong islands of parallel marine-freshwater divergence. This was also true for the
473 Atlantic when the number of QTL was low (3 or 6 QTL per chromosome) and when trans-
474 oceanic gene flow was high (N_em=5). However, when number of QTL was high (9 per
475 chromosome) and trans-oceanic gene flow was low (N_em=1), only one chromosome in the
476 Atlantic (chromosome 6) showed strong marine-freshwater divergence (Fig. 5a). The mean
477 squared correlation coefficient of F_{ST} (measured in sliding windows of 20 SNPs) between
478 replicates was stronger in the Pacific (r^2=0.42-0.57) than in the Atlantic (r^2=0.09-0.42,
479 Supplementary Table 4). This translates in a higher repeatability of genome-wide parallel
480 differentiation island across the replicates in the Pacific than in the Atlantic. The weakest
481 correlation was observed when trans-oceanic gene flow was low and QTL density was high (9
482 QTL per chromosome; r^2 = 0.09).

483

484

485

Discussion

Using genome-wide SNP data from a comprehensive global sampling of marine and freshwater ecotypes, we demonstrated that only a small proportion of the dataset (13 LD-clusters comprising 0.208% of the dataset) show consistent patterns of parallel ecotype differentiation at the global scale (i.e. in at least two freshwater populations from different biogeographic regions). Two large LD-clusters (comprising 2.149% of the dataset) separated freshwater individuals from Eastern Pacific from all other individuals, while no comparable cluster could be detected for any other geographic region. This shows that parallel evolution in the three-spined stickleback is much more pervasive in the Eastern Pacific than anywhere else in the world. Indeed, marine-freshwater divergence in the Eastern Pacific is even stronger than geographic structuring between the Pacific and Atlantic Oceans (LD-clusters separating freshwater individuals from the Eastern Pacific comprised five times as many loci than the LD-cluster reflecting geographic structuring between Pacific and Atlantic Oceans). With simulations, we demonstrate that this pattern could be explained by the stochastic loss of low frequency freshwater-adapted alleles (in the sea) during range expansion from the Eastern Pacific. This loss was the strongest when trans-oceanic gene flow was low and when the number locally adapted QTL in the Eastern Pacific was high, particularly in low recombination regions where genomic islands of parallel ecotype divergence are expected to occur (Roesti et al. 2014). Supervised F_{ST} genome scans using different sampling strategies demonstrated that a sampling scheme with a large proportion of Eastern Pacific populations could overestimate the number of loci involved in parallel evolution. In the following, we discuss the biological factors potentially contributing to the differences in the degree of marine-freshwater parallelism between Eastern Pacific and non-Eastern Pacific populations of the three-spined stickleback.

Geographic heterogeneity in standing genetic variation

The “transporter hypothesis” (Schluter & Conte, 2009) postulates that three-spined sticklebacks have repeatedly colonized and adapted to freshwater environments via selection on standing genetic variation derived from large marine populations, where recurrent gene

516 flow from previously colonised freshwater habitats maintains freshwater-adapted alleles. This
517 implicitly assumes that a large pool of locally adapted alleles has accumulated during a long
518 period of time, as gene-flow is expected to spread potentially beneficial mutations across
519 demographically independent populations (Johannesson et al., 2010; Kemppainen et al.,
520 2011). In support of this hypothesis, it has been shown that haplotypes repeatedly used in
521 freshwater adaptation are identical by descent (Colosimo et al., 2005; Roesti et al., 2014) and
522 old, averaging six million years (My) with some as old as 15 My (Colosimo et al., 2005, Roesti
523 et al., 2004, Nelson and Cresko, 2018). These studies analysed populations from the Eastern
524 Pacific region, which represents the oldest and most ancestral marine population (Fang et al.,
525 2018; 2019) where three-spined sticklebacks are thought to have persisted since the split
526 from their close relative, the nine-spined stickleback (*Pungitus pungitus*), approximately 26
527 Mya (Matschiner et al., 2011; Meynard et al., 2012; Betancu-R et al., 2015; Sanciangco et al.,
528 2016; Varadharajan et al., 2019). However, populations in the Western Pacific and the
529 Atlantic are much younger, as they were colonized from the Eastern Pacific during the late
530 Pleistocene (36.9-346.5 kya; Fang et al., 2018, 2019). Furthermore, there is no evidence for
531 trans-oceanic admixture in previous phylogenetic studies (Fang et al. 2018; 2019) following
532 the split of Pacific and Atlantic clades, and there are no extant populations of three-spined
533 sticklebacks in the arctic Russia between the Kara Sea and the Eastern Siberian Sea. Thus,
534 the spread of freshwater-adapted alleles from the Eastern Pacific to elsewhere via migration
535 through the Bering Strait is unlikely, and has probably not occurred in recent times. Our
536 simulations show that following colonisation of freshwater populations from the sea, the
537 accessibility of freshwater-adapted alleles, which is a function of colonisation history, the
538 number of QTL under selection and recombination rate variation, largely determine the
539 number of loci that show consistent marine-freshwater divergence. Thus, the lower levels of
540 parallelism at the genetic level in European populations could at least in part be explained by
541 access to a smaller pool of standing genetic variation relative to Eastern Pacific populations
542 as a result of a range expansion.

543 Under the hypothesis that the colonisation of the Atlantic from the Pacific involved a limited
544 number of founder individuals, genetic diversity (expected heterozygosity, H_E, and nucleotide
545 diversity, π) is expected decrease with distance from the source population from which the
546 range expansion started (Ramachandran et al., 2005). We found no evidence to support this

hypothesis (Supplementary Fig. 8 a,b). However, since freshwater-adapted alleles occur in the sea only at low frequencies (and are selected against in the sea), they are less likely to spread to new geographic regions than neutral alleles (Halliburton and Halliburton, 2004; Hyten et al., 2006), so the loss of genetic diversity is expected to be less severe for neutral than freshwater adapted alleles..

Another possible explanation for the lack of parallel islands of ecotype divergence in the Atlantic could be stronger spatial genetic structure in marine populations outside of the Eastern Pacific causing heterogeneity in standing genetic variation available for freshwater adaptation, which was not tested in our simulations. Limited gene flow among marine populations may increase the chance that freshwater-adapted alleles are stochastically lost in some of the sub-populations, thus resulting in smaller and more heterogeneous pools of freshwater-adapted alleles. While there is indeed a significant IBD in our Atlantic marine populations (Supplementary Fig. 8e), spatial structure in the Eastern Pacific seems to be even stronger (Supplementary Fig. 8; data from Morris et al., 2018). Thus, we find no support for the hypothesis that differences in the degree of spatial genetic structuring among marine populations within regions explain the observed patterns.

Based on results of this and earlier studies, there is some evidence for parallel evolution in Europe (Ferchaud & Hansen 2016, Liu et al. 2018, Terekhanova et al., 2014, 2019), though such patterns are locally restricted and involve fewer genomic regions than in the Eastern Pacific. Ferchaud and Hansen (2016) and Liu et al. (2018), reported little parallelism in genomic regions involved in local adaptation, most notably in Denmark and Greenland, compared to the Eastern Pacific. Similarly, Terekhanova et al. (2014) identified a total of 19 distinct genomic regions showing consistent marine-freshwater divergence in the White Sea (later refined to 21, Terekhanova et al. 2019), all of which were also identified as differentiation islands between marine and freshwater ecotypes in the Eastern Pacific (Jones et al., 2012, Hohenlohe et al., 2010). Thus, while local adaptation in all geographic regions indeed likely stem from standing genetic variation originating from the Eastern Pacific (sensu the transporter hypothesis), only Eastern Pacific populations show extensive parallel evolution over a larger geographic area. These results are entirely consistent with our analyses, where

578 no LD-cluster showed parallel marine-freswater divergence exclusively among non-Eastern
579 Pacific samples.

580 Another likely explanation for the lack of globally shared genetic parallelism is heterogeneity
581 in selective regimes among freshwater habitats, both between Atlantic and Eastern Pacific
582 Oceans and between different geographic areas in the Atlantic (DeFaveri et al. 2011).
583 However, there is currently no data to assess to what extent differences in selection regimes
584 could contribute to the observed heterogeneity in parallel patterns of marine-freshwater
585 divergence at the trans-oceanic scale. A recent simulation study showed that the probability
586 of genetic parallelism from standing genetic variation rapidly declines as selection starts to
587 change from fully parallel (optima angle of 0°) to divergent (optima angle of 180°), especially
588 when a large number of traits affect fitness (Thompson et al., 2018). For example, populations
589 adapting to optima separated by an angle of just 33° might have only 50% of shared
590 beneficial alleles, even if they have access to the same pool of genetic variation. However, in
591 this model adaptation proceeded via the sorting of naive alleles and not via alleles that are
592 "pretested" by selection (as under the transporter hypothesis) when parallel evolution is
593 particularly likely (Thompson et al., 2018).

594

595 **Accessibility to ancestral variation and recombination rate variation could explain**
596 **"differentiation islands"**

597 In the simulated data, correlations of genome-wide patterns of genetic differentiation between
598 independent replicate simulations were strong for Pacific populations (Supplementary Table
599 4). This demonstrates that – given that the same set of locally adapted alleles are available
600 from standing genetic variation, the underlying recombination rate variation and the
601 population demographic history and selection regimes are the same – evolution of local
602 adaptation can be highly repeatable as also demonstrated by earlier empirical work (e.g.
603 Bassham et al. 2018). According to the divergence hitchhiking model (Feder and Nosil, 2010)
604 clustering of several QTL in low recombination regions should reduce gene flow between
605 ecotypes and thus extend LD across larger genomic regions. This is consistent with our
606 simulations where we were only able to detect genomic islands when several of these QTL
607 were clustered in the same low recombination region. Under such conditions, clear

608 differentiation islands between ecotypes were detected, in agreement with empirical data
609 (Roesti et al., 2014) showing that differentiation islands between freshwater and marine
610 sticklebacks tend to be located in genomic regions of low recombination (for example close to
611 centromeres). Our results contrast with the results of Flaxman et al. (2013) who, using
612 simulations based on simple genetic maps without recombination modifiers, concluded that
613 under most conditions, divergence hitchhiking is unlikely to significantly contribute to the
614 generation of differentiation islands.

615 Somewhat surprisingly, we also noted that, while genomic islands of parallel ecotype
616 divergence was most likely when several QTL clustered in the same low recombination
617 region, these were also the QTL where the frequency of the freshwater-adapted allele
618 showed the lowest frequencies in the sea (and thus were least likely to spread to Atlantic
619 during colonisation from Pacific). Since QTL in the low recombination region are less likely to
620 be separated by recombination when freshwater-adapted individuals migrate to the sea, it is
621 reasonable to assume that the selection pressure against these “haplotypes” in the sea is
622 stronger. However, this is not consistent with the empirical data showing that the genomic
623 regions most likely to show global parallel ecotype divergence are inversions, where
624 recombination in heterokaryotypes is particularly restricted. Our simulations assume that
625 freshwater-adapted alleles are selected against in the sea (and is equal for all QTL) while in
626 reality, selection against some of the "freshwater haplotypes/karyotypes” in the sea may be
627 weak or even absent allowing them to easily spread during range expansions. Consistent with
628 this hypothesis, in PCAs based on loci from LD-clusters corresponding to inversions (LD-
629 clusters 6, 22 and 29) several marine individuals also cluster with the freshwater individuals
630 (Fig. 2 d,h), indicating frequent occurrence of the “freshwater karyotypes” in the sea. Indeed,
631 Terekhanova *et al*. (2019) found that the genomic regions most commonly involved in local
632 adaptation in multiple independent freshwater populations were also those with the highest
633 frequencies of these haplotypes in the sea. This is in agreement with our simulation study that
634 predicts that the strongest limiting factor for the transporter hypothesis in the Atlantic is the
635 frequency of the freshwater adapted alleles in the sea.

636 A difference between LDna and self-organizing map-based iterative Hidden Markov Model
637 (SOM/HMM) used by Jones et al. (2012) to detect genomic islands of marine-freshwater

638 divergence is that SOM/HMM finds regions of the genome that support a given set of
639 commonly found tree topologies in the data, where the loci may or may not be in high LD with
640 each other, whereas LDna exclusively depends on LD and treats each locus individually.
641 Since our three-step LDna approach firstly only keeps correlated loci (above a given
642 threshold, Supplementary Information 1) within non-overlapping windows within
643 chromosomes, LDna is biased to detect LD-clusters where loci are in LD with at least one
644 physically adjacent locus within a window (of 100 SNPs). Thus, LDna is particularly sensitive
645 to detecting inversions, and may underestimate the number genomic regions involved in
646 parallel genomic divergence globally, if LD is not sufficiently high, or single loci are in LD
647 across regions spanning more than 100 SNPs in the data (on average corresponding to ~
648 15kb). Nevertheless, next to the largest LD-clusters that reflected ecotype differentiation
649 exclusively in the Eastern Pacific and the inversions, we identified most of the F_{ST} peaks in
650 the global dataset (Supplementary Fig. 3), thus demonstrating that our method (and
651 parameter settings used for the LDna) is sensitive to the main LD-signals in the data. Most
652 notably, we also found several genomic regions involved in parallel marine-freshwater
653 divergence globally that were not apparent in the F_{ST} analyses (e.g. the marine-freshwater
654 divergence only involved a small proportion of the samples in the total data set),
655 demonstrating the main strength of LDna. This cannot thus explain the complete lack of LD-
656 clusters that differentiate between ecotypes outside of the Eastern Pacific. Furthermore our
657 results are entirely consistent with the lack of large regions of high F_{ST} in the global dataset
658 (with a bias towards non-Eastern populations; Fig. 3 b), the numerous regional studies that
659 show much less parallel evolution outside of the Eastern Pacific and the fact that also in
660 Jones et al. (2012) much stronger marine-freshwater divergence was found within the Eastern
661 Pacific than at the global scale (see Introduction). Therefore, the evidence for a strong
662 discrepancy of marine-freshwater divergence between the Eastern and non-Eastern Pacific
663 populations is indeed substantial.

664 Jones et al. (2012) described ~200 genomic regions involved in marine-freshwater
665 divergence globally, and we know that this number could much higher in the Eastern Pacific
666 where ecotype divergence is the strongest. Furthermore, islands do not occur on all
667 chromosomes (presumably because these chromosomes do not contain loci for freshwater
668 adaptation), thus the number of QTL responsible for marine-freshwater divergence in the

669 Eastern Pacific could be substantial in the most QTL-dense chromosomes (e.g. Chromosome
670 IV, where global marine-freshwater divergence is the strongest). In the simulations, we
671 included up to nine QTL per chromosome, with up to six QTL within a single low
672 recombination region within a chromosome (Chr. 3 in the simulations, Fig. 5 a, b). While this
673 might seem like an exceptionally high number, this is nevertheless consistent with empirical
674 data.

675

676 **Are three-spined sticklebacks a representative model to study parallel evolution?**

677 Since the pattern of parallel genetic differentiation between marine and freshwater stickleback
678 ecotypes in the Eastern Pacific are in stark contrast to what is seen across other parts of the
679 species distribution range, one is posed to ask how representative are the results from the
680 Eastern Pacific stickleback studies with respect to parallel evolution more generally. A recent
681 review of parallel evolution suggests that even dramatic phenotypic parallelism can be
682 generated by a continuum of parallelism at the genetic level (Bolnick et al., 2018). For
683 instance, the coastal ecotypes of *Senecio lautus* exhibit only partial reuse of particular QTL
684 among replicate populations (Roda et al., 2017), and genetic redundancy frequently underlies
685 polygenic adaptation in *Drosophila* (Barghi et al., 2019). Similarly, Conte et al. (2015) studied
686 the extent of QTL reuse in parallel phenotypic divergence of limnetic and benthic three-spined
687 sticklebacks (*Gasterosteus aculeatus*) in Paxton and Priest lakes, and found that 76% of 42
688 phenotypic traits diverged in the same direction, whereas only 49% of underlying QTL
689 evolved in parallel in both lakes. For highly parallel traits in two other pairs of benthic-limnetic
690 sticklebacks, only 32% of the underlying QTL were reported to be shared (Conte et al., 2012).
691 Finally, using F_{ST} outliers to detect putative genomic targets of selection Kautt et al. (2012,
692 cichlid fishes), Le Moan et al. (2016, anchovy) and Westram et al. (2014, periwinkles),
693 showed that phenotypically very similar populations often share only a small proportion of
694 their F_{ST} outliers. Thus, these studies are also in stark contrast to the original conclusions of
695 widespread genetic parallelism in three-spined sticklebacks.

696 One exception that seems more general across taxa is the repeated involvement of
697 chromosomal inversions in parallel evolution. Chromosomal inversions could store standing
698 variation as balanced polymorphism and distribute it to fuel parallel adaptation (Morales et al.,

699 2019). For instance, the same Chr. I inversion involved in global marine-freshwater
700 divergence in three-spined sticklebacks (Jones et al., 2012, Terekhanova et al., 2014, 2019,
701 Liu et al., 2018, this study) also differentiates stream and lake ecotypes in the lake Constance
702 basin in Central Europe (Roesti et al., 2015). Two other clear examples where most genetic
703 differentiation between ecotypes at larger geographic scales is partitioned into inversions
704 come from monkey flowers (*Mimulus guttatus*; Twyford and Friedman 2015) and marine
705 periwinkles (*Littorina saxatilis*; Faria et al., 2018; Westram et al., 2018).

706

707 **Conclusions**

708 Our results demonstrate that genetic parallelism in the three-spined stickleback model system
709 is in fact not as pervasive as some earlier papers focusing on Eastern Pacific populations
710 have led us to believe. Our analysis of geographically more comprehensive data, with similar
711 and less assumption-burden methods as used in earlier studies, show that the extraordinary
712 genetic parallelism observed in the Eastern Pacific Ocean is not detectable elsewhere in the
713 world (e.g. Atlantic Ocean, Western Pacific Ocean). Hence, the focus on Eastern Pacific has
714 generated a perception bias – what goes on there, does not actually apply to the rest of the
715 world. Furthermore, our simulations show that the spread of freshwater adapted alleles can
716 easily be hampered if the colonisation of the Atlantic (from the Pacific) was limited, in
717 particular for QTL clustered in low recombination regions; i.e. the ones that are most likely to
718 result in parallel islands of ecotype differentiation. Therefore, geographic differences in
719 incidence and pervasiveness of parallel evolution in three-spined sticklebacks likely stem from
720 geographic heterogeneity in access to, and amount of, standing genetic variation, which in
721 turn has been influenced by historical population demographic factors. Hence, while striking
722 genome-wide patterns of genetic parallelism exist (e.g. in Eastern Pacific sticklebacks), the
723 conditions under which such pattern can occur may be far from common, perhaps even
724 exceptional.

725

726 **Acknowledgements**

We are grateful to the following people who 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. Our research was supported by grants from Academy of Finland (250435, 263722, 265211 and 1307943 to JM and grant 316294 to PM), the Finnish Cultural Foundation (grant 00190489 to PK) and the Chinese Scholarship Council (grant 201606270188 to BF).

References

Arendt, J., Reznick, D., 2008. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends Ecol Evol 23, 26-32.

Barghi, N., Tobler, R., Nolte, V., Jakšić, A.M., Mallard, F., Otte, K.A., Dolezal, M., Taus, T., Kofler, R., Schlötterer, C., 2019. Genetic redundancy fuels polygenic adaptation in Drosophila. PLoS Biol 17, e3000128.

Bassham, S., Catchen, J., Lescak, E., von Hippel, F.A., Cresko, W.A., 2018. Repeated Selection of Alternatively Adapted Haplotypes Creates Sweeping Genomic Remodeling in Stickleback. Genetics 209, 921-939.

Bell, M.A., Foster, S.A., 1994. The evolutionary biology of the threespine stickleback. Oxford University Press.

Betancur-R, R., Ortí, G., Pyron, R.A., 2015. Fossil-based comparative analyses reveal ancient marine ancestry erased by extinction in ray-finned fishes. Ecol Lett 18, 441-450.

Bolnick, D.I., Barrett, R.D.H., Oke, K.B., Rennison, D.J., Stuart, Y.E., 2018. (Non)Parallel Evolution. Annu. Rev. Ecol. Evol. Syst. 49, 303-330.

Chan, Y.F., Marks, M.E., Jones, F.C., Villarreal, G., Jr., Shapiro, M.D., Brady, S.D., Southwick, A.M., Absher, D.M., Grimwood, J., Schmutz, J., Myers, R.M., Petrov, D., Jonsson, B., Schluter, D., Bell, M.A., Kingsley, D.M., 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327, 302-305.

Colosimo, P.F., Hosemann, K.E., Balabhadra, S., Villarreal, G., Jr., Dickson, M., Grimwood, J., Schmutz, J., Myers, R.M., Schluter, D., Kingsley, D.M., 2005. Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science 307, 1928-1933.

Conte, G.L., Arnegard, M.E., Best, J., Chan, Y.F., Jones, F.C., Kingsley, D.M., Schluter, D., Peichel, C.L., 2015. Extent of QTL reuse during repeated phenotypic divergence of sympatric threespine stickleback. Genetics 201, 1189-1200.

765 Conte, G.L., Arnegard, M.E., Peichel, C.L., Schluter, D., 2012. The probability of genetic
766 parallelism and convergence in natural populations. Proc. Biol. Sci. 279, 5039-5047.

767 DeFaveri, J., Jonsson, P.R., Merilä, J., 2013. Heterogeneous Genomic Differentiation in
768 marine threespine sticklebacks: adaptation along an environmental gradient. Evolution 67,
769 2530-2546.

770 DeFaveri, J., Merilä, J., 2013. Variation in age and size in Fennoscandian three-spined
771 sticklebacks (Gasterosteus aculeatus). PLoS ONE 8, e80866.

772 DeFaveri, J., Shikano, T., Shimada, Y., Goto, A., Merilä, J., 2011. Global analysis of genes
773 involved in freshwater adaptation in threespine sticklebacks (*Gasterosteus aculeatus*).
774 Evolution 65, 1800-1807.

775 Defaveri, J., Shikano, T., Shimada, Y., Merila, J., 2013. High degree of genetic differentiation
776 in marine three-spined sticklebacks (*Gasterosteus aculeatus*). Mol Ecol 22, 4811-4828.

777 Fang, B., Merila, J., Ribeiro, F., Alexandre, C.M., Momigliano, P., 2018. Worldwide phylogeny
778 of three-spined sticklebacks. Mol Phylogenet Evol 127, 613-625.

779 Fang, B., Merilä, J., Matschiner, M., Momigliano, P., 2019. Estimating uncertainty in
780 divergence times among three-spined stickleback clades using the multispecies coalescent.
781 Mol Phylogenet Evol, 106646.

782 Faria, R., Chaube, P., Morales, H.E., Larsson, T., Lemmon, A.R., Lemmon, E.M., Rafajlović,
783 M., Panova, M., Ravinet, M., Johannesson, K., Westram, A.M., Butlin, R.K., 2018. Multiple
784 chromosomal rearrangements in a hybrid zone between Littorina saxatilis ecotypes. Mol Ecol
785 00, 1-19.

786 Feder, J.L., Nosil, P., 2010. The efficacy of divergence hitchhiking in generating genomic
787 islands during ecological speciation. Evolution 64, 1729-1747.

788 Ferchaud, A.L., Hansen, M.M., 2016. The imPacifict of selection, gene flow and demographic
789 history on heterogeneous genomic divergence: three-spine sticklebacks in divergent
790 environments. Mol Ecol 25, 238-259.

Feulner, P.G., Chain, F.J., Panchal, M., Eizaguirre, C., Kalbe, M., Lenz, T.L., Mundry, M., Samonte, I.E., Stoll, M., Milinski, M., Reusch, T.B., Bornberg-Bauer, E., 2013. Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. Mol Ecol 22, 635-649.

Flaxman, S.M., Feder, J.L., Nosil, P., 2013. Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. Evolution 67, 2577-2591.

Fox, E.A., Wright, A.E., Fumagalli, M., Vieira, F.G., 2019. ngsLD: evaluating linkage disequilibrium using genotype likelihoods. Bioinformatics.

Gibson, G., 2005. The synthesis and evolution of a supermodel. Science 307, 1890-1891.

Haldane, J.B.S., 1919. The combination of linkage values, and the calculation of distance between the loci of linked factors. J Genet 8, 299-309.

Halliburton, R., Halliburton, R., 2004. Introduction to population genetics. Pearson/Prentice Hall Upper Saddle River, NJ.

Hedrick, P.W., 2007. Sex: differences in mutation, recombination, selection, gene flow, and genetic drift. Evolution 61, 2750-2771.

Hendry, A.P., Peichel, C.L., Matthews, B., Boughman, J.W., Nosil, P., 2013. Stickleback research: the now and the next. Evol Ecol Res 15, 111-141.

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theor Appl Genet 38, 226-231.

Hohenlohe, P.A., Bassham, S., Currey, M., Cresko, W.A., 2012. Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. Philos Trans R Soc Lond B Biol Sci 367, 395-408.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A., Cresko, W.A., 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet 6, e1000862.

816 Hohenlohe, P.A., Magalhaes, I.S., 2019. The Population Genomics of Parallel Adaptation:
817 Lessons from Threespine Stickleback. In: . Population Genomics. Springer, Cham

818 Hu, A., Meehl, G.A., Otto-Bliesner, B.L., Waelbroeck, C., Han, W., Loutre, M.-F., Lambeck,
819 K., Mitrovica, J.X., Rosenbloom, N., 2010. Influence of Bering Strait flow and North
820 Atlanticcirculation on glacial sea-level changes. Nature Geoscience 3, 118-121.

821 Hubbard, T., Andrews, D., Cáccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L.,
822 Coates, G., Cox, T., Cunningham, F., 2005. Ensembl 2005. Nucleic Acids Res 33, D447-
823 D453.

824 Hyten, D.L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R.L., Costa, J.M., Specht, J.E.,
825 Shoemaker, R.C., Cregan, P.B., 2006. ImPacificts of genetic bottlenecks on soybean genome
826 diversity. Proceedings of the National Academy of Sciences 103, 16666-16671.

827 Johannesson, K., Panova, M., Kemppainen, P., Andre, C., Rolan-Alvarez, E., Butlin, R.K.,
828 2010. Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of
829 speciation. Philos Trans R Soc Lond B Biol Sci 365, 1735-1747.

830 Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R.,
831 Pirun, M., Zody, M.C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson,
832 M.C., Myers, R.M., Miller, C.T., Summers, B.R., Knecht, A.K., Brady, S.D., Zhang, H., Pollen,
833 A.A., Howes, T., Amemiya, C., Broad Institute Genome Sequencing, P., Whole Genome
834 Assembly, T., Baldwin, J., Bloom, T., Jaffe, D.B., Nicol, R., Wilkinson, J., Lander, E.S., Di
835 Palma, F., Lindblad-Toh, K., Kingsley, D.M., 2012. The genomic basis of adaptive evolution in
836 threespine sticklebacks. Nature 484, 55-61.

837 Kautt, A.F., Elmer, K.R., Meyer, A., 2012. Genomic signatures of divergent selection and
838 speciation patterns in a ‘natural experiment’, the young parallel radiations of Nicaraguan
839 crater lake cichlid fishes. Mol Ecol 21, 4770-4786.

840 Kemppainen, P., Knight, C.G., Sarma, D.K., Hlaing, T., Prakash, A., Maung Maung, Y.N.,
841 Somboon, P., Mahanta, J., Walton, C., 2015. Linkage disequilibrium network analysis (LDna)

gives a global view of chromosomal inversions, local adaptation and geographic structure. Mol. Ecol. Resour. 15, 1031-1045.

Kitano, J., Ross, J.A., Mori, S., Kume, M., Jones, F.C., Chan, Y.F., Absher, D.M., Grimwood, J., Schmutz, J., Myers, R.M., 2009. A role for a neo-sex chromosome in stickleback speciation. Nature 461, 1079.

Korneliussen, T.S., Albrechtsen, A., Nielsen, R., 2014. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics 15, 356.

Le Moan, A., Gagnaire, P.A., Bonhomme, F., 2016. Parallel genetic divergence among coastal–marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. Mol Ecol 25, 3187-3202.

Lescak, E.A., Bassham, S.L., Catchen, J., Gelmond, O., Sherbick, M.L., von Hippel, F.A., Cresko, W.A., 2015. Evolution of stickleback in 50 years on earthquake-uplifted islands. Proc NAtlantic Acad Sci U S A 112, E7204-7212.

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987-2993.

Li, H., Durbin, R.J.B., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589-595.

Li, Z., Kemppainen, P., Rastas, P., Merila, J., 2018. Linkage disequilibrium clustering-based approach for association mapping with tightly linked genomewide data. Mol. Ecol. Resour. 18, 809-824.

Liu, S., Ferchaud, A.L., Gronkjaer, P., Nygaard, R., Hansen, M.M., 2018. Genomic parallelism and lack thereof in contrasting systems of three-spined sticklebacks. Mol Ecol 27, 4725-4743.

Matschiner, M., Hanel, R., Salzburger, W., 2011. On the origin and trigger of the notothenioid adaptive radiation. PLoS ONE 6, e18911.

McKinnon, J.S., Rundle, H.D., 2002. Speciation in nature: the threespine stickleback model systems. Trends Ecol Evol 17, 480-488.

Meiri, M., Lister, A.M., Collins, M.J., Tuross, N., Goebel, T., Blockley, S., Zazula, G.D., van Doorn, N., Dale Guthrie, R., Boeskorov, G.G., Baryshnikov, G.F., Sher, A., Barnes, I., 2014. Faunal record identifies Bering isthmus conditions as constraint to end-Pleistocene migration to the New World. Proc. Biol. Sci. 281, 20132167.

Meynard, C.N., Mouillot, D., Mouquet, N., Douzery, E.J.P., 2012. A phylogenetic perspective on the evolution of mediterranean teleost fishes. PLoS ONE 7, e36443.

Morris, M.R.J., Bowles, E., Allen, B.E., Jamniczky, H.A., Rogers, S.M., 2018. Contemporary ancestor? Adaptive divergence from standing genetic variation in Pacific marine threespine stickleback. BMC Evol Biol 18, 113.

Natri, H.M., Shikano, T., Merilä, J., 2013. Progressive recombination suppression and differentiation in recently evolved neo-sex chromosomes. Mol Biol Evol 30, 1131-1144.

Nelson, T.C., Cresko, W.A., 2018. Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. Evol Lett 2, 9-21.

Neuenschwander, S., Hospital, F., Guillaume, F., Goudet, J., 2008. quantiNemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. Bioinformatics 24, 1552-1553.

Orti, G., Bell, M.A., Reimchen, T.E., Meyer, A., 1994. Global survey of mitochondrial DNA sequences in the threespine stickleback: evidence for recent migrations. Evolution 48, 608-622.

Pujolar, J.M., Ferchaud, A.L., Bekkevold, D., Hansen, M.M., 2017. Non-parallel divergence across freshwater and marine three-spined stickleback Gasterosteus aculeatus populations. J Fish Biol 91, 175-194.

Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W., Cavalli-Sforza, L.L., 2005. Support from the relationship of genetic and geographic distance in

893 human populations for a serial founder effect originating in Africa. Proceedings of the National
894 Academy of Sciences 102, 15942-15947.

895 Roda, F., Walter, G.M., Nipper, R., Ortiz-Barrientos, D., 2017. Genomic clustering of adaptive
896 loci during parallel evolution of an Australian wildflower. Mol Ecol 26, 3687-3699.

897 Roesti, M., Gavrilets, S., Hendry, A.P., Salzburger, W., Berner, D., 2014. The genomic
898 signature of parallel adaptation from shared genetic variation. Mol Ecol 23, 3944-3956.

899 Roesti, M., Kueng, B., Moser, D., Berner, D., 2015. The genomics of ecological vicariance in
900 threespine stickleback fish. Nature Communications 6, 8767.

901 Roesti, M., Moser, D., Berner, D., 2013. Recombination in the threespine stickleback
902 genome—patterns and consequences. Mol Ecol 22, 3014-3027.

903 Sanciangco, M.D., Carpenter, K.E., Betancur-R, R., 2016. Phylogenetic placement of
904 enigmatic percomorph families (Teleostei: Percomorphaceae). Mol Phylogenet Evol 94, 565-
905 576.

906 Schaffner, S.F., 2004. The X chromosome in population genetics. Nat Rev Genet 5, 43.

907 Schluter, D., Conte, G.L., 2009. Genetics and ecological speciation. Proc Nat Acad Sci U S A
908 106 Suppl 1, 9955-9962.

909 Stern, D.L., 2013. The genetic causes of convergent evolution. Nat Rev Genet 14, 751-764.

910 Terekhanova, N.V., Logacheva, M.D., Penin, A.A., Neretina, T.V., Barmintseva, A.E.,
911 Bazykin, G.A., Kondrashov, A.S., Mugue, N.S., 2014. Fast evolution from precast bricks:
912 genomics of young freshwater populations of threespine stickleback *Gasterosteus aculeatus*.
913 PLoS Genet 10, e1004696.

914 Terekhanova NV, Barmintseva AE, Kondrashov AS, Bazykin GA, Mugue NS. 2019.
915 Architecture of parallel adaptation in ten lacustrine threespine stickleback populations from
916 the White Sea area. Genome Biology and Evolution 11:2605-2618.

917 Thompson, K.A., Osmond, M.M., Schluter, D., 2018. Parallel genetic evolution and speciation
918 from standing variation. Evolution Letters 0-0, 1-13.

919 Twyford, A.D., Friedman, J., 2015. Adaptive divergence in the monkey flower *Mimulus*
920 *guttatus* is maintained by a chromosomal inversion. Evolution 69, 1476-1486.

921 Varadharajan, S., Rastas, P., Löytynoja, A., Matschiner, M., Calboli, F.C.F., Guo, B.,
922 Nederbragt, A.J., Jakobsen, K.S., Merilä, J., 2019. Genome sequencing of the nine-spined
923 stickleback (*Pungitius pungitius*) provides insights into chromosome evolution. BioRxiv,
924 741751.

925 Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population
926 structure. Evolution 38, 1358-1370.

927 Westram, A., Galindo, J., Alm Rosenblad, M., Grahame, J., Panova, M., Butlin, R.J.M.e.,
928 2014. Do the same genes underlie parallel phenotypic divergence in different *Littorina*
929 *saxatilis* populations? 23, 4603-4616.

930 Westram, A.M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., Ravinet, M.,
931 Blomberg, A., Mehlig, B., Johannesson, K., 2018. Clines on the seashore: The genomic
932 architecture underlying rapid divergence in the face of gene flow. Evolution Letters 2, 297-
933 309.

934 Yoshida, K., Makino, T., Yamaguchi, K., Shigenobu, S., Hasebe, M., Kawata, M., Kume, M.,
935 Mori, S., Peichel, C.L., Toyoda, A., Fujiyama, A., Kitano, J., 2014. Sex Chromosome Turnover
936 Contributes to Genomic Divergence between Incipient Stickleback Species. PLoS Genet 10,
937 e1004223.

938 Östlund-Nilsson, S., Mayer, I., Huntingford, F.A., 2006. Biology of the three-spined
939 stickleback. CRC Press.
940
941 Morales HE, Faria R, Johannesson K, Larsson T, Panova M, Westram AM, Butlin R. 2018.
942 Genomic architecture of parallel ecological divergence: beyond a single environmental
943 contrast. bioRxiv:447854.

FIGURES

Figure 1 | Marine-freshwater divergence in datasets with different geographic sampling. The four sampling schemes contrasting freshwater and marine individuals with equal sample sizes are illustrated as follows (a) Dataset 1: Eastern Pacific freshwater individuals against global sampling of marine individuals. (b) Dataset 2: half of freshwater individuals are from Eastern Pacific with the remaining individuals randomly sampled from individuals elsewhere. (c) Dataset 3: random sampling of global freshwater individuals against random global sample of marine individuals. (d) A negative control (Dataset 4) where half of the freshwater individuals are from the Western Pacific with the remaining individuals randomly sampled from individuals elsewhere. Genome-wide F_{ST} was calculated 100 times from random sampling under the three sampling schemes, and the mean F_{ST} values are presented on the right side. Red and green lines represent moving averages of 95%-quantiles and medians in windows of 100 SNPs with a step size of 50 SNPs, thus "islands of differentiation" are indicated by a large difference between the 95%-quantile and the median indicating excess of highly differentiated loci.

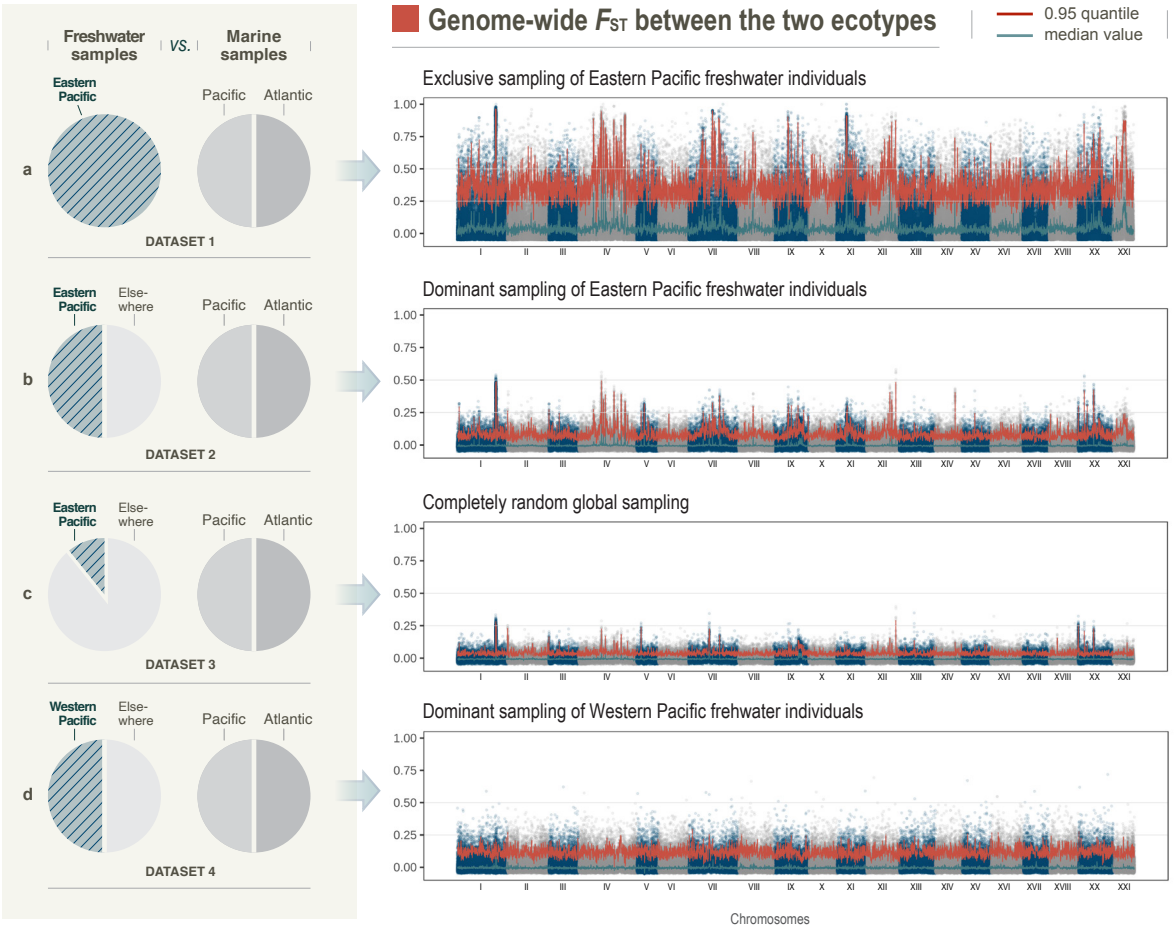

Figure 2 | Linkage Disequilibrium network analysis (LDna). (a-i) Nine main LD-clusters of loci identified by LDna. In each plot, the upper depicts the position of the clustered loci across the genome (x-axis) and for each locus, mean LD (r^2) for each locus (relative to all other loci in its cluster; high values indicates strongly connected loci) on the y-axis. The lower presents visualization of principal component analysis (PCA) showing population divergence based on loci from the LD-clusters. The seven different colours represent the geographic origin of populations. Solid and open circles respectively refer to freshwater and marine ecotypes. All identified 29 LD-clusters and corresponding information are listed in Supplementary Fig. 3 and Supplementary Table 2. (j) Map of sampled populations with colours matching the PCA results. A flat sampling map is given in the Supplementary Fig. 1.

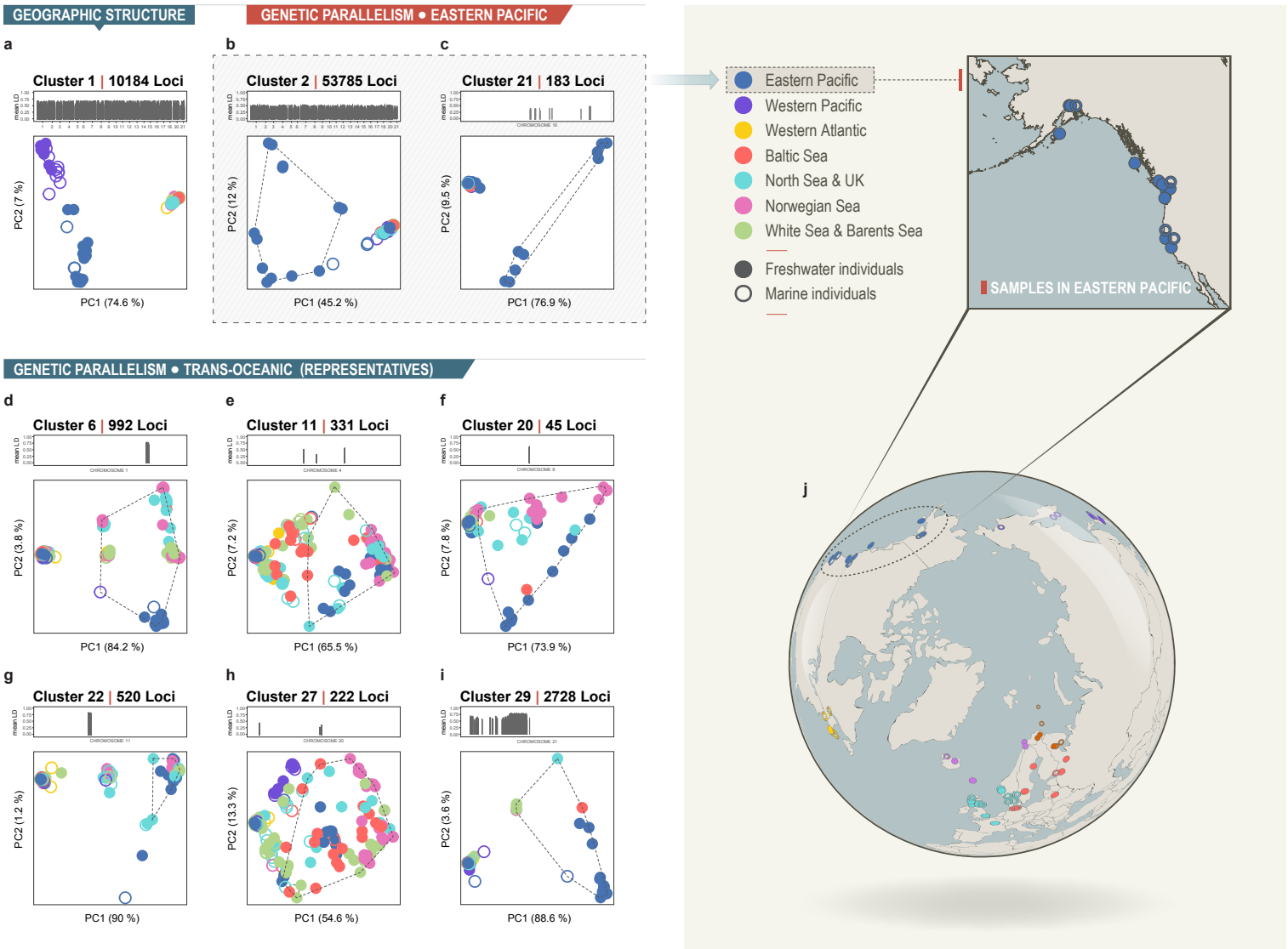

Figure 3 | Genetic parallelism identified by the supervised and unsupervised method. (a) The figure displays loci from the 15 LD-clusters identified by LDna, representing the parallel genomic loci of marine-freshwater divergence across either the Eastern Pacific or the global freshwater populations. The x-axis refers to the genome-wide F_{ST} based on the random sampling of freshwater individuals globally (*viz.* [b] and Fig. 1c). The y-axis represents the genome-wide F_{ST} derived when half of the individuals are from the Eastern Pacific (Fig. 1b). A figure displaying the remaining LDna loci as well as all non-LDna loci across the genome can be found in Supplementary Fig. 2. (b) The genomic position of the parallel genetic variants across the global (trans-oceanic) freshwater populations (13 LD-clusters of lower panel of [a]). The genome-wide F_{ST} (dots in grey colour) is based on the random sampling of freshwater individuals globally (*viz.* Fig. 1c). The LD-clusters are distinguished by corresponding colour at (a).

Figure 4 | Ecological genetics in simulated data. (a-e) A schematic of the demographic scenario used for the simulations that is consistent with the "transporter hypothesis" of Conte & Schluter (2009) in the Eastern Pacific. (a) Initial local adaption of the freshwater populations in the Pacific. (b) The colonisation of stickleback populations from the Pacific to the Atlantic. (c) Geographic isolation between the two oceans. (d) Extinction of lakes during last glacial period (LGP) with the survival of refugee populations. (e) The post-glacial colonisation of the new freshwater populations. (f) Frequency of selected (freshwater-adapted) alleles in the new established freshwater populations through generations at high and low levels of trans-oceanic gene flow and three levels of number of locally adapted traits. (g) Frequency of selected alleles in the marine populations through generations at high and low levels of trans-oceanic gene flow and three different numbers of QTL per chromosome.

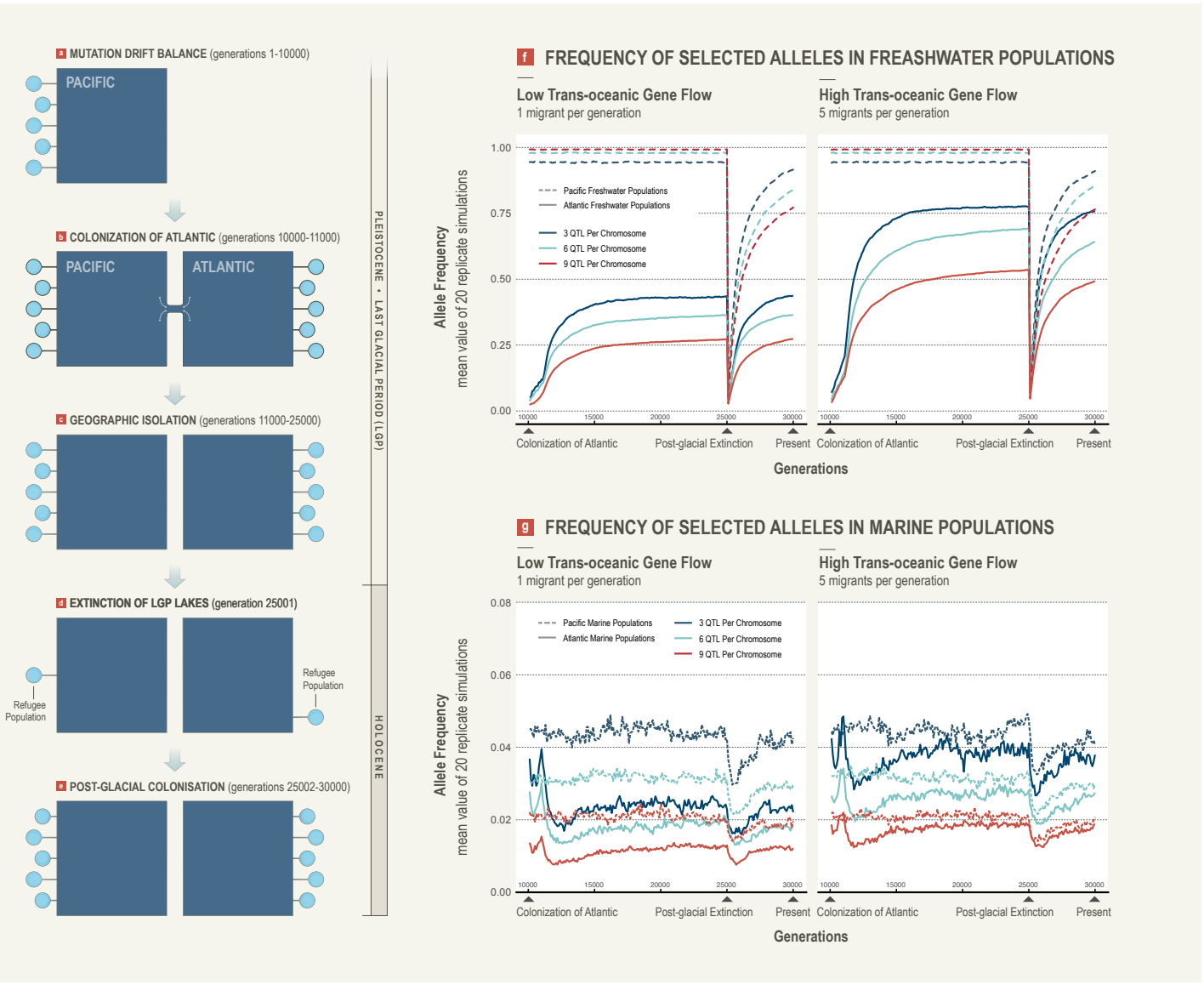

Figure 5 | Genomic differentiation in simulated data. Six plots demonstrate the results of genome-wide marine-freshwater divergence (F_{ST}) under six demographic scenarios at the end of simulations (generation 30,000). The F_{ST} is summarized by the mean value from 20 replicate simulations. The white and shaded background represents “High” or ”Low” recombining regions, corresponding to chromosome arms or centromeres, respectively. The positions of QTL (3, 6 or 9 QTL per Chromosome 1-8) are fixed across simulation replicates. The crosses (+) and black dots (•), at the base in each plot, represent the QTL in high and low recombination genomic regions respectively.