On the maintenance of standing genetic variation by migration-selection balance

Kung-Ping Lin[1], Felicity Jones[2], Sean Rogers[34], David Kingsley[5] and Dolph Schluter[1]
[1]Department of Zoology, University of British Columbia
[2]Groningen Institute for Evolutionary Life Sciences, University of Groningen
[3]Department of Biological Sciences, University of Calgary
[4]Department of Biological Sciences, Bamfield Marine Sciences Centre
[5]Department of Developmental Biology, Stanford University

1. Introduction

When a population encounters a new environment, adaptation can arise from novel or standing genetic variation (SGV). Adaptation using SGV is thought to be faster than that using new mutation, primarily because the population do not need to wait for new advantageous mutations to arise and the standing variants are already at a higher frequency compared to a newly arisen variant (Barrett & Schluter, 2008; Hedrick, 2013). However, the variants beneficial in the new environment are often deleterious in the ancestral environment and are therefore prone to be purged by negative selection before the environmental change. Yet many natural populations have been shown to adapt rapidly to new environments using ancient alleles segregating in ancestral populations, leading to the unresolved question of how they maintain such SGV (Han et al., 2017; Jones, Grabherr, et al., 2012; Lai et al., 2019). In addition, understanding the mechanism maintaining SGV is also crucial for the conservation effort to keep the adaptive potential of populations amid rapid environmental changes nowadays.

Various hypotheses have been proposed to explain how populations maintain SGV. For instance, SGV can be maintained by a migration-selection balance, whereby the variants are selected against yet continuously enter an ancestral population via gene flow from other populations already adapted to the new environment (Barrett & Schluter, 2008). Under such balance, deleterious alleles can be kept at a certain (often low) frequency in the ancestral population (Galloway et al., 2020). Once the ancestral population encounters the new environment, the alleles will then be favored by selection, fueling a rapid adaptation. This hypothesis was partially supported by geographic signals of a few genes, patterns of recombination rates and simulation results (Galloway et al., 2020; Roberts Kingman et al., 2021; Schluter & Conte, 2009). These studies confirm that SGV are critical for rapid adaptation and clarify the mechanism of how ancient alleles reassemble in newly established populations, yet the source of SGV and how the ancestral populations maintain it remains debated (Barrett & Schluter, 2008; Guerrero & Hahn, 2017; Haenel et al., 2022; Kirch, 2025).

In this study, we aim to test migration-selection balance as the fundamental mechanism by which populations maintain potentially deleterious SGV for future adaptation. We focus on threespine sticklebacks (*Gasterosteus aculeatus*), an excellent study system for which SGV has been implicated in fueling its rapid adaptation. The threespine stickleback species complex inhabits most of the northern hemisphere and has repeatedly diverged into freshwater forms from the ancestral marine forms after the glacial retreat in late Pleistocene (~12,000 years ago; Bell & Foster, 1994). At a rapid rate, freshwater sticklebacks around the world evolved remarkably similar freshwater-adapted morphological traits using similar genetic elements. Such genetic parallelism was first demonstrated in *Eda* gene (Colosimo et al., 2005), and later in other genes and loci across the stickleback chromosomes (Jones, Grabherr, et al., 2012; Miller et al., 2007; Roberts Kingman et al., 2021). The repeatedly used genetic elements, along with the remarkable speed of the adaptation, suggest that the parallel adaption to freshwater habitats was likely fueled

by SGV. Moreover, the freshwater-adapted alleles from the parallel adaptation regions were estimated to diverge from the ancestral marine-adapted alleles more than two million years ago, far predating the divergence time of the freshwater lineages from their marine ancestors (<12,000 years ago; Colosimo et al., 2005; Roberts Kingman et al., 2021). A low frequency (~1%) of freshwater alleles can also be found in present-day marine populations, mostly in a heterozygous form (Colosimo et al., 2005; Roberts Kingman et al., 2021). The critical role of SGV in the rapid parallel adaptation of threespine sticklebacks was well supported.

What remains relatively unknown is the mechanism maintaining the low-frequency freshwater alleles as deleterious SGV in the ancestral marine populations. We tested whether such SGV can be maintained by the balance between gene flow from freshwater to marine populations and negative selection in sea. Focusing on the stickleback populations along the coast of northeast Pacific Ocean, we tested the following predictions if migration-selection balance prevails: (1) the degrees of gene flow from freshwater to marine populations should be significant, (2) the freshwater alleles in marine populations should be deleterious and (3) the freshwater alleles should not disperse far once they entered the marine environment and were exposed to negative selection (Figure 1). We further investigated negative frequency-dependent selection with and without gene flows as alternative mechanisms for explaining the distribution of dispersal distances of freshwater alleles in marine environment. This is one of the first studies to test migration-selection balance as a mechanism maintaining SGV that is beneficial in novel but deleterious for current environments in a natural population using genomic evidence.

2. Material and Methods

2.1. Sample collections

To identify the rare freshwater variants in the marine stickleback populations, we collected a total of 904 threespine sticklebacks from the marine, brackish and freshwater bodies across the coast of northeast Pacific (Figure 1 & Table S1). Among the samples, 620 of them were sampled from marine environments and 284 from freshwater environments. Samples were divided into 10 geographical divisions from north to south based on their geographical location and genetic structure: North Alaska (NAK), South Alaska (SAK), Haida Gwaii (HDG), Koeye River (KER), Strait of Georgia (SOG), West Vancouver Island (WVI), Washington Coast (WAC), Oregon (ORE), North California (NCA) and South California (SCA). A total of 585 samples were collected and sent for sequencing for this study, while the other 319 samples were published by previous literatures (Bolnick et al., 2008; Chain et al., 2014; Marques et al., 2018; Morris et al., 2018; Roberts Kingman et al., 2021; Turba et al., 2022). We also included two closely related outgroups to this study, the ninespine stickleback (*Pungitius pungitius*; sequence downloaded from (White et al., 2015)) and the blackspotted stickleback (Gasterosteus wheatlandi; sequence downloaded from (Yoshida et al., 2014)).

2.2. DNA extraction, library preparation and next-generation sequencing

DNA was extracted from the tail fins of sticklebacks using DNeasy® Blood & Tissue Kit (Qiagen, Hilden, Germany). Whole-genome libraries were prepared using Nextera XT DNA Library Preparation Kit (Illumina, Inc, San Diego, USA) and sent to Genome Quebec for 150bp paired-end sequencing on an Illumina Hi-SeqX.

2.3. Variant and invariant calling and their quality controls

FASTQ files were aligned against threespine stickleback reference genome v. 5 (Peichel et al., 2020) using BWA-MEM 0.7.18 (Li, 2013). The aligned reads were sorted using GATK 4.4.0.0 SortSam (Auwera & O'Connor, 2020). Their base quality scores were recalibrated and duplicated reads were removed using GATK BaseRecalibrator and MarkDuplicates. Single nucleotide polymorphisms (SNPs) and invariants were then called and extracted as VCF files using GATK HaplotypeCaller, GenotypeGVCFs and SelectVariants (Poplin et al., 2018). During the above processes, a set of known SNPs from 21 marine and freshwater sticklebacks in (Jones, Chan, et al., 2012; Jones, Grabherr, et al., 2012) were provided as a reference panel to assist the variant calling. All referential SNPs had at least 8x read depth. Their genomic positions were lifted over from stickleback reference genome v.1 to v.5 using the UCSC Genome Browser liftOver command line tool and the chain file provided on the stickleback genome browser website (https://stickleback.genetics.uga.edu/).

We dropped SNPs with more than three alleles, SNPs within three base pairs (bp) from an indel and SNPs in the known repeat regions on the stickleback genome, using custom R scripts and the RepeatMasker file downloaded from the stickleback genome browser website. SNPs with their QD < 2, FS > 60, SOR > 3, MQ < 40, MQRankSum < -4, ReadPosRankSum < -4 and ExcessHet > 54.69 were removed using GATK VariantFiltration, following the recommended practice of GATK (Van der Auwera et al., 2013) with some slight adjustments based on the observed score distributions. Since GATK assigned the missing genotypes as 0/0, we re-assigned the missing genotypes to ./. if the their GQ < 10. For the invariants, we set the loci with DP < 1 to missing genotypes using bcftools 1.19 with plugin +setGT (Danecek et al., 2021).

Potentially contaminated and duplicated samples were discarded using thresholds of per-individual missing rate > 0.7, inbreeding coefficient < -0.1 and KING-robust kinship estimator > 0.354 (Manichaikul et al., 2010). Individual-wise missingness and inbreeding coefficient were calculated using vcftools 0.1.16 (Danecek et al., 2011) and kingship estimator was calculated using PLINK 2.0 (Purcell & Chang, n.d.). Individuals that did not pass through the filters were removed using vcftools 0.1.16.

2.4. Identification of marine-freshwater divergent regions (EcoPeaks) and non-divergent regions (EcoTroughs)

To search for the freshwater-adapted haplotypes in marine populations, we first identified genomic regions that are highly divergent between marine and freshwater individuals. These regions were called "EcoPeaks" hereafter, following the terminology of (Roberts Kingman et al., 2021). We used a modified version of cluster separation score (CSS) to define the degree of marine-freshwater divergence as the score quantifies the averaged genetic distance between the two ecotypes weighted by the within-ecotype variances (Jones, Grabherr, et al., 2012). CSS was calculated as the following formula for a given genomic window:

$$CSS = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} s_{i,j}}{mn} - (\frac{1}{m+n})(\frac{2\sum_{i=1}^{m-1} s_{i,i+1}}{m-1} + \frac{2\sum_{j=1}^{n-1} s_{j,j+1}}{n-1}),$$

where s is the total genetic differences in the window between a pair of diploid individuals, i and j are the marine and freshwater individuals, respectively, and m and n are the corresponding group sizes. We calculated CSS for a 10000 base pairs (bp) window sliding across all chromosomes with 2500 bp gaps using a custom python script.

To define the regions with high and low CSS, i.e., high and low marine-freshwater divergence, we utilized a hidden Markov model (HMM) with Gaussian emissions. The HMM was built with three states (low, medium and high CSS) where the means and variances of each state were estimated through training. We trained the model 100 times with the concatenated window-wise CSS on 23 chromosomes excluding chrY. The model with the highest log probability was then used to predict the states for each window. Windows assigned with the high and low-CSS states were designated as the raw "EcoPeaks" and "EcoTroughs". The model was built and ran using custom python scripts with the function *GaussianHMM* in python package *hmmlearn*.

We defined northern and southern sets of raw EcoPeaks and EcoTroughs using the above methods. The northern EcoPeak was defined using the marine and freshwater individuals north of Washington state (geographical divisions NAK, SAK, HDG, KER, SOG, WVI and WAC), while the southern EcoPeak was defined using the marine individuals north of Washington state and freshwater individuals south of Oregon state (geographical divisions ORE, NCA and SCA). For the northern EcoPeaks, we down-sampled the freshwater individuals in SOG as the sample number in this division is larger than other regions (Figure 2A). To reduce the sample size in SOG, we removed all benthic sticklebacks and reduced the number of individuals from each lake population to one.

2.5. Identification of marine-freshwater divergent SNPs (EcoSNPs) and EcoPeak filtering

To further filter the EcoPeaks and define the types of haplotypes on them, we identify SNPs with high marine-freshwater divergence in the raw EcoPeaks (hereafter, EcoSNPs). For each SNP in the raw EcoPeaks, we defined the two alleles as either marine or freshwater based on their frequencies in the marine and freshwater groups. A SNP was defined as an EcoSNP if its freshwater allele frequency in the freshwater group and its marine allele frequency in the marine

group were both higher than 0.8, while the missing rates in both groups must be lower than 0.3. We used this hard allele frequency cutoff instead of other statistics because it is a straightforward and fair method to define highly divergent sites for both northern and southern EcoPeaks (a cutoff based on the *p*-values of Fisher exact test, for example, is hard to be adjusted to fit both EcoPeaks given their uneven marine/freshwater group sizes). After obtaining the EcoSNPs in both northern and southern EcoPeaks, EcoPeaks with <10 EcoSNPs on them were dropped. Lastly, the intersecting regions of the filtered EcoPeak-N and EcoPeak-S were designated as the common EcoPeaks across the coast of northeast Pacific. We also identified common EcoTroughs as the intersecting regions between the northern and southern regions of low marine-freshwater divergence. The above analyses were done using custom python scripts and *BEDtools 2.31.0* (Quinlan & Hall, 2010).

2.6. Constructing unfolded site frequency spectrum

To construct the unfolded site frequency spectrum (SFS) for the gene flow measurement and demographic history inference, we defined the two alleles for the SNPs in the common EcoTroughs as either ancestral or derived using two closely related outgroup individuals: the ninespine stickleback (*P. pungitius*) and the blackspotted stickleback (*G. wheatlandi*). For each SNP, we set the allele to ancestral (REF allele in VCF file) if it is called homozygous in both outgroups, or homozygous in one outgroup and missing in the other outgroup using a custom python script. SNPs with alleles failed to be sorted into ancestral/derived or with QUAL < 30 or DP < 3 were removed. For the invariants, we removed the loci that were missing in either outgroup using bcftools 1.19.

We constructed two-dimensional joint SFS (JSFS) from nine pairs of populations, including three pairs between freshwater and marine populations (MudLake, Little Campbell River LCR and BigRiver) and six pairs among the marine populations (MudLake, FishCreek, Koeye, LCR, CoosBay, BigRiver and Elkhorn), forming a stepping-stone demographic network (Figure 4A). Using SNPs within the common EcoTroughs excluding the sex chromosome chrXIX, we first removed the SNPs that were in linkage disequilibrium (LD) to each other using *PLINK 2.0* --indep-pairwise with a window size of 50 SNPs, a step size of 5 and a r^2 threshold of 0.2. For each pair of populations, the total number of nucleotides is then calculated as the number of filtered SNPs plus the number of filtered invariants scaled down by the filtering ratio of the LD pruning of SNPs. The unfolded JSFS of all 10 pairs were than constructed using *easySFS* (by Isaac Overcast, https://github.com/isaacovercast/easySFS; Gutenkunst et al., 2009) with the LD-pruned SNPs of the given individuals, the calculated total number of nucleotides and the projection values selected based on the strategy of maximizing the number of SNPs (Table S3).

2.7. Gene flow estimation using SFS-based demographic inference

Fastsimcoal 2.7 was used to estimate the degree of gene flow and the underlying demographic histories for each JSFS constructed for a pair of stickleback populations (Excoffier et al., 2013, 2021). We tested four underlying demographic models: 1Pop (a single panmictic population), 2Pops (an ancestral population split into two subpopulations), 2PopsMig (2Pops with constant gene flow) and 2PopsDifMig (2Pops with two different gene flows; Figure 4B). We set the minimum observed JSFS entries for likelihood calculation (-C) to 10, the number of simulations to perform (-n) to 100000, the number of ECM cycles (-L) to 40 and the mutation rate to 6.8e-8 (Roesti et al., 2015). For each model, we performed the parameter estimation for

100 times and select the run with highest log likelihood to represent the model using a script written by Joana Meier (downloaded from https://raw.githubusercontent.com/speciationgenomics/scripts/master/fsc-selectbestrun.sh). The Akaike information criterion (AIC) of the models were then calculated and compared using a custom python script. Best-fit models were then selected for each pair of populations as the model with the lowest AIC value. For the best-fit model in each population pair, we performed 100 parametric bootstraps to construct the bootstrap distributions for each demographic parameter, with 200000 loci simulated.

2.8. Phasing and defining the types of haplotypes for each EcoPeak

To better identify the rare freshwater haplotypes in the sea, we phased the sequences in the common EcoPeaks to separate the heterozygous marine and freshwater haplotypes. For each common EcoPeak region with heterozygous marine and freshwater haplotypes, we assumed that no recombination had occurred between the two haplotypes. Specifically, we created a VCF file as the reference panel to guild the phasing program, where each heterozygous EcoSNP was manually phased by putting the marine allele at the first haplotype and the freshwater allele at the second. Practically this was achieved by replacing the "/" symbols with the "|" symbols with the marine allele in the first character for each EcoSNP's genotypes in VCF files, using a custom python script. We used *SHAPEIT 5.1.1* to phase the common EcoPeak regions on every chromosome, along with the reference panel we created using the --scaffold tag (Hofmeister et al., 2023).

The linakge maps measured by (Glazer et al., 2015) were also provided to help with the phasing process. Recombination rates measured from a cross between a pair of marine and freshwater sticklebacks from Fishtrap Creek (Washington state) were downloaded from the literature. The coordinates of the SNP bins were first transferred into BED file and lift overed from v. 3 to v. 5 reference genome. The recombination rates were then assigned to the midpoints of each bin to make the MAP file. Note that we only assumed complete linkage within but not among the EcoPeak. Since imputation tended to introduce opposite allele types with other EcoSNPs on the same haplotype, and there are no current ways to turn off the imputation in *SHAPEIT 5*, we replaced all imputed allele back to missing values using a custom python script.

After the phasing, we defined the two haplotypes on each common EcoPeak to be marine, freshwater or undetermined. The types of each haplotype were defined by the proportion of the allele types on every EcoSNP carried by the haplotype. Haplotype were defined as marine, freshwater or undetermined if the proportion of marine alleles was higher than 0.8, lower than 0.2 or in between, respectively, after taking out missing alleles. Such criterium was determined by observing the marine allele proportion distribution on every common EcoPeaks (Figure S2). Haplotypes to the north of Washington and to the south of Oregon were defined using the northern and southern EcoSNPs, respectively. The above analyses were done using custom python scripts.

2.9. Measuring freshwater haplotype frequencies and selection coefficients in marine populations

We measured the frequencies of marine, freshwater and undetermined haplotypes in the marine populations in each geographical division for every common EcoPeaks. The mean frequencies were calculated and plotted on the map using custom python scripts with package GeoPandas. For the three population pairs that we obtained estimates of migration rates from

freshwater into marine populations (MudLake, LCR and BigRiver), we calculated the selection coefficients applying on the freshwater haplotype in the marine populations as

$$\hat{s} = \frac{\sum_{i=1}^{k} m_i(q_i - q_m)}{q_m(1 - q_m)},$$

where k is the number of source populations (can be marine or freshwater) providing the freshwater haplotypes, m_i is the estimated migration rate, q_i is the observed frequency of freshwater haplotype in the source population and q_m is the observed frequency of freshwater haplotype in the sink marine population. The source populations for the target marine populations were its adjacent populations on the stepping-stone network (Figure 4A). Note that the estimated migration rates were from the gene flow measurements among population pairs, while the observed frequencies were calculated from all populations in the same geographical division to reduce sampling error due to the rareness of freshwater haplotype in some marine populations.

2.10. Principal component analysis (PCA) for EcoTroughs and freshwater haplotypes

PCA was performed firstly on the EcoTroughs (quasi-neutral regions) excluding chrXIX for all marine individuals using unphased sequences using *PLINK 2.0*, with SNPs that were in LD removed by --indep-pairwise with 50 SNPs window size, a step size of 5 and a r^2 threshold of 0.5. We also performed PCA for the phased freshwater haplotypes in both marine and freshwater populations on a given common EcoPeak using python package *scikit-learn*. Missing genotypes were permutated using *scikit-learn SimpleImputer* with the mean method.

2.11. Predicting source populations of freshwater haplotypes in marine populations

To identify the most likely freshwater source for each freshwater haplotype in the marine populations, we first transformed the sampling locations of both marine and freshwater populations to 1-dimensional coordinates along the northeast Pacific coast. We used the most southern haplotype location in our dataset (El Rosario, 30.041° N 115.788° W) as the base point and sorted the sample locations into five waypoint groups (Table S1), separated by four waypoints: Cape Mendocino (40.439° N 124.411° W), Neah Bay (48.387° N 124.729° W), Yakutat Bay (59.540° N 139.863° W) and Perl Island (59.145° N 151.697° W). The coastline coordinates of a sample location were then defined as the geodesic distance of the connecting line from the sample location to southern basepoint, passing through the waypoints if they belonged to different waypoint groups (Figure S4). Geodesic distances were calculated using python package *GeoPy*.

Freshwater haplotypes in the marine populations (hereafter, sea haplotypes) were compared to the freshwater haplotypes in all freshwater populations (hereafter, land haplotypes). For each comparison, we estimated their genetic similarity based on identity-by-state (IBS). The comparison will receive a similarity score of one if the two haplotype sequences were identical, and zero if completely different. For each sea haplotype, IBS values from nearby freshwater populations were aggregated into 14 land groups to mitigate the effects of small sample sizes in some freshwater populations. The aggregated IBS values were then sorted from south to north based on the mean coastline coordinates of the land groups. We then fitted the sorted IBS values with a unimodal monotone regression with haplotype numbers in each land group as weights to identify the peak with the highest similarity using a custom R script with package *monotone*. The predicted peak was supposed to locate at the land group where the sea haplotype shared the most similarity with and therefore the most likely source of the sea haplotype. To mitigate bias by

sampling scheme, such prediction was only made for the sea haplotypes where freshwater populations were also sampled in the same geographic division (NAK, HDG, SOG, WVI, NCA and SCA).

2.12 Estimating the dispersal distances of freshwater haplotypes in the sea

Since our IBS prediction method gave a one-to-one prediction of a source land population for each sea haplotype, the prediction results could be noisy due to overconfidence. Therefore, we transformed the prediction result to probabilities of each freshwater population being the true source for a given sea haplotype. First, we performed a "leave-one-haplotype-out" cross validation by predicting the "sources" of every land haplotype excluding the exact haplotypes themselves, using the same IBS-monotone method. For each EcoPeak, the prediction results were then summarized by a confusion matrix C, where rows and columns represented the true and predicted land groups, respectively, while the entries recorded the prediction results as probabilities summed to one in each row. The overall performance of the prediction method for each EcoPeak can be summarized as the proportion of diagonal entries in C, i.e., the accuracy of the confusion matrix. Finally, the transformed probabilities of each land group being the source for the target sea haplotype can be obtained by

$$T = C^{-1} \times E,$$

where E was a vector summed to 1 for the empirical prediction result, where 1 indicates the predicted source group and 0 the others.

For each transformed probability, the "dispersal distances" of the sea haplotype can be calculated as the coastline coordinates of sampling location minus that of the predicted land group. Such distances were then weighted by (1) the transformed source probability, (2) the accuracy of C for the EcoPeak and (3) the reciprocal of total number of marine fish in the geographic division where the sea haplotype was sampled. The weighted dispersal distances were then sorted by their source geographic divisions and visualized as histograms or fitted normal distributions for. We referred to the distributions of the dispersal distances as the "migration kernels" of freshwater haplotypes in the sea. The above analyses were done using custom python scripts.

2.13 Expected migration kernel under migration-selection balance and negative frequency-dependent selection

To understand the expected dispersal distances of sea haplotypes, we simulated the stepping-stone demographic network using *SLiM 4.0.1* (Haller & Messer, 2023) with seven marine and freshwater populations representing MudLake, FishCreek, Koeye, LCR, CoosBay, BigRiver and Elkhorn with empirically measured population sizes and migration rates (for freshwater-marine migrations we used effective migration rates $m{\cdot}f$ where f was the freshwater frequency in the freshwater geographic division, measured or assumed; Table S6). We simulated a single genetic locus in diploid organisms with two alleles: allele F fixed in freshwater and allele M fixed in marine populations, assuming no spontaneous mutations and recombination. Once an allele F migrated to a marine population, it was exposed to a constant negative selection with selection coefficient s, which was empirically measured for the marine population or interpolated (Table S6). We ran the simulations 500 times for 8000 generations and recorded the source populations for every allele F in marine populations at generation 1000, 2000, 4000 and 8000. The expected distributions of migration kernel at a certain generation can then be constructed by calculating the dispersal distances based on the coastline coordinates of the

marine populations in the stepping-stone network. We can then estimate the *p*-values of the observed means and standard deviations of migration kernels based on the simulated kernels.

To test the alternative scenario of balancing selection with gene flow, we constructed the same model using SLiM but with negative frequency-dependent selection imposing on allele F in marine populations, where selection coefficients were

$$s = 1 + q - f,$$

where q was the observed frequency of freshwater haplotypes in the marine geographic division and f was the freshwater allele frequency in the simulated marine population. We also tested another alternative scenario with pure balancing selection without any freshwater-to-marine gene flow. In such model, freshwater populations in the stepping-stone network were removed while the marine populations were introduced with 100 freshwater individuals at the beginning of the simulation. We defined the pseudo sources of each freshwater haplotype to be the adjacent (now removed) freshwater populations of the marine populations where they were introduced. The sources for each freshwater haplotypes were again recorded at generation 1000, 2000, 4000 and 8000 and the expected migration kernels were constructed.

2.14 Multiple linear regression on the standard deviation of migration kernel

We tested the number of SNPs, accuracy of the confusion matrix and frequencies of freshwater haplotypes in the northern and southern seas as the explanatory variables of the standard deviations of the migration kernels for the common EcoPeak. The regression was performed using python package *statsmodels* on all common EcoPeaks and common EcoPeaks with ≥ 0.7 accuracy. We also plotted the simple linear regression plot with confidence intervals between the standard deviation and the frequencies for visualization using python package *seaborn*.

3. Results

3.1. Sample collection, SNP calling and quality control

We obtained a total of 30148991 single nucleotide polymorphisms (SNPs) from whole-genome resequencing data of 900 threespine sticklebacks from the marine (619) and freshwater (281) habitats along the northeastern Pacific coast, after all quality control procedures (Figure 2A; Table S1). Samples were grouped into 10 geographical divisions from north to south based on their geographical proximity and genetic structure: North Alaska (NAK), South Alaska (SAK), Haida Gwaii (HDG), Koeye River (KER), Strait of Georgia (SOG), West Vancouver Island (WVI), Washington Coast (WAC), Oregon (ORE), North California (NCA) and South California (SCA).

3.2. Identifying regions with high and low marine-freshwater divergence

To identify the freshwater haplotypes in the marine stickleback populations, we first defined the regions with high marine-freshwater divergence. We used a modified version of cluster separation score (CSS) as an estimator for parallel marine-freshwater divergence within genomic windows (Jones, Grabherr, et al., 2012; Roberts Kingman et al., 2021). A high CSS indicates a large average genetic distance between marine and freshwater individuals, controlling for within-group variation, and thus suggests strong divergence. We calculated CSS in sliding windows (25000 bp with 10000 bp steps) across the stickleback genome, excluding the Y chromosome. Each window was classified as low, intermediate, or high divergence using a hidden Markov model with three hidden states based on CSS. Windows classified as high and low divergences were designated as raw "EcoPeaks" and "EcoTroughs," respectively.

Because that marine populations south of Oregon showed strong introgression with freshwater haplotypes (Figure 5B) and that freshwater haplotypes clustered into northern (north of Washington) and southern (south of Oregon) groups (Figure 6A), we defined two separate sets of EcoPeaks and EcoTroughs. The north set was defined using northern marine and northern freshwater populations, while the south set was defined using northern marine and southern freshwater populations (Figure S1A). We further filtered both sets by excluding regions with fewer than 10 marine-freshwater divergent SNPs (EcoSNPs), defined by an allele frequency cutoff of 0.8. As a result, EcoPeaks were defined using two criteria: (1) a window-wise separation score based on between- and within-group genetic distances and (2) SNP-wise scores based on allele frequency thresholds. The north and south EcoPeaks showed moderate similarity, with 61.5% of regions overlapping (Figure S1B), indicating that while some freshwater-adapted alleles are unique to either northern or southern populations, most of them are shared.

By intersecting the north and south EcoPeaks and retaining regions with at least 10 EcoSNPs in both sets, we identified 36 common EcoPeaks, covering 16.5 Mb, 3.61% of the genome. These regions encompassed three known inversions and genes associated with marine-freshwater parallel adaptation such as *Eda* (Table 1 & S2; Figure 3 & S1). Similarly, intersecting north and south EcoTroughs yielded 1652 common EcoTroughs, spanning 85.4 Mb, 18.7% of the genome (Table 1 & Figure S1). These common EcoPeaks and EcoTroughs thus represent, respectively, the universally divergent and quasi-neutral loci between marine and freshwater populations along the northeastern Pacific coast. Our common EcoPeaks overlap moderately (29.4%) with the Northeast Pacific-specific EcoPeaks reported by Roberts Kingman et al. (2021), likely because their study used only northern marine and freshwater populations and relied on *p*-value-based methods (Figure S1C).

3.3. Gene flows are significant from freshwater to marine populations

Under a migration-selection balance, gene flow from freshwater populations into marine populations is expected to be significant. To test this prediction, we measured freshwater-to-marine gene flow in three marine-freshwater population pairs: Mud Lake, Little Campbell River (LCR), and Big River. These pairs were selected based on large sample sizes and geographic proximity (freshwater populations needed to be upstream of their marine counterparts). We also measured gene flow among seven marine populations: Mud Lake, Fish Creek, Koeye, LCR, Coos Bay, Big River, and Elkhorn, which were selected based on large sample sizes and genetic structure (marine populations formed clusters in principal component analysis, PCA, based on common EcoTroughs; Figure 2B). Together, these populations formed a stepping-stone demographic network along the northeastern Pacific coast (Figure 4A). We constructed the joint site-frequency spectrum (JSFS) between the nine population pairs in the network using the common EcoTroughs. Gene flow was estimated from the JSFS using *fastsimcoal2* for each population pair under four demographic models (Figure 4B; Excoffier et al., 2013, 2021). For all pairs, the best-fit models were always those that included gene flow (Models III or IV; Figure 4A; Table 2 & S3), indicating that gene flow was significant in all comparisons, including from freshwater into marine populations that potentially carrying freshwater alleles. Interestingly, the degree of recent gene flow from freshwater to marine populations was much lower in Alaska compared to British Columbia and California, although still statistically significant (Figure 4C). In addition, the levels of recent gene flow among marine populations were also relatively low, suggesting that marine stickleback populations along the northeastern Pacific coast form distinct genetic structures in the ocean, consistent with the assuming stepping-stone network model (Table S3).

3.4 Freshwater haplotypes are deleterious in the sea but less deleterious in the south

The second prediction of migration-selection balance is that freshwater haplotypes should be deleterious in marine environments. To test this prediction, we first identified rare heterozygous freshwater haplotypes in marine populations by phasing sequences in each common EcoPeak independently using *SHAPEIT 5*, with EcoSNPs as reference panels (Hofmeister et al., 2023). Haplotypes carrying 80% or more freshwater alleles at EcoSNP sites were classified as freshwater haplotypes (Figure 5A & S2). We found that the frequencies of freshwater haplotypes in marine populations, grouped by geographic divisions, decreased from south to north along the northeastern Pacific coast (Figure 5B & Table 3), while frequencies in freshwater populations remained consistently high (Figure S3 & Table S4). Assuming that the observed frequencies are under equilibrium between selection and gene flow, we estimated the selection coefficient $\hat{s}$ acting on freshwater haplotypes in marine populations in three geographic divisions: NAK, SOG, and NCA, based on empirically measured freshwater haplotype frequencies and migration rates among freshwater and marine populations. We found that selection against freshwater haplotypes in the sea was strongest in the north and nearly neutral in the south (Table 3). This pattern aligned with results from the ancestry-heterozygosity triangular plot (Figure 5C) and the distribution of the number of freshwater haplotypes per individual (Figure 5D), where southern marine sticklebacks carried more freshwater haplotypes than their northern counterparts. In the north, freshwater haplotypes in marine individuals were mostly backcrosses with low freshwater ancestry and heterozygosity, suggesting that individuals carrying many deleterious freshwater alleles could not survive beyond the F1 generation. In contrast, many freshwater haplotypes in southern marine populations were homozygous and

carried high freshwater ancestry, implying that these haplotypes are under weak selection or even effectively neutral in the sea (Figure 5C).

3.5 Predicting the sources and measuring the dispersal distances of freshwater haplotypes in the sea

The third prediction of migration-selection balance is that freshwater haplotypes in the marine environment should originate from nearby freshwater populations and their dispersal distances should be short. To test this, we first performed a principal component analysis (PCA) on phased freshwater haplotypes found in marine populations (hereafter, sea haplotypes) and in freshwater populations (hereafter, land haplotypes), to visualize whether sea and land haplotypes clustered according to their geographic locations. As an example, we found that sea haplotypes clustered with land haplotypes from the same geographic divisions on the EcoPeak containing the chromosome I inversion, particularly in the southern regions (Figure 6A).

To further quantify how far freshwater haplotypes had traveled in the sea, we predicted the most likely source location on the land for each sea haplotype. Each sea haplotype sequence from the common EcoPeaks was compared to all land haplotype sequences to calculate an identity-by-state (IBS) score, reflecting genetic similarity. For each sea haplotype, IBS scores were aggregated into 14 land groups (each consisting of nearby freshwater populations) and sorted from south to north based on their one-dimensional coordinates along the northeastern Pacific coastline (Figure S4). We then applied unimodal monotonic regression to the sorted IBS scores, with the peak of highest genetic similarity indicating the predicted source location of the sea haplotype (Figure 6B). We evaluated the performance of this source prediction method across all common EcoPeaks using the accuracies of the confusion matrices derived from leave-one-haplotype-out cross-validation. EcoPeaks with a higher number of SNPs generally showed greater prediction accuracy (Figure S5 & Table S2).

We selected common EcoPeaks with prediction accuracy ≥ 0.7 to estimate the dispersal distances of sea haplotypes. This subset included seven EcoPeaks, which together contained two known inversions and the *Eda* gene (Figure S5). The dispersal distance of each sea haplotype in these regions was calculated as the coastal distance between its sampling location and its predicted source location, weighted by prediction accuracy, source probabilities (adjusted by the confusion matrices), and the number of marine samples in each geographic division. The resulting distributions of dispersal distances—grouped by source geographic divisions—represent how far freshwater haplotypes travel once they enter the sea. We refer to these distributions as the "migration kernels." We found that the mean dispersal distances were within 250 km from the source for most regions, except for Alaska and Haida Gwaii (Figure 6C & S6; Table 4 & S5), where freshwater haplotypes traveled significantly farther, reaching as far south as the Strait of Georgia with average distances over 2,000 km. The standard deviations of the migration kernels were within 500 km, also except for Alaska. Additionally, we constructed migration kernels using all 36 common EcoPeaks, which produced more extreme mean distances and wider standard deviations (Figure S6 & Table S5).

3.6 Migration-selection balance can explain how far a freshwater haplotype travels in the sea

We performed simulations under a pure migration-selection balance scenario to test whether the observed migration kernels could be explained by this hypothesis. The simulations were based on the stepping-stone demographic network, incorporating previously estimated migration rates and selection coefficients acting on freshwater haplotypes in marine populations.

Simulations were conducted using *SLiM* (Figure 4A & Table S6; Haller & Messer, 2023). We tracked freshwater alleles present in marine populations and recorded their original land sources at generations 1000, 2000, 4000 and 8000. Expected migration kernels under the migration-selection balance were constructed using the coastline coordinates of marine populations in the stepping-stone network (Mud Lake, Fish Creek, Koeye, LCR, Coos Bay, Big River and Elkhorn) at generation 4000. This generation was chosen based on empirically estimated divergence times among the stepping-stone populations (Table S3) and the point at which changes in freshwater allele frequencies in marine populations had stabilized (Figure S7). We found that the simulated migration kernels closely matched the observed means and standard deviations of the empirical kernels ($p \geq 0.05$), except for those from Alaska and the mean for South California (Figure 6C & D; Table S5). We also tested alternative hypotheses involving negative frequency-dependent selection with and without freshwater-to-marine gene flow. The scenario with frequency-dependent selection with gene flow produced migration kernels similar to those under pure migration-selection balance. In contrast, the scenario without gene flow only fit the kernel from South California (Figure 6D; Table S5).

3.7 Positive but insignificant correlation between the dispersal distance of freshwater haplotypes in the sea and their frequency

If standing genetic variation (SGV) is maintained by a balance between gene flow and forms of negative selection, we would expect freshwater haplotypes under weaker selection to disperse farther than those under stronger selection. To test this prediction, we examined the relationship between the standard deviation of migration kernels and the frequency of freshwater haplotypes in both northern and southern marine populations using multiple linear regression. We found a positive correlation between freshwater haplotype frequency and dispersal distance in both regions (Table S7 & Figure S8). However, this correlation was only significant for freshwater haplotypes in the southern sea when using the subset of common EcoPeaks with prediction accuracy ≥ 0.7 ($p = 0.092$).

4. Discussions

How populations maintain standing genetic variation (SGV) has long been considered a paradox, as such variation is often assumed to be deleterious and subject to removal by natural selection (Yeaman, 2015). In efforts to resolve this puzzle, the maintenance of SGV through migration-selection balance has emerged as a leading explanation (Galloway et al., 2020; Schluter & Conte, 2009). However, alternative hypotheses and contrasting evidence have also been proposed (Guerrero & Hahn, 2017; Haenel et al., 2022; Kirch, 2025). In this study, we investigated the mechanisms driving the maintenance of freshwater alleles as SGV in marine threespine sticklebacks along the northeastern Pacific coast. Our results show that gene flow plays a critical role in shaping the observed SGV patterns in marine sticklebacks, with natural selection acting as a counterbalancing force. The specific form of the selection acting on SGV, however, remain unresolved.

4.1 Stickleback populations are highly structured in the sea

Marine sticklebacks, often considered proxies for the ancestral populations in the threespine stickleback system, have traditionally been viewed as a single panmictic lineage due to the absence of clear physical barriers and their migration capability (Bell, 1976; Bell & Foster, 1994; Galloway et al., 2020). However, such assumption had been challenged by recent studies (Catchen et al., 2013; Morris et al., 2018). Consistent with these findings, we detected significant population structure among marine sticklebacks along the northeastern Pacific coast, with limited gene flow measured between populations (Figures 2B & 4A). This genetic structuring plays an important role in shaping the geographic distribution of freshwater alleles within marine populations, suggesting that population connectivity is a key factor influencing the persistence and spatial dynamics of SGV in ancestral populations.

4.2 Significant freshwater-to-marine gene flow in threespine sticklebacks and its latitudinal gradient

A key prerequisite of the migration-selection balance hypothesis is the presence of gene flow that transports SGV from derived (freshwater) populations back into ancestral (marine) populations. Although various hybrid zones have been documented between anadromous and freshwater forms of sticklebacks (Hagen, 1967; Hay & McPhail, 2000; Jones et al., 2006), whether gene flow between them is significant has not been directly tested to our knowledge. Here we demonstrate that gene flow between freshwater and marine populations is indeed significant, and that its magnitude varies along a latitudinal gradient (Figure 4 & Table 2). Specifically, freshwater-to-marine gene flow is much lower in Mud Lake, Alaska, compared to Little Campbell River in British Columbia and Big River in California (Figure 4C). This finding is consistent with previous field observations reporting no hybrids between anadromous and freshwater resident sticklebacks in Mud Lake (Karve et al., 2008). In contrast, we detected high levels of gene flow from marine to freshwater populations in Alaska (Figure 4A & Table S3). This asymmetry suggests that reproductive isolation may be stronger in the freshwater-to-marine direction than in the opposite direction. One potential caveat is that our estimates of gene flow might be artificially elevated if balancing selection maintains freshwater alleles in the marine environment—an alternative mechanism for the persistence of SGV (Smith & Hahn, 2024). While the extreme case that freshwater alleles are maintained entirely by balancing selection with no freshwater-to-marine gene flow seems unlikely (given that most freshwater alleles found

in marine populations trace back to nearby freshwater sources; Figure 6), a model combining both balancing selection and gene flow remains plausible (Figure S6).

4.3 Negative selection on freshwater haplotypes in the sea and its latitudinal gradient

Although we observed low freshwater-to-marine gene flow in Alaska, the estimated selection coefficient for freshwater alleles in the sea here was still high, indicating strong negative selection against them. For marine alleles in Alaskan freshwater populations, we measured even stronger selection, consistent with previous studies showing that cold temperatures and low salinity pose greater challenges for marine sticklebacks in freshwater than vice versa (Table S4; Barrett et al., 2010; Gibbons et al., 2017). In contrast, we found near-zero selection coefficients for freshwater alleles in southern marine populations, suggesting that these alleles are nearly neutral in those environments (Table 3). Collectively, our results reveal a latitudinal gradient of decreasing negative selection against freshwater alleles in the sea from north to south.

This gradient may reflect unique environmental conditions in southern marine habitats, where freshwater alleles are found at higher frequencies and many individuals exhibit freshwater-like low-plate morphology (Table S1). This pattern aligns with previous research documenting high frequencies of freshwater alleles and lateral plate phenotypes in California's marine stickleback populations (Baumgartner & Bell, 1984; Colosimo et al., 2005; Des Roches et al., 2020). A likely explanation is that southern estuarine environments, characterized by more stable water bodies due to reduced precipitation, higher temperatures and greater drought severity, favor freshwater alleles (Des Roches et al., 2020). Indeed, a large proportion of our marine samples south of Oregon were collected from saltwater marshes, where salinity is known to fluctuate (Table S1; Morris et al., 2018). The neutral-like selection observed for freshwater alleles in these areas may indicate that such alleles are beneficial, or at least not harmful, for adaptation to these complex habitats. However, it remains unclear whether pure marine/anadromous forms exist in the southern California marine ecosystem or whether they were simply not sampled. Overall, our findings suggest that the frequency of freshwater alleles in marine populations as SGV is shaped by a balance between freshwater-to-marine gene flow and selection that varies along the northeastern Pacific coast. In Alaska, low gene flow and strong negative selection lead to low frequencies of freshwater alleles in the sea. In California, moderate gene flow combined with weak or neutral selection results in higher frequencies. The major transition zone appears to lie somewhere between Washington and Oregon.

4.4 Migration-selection balance explains the dispersal distances of freshwater haplotypes in the sea in BC and California

If SGV in the ancestral population is truly maintained by a balance between gene flow and negative selection, a key prediction is that most freshwater haplotypes in marine populations should originate from nearby derived (freshwater) populations and should not disperse far. To test this, we predicted the source of each freshwater haplotype found in marine individuals along the coast and computed their dispersal distance distributions, referred to as the migration kernels. We compared the empirical estimates of the means and standard deviations to the expected migration kernels based on simulations built from three different mechanisms maintaining SGV. These simulations allowed us to control for potential artifacts such as regression to the mean and enabled more interpretable estimates. The three hypotheses tested were: (1) migration-selection

balance, (2) negative frequency-dependent selection (NFDS) with freshwater-to-marine gene flow and (3) NFDS without gene flow.

For migration kernels originating from the Strait of Georgia and North California (central areas of the coastline), both the means and standard deviations can be explained by either migration-selection balance or NFDS with freshwater-to-marine gene flow. This result is perhaps unsurprising, as SGV under NFDS is typically slightly deleterious due to the continuous influx of maladaptive alleles from nearby derived populations via gene flow. In fact, the behavior of these two mechanisms is so similar that the dynamics of migration kernels generated by them over simulated generations are nearly indistinguishable (Figure S9). As a result, it is difficult to distinguish between these two mechanisms based solely on the dispersal distances of SGV. To further evaluate the role of gene flow, we tested a hypothetical scenario involving NFDS without freshwater-to-marine gene flow. The simulated migration kernels under this model fit the observed data poorly, producing significantly broader distributions and more extreme mean distances (Figure 6D & Table S5). It is worth noting that to some degree our NFDS model simulates the alternative hypothesis that SGV is maintained by reduced purifying selection when its frequency is low (Haenel et al., 2022). The results shown here therefore suggest that such hypothesis can be valid only with there is significant amount of gene flow from the derived to ancestral populations.

For migration kernels from Southern California (southern areas of the coastline), we found that the observed mean dispersal distance is larger than the expected means under both migration-selection balance and NFDS with gene flow, while the standard deviations did not differ significantly. Interestingly, both the mean and standard deviation fit better under the NFDS model without gene flow (Figure 6D & Table S5). These results may suggest that balancing selection plays a stronger role in the south and/or that freshwater-to-marine gene flow is more limited in this region. However, such interpretation should be made with caution. In Southern California, we lacked direct estimates of gene flow and instead relied on migration rates inferred from assumed selection coefficients in the sea (Table S6). Nevertheless, our findings suggest that the selective pressure on freshwater alleles in the southern marine environment may be weaker than the constant negative selection assumed under pure migration-selection balance. As a result, gene flow among marine populations may play a relatively greater role in shaping the observed patterns of SGV in this region.

The most unexpected result came from the migration kernel originating in Alaska, where the observed kernel did not align with any of the three tested hypotheses. Specifically, freshwater haplotypes appeared to disperse much farther south than expected and exhibited substantially greater variation (Figure 6C & D). A similar pattern was observed for the migration kernel from Haida Gwaii, although no simulated kernel was available for direct comparison in that case (Figure S6 & Table S5). However, this pattern may largely reflect an artifact arising from the limited sampling of freshwater haplotypes in the northern Alaskan sea that only one effective haplotype had a source probability ≥ 0.1 for both North Alaska and Haida Gwaii (Figure S10). As we defined the migration kernels in terms of the absolute number of freshwater haplotypes rather than their frequencies, regions with more extensive marine sampling, such as the Strait of Georgia, contribute more heavily to the overall kernel and potentially exacerbate the influence of prediction errors in these regions.

4.5 Freshwater haplotypes travel further in the sea if they are less deleterious

If migration-selection balance is indeed the mechanism maintaining SGV in marine threespine stickleback populations, we would expect them to disperse farther when subjected to weaker negative selection. To test this prediction, we examined the correlation between the standard deviations of migration kernels and the frequencies of freshwater haplotypes in the northern and southern seas, using haplotype frequency as a proxy for the strength of selection. Consistent with our prediction, we found positive correlations between the standard deviations of dispersal distances and the frequencies of freshwater haplotypes, using both high-accuracy EcoPeaks and the full set of EcoPeaks (Figure S8 & Table S7). However, statistical significance was only observed for the southern sea when using the subset of high-accuracy EcoPeaks. The lack of significance in other cases may be due to the small sample size (only seven high-accuracy EcoPeaks) and the potential non-linear relationship between haplotype frequency and dispersal distance. Future analyses with more accurate prediction methods and samples would help address these limitations.

5. Conclusions

Standing genetic variation (SGV) plays a critical role in rapid and parallel adaptation but is often difficult to maintain in natural populations. In this study, we tested whether migration-selection balance can explain the persistence of freshwater alleles as SGV in marine threespine sticklebacks. Our findings support this hypothesis and reveal the following: (1) gene flow is a major mechanism introducing the derived SGV into ancestral populations, (2) derived SGV is maladaptive in the ancestral populations but with a latitudinal gradient and (3) dispersal distances of SGV in the sea reflect the balance between gene flow and selection, with gene flow playing a larger role in the south and selection dominating in the north. While migration-selection balance provides a strong framework, the precise form of selection remains unresolved and warrants further investigation, particularly for the southern Californian marine environments.

6. Acknowledgement

XXX

Tables

Table 1 statistics of the inferred regions of high and low marine-freshwater divergence excluding chrY and chrM.

Genomic regions	# regions	Total bases (%)	Median size
Common EcoPeak	36	16455000 (3.61%)	342500
Roberts Kingman EcoPeak*	198	25993491 (5.70%)	80513
Common EcoTrough	1652	85412500 (18.72%)	30000

*North East pacific specific EcoPeak from (Roberts Kingman et al., 2021)

Table 2 best estimates and 95% bootstrap confidence intervals of demographic parameters in nine population pairs across the northeast Pacific coast.

Population 1	Population 2	Best model	N_{e1}	N_{e2}	$m_{1\to2}$	$m_{2\to1}$
MudLake marine	MudLake freshwater	IV	13176 [12980, 13441]	313 [293, 356]	9.45e-3 [8.21e-3, 1.02e-2]	3.13e-5 [1.62e-5, 8.10e-5]
LCR* marine	LCR freshwater	IV	6560 [6442, 6665]	436 [425, 482]	2.82e-4 [3.67e-5, 4.96e-4]	1.08e-3 [8.16e-4, 1.30e-3]
BigRiver marine	BigRiver freshwater	IV	12408 [11981, 12698]	1922 [1684, 2163]	3.85e-3 [3.29e-3, 4.59e-3]	1.15e-3 [9.59e-4, 1.30e-3]
MudLake marine	FishCreek marine	IV	13569 [13108, 13904]	3040 [2965, 3402]	9.59e-3 [8.32e-3, 1.01e-2]	8.73e-4 [6.72e-4, 1.20e-3]
FishCreek marine	Koeye marine	IV	9328 [9142, 9551]	2521 [2454, 2625]	7.72e-5 [2.53e-05, 3.35e-4]	4.92e-5 [3.01e-5, 2.70e-4]
Koeye marine	LCR marine	III	2958 [2925, 3088]	9833 [9637, 9983]	3.04e-4 [2.80e-4, 3.36e-4]	1.50e-3 [1.41e-3, 1.51e-3]
LCR marine	CoosBay marine	IV	8281 [8115, 8460]	10489 [10258, 10724]	7.75e-4 [6.68e-4, 8.19e-4]	7.97e-4 [7.37e-4, 8.57e-4]
CoosBay marine	BigRiver marine	IV	12799 [12489, 13205]	8101 [7785, 8544]	1.23e-3 [1.15e-3, 1.34e-3]	5.24e-4 [4.63e-4, 5.66e-4]
BigRiver marine	Elkhorn marine	IV	13674 [13174, 13829]	2546 [2494, 2669]	6.02e-4 [5.40e-4, 6.53e-4]	9.58e-5 [4.80e-5, 1.12e-4]

*Little Campbell River

Table 3 frequencies and selection coefficients of freshwater haplotypes in marine populations in different geographic divisions

Geographic division	# haplotypes	Frequency	$\hat{s}^*$
NAK	96	0.0026	0.0496
SAK	53	0.1174	-
KER	45	0.0531	-
SOG	127	0.0468	0.0273
WVI	50	0.0736	-
ORE	84	0.4335	-
NCA	112	0.5246	0.0004
SCA	51	0.5648	-

*$\hat{s}$ was calculated using frequencies of freshwater haplotypes and migration rates between freshwater and marine and among marine populations (Table 2).

Table 4 dispersal distances of freshwater haplotypes in marine environment from their predicted land sources in each geographic division using high-accuracy common EcoPeaks

Source geographic division	# effective haplotypes+	Mean (km)	Standard deviation (km)
NAK	7	-2239.6	727.9
SOG	83	-229.7	488.7
NCA	100	11.3	178.9
SCA	60	230.4	257.1

+Haplotypes with ≥ 0.1 source probabilities

Figures

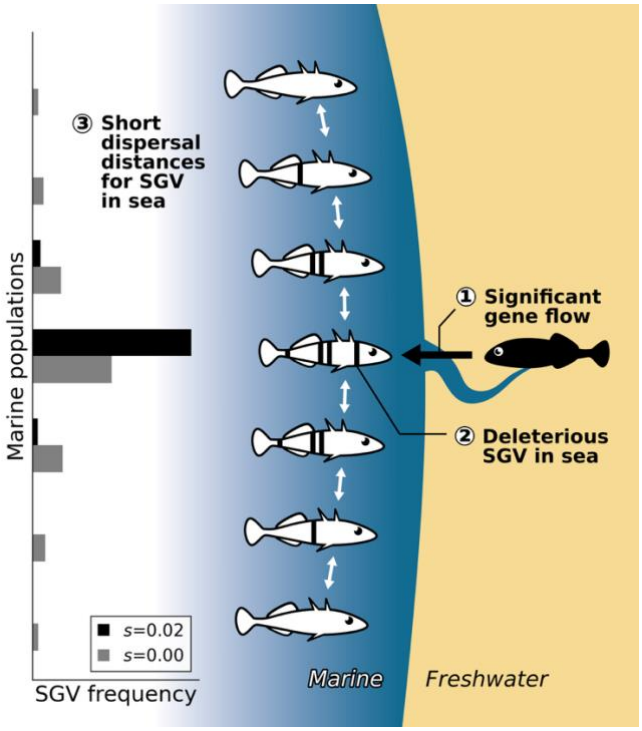

Figure 1 predictions of migration-selection balance as the mechanism maintaining standing genetic variation (SGV) in marine threespine sticklebacks. Under migration-selection balance hypothesis, freshwater alleles are maintained in marine populations as SGV by gene flows and negative selection. The hypothesis predicts (1) significant amount of gene flow from freshwater into marine popualtions, (2) freshwater alleles are deleterious in the marine environment and (3) the dispersal distances of freshwater alleles should be relatively short once entering the sea. Histogram on the left shows the expected distributions of the frequencies of freshwater alleles in the seven marine populations with different selection coefficients using 100 SLiM simulations lasting 4000 generations. The marine populations are connected by gene flows with migration rates (m) = 10^{-3} (white arrows) while freshwater alleles continuously arrive at the middle marine population also with m = 10^{-3} (black arrow).

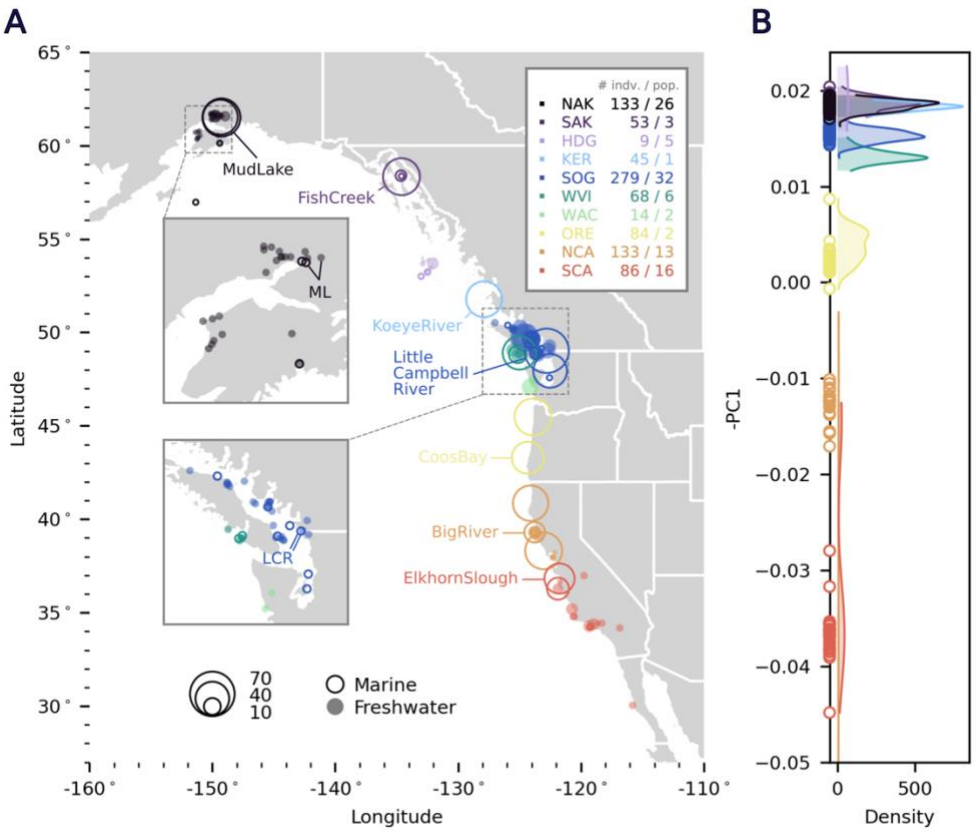

Figure 2 sample locations and the marine population structure along the coast of northeast Pacific. (A) The locations of the sample populations and their corresponding geographic divisions. The populations for gene flow measurements for the stepping-stone demographic network were labeled with texts. Map data was downloaded from (Commission for Environmental Cooperation (CEC), 2022). (B) Principal component analysis result of common EcoTroughs (quasi-neutral regions) of all marine populations. Marine populations used for gene flow measurement were marked with circles. The plot was zoomed in for better visualization.

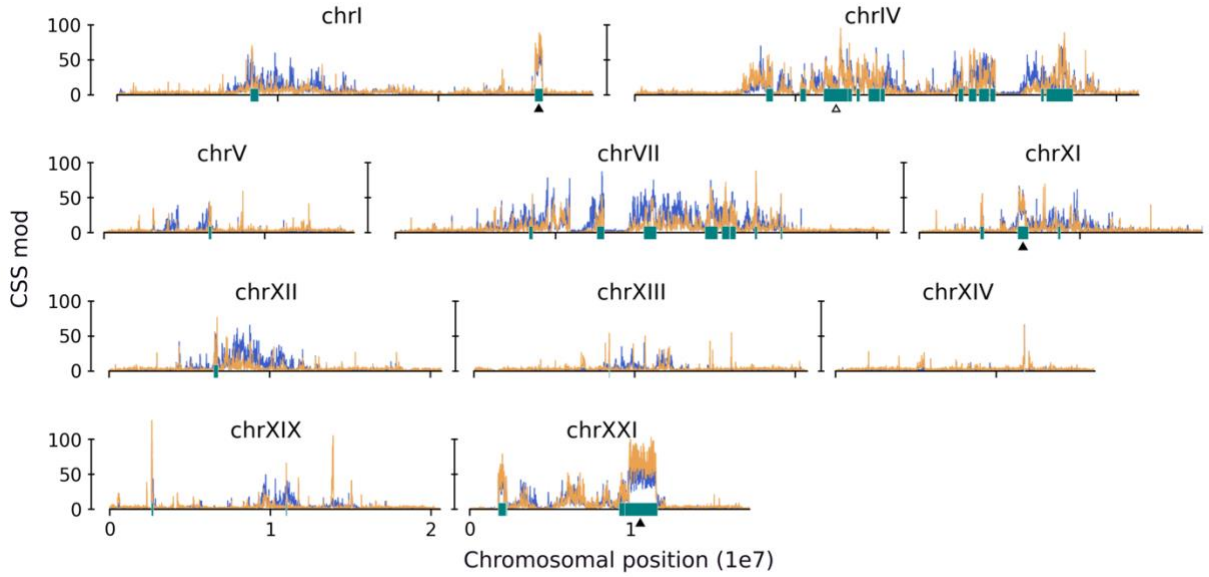

Figure 3 genomic regions of high marine-freshwater divergence (EcoPeaks) on threespine stickleback chromosomes. Window-wise cluster separation scores (CSS) were calculated between the marine fish north of Washington and the freshwater fish north of Washington (blue), and between the marine fish north of Washington and freshwater fish south of Oregon (orange). Teal ribbons indicate the locations of common EcoPeaks defined by the intersected regions of north and south EcoPeaks, based on the north and south CSS, respectively. Common EcoPeaks covering known inversion regions and *Eda* gene were marked by solid and hollowed black triangles, respectively.

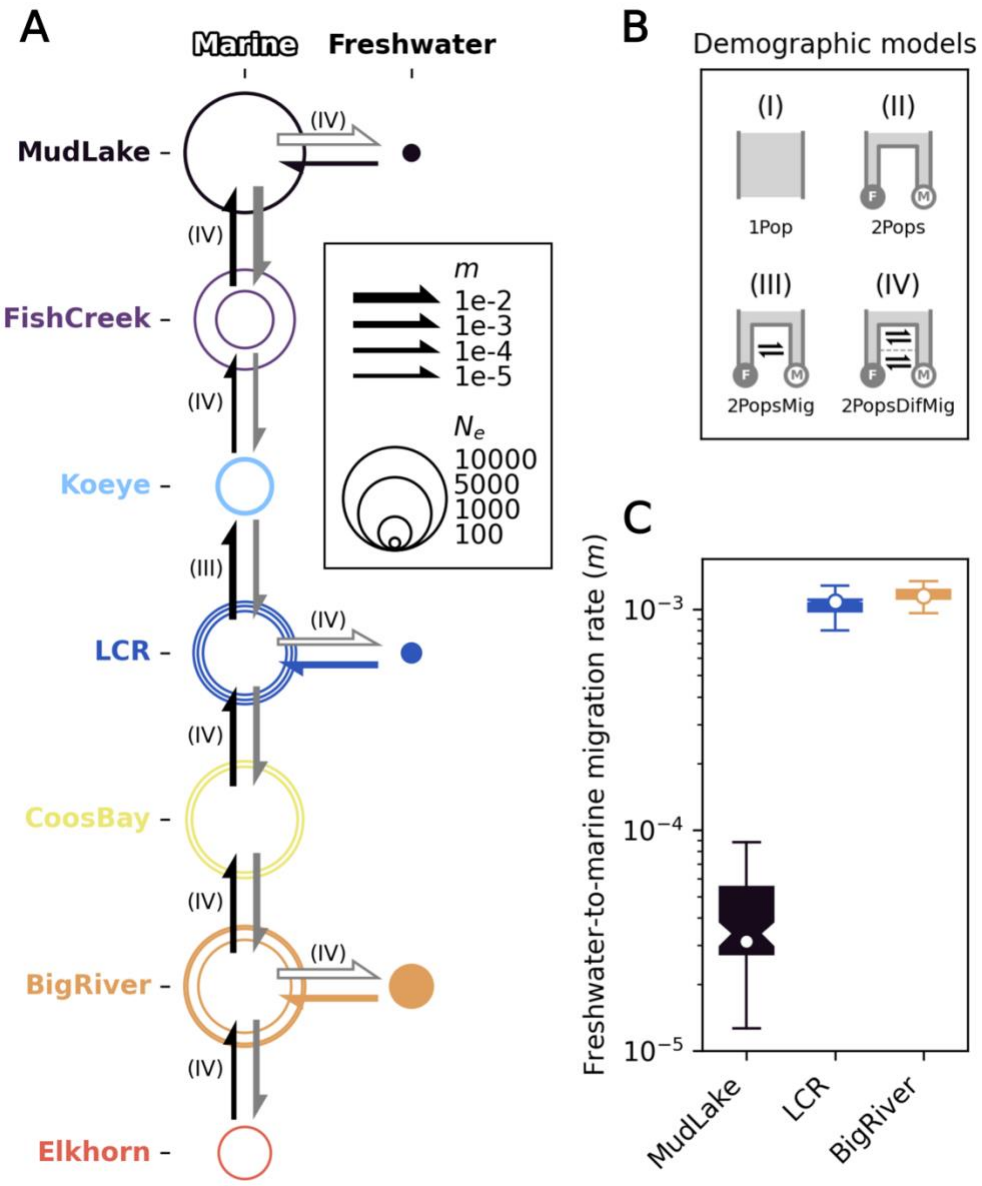

Figure 4 estimates of the degree of gene flow from site-frequency spectrum-based demographic modeling. (A) Diagram of the estimates of demographic parameters from the modeling results of 10 population pairs in a stepping-stone demographic network. For each pair, the estimates of most recent migration rates (thickness of the arrows) and effective population sizes (circle sizes) from the best demographic models (letters in parentheses) were shown. (B) The four demographic models tested include Model I: a single panmictic population, Model II: an ancestral population split into two subpopulations, Model III: Model II with continuous gene flow, Model IV: Model II with different gene flow rates at two different time periods. (C) The estimated most recent migration rates from freshwater to marine populations in three population pairs. Dots are the estimated values and the box plots indicate the bootstrap distributions.

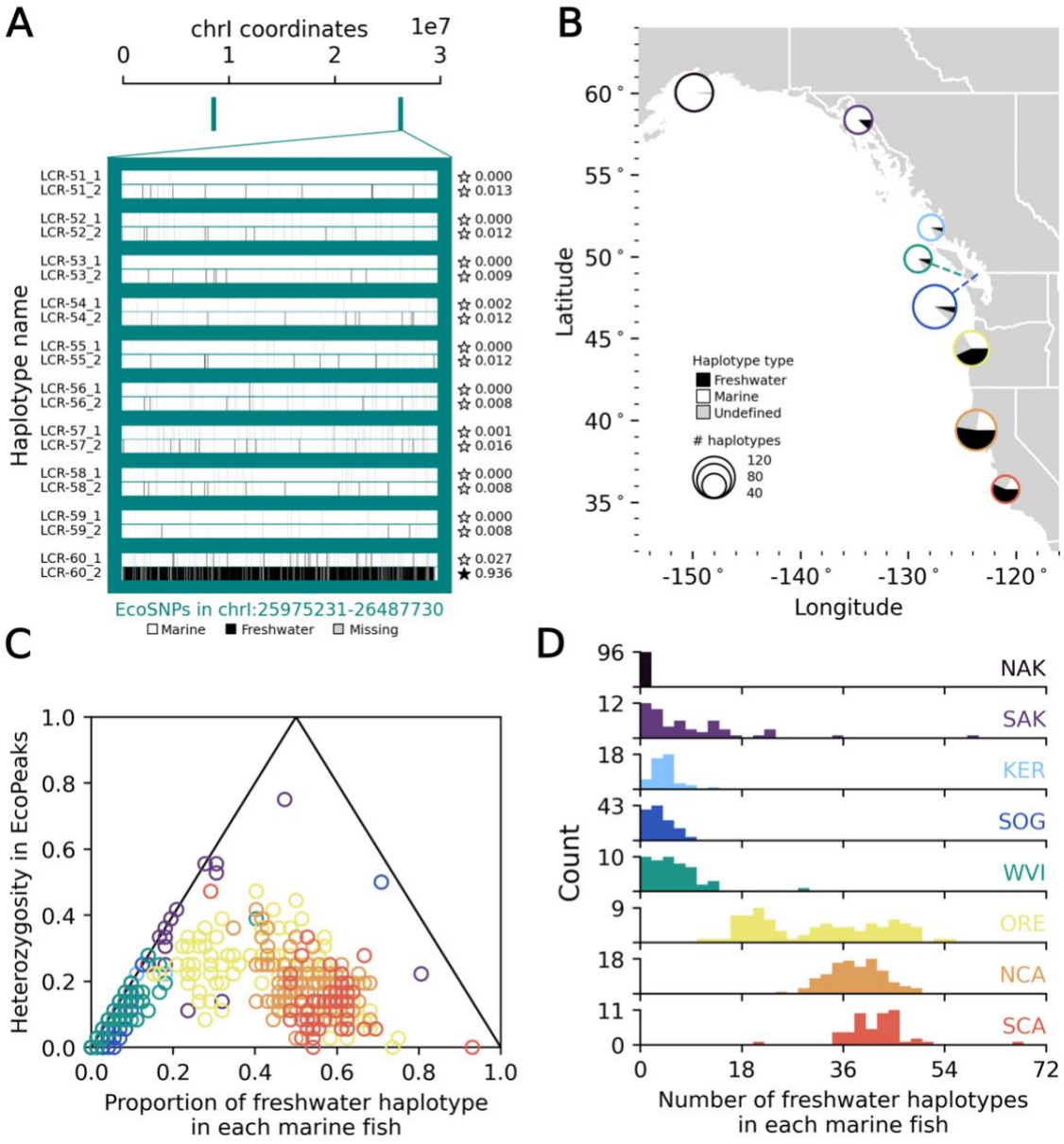

Figure 5 frequencies and hybrid indices of freshwater haplotypes in marine populations. (A) Upper panel shows all common EcoPeaks on chrI. Lower panel shows the types of alleles for all EcoSNPs on the second chrI EcoPeak (covering inversion region) on the haplotypes for a subset of marine individuals from Little Campbell River (LCR). The stars on the right denote the haplotype types defined by their proportions of freshwater alleles (shown in texts). (B) Frequencies of types of haplotypes in marine populations across all geographic divisions in northeast Pacific. (C) Heterozygosity and freshwater haplotype proportion for all common EcoPeaks in each marine fish. (D) Distribution of the numbers of freshwater haplotypes for each marine fish in each geographic division. Haida Gwaii (HDG) marine individuals were not shown due to small sample size.

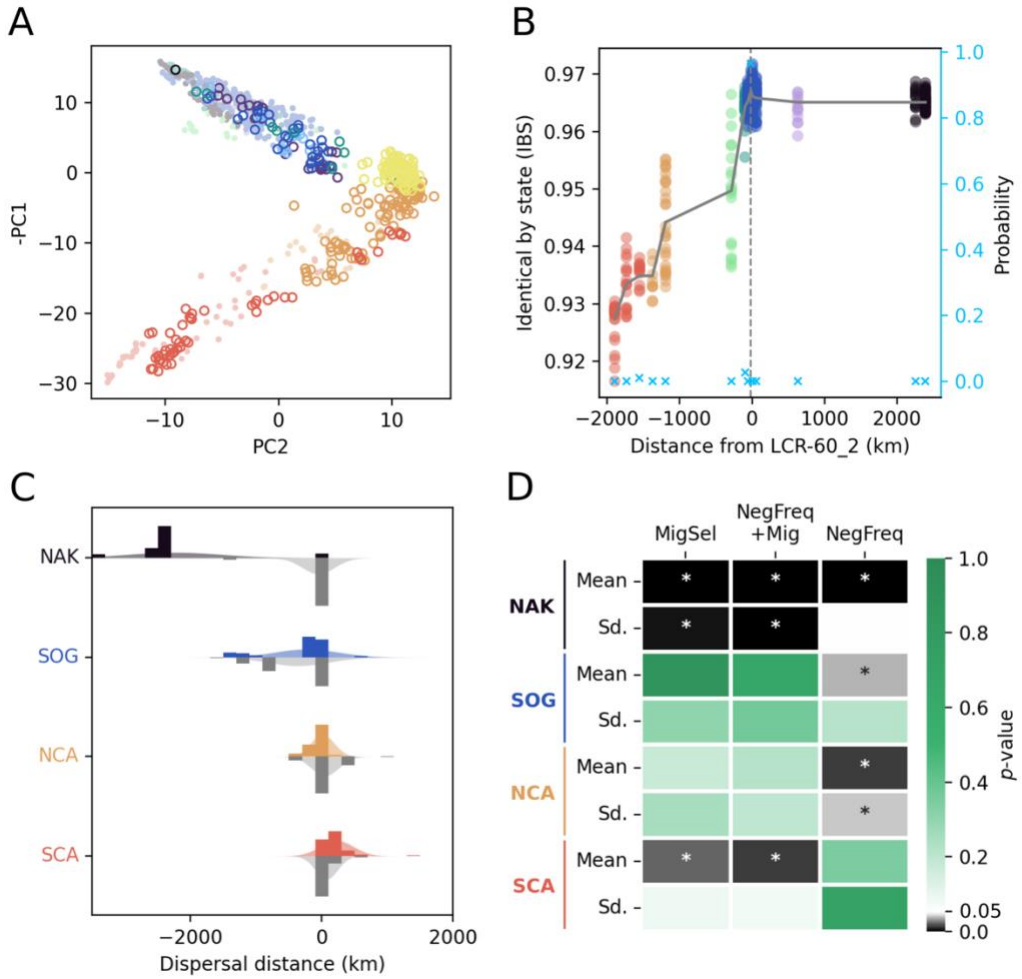

Figure 6 predicting the sources of freshwater haplotypes and their dispersal distances in the sea. (A) Principal component analysis (PCA) result of all freshwater haplotypes on a phased EcoPeak covering an inversion region on chrI. Freshwater haplotypes sampled in the sea (circles) generally group with that on the land (pale solid dots) in the same geographic regions (colored same as Figure 1). (B) The source location of a target freshwater haplotype in the sea (LCR-60_2 here, also see Figure 5A) was predicted by the peak of the fitted (grey) line of a unimodal monotone regression on the identity-by-state (IBS) scores comparing the sea haplotype with freshwater haplotypes on the land sorted from south (-) to north (+) by their coordinates across the coastline (dots). The prediction was then adjusted to probabilities across the groups using a leave-one-out cross validation (cyan crosses). (C) The migration kernels showing the distributions of dispersal distances of freshwater haplotypes in sea on seven EcoPeaks with high prediction accuracies sorted by their source geographic divisions. Colored and grey distributions indicate the observed distributions with fitted Gaussians and expected distributions under migration-selection balance in SLiM simulations, respectively. (D) The *p*-values of observed means and standard deviations of the migration kernels comparing to the expected kernels under pure migration-selection balance (MigSel) and negative frequency-dependent selection with and without freshwater-to-marine gene flows (NegFreq+Mig and NegFreq, respectively). Values ≤ 0.05 were marked with asterisks.

Literature cited

Auwera, G. A. V. der, & O'Connor, B. D. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, Inc.

Barrett, R. D. H., Paccard, A., Healy, T. M., Bergek, S., Schulte, P. M., Schluter, D., & Rogers, S. M. (2010). Rapid evolution of cold tolerance in stickleback. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1703), 233–238. https://doi.org/10.1098/rspb.2010.0923

Barrett, R. D. H., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology and Evolution*, *23*(1), 38–44. https://doi.org/10.1016/j.tree.2007.09.008

Baumgartner, J. V., & Bell, M. A. (1984). Lateral Plate Morph Variation in California Populations of the Threespine Stickleback, Gasterosteus Aculeatus. *Evolution*, *38*(3), 665–674. https://doi.org/10.1111/j.1558-5646.1984.tb00333.x

Bell, M. A. (1976). Evolution of Phenotypic Diversity in Gasterosteus Aculeatus Superspecies on the Pacific Coast of North America. *Systematic Biology*, *25*(3), 211–227. https://doi.org/10.2307/2412489

Bell, M. A., & Foster, S. A. (Eds.). (1994). *The evolutionary biology of the threespine stickleback*. Oxford University Press.

Bolnick, D. I., Caldera, E. J., & Matthews, B. (2008). Evidence for asymmetric migration load in a pair of ecologically divergent stickleback populations. *Biological Journal of the Linnean Society*, *94*(2), 273–287. https://doi.org/10.1111/j.1095-8312.2008.00978.x

Catchen, J., Bassham, S., Wilson, T., Currey, M., O'Brien, C., Yeates, Q., & Cresko, W. A. (2013). The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular Ecology*, *22*(11), 2864–2883. https://doi.org/10.1111/mec.12330

Chain, F. J. J., Feulner, P. G. D., Panchal, M., Eizaguirre, C., Samonte, I. E., Kalbe, M., Lenz, T. L., Stoll, M., Bornberg-Bauer, E., Milinski, M., & Reusch, T. B. H. (2014). Extensive Copy-Number Variation of Young Genes across Stickleback Populations. *PLOS Genetics*, *10*(12), e1004830. https://doi.org/10.1371/journal.pgen.1004830

Colosimo, P. F., Hosemann, K. E., Balabhadra, S., Jr, G. V., 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*(March), 1928–1933. https://doi.org/10.1126/science.1107239

Commission for Environmental Cooperation (CEC). (2022). *North American Atlas – Political Boundaries* (Version 3.0) [Dataset]. http://www.cec.org/north-american-environmental-atlas/political-boundaries-2021/

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Des Roches, S., Bell, M. A., & Palkovacs, E. P. (2020). Climate-driven habitat change causes evolution in Threespine Stickleback. *Global Change Biology*, *26*(2), 597–606. https://doi.org/10.1111/gcb.14892

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*(10). https://doi.org/10.1371/journal.pgen.1003905

Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). fastsimcoal2: Demographic inference under complex evolutionary scenarios. *Bioinformatics (Oxford, England)*, *37*(24), 4882–4885. https://doi.org/10.1093/bioinformatics/btab468

Galloway, J., Cresko, W. A., & Ralph, P. (2020). A Few Stickleback Suffice for the Transport of Alleles to New Lakes. *G3 Genes|Genomes|Genetics*, *10*(2), 505–514. https://doi.org/10.1534/g3.119.400564

Gibbons, T. C., Rudman, S. M., & Schulte, P. M. (2017). Low temperature and low salinity drive putatively adaptive growth differences in populations of threespine stickleback. *Scientific Reports*, *7*(1), 16766. https://doi.org/10.1038/s41598-017-16919-9

Glazer, A. M., Killingbeck, E. E., Mitros, T., Rokhsar, D. S., & Miller, C. T. (2015). Genome Assembly Improvement and Mapping Convergently Evolved Skeletal Traits in Sticklebacks with Genotyping-by-Sequencing. *G3 Genes|Genomes|Genetics*, *5*(7), 1463–1472. https://doi.org/10.1534/g3.115.017905

Guerrero, R. F., & Hahn, M. W. (2017). Speciation as a sieve for ancestral polymorphism. *Molecular Ecology*, *26*(20), 5362–5368. https://doi.org/10.1111/mec.14290

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*(10). https://doi.org/10.1371/journal.pgen.1000695

Haenel, Q., Guerard, L., MacColl, A. D. C., & Berner, D. (2022). The maintenance of standing genetic variation: Gene flow vs. selective neutrality in Atlantic stickleback fish. *Molecular Ecology*, *31*(3), 811–821. https://doi.org/10.1111/mec.16269

Hagen, D. W. (1967). Isolating Mechanisms in Threespine Sticklebacks (Gasterosteus). *Journal of the Fisheries Research Board of Canada*, *24*(8), 1637–1692. https://doi.org/10.1139/f67-138

Haller, B. C., & Messer, P. W. (2023). SLiM 4: Multispecies Eco-Evolutionary Modeling. *The American Naturalist*, *201*(5), E127–E139. https://doi.org/10.1086/723601

Han, F., Lamichhaney, S., Grant, B. R., Grant, P. R., Andersson, L., & Webster, M. T. (2017). Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Research*, *27*(6), 1004–1015. https://doi.org/10.1101/gr.212522.116

Hay, & McPhail. (2000). COURTSHIP BEHAVIOUR OF MALE THREESPINE STICKLEBACKS (GASTEROSTEUS ACULEATUS) FROM OLD AND NEW HYBRID ZONES. *Behaviour*, *137*(7–8), 1047–1063. https://doi.org/10.1163/156853900502420

Hedrick, P. W. (2013). Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, *22*(18), 4606–4618. https://doi.org/10.1111/mec.12415

Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S., & Delaneau, O. (2023). Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nature Genetics*, *55*(7), 1243–1249. https://doi.org/10.1038/s41588-023-01415-w

Jones, F. C., Brown, C., Pemberton, J. M., & Braithwaite, V. A. (2006). Reproductive isolation in a threespine stickleback hybrid zone. *Journal of Evolutionary Biology*, *19*(5), 1531–1544. https://doi.org/10.1111/j.1420-9101.2006.01122.x

Jones, F. C., Chan, Y. F., Schmutz, J., Grimwood, J., Brady, S. D., Southwick, A. M., Absher, D. M., Myers, R. M., Reimchen, T. E., Deagle, B. E., Schluter, D., & Kingsley, D. M. (2012). A Genome-wide SNP Genotyping Array Reveals Patterns of Global and Repeated Species-Pair Divergence in Sticklebacks. *Current Biology*, *22*(1), 83–90. https://doi.org/10.1016/j.cub.2011.11.045

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Karve, A. D., von Hippel, F. A., & Bell, M. A. (2008). Isolation between sympatric anadromous and resident threespine stickleback species in Mud Lake, Alaska. *Environmental Biology of Fishes*, *81*(3), 287–296. https://doi.org/10.1007/s10641-007-9200-2

Kirch, M. (2025). *The Role of Standing Genetic Variation in Rapid Adaptation-Insights from Ancient and Large-scale Contemporary Threespine Stickleback Genomes*. Universität Tübingen.

Lai, Y.-T., Yeung, C. K. L., Omland, K. E., Pang, E.-L., Hao, Y., Liao, B.-Y., Cao, H.-F., Zhang, B.-W., Yeh, C.-F., Hung, C.-M., Hung, H.-Y., Yang, M.-Y., Liang, W., Hsu, Y.-C., Yao, C.-T., Dong, L., Lin, K., & Li, S.-H. (2019). Standing genetic variation as the predominant source for adaptation of a songbird. *Proceedings of the National Academy of Sciences*, *116*(6), 2152–2157. https://doi.org/10.1073/pnas.1813597116

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (arXiv:1303.3997). arXiv. https://doi.org/10.48550/arXiv.1303.3997

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Marques, D. A., Jones, F. C., Di Palma, F., Kingsley, D. M., & Reimchen, T. E. (2018). Experimental evidence for rapid genomic adaptation to a new niche in an adaptive radiation. *Nature Ecology & Evolution*, *2*(7), Article 7. https://doi.org/10.1038/s41559-018-0581-8

Miller, C. T., Beleza, S., Pollen, A. A., Schluter, D., Kittles, R. A., Shriver, M. D., & Kingsley, D. M. (2007). Cis-Regulatory Changes in Kit Ligand Expression and Parallel Evolution of Pigmentation in Sticklebacks and Humans. *Cell*, *131*(6), 1179–1189. https://doi.org/10.1016/j.cell.2007.10.055

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 Evolutionary Biology*, *18*(1), 113. https://doi.org/10.1186/s12862-018-1228-8

Peichel, C. L., McCann, S. R., Ross, J. A., Naftaly, A. F. S., Urton, J. R., Cech, J. N., Grimwood, J., Schmutz, J., Myers, R. M., Kingsley, D. M., & White, M. A. (2020). Assembly of the threespine stickleback Y chromosome reveals convergent signatures of sex chromosome evolution. *Genome Biology*, *21*(1), 177. https://doi.org/10.1186/s13059-020-02097-x

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Auwera, G. A. V. der, Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2018). *Scaling accurate genetic variant discovery to tens of thousands of samples* (p. 201178). bioRxiv. https://doi.org/10.1101/201178

Purcell, S., & Chang, C. (n.d.). *PLINK 2.0* [Computer software]. www.cog-genomics.org/plink/2.0/

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Roberts Kingman, G. A., Vyas, D. N., Jones, F. C., Brady, S. D., Chen, H. I., Reid, K., Milhaven, M., Bertino, T. S., Aguirre, W. E., Heins, D. C., Hippel, F. A. von, Park, P. J., Kirch, M., Absher, D. M., Myers, R. M., Palma, F. D., Bell, M. A., Kingsley, D. M., & Veeramah, K. R. (2021). Predicting future from past: The genomic basis of recurrent and rapid stickleback evolution. *Science Advances*, *7*(25), eabg5285. https://doi.org/10.1126/sciadv.abg5285

Roesti, M., Kueng, B., Moser, D., & Berner, D. (2015). The genomics of ecological vicariance in threespine stickleback fish. *Nature Communications*, *6*(1), 8767. https://doi.org/10.1038/ncomms9767

Schluter, D., & Conte, G. L. (2009). Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the United States of America*, *106 Suppl*(Supplement 1), 9955–9962. https://doi.org/10.1073/pnas.0901264106

Smith, M. L., & Hahn, M. W. (2024). Selection leads to false inferences of introgression using popular methods. *Genetics*, *227*(4), iyae089. https://doi.org/10.1093/genetics/iyae089

Turba, R., Richmond, J. Q., Fitz-Gibbon, S., Morselli, M., Fisher, R. N., Swift, C. C., Ruiz-Campos, G., Backlin, A. R., Dellith, C., & Jacobs, D. K. (2022). Genetic structure and historic demography of endangered unarmoured threespine stickleback at southern latitudes signals a potential new management approach. *Molecular Ecology*, *31*(24), 6515–6530. https://doi.org/10.1111/mec.16722

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, *43*(1), 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

White, M. A., Kitano, J., & Peichel, C. L. (2015). Purifying Selection Maintains Dosage-Sensitive Genes during Degeneration of the Threespine Stickleback Y Chromosome. *Molecular Biology and Evolution*, *32*(8), 1981–1995. https://doi.org/10.1093/molbev/msv078

Yeaman, S. (2015). Local Adaptation by Alleles of Small Effect. *The American Naturalist*, *186*(S1), S74–S89. https://doi.org/10.1086/682405

Yoshida, K., Makino, T., Yamaguchi, K., Shigenobu, S., Hasebe, M., Kawata, M., Kume, M., Mori, S., Peichel, C. L., Toyoda, A., Fujiyama, A., & Kitano, J. (2014). Sex Chromosome Turnover Contributes to Genomic Divergence between Incipient Stickleback Species. *PLOS Genetics*, *10*(3), e1004223. https://doi.org/10.1371/journal.pgen.1004223

Supplementary materials for

On the maintenance of standing genetic variation by migration-selection balance

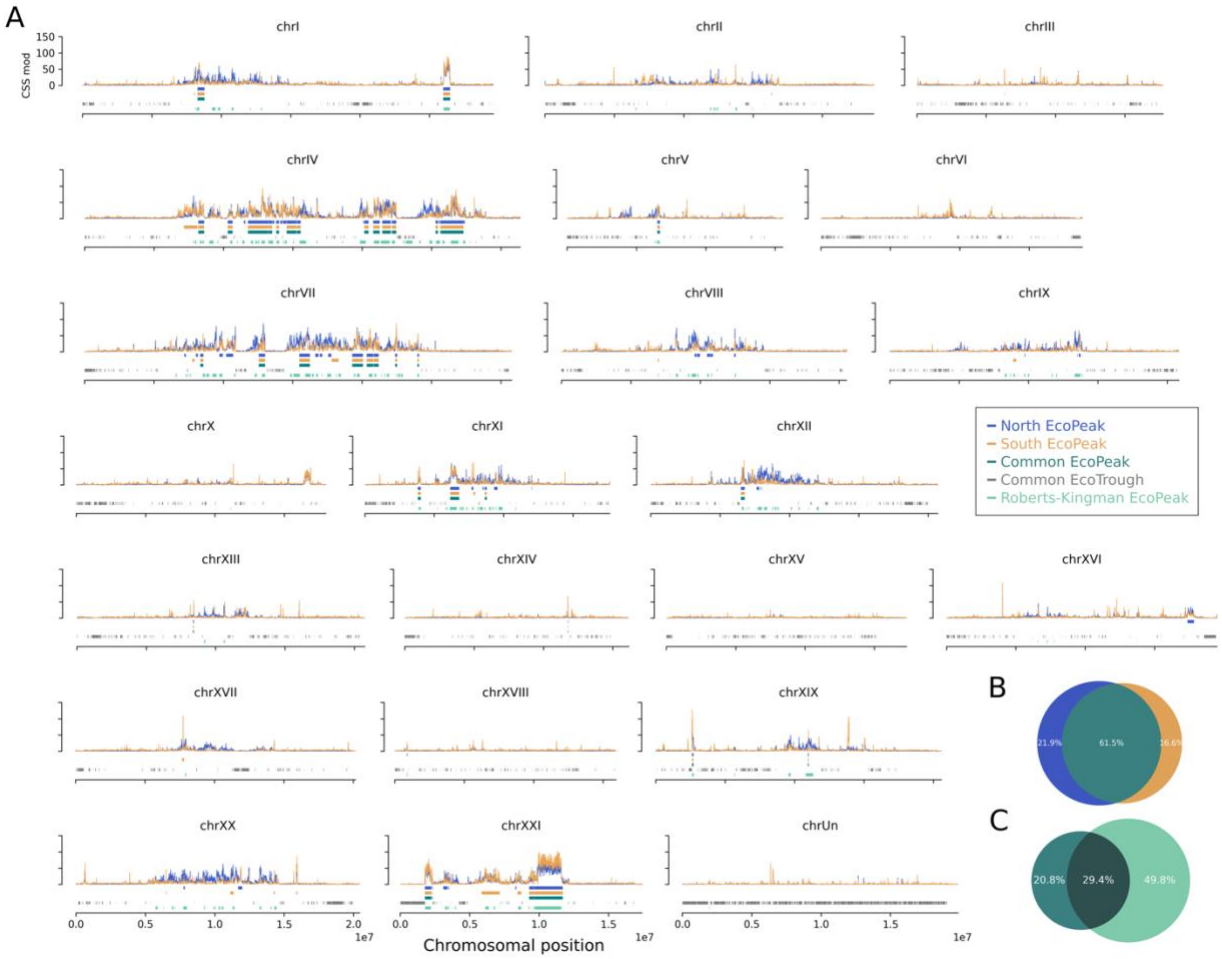

Figure S1 all EcoPeaks and EcoTroughs defined in this study. (A) CSS and locations of EcoPeaks and EcoTroughs defined using the north and south sets, along with the comparison with the pacific specific EcoPeaks from Roberts Kingman et al. (2021). (B) Venn diagram showing the percentage of base pairs covered in north and south EcoPeaks and their overlapping regions (common EcoPeaks). (C) Venn diagram showing the percentage of base pairs covered in the common and Roberts Kingman North East pacific specific EcoPeaks and their overlapping regions.

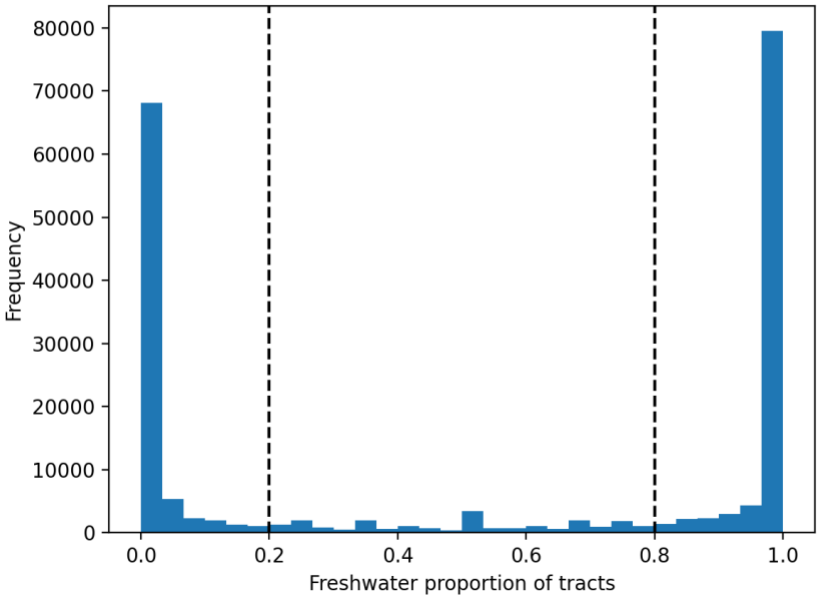

Figure S2 proportion of freshwater alleles in each phased haplotype. Haplotypes with frequencies higher than 0.8 and lower than 0.2 were defined as freshwater and marine haplotypes, respectively.

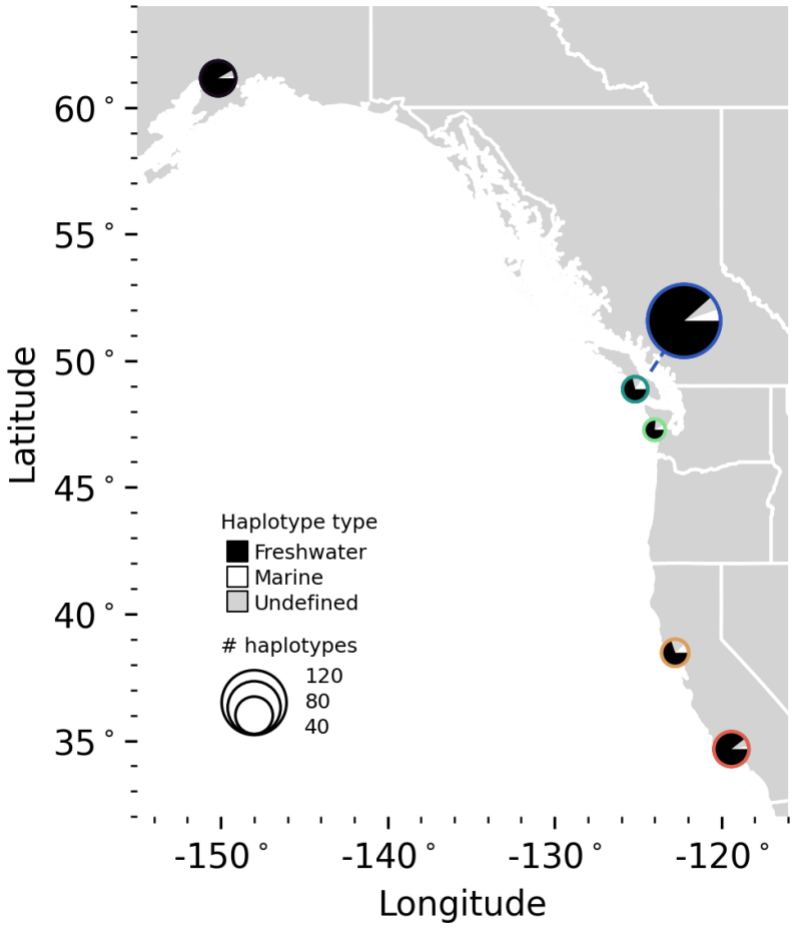

Figure S3 Frequencies of types of haplotypes in freshwater populations in northeast pacific coast.

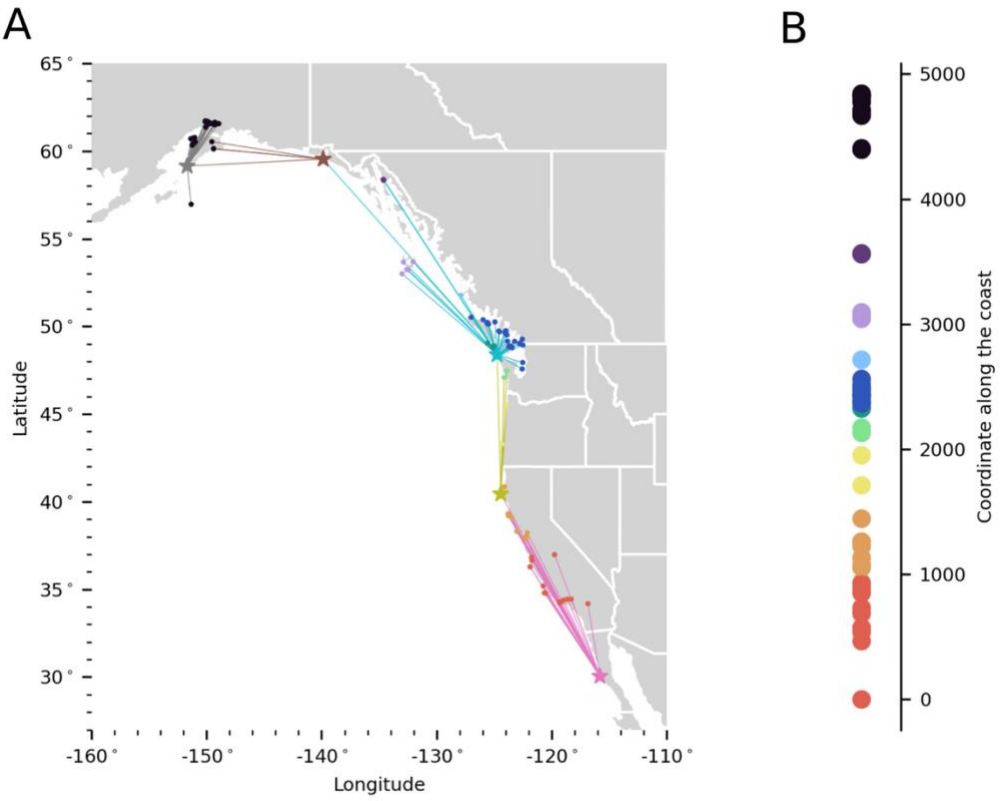

Figure S4 geographic coordinates along the northeast pacific coast. (A) Sampling locations of each population (dots with color scheme same as Figure 2) and their connection lines through the four waypoints (stars) to the most southern base point El Rosario (pink star). Notice that the lines may not represent the true geodesic distances. (B) The transformed 1-dimensional coordinates for each population along the coast.

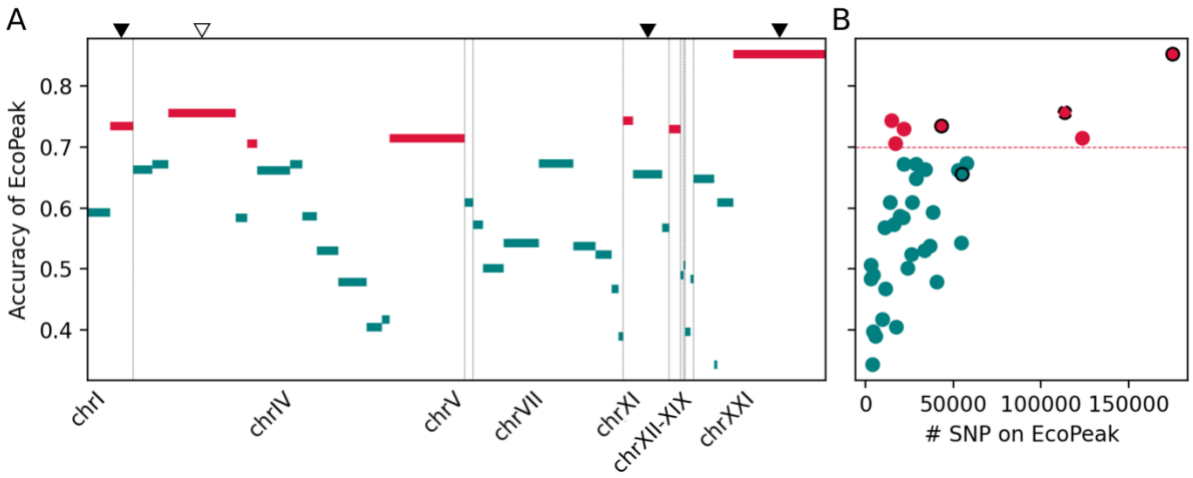

Figure S5 accuracy of the confusion matrix on each common EcoPeak. Confusion matrix accuracies for all common EcoPeaks were compared with their (A) lengths (base pairs) and (B) number of SNPs carried. EcoPeaks with >= 0.7 accuracies were painted with red color. EcoPeaks covering known inversions and *Eda* gene were indicated by solid and dashed black edges, respectively.

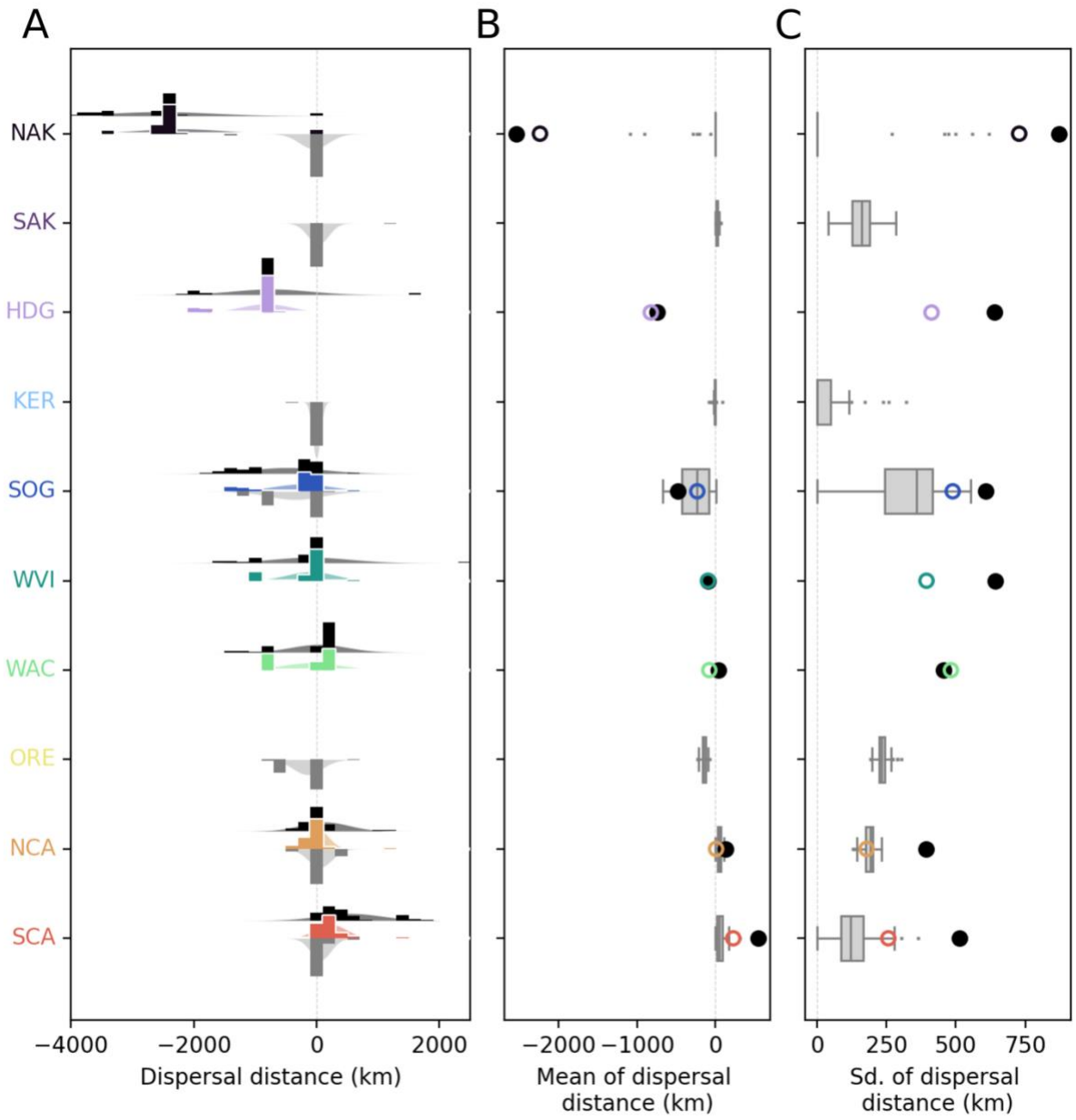

Figure S6 observed and expected migration kernels under migration-selection balance. (A) Migration kernels were constructed using common EcoPeaks with >= 0.7 accuracies (colored), all common EcoPeaks (black in background) and simulations under pure migration-selection balance at generation 4000 (grey). Their means and standard deviations were compared in (B) and (C).

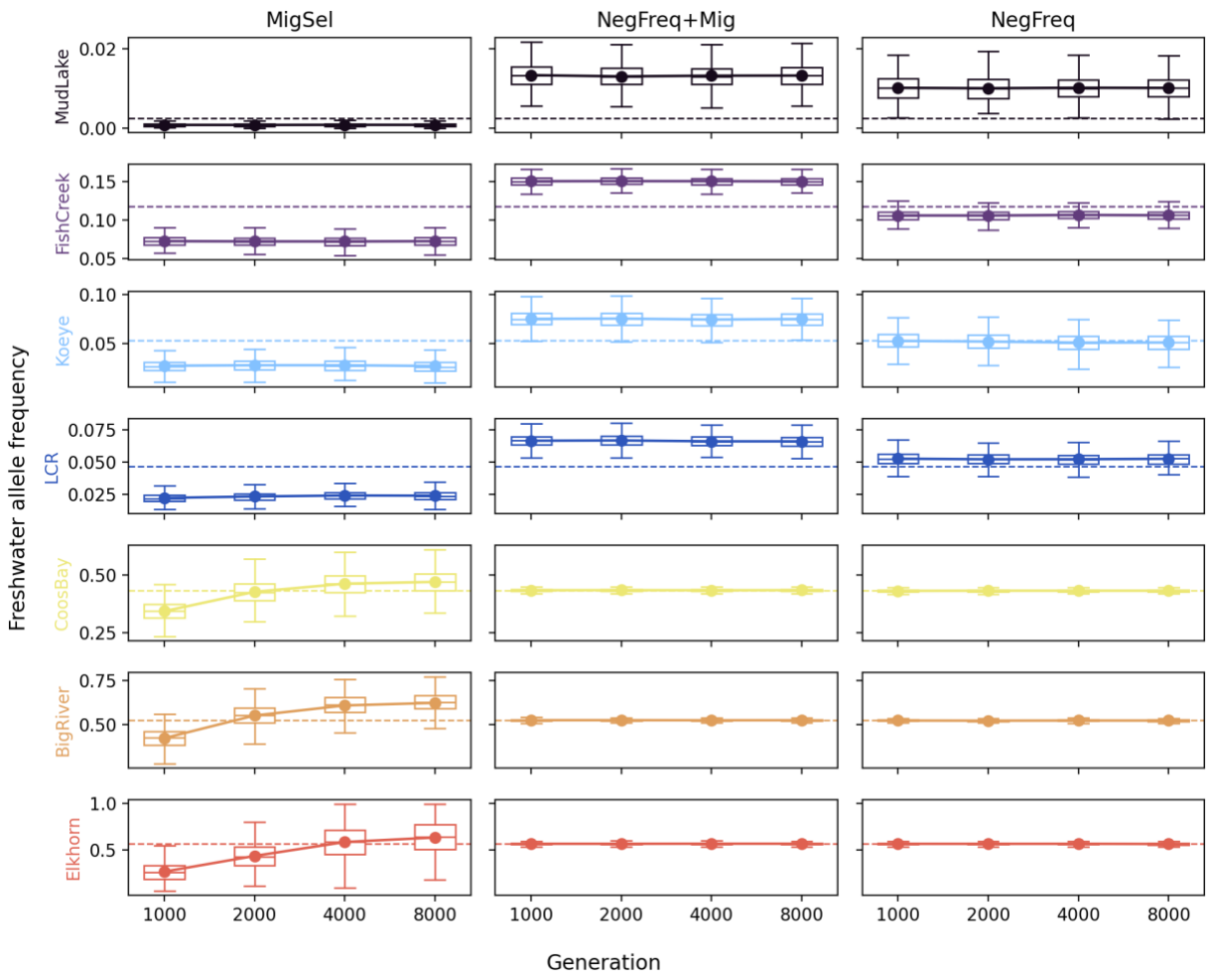

Figure S7 changes of freshwater alleles in simulated marine populations under migration-selection balance and negative frequency-dependent selection with and without freshwater-to-marine gene flow. Dashed lines and box plots represent the empirical measured and simulated frequencies, respectively.

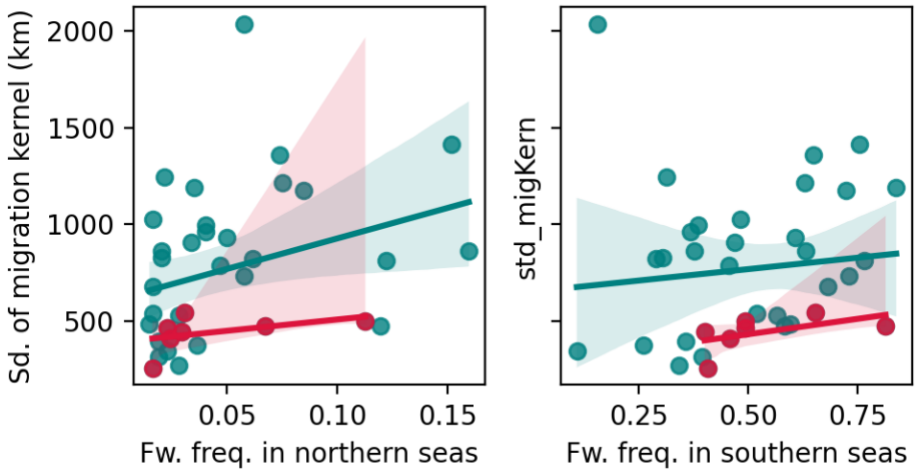

Figure S8 linear regression between the standard deviations of observed migration kernels and the frequencies of freshwater haplotypes in northern and southern seas. Red colors indicate the common EcoPeaks with accuracy ≥ 0.7 while teal colors indicate other common EcoPeaks.

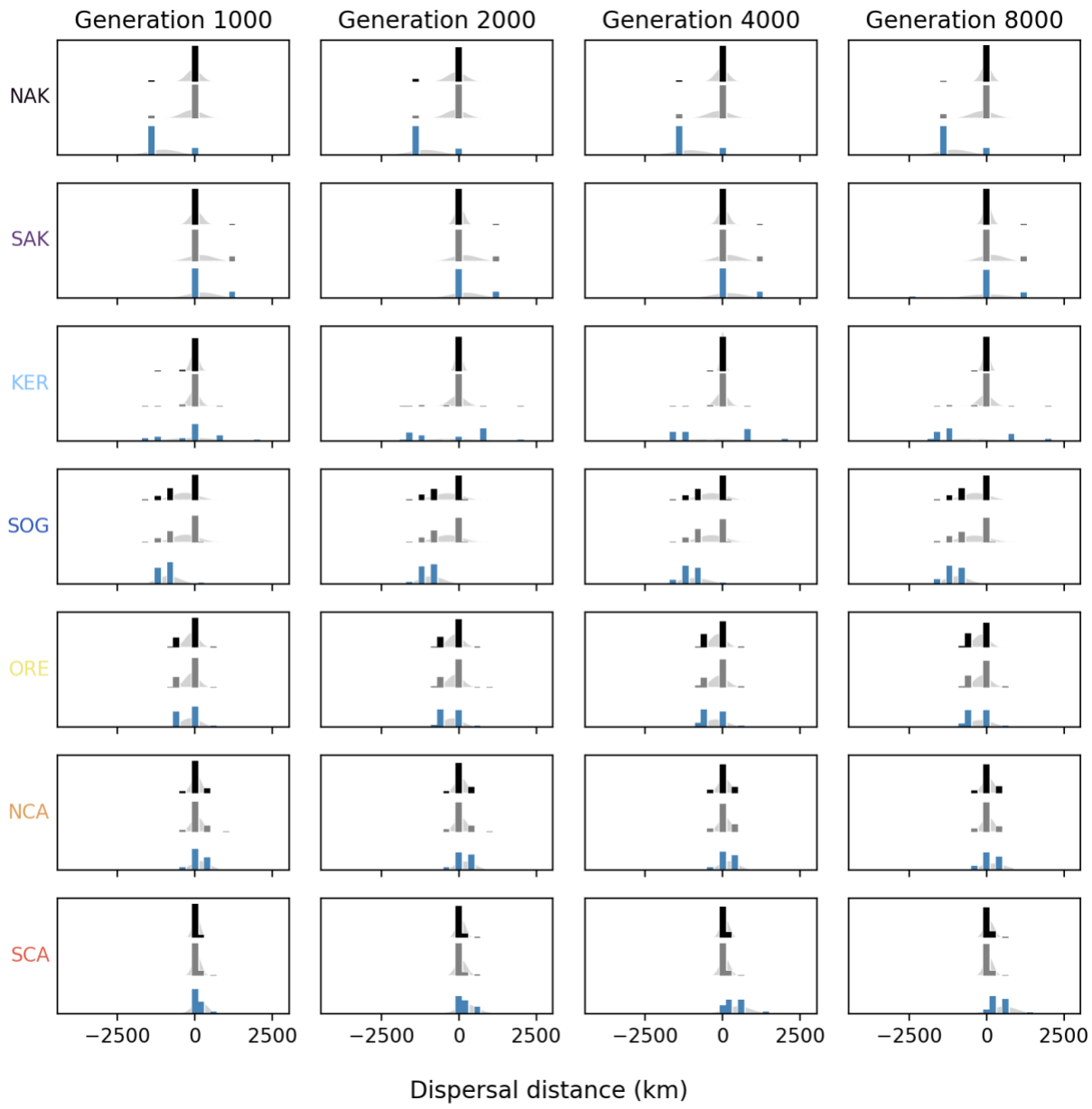

Figure S9 simulated migration kernels with different hypotheses maintaining SGV. Migration kernels were simulated based on migration-selection balance (black), negative frequency-dependent selection with (grey) and without freshwater-to-marine gene flow (blue). The fitted normal distributions are colored with light grey in the background.

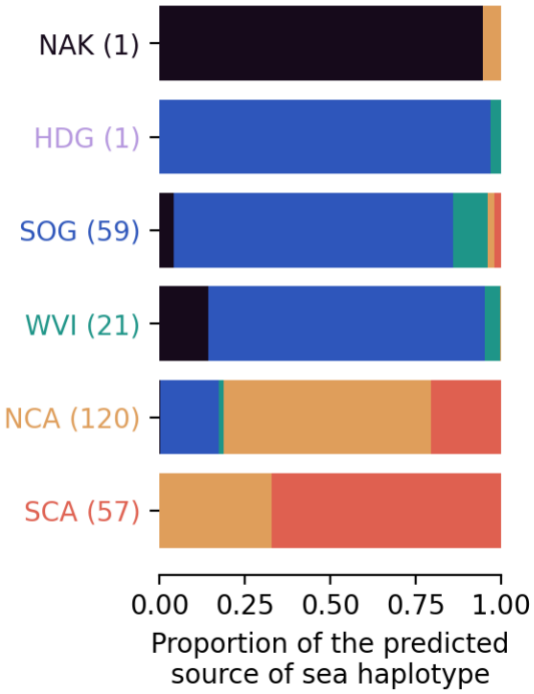

Figure S10 proportion of the geographic division of predicted source for freshwater haplotypes in the sea. Color scheme same as Figure 2. Numbers of haplotypes with ≥ 0.1 source probabilities in each geographic division are labeled in the parentheses. Only EcoPeaks with ≥ 0.7 are included.