

# Genome-wide Analysis Reveals High Genomic Diversity and Panmixia in Bay Pipefish (*Syngnathus leptorhynchus*) from Coos Bay, Oregon

Mark C. Currey<sup>1</sup> , Micah A. Woods<sup>1</sup> , Vithika Goyal<sup>1</sup>, Susan Bassham<sup>1</sup>, Hope M. Healey<sup>1,2</sup>, and William A. Cresko<sup>1,3</sup> 

**Running Head:** High genomic diversity and panmixia in bay pipefish from Coos Bay, Oregon

<sup>1</sup>Institute of Ecology and Evolution, University of Oregon, 5289 University of Oregon, Eugene, Oregon 97403-5289, USA; ORCID: (M.C.C.) 0000-0003-3688-7208; (M.A.W.) 0009-0004-2156-3352; (V.G.) 0009-0009-8601-8474; (S.B.) 0000-0002-7309-2095; (H.M.H.) 0000-0001-9978-1553; and (W.A.C.) 0000-0002-3496-8074; Email: (M.C.C.) [mcurrey@uoregon.edu](mailto:mcurrey@uoregon.edu); (M.A.W.) [woods.micah.a@gmail.com](mailto:woods.micah.a@gmail.com); (V.G.) [vithgoy@gmail.com](mailto:vithgoy@gmail.com); (S.B.) [sbassham@uoregon.edu](mailto:sbassham@uoregon.edu); (H.M.H.) [hhealey@uoregon.edu](mailto:hhealey@uoregon.edu); and (W.A.C.) [cresko@uoregon.edu](mailto:cresko@uoregon.edu).

<sup>2</sup>Knight Campus Graduate Internship Program, University of Oregon, Eugene, Oregon 97403, USA.

<sup>3</sup>Department of Bioengineering, Knight Campus, University of Oregon, Eugene, Oregon 97403, USA.

 Present address: Department of History and Philosophy of Science, University of Cambridge, Cambridge, United Kingdom.

 Send correspondence to this address (M.C.C.).

 Send correspondence to this address (W.A.C.).

# ABSTRACT

The over 300 species of fishes in the charismatic family Syngnathidae (seahorses, pipefish, pipehorses, and seadragons) exhibit remarkable morphological and reproductive adaptations. Most syngnathid species inhabit shallow coastal regions where they are vulnerable to human activities, including habitat degradation and collection for use in traditional medicine and the aquarium trade. A gap remains in our understanding of within-species genetic diversity and population structure for most of the species in this family—information that is crucial for establishing baselines to detect future human-driven changes. Here, we investigate fine-scale population structure and genomic diversity in the bay pipefish (*Syngnathus leptorhynchus*), one of the relatively few temperate lineages within the Syngnathidae. Bay pipefish are habitat specialists that are closely associated with eelgrass (*Zostera marina*) along the Pacific coast of North America. Due to the patchy and changing nature of eelgrass habitat, bay pipefish provide an instructive case for using population genomics to explore patterns of connectivity in a dynamic coastal environment. We sampled 116 individuals from two locations separated by approximately 17 km in the Coos Bay estuary (Oregon, USA) and used restriction site-associated DNA sequencing (RAD-seq) to discover high and consistent genetic diversity at both sampling locations, with no indication of inbreeding. Genome-wide patterns of genetic variation showed minimal differentiation between the sampling sites, with levels of divergence indistinguishable from those expected under random sampling. Linkage disequilibrium was low and declined rapidly, and estimates of effective population size indicate a large, well-connected population within Coos Bay. These results establish a crucial genomic baseline for understanding how bay pipefish populations may respond to ongoing changes in their eelgrass habitat and offer a genomic framework for conservation and comparative studies across the Syngnathidae.

# INTRODUCTION

Understanding how habitat specialists maintain genetic connectivity across fragmented or dynamic habitats is critical for predicting their responses to environmental change, particularly in marine systems where connectivity patterns remain poorly understood for many taxa (Cowen and Sponaugle, 2009; Lowe and Allendorf, 2010). The family Syngnathidae, which includes over 300 species of pipefish, seahorses, pipehorses, and seadragons, is characterized by specialized morphology and strong habitat associations (Dawson, 1985; Mobley et al., 2011). Most syngnathid species are adapted to shallow coastal and estuarine environments that are increasingly impacted by outcomes from human activities, including habitat degradation, pollution, and climate-driven shifts in temperature and salinity (Foster and Vincent, 2004). Many of these species also face direct human pressures through harvesting for traditional medicine, the aquarium trade, and scientific research (Vincent et al., 2011).

Despite their ecological, scientific, and cultural importance, within-species patterns of genetic diversity, connectivity, and population structure remain poorly characterized for much of the Syngnathidae (Mobley et al., 2011; Vincent et al., 2011). These knowledge gaps limit our ability to assess population resilience to environmental change and exploitation. Establishing robust baseline measurements of genomic diversity and spatial genetic structure is therefore critical for detecting future shifts in population status and for informing effective conservation and management strategies for syngnathid and other vulnerable fish taxa. Furthermore, the small number of population genetic studies in syngnathids have revealed variable findings. Specialized habitat dependence and limited dispersal have been linked to fine-scale genetic structure in many syngnathids (Teske et al., 2003; Lourie et al., 2005; Stiller et al., 2017; Bertola et al., 2020), though patterns vary considerably. Some species show strong population structure at small spatial scales (Partridge et al., 2012), while others exhibit surprisingly broad genetic connectivity

despite being distributed over large geographic distances (Mobley et al., 2010; Braga Goncalves et al., 2017; Woodall et al., 2018). These contrasting patterns suggest that connectivity in habitat-specialist syngnathids depends on complex interactions between life history traits, habitat configuration, and species-specific dispersal behavior.

The bay pipefish (*Syngnathus leptorhynchus*) provides an exemplary syngnathid system for examining population genetic patterns. Despite having one of the broadest distributions among *Syngnathus* species—from Bahía Santa María, Baja California to Prince William Sound, Alaska (Fritzsche, 1980; Orsi et al., 1991)—bay pipefish are strongly associated with eelgrass (*Zostera marina*) throughout their life cycle (Adams, 1976; Fritzsche, 1980). Previous studies of bay pipefish have examined population structure at different spatial scales, with findings that are complementary but leave some uncertainty about the factors governing gene flow. Broad-scale phylogeographic work along the west coast of North America, spanning hundreds to thousands of kilometers, documented clear genetic breaks among distant coastal regions (Louie, 2003; Wilson, 2006) whereas fine-scale analyses in British Columbia's Barkley Sound detected measurable divergence ( $F_{ST} = 0 - 0.017$ ) among populations separated by a few to tens of kilometers, with seascape complexity restricting gene flow and creating distinct genetic clusters (de Graaf, 2006). These studies leave open the question of whether fine-scale structure could develop in topographically simple, continuous estuaries where physical barriers to dispersal are largely absent. As eelgrass habitat specialists with limited swimming ability, bay pipefish could exhibit restricted dispersal even in continuous habitats, yet the dynamic and patchy nature of eelgrass meadows could alternatively promote high connectivity through repeated colonization and mixing.

Previous bay pipefish studies relied on mitochondrial or microsatellite markers, limiting inferences about genome-wide diversity and genetic connectivity. Here, we use restriction site-associated DNA sequencing (RAD-seq), enabled by the recent assembly of a reference genome, to examine fine-scale population structure and genetic diversity genome-wide in bay pipefish within the Coos Bay estuary, Oregon—a continuous estuarine system where topographic barriers are minimal. We sampled individuals from two locations approximately 17 km apart and characterized genetic variation in thousands of single nucleotide polymorphisms (SNPs) to assess patterns of population structure, genetic diversity, kinship, and effective population size. We assessed whether bay pipefish exhibit (1) fine-scale genetic structure reflecting limited dispersal and habitat specialization, as observed in more topographically complex regions, or (2) panmixia reflecting the continuous nature of an estuarine environment that has potentially high connectivity within eelgrass meadows.

This study establishes a baseline understanding of the genetic diversity and structure of bay pipefish in Coos Bay and offers a point of reference for future monitoring and conservation efforts, as well as a useful comparison for other estuarine systems along the west coast. More broadly, this work contributes to a growing understanding of how spatial scale, habitat continuity, and life history traits shape genetic structure in syngnathid fishes and informs future conservation and sampling strategies in estuarine environments.

## MATERIALS AND METHODS

### *Pipefish collections*

Bay pipefish were collected by beach seine in the summer of 2023 from two sites in the Coos Bay estuary: “Causeway” and Valino Island (Fig. 1 and Table 1). At each site, beach seines

were pulled through areas containing eelgrass beds. We collected 50 adult fish from the Causeway location and 66 from Valino Island. Fish were placed in cooled, aerated seawater for transport until euthanized with tricaine methane sulfonate (MS-222) and immediately preserved in 95% ethanol. GPS coordinates of sampling locations were obtained from Google Earth (Table 1). All research involving vertebrate animals was approved by the University of Oregon Institutional Animal Care and Use Committees under protocol number AUP-20-23. Fish were collected under Oregon Department of Fish and Wildlife scientific taking permit number 26987.

#### ***DNA extraction and quality control***

Genomic DNA was purified from fin and other body tissues using a column-based DNeasy Blood & Tissue Kit (Qiagen) or by tissue digestion in Qiagen ATL buffer with proteinase K followed by DNA isolation using Omega Tech paramagnetic beads which were employed for high-throughput processing of samples. To avoid potential sampling bias, samples from each location were processed using both extraction methods. All samples underwent an RNaseA treatment. Purified DNA was quantified using a Qubit fluorometer and dsDNA Broad Range Assay Kit. A subset of samples was assessed for DNA purity using a NanoDrop spectrophotometer, and for integrity *via* agarose gel electrophoresis. Both assessments indicated high-quality DNA with minimal contamination. DNA was then normalized in concentration to ensure equal representation of each individual in the RAD-seq library.

#### ***RAD library construction and genotyping***

RAD-seq libraries were constructed for 116 fish using the restriction endonuclease *SbfI*-HF (NEB), following established protocols (Catchen et al., 2013a). Sequencing was performed on an

Illumina NovaSeq 6000 platform with a sp100 cycle, generating 135-nucleotide reads, including inline 7 nt barcodes. Raw reads were de-multiplexed and filtered by quality score using default settings in the process\_radtags module from the Stacks v2.68 pipeline (Catchen et al., 2013b). Briefly, the reads were first checked for intact barcodes and RAD cut sites and then demultiplexed. Next, a sliding window approach (default window size of 15% of read length) was used to assess average quality scores, discarding reads where the score dropped below a Phred score of 10 (corresponding to 90% probability of base call accuracy). Sequencing and quality filtering yielded approximately 500 million initial reads with a little over 373 million sequences retained after passing quality filters (Supplemental Table 1). These high-quality reads were then aligned to the *Syngnathus leptorhynchus* genome (GenBank assembly accession: GCA\_XXXXXXXXX, to be assigned upon release) using GSNAP version 2024-05-07 (Wu and Watanabe, 2005), allowing for up to five nucleotide mismatches and gap lengths of two nucleotides. Only reads with unique alignments were retained.

Reads were then processed through the Stacks pipeline (modules: ref\_map.pl and populations) to construct a catalog and to call genotypes. Individual samples were filtered using VCFtools (Danecek et al., 2011) based on the following conservative, quality-focused criteria. To ensure that only high-confidence genotypes were retained, samples could not have a missing data rate >12%, exhibit excess heterozygosity or high inbreeding ( $F < -0.10$  or  $F > 0.25$ ), or show a combination of >10% missing data and moderate heterozygosity deviations ( $F < -0.05$  or  $F > 0.20$ ). This combined filtering removed 36 individuals in total—14 from the Causeway site and 22 from the Valino Island site—leaving 36 individuals from Causeway and 44 from Valino Island for downstream analyses.

A whitelist (i.e., a list of locus IDs to retain in downstream analyses) of loci including reads from only the first 22, which are likely the chromosomes in this species, was then used in a second populations run. Separate sampling locations were retained in the data structure to enable population structure analyses. For their inclusion, we required that loci be present in at least 80% of individuals within each population ( $-r\ 0.8$ ), they be present in at least 80% of all individuals across the study ( $-R\ 0.8$ ), have a minimum minor allele frequency of 0.05 ( $--min-maf\ 0.05$ ), and, using the  $--write-single-snp$  flag, only one SNP per RAD read was retained. This resulted in 6,074 variant sites (Supplemental Table 2). A third populations run was conducted using individuals from both sampling locations combined. The same filtering criteria were applied, except the  $-r$  flag was omitted because we collapsed the populations into one. This resulted in 6,310 variant sites (Supplemental Table 2). These two whitelisted sets represent average genome-wide marker densities of  $\sim 20$  and  $\sim 21$  markers per megabase, respectively.

### ***Population genetic analysis of diversity and structure***

The populations program within the Stacks framework was used to calculate per-locus and genome-wide averages of nucleotide diversity ( $\pi$ ), heterozygosity,  $F_{IS}$ , and  $F_{ST}$ .

To assess the significance of genome-wide genetic differentiation, we applied two complementary statistical approaches that test distinct null hypotheses. First, permutation testing evaluated whether the observed genome-wide  $F_{ST}$  exceeded expectations under panmixia (no population structure). Individuals were randomly reassigned to the two sampling locations while preserving the original sample sizes, and genome-wide  $F_{ST}$  was recalculated for each permutation. This procedure was repeated 100 times to generate a null distribution of  $F_{ST}$  values. Statistical significance was calculated using the formula  $p = (r + 1) / (n + 1)$ , where  $r$  is the

number of permuted  $F_{ST}$  values greater than or equal to the observed value, and  $n$  is the number of permutations.

Second, bootstrap resampling assessed whether the observed genome-wide  $F_{ST}$  could be explained by sampling variance among loci. From the quality-filtered dataset of 6,074 SNPs, we generated 1,000 bootstrap replicates by randomly sampling 1,000 loci with replacement. For each replicate, genome-wide  $F_{ST}$  was recalculated using the same parameters as the original analysis in Stacks. The resulting distribution was summarized by its mean, median, standard deviation, and 95% confidence interval. The percentile rank of the observed  $F_{ST}$  within this distribution was used to assess whether the observed value fell outside the range expected from sampling variance alone.

Output files from the Stacks populations program were generated under two grouping scenarios. First, we analyzed the two sampling sites separately to test for population structure, calculate pairwise divergence ( $F_{ST}$ ), and assess kinship patterns. Second, because preliminary analyses indicated panmixia (see Results), we combined both sites into a single population for analyses that assume random mating: genome-wide linkage disequilibrium (LD) patterns and effective population size ( $N_e$ ) estimation.

To visualize and investigate the major axes of genetic variation, a principal component analysis (PCA) was performed using Genodive v3.04 software (Meirmans, 2020) employing a covariance method on individual genotype data from the two collecting locations (Causeway and Valino Island). The analysis utilized the populations.structure file generated by Stacks (Catchen et al., 2013b), which contained all 6,074 markers, to calculate genetic principal components to assess population differentiation. Principal component plots were generated in R v4.4.0 (R Core

Team, 2024) using ggplot2 v3.5.2 (Wickham, 2016). The first two components were visualized with their variance explained, and a density plot of PCA scores was also produced using ggplot2.

Population genetic structure was investigated using STRUCTURE v2.3.4 (Pritchard et al., 2000), and the same populations.structure file as in the PCA analysis, to assess the number of distinct genetic clusters among individuals from the two collecting locations. For each value of  $K$  (number of genetic groupings), five independent runs were performed using 100,000 burn-in steps followed by 100,000 Markov Chain Monte Carlo replicates. The range of  $K$  values tested was 1 to 2 to visualize genetic clustering. Results from multiple runs were processed and visualized using CLUMPAK (Kopelman et al., 2015) to assign individuals to genetic clusters.

Genome-wide patterns of differentiation were assessed using pairwise  $F_{ST}$  estimates calculated by the populations module of Stacks for all retained SNPs between the two collecting locations. To visualize genomic patterns of differentiation,  $F_{ST}$  values were plotted against genomic position using both individual SNP point estimates and smoothed trend lines to identify genomic regions of elevated or reduced differentiation. All visualizations were created using R.

### ***Kinship Estimation***

To assess relatedness among individuals, pairwise kinship coefficients were estimated using the Method of Moments (MoM) in the R package SNPRelate v1.38.1 (Zheng et al., 2012). Prior to analysis, VCFs from Stacks were filtered to retain only biallelic SNPs with  $\leq 25\%$  missing data. Linkage disequilibrium (LD) pruning was applied using a 10 kb sliding window and an LD threshold of  $r^2 < 0.2$  to ensure marker independence. Kinship was calculated for all individual pairs (630 within Causeway, 946 within Valino Island, and 1,584 between locations; total = 3,160 pairs). Results were formatted as symmetric matrices and pairwise tables for

downstream analysis. Kinship values were interpreted using standard thresholds (e.g.,  $\geq 0.45$ : identical;  $\geq 0.22$ : parent-offspring/full siblings;  $\geq 0.11$ : half-siblings/grandparent-grandchild;  $\geq 0.055$ : first cousins;  $\geq 0.027$ : second cousins;  $> 0$ : distant relatives;  $\leq 0$ : unrelated). Heatmaps and summary plots were used to visualize relatedness patterns within and between sites using the pheatmap v1.0.13 package in R (Kolde, 2025).

### ***Genome-wide Linkage Disequilibrium and LD Decay Analysis***

Genome-wide linkage disequilibrium (LD) was calculated using PLINK2.0 (Chang et al., 2015) with the flags --r2-phased, cols=+dprimeabs, --inter-chr, and --ld-window-r2 0 to generate both  $r^2$  and  $D'$  values for all SNP pairs. To visualize LD, values were aggregated into 1 Mb genomic bins using maximum aggregation for  $r^2$  and mean aggregation for  $D'$ . Symmetric heatmaps displaying 287 bins (covering 91% of the genome) across 22 chromosomes were generated using ggplot2.

LD decay was examined using two complementary approaches: a fine-scale analysis ( $\leq 500$  kb with 10 kb bins) to capture local recombination patterns, and a broad-scale analysis ( $\leq 5$  Mb with 100 kb bins) to reveal longer range trends. For each distance bin, mean LD values were calculated, excluding bins containing fewer than five SNP pairs and plotted along with raw pairwise LD values (as transparent points) and binned means (as solid lines and points), illustrating the underlying data distribution alongside smoothed trends. Plots were created with R and ggplot2.

The correlation between  $r^2$  and  $D'$  was calculated across all intra-chromosomal SNP pairs within 5 Mb using Pearson correlation. Background LD levels were defined as the mean  $r^2$  and  $D'$  values for SNP pairs separated by  $> 1$  Mb, representing the baseline level of residual linkage

disequilibrium. Decay to background levels was determined using fine-scale 2 kb bins up to 200 kb distance, identifying the first distance bin where LD values dropped to within 10% of the background threshold. Distance-controlled relationships were assessed using linear regression models: (1)  $r^2 \sim D'$ , (2)  $r^2 \sim D' + \log(\text{distance})$ , and (3)  $r^2 \sim D' \times \log(\text{distance})$ , with model improvements evaluated using ANOVA. Partial correlation between  $r^2$  and  $D'$  was calculated while controlling for log-transformed physical distance using the ppcor v1.1 (Kim, 2015) package in R.

Potential regions of elevated LD on chromosome 1 (Chr 1) were identified using sliding 1 Mb windows. For each window containing  $\geq 10$  SNP pairs, mean  $r^2$  and  $D'$  values were calculated and compared to genome-wide thresholds defined as the mean plus two standard deviations of values from all other chromosome groups. Windows exceeding either threshold were considered elevated, and contiguous elevated windows defined region boundaries. LD magnitude comparisons between the identified elevated region and genome-wide averages (excluding all Chr 1 data) were expressed as fold-changes and absolute differences to quantify the degree of elevation.

All analyses were restricted to intra-chromosomal SNP pairs to avoid confounding effects from inter-chromosomal associations, and statistical significance was assessed at  $\alpha = 0.05$ . Analyses were conducted using data.table v1.17.8, dplyr v2.5.1, and ggplot2 v3.5.2 (Wickham et al., 2023; Barrett et al., 2025).

### ***Contemporary Effective Population size ( $N_e$ )***

Contemporary effective population size ( $N_e$ ) was estimated using two complementary approaches to provide cross-validation and increase confidence in the results. These methods

capture different aspects of population genetic variation: the allele frequency variance method estimates  $N_e$  from the distribution of allele frequencies across loci, while the linkage disequilibrium (LD) method relies on non-random associations between alleles at different loci. Because the two methods are based on independent principles and assumptions—one focusing on allele frequency variance and the other on inter-locus associations—they provide robust and complementary insights into contemporary  $N_e$ .

To ensure independence among markers, linkage disequilibrium (LD) pruning was performed using PLINK v2.0 with the `--indep-pairwise 50 10 0.2` algorithm, which examines SNPs in sliding windows of 50 markers, shifts by 10 SNPs per iteration, and removes one SNP from each pair exhibiting LD with  $r^2 > 0.2$ . This retained 5,676 independent SNPs from the original dataset for all  $N_e$  analyses.

$N_e$  was first estimated using the allele frequency variance method described by Nei and Tajima (Nei and Tajima, 1981), which calculates effective population size based on the variance in allele frequencies across loci. Specifically,  $N_e$  is estimated using the formula  $N_e = n / (2 \times \text{Var}(p))$ , where  $n$  is the sample size, and  $\text{Var}(p)$  is the variance in minor allele frequencies. To quantify uncertainty, 1,000 bootstrap replicates were performed by resampling loci with replacement from the pruned SNP dataset, and 95% and 90% confidence intervals were calculated from the resulting distribution of  $N_e$  values. Analysis was implemented in R using custom scripts with the `data.table` package (Barrett et al., 2025).

A second estimate of  $N_e$  was obtained using the linkage disequilibrium method implemented in NeEstimator v2 (Do et al., 2014), which infers effective population size based on correlations between alleles at different loci. To evaluate sensitivity to rare alleles,  $N_e$  was estimated using multiple critical allele frequency thresholds ( $P_{\text{crit}} = 0.05, 0.02, \text{ and } 0.01$ ), with

Pcrit = 0.02 selected as the primary estimate following standard recommendations for balanced precision and bias. Genotype data were converted from PLINK to GenePop format using custom R scripts, and parametric 95% confidence intervals were calculated from the standard errors reported by NeEstimator.

## RESULTS

### *Notable genetic variation within and among collecting locations*

Of the 6,074 variant sites identified in Stacks, none carried alleles unique to the Causeway collection. Only one variant site had an allele private to the Valino Island collection, which likely reflects sampling variation rather than a fixed presence/absence polymorphism at that locus. Genome-wide averages for the Causeway site were as follows: expected heterozygosity ( $H_e$ ) 0.293, observed heterozygosity ( $H_o$ ) 0.257, nucleotide diversity ( $\pi$ ) 0.298, and inbreeding coefficient ( $F_{IS}$ ) 0.130. For the Valino Island site, the corresponding values were  $H_e$  0.297,  $H_o$  0.259,  $\pi$  0.301, and  $F_{IS}$  0.136. When both locations were analyzed together, genome-wide averages were  $H_e$  0.300,  $H_o$  0.256,  $\pi$  0.302, and  $F_{IS}$  0.149 (Supplemental Table 2).

### *Little to no observed population structure between collecting locations*

PCA of 80 individuals revealed that the first two principal components captured 2.0% and 1.6% of the total genetic variance (Fig. 2, Supplemental Figure 1). Individuals from the two sampling locations largely overlapped in PCA space, with no clear clustering by site, indicating minimal genetic differentiation across the 17 km separating the habitats. STRUCTURE analysis revealed a similar pattern, with each individual showing nearly equal assignment probabilities to two hypothesized populations ( $K = 2$ ), regardless of their collection site (Fig. 3). Despite using

thousands of genomic markers—which provide substantially greater resolution than previous analyses using relatively fewer microsatellite markers—no spatial genetic structure was detected. The lack of clear assignment of individuals to distinct genetic clusters further supports the conclusion that the individuals collected from the two locations represent a single, well-mixed population rather than genetically differentiated groups.

Genome-wide genetic divergence between the two collection sites was extremely low, with  $F_{ST} = 0.0074$ . In addition, there are no discernible regions of population separation along the genome (Fig. 04). To distinguish genuine population structure from sampling artifacts, we applied complementary permutation testing and bootstrap resampling approaches that examine different sources of variance in population genetic estimates.

Permutation testing evaluated whether the observed differentiation exceeded that expected under random population assignment. Across 100 permutation replicates, 99 permuted  $F_{ST}$  values met or exceeded the observed value, yielding a p-value of 0.990 and indicating that the observed differentiation could readily arise by chance (Supplemental Figure 2). Bootstrap resampling assessed whether differentiation could result from random sampling of loci. Using 1,000 replicates, each of which sampled 1,000 loci from the 6,074 SNP dataset, the bootstrap distribution had a mean  $F_{ST}$  of  $0.0074 \pm 0.0005$  (95% CI: 0.0065–0.0083). The observed  $F_{ST}$  (0.0074) fell at the 49.3rd percentile, differing from the bootstrap mean by only 0.000008, demonstrating that observed differentiation lies well within expectations under random sampling of loci (Supplemental Figure 3). Together, these results indicate that the apparent genetic differentiation reflects sampling variance rather than systematic population structure.

These combined metrics provide robust evidence of panmixia between the two collecting locations. Further analysis treated the locations as one population.

# ***Large, Well-Mixed Population Exhibiting Low Kinship, Rapid LD Decay, and High Genetic Diversity***

Kinship analysis indicated that most individuals were unrelated within and between sampling locations, with only occasional distant relationships (second-cousin level, kinship coefficient  $\approx 0.027$ ) (Supplemental Figure 4). At Valino Island, most coefficients were zero, with a few up to  $\sim 0.028$ . Causeway showed a similar pattern, though slightly more frequent, reaching  $\sim 0.036$ . Between locations, most pairs were unrelated, with a small subset exceeding the 0.027 threshold, indicating limited but detectable familial connectivity.

Genome-wide linkage disequilibrium (LD) analysis showed generally low levels of LD. Analysis of the relationship between LD metrics (both  $r^2$  and  $D'$ ) revealed a moderate overall correlation ( $r = 0.308$ ,  $p < 0.001$ ) across all intra-chromosomal SNP pairs within 5 Mb (Supplemental Figure 5 A). LD decayed rapidly outside of this interval, reaching background levels within 10.9 kb for  $r^2$  and 7.0 kb for  $D'$  (Supplemental Figure 5 B&C), indicating highly effective recombination across the bay pipefish genome.

A notable exception to this genome-wide pattern was a region on the distal end of chromosome 1 spanning the region from approximately 4 to 7 Mb (Fig. 5, A&B). This region exhibited substantially elevated linkage disequilibrium compared to genome-wide averages, with mean  $r^2$  values  $\sim 6$  times higher (0.075 vs 0.012) and mean  $D'$  values 1.78 times higher (0.550 vs 0.309) across 1,176 SNP pairs, consistent with reduced effective recombination in this 3 Mb segment. The elevated LD could reflect a misassembled region potentially due to repetitive sequence, a structural polymorphism, or could indicate the presence of a sex-determining region

(McKinney et al., 2020). Further investigation is needed to distinguish between these possibilities.

Estimates of contemporary effective population size ( $N_e$ ) varied across two complementary methods. The allele frequency variance method produced the most precise estimate:  $N_e = 2,265$  (95% CI: 2,207 - 2,329), based on 1,000 bootstrap replicates of 5,676 LD-pruned SNPs. The linkage disequilibrium method yielded a substantially higher estimate,  $N_e = 36,282$  (95% CI: 11,230 -  $\infty$ ), using 2,785 polymorphic markers and the recommended  $P_{crit}$  threshold of 0.02.

These discrepancies reflect differences in methodological assumptions, and the temporal scales each approach captures. The variance method is sensitive to recent breeding dynamics, while the elevated LD-based estimate may reflect historical demographic events such as population expansion or ongoing gene flow from a larger metapopulation. Despite differences in magnitude, both approaches suggest that bay pipefish maintain an effective population size in the thousands, reinforcing the conclusion of a genetically diverse and panmictic population.

## DISCUSSION

While the number of high-quality reference genome assemblies available for syngnathids is growing (Ramesh et al., 2023; Scott-Somme et al., 2023; Wolf et al., 2024), population genomic assessments are lagging and are still needed to establish baselines for understanding connectivity and resilience in these vulnerable species. This is particularly true in estuarine systems like Coos Bay that are experiencing ongoing and increasing human influence. Given the limited population genomic information available for most syngnathid species, including bay

pipefish, our study helps fill a key gap in understanding baseline patterns of genetic diversity and connectivity in this group.

In service of this goal, we produced the first genomic-scale assessment of genetic diversity, population structure, kinship, and effective population size in bay pipefish from Coos Bay, Oregon. Across multiple analyses, we observed patterns consistent with a large, well-mixed, and genetically cohesive population. Specifically, our results revealed high genetic diversity, minimal inbreeding, little to no population structure between sites, very low levels of relatedness among individuals, and low genome-wide linkage disequilibrium—together pointing to effective dispersal and gene flow within this estuarine system.

Population structure and kinship analyses collectively indicate that bay pipefish within Coos Bay are a single, well-mixed population. Individuals from the two sites showed no genetic clustering in multivariate analyses, and the very low genome-wide  $F_{ST}$  (0.0074) was statistically indistinguishable from random variation, suggesting that gene flow effectively prevents differentiation across the sites. Kinship analysis reinforced this conclusion, revealing that most individuals were unrelated, with no parent-offspring or full-sibling pairs detected and only occasional distant relationships observed. This pattern implies effective dispersal mechanisms or high population densities that limit the co-occurrence of close relatives. Large effective population size estimates further supported this interpretation: while the two methods of allele frequency variance and linkage disequilibrium differed in magnitude ( $N_e \sim 2,265$  vs.  $\sim 36,000$ ), both indicate that thousands of individuals contribute to reproduction, consistent with a demographically robust population capable of maintaining genetic connectivity across continuous estuarine habitat.

Bay pipefish have also shown variable genetic patterns across their range, depending on spatial scale and habitat context. Mitochondrial studies along the Pacific coast have suggested high gene flow among Oregon and California populations, though limited sample sizes constrained fine-scale resolution (Louie, 2003). A broader study identified significant genetic divergence between populations from Alaska, Washington, Oregon, and California, indicating that population structure emerges at larger spatial scales (Wilson, 2006). At finer scales, a study in Barkley Sound, British Columbia—where sites were separated by 9–83 km—found a mix of strong genetic divergence and panmixia among localities, shaped by habitat continuity and physical barriers like fjords (de Graaf, 2006). That study estimated a genetic neighborhood size of 40–60 km for bay pipefish, within which gene flow maintains genetic cohesion.

The two sites from which we sampled bay pipefish are separated by approximately 17 km of mostly continuous eelgrass habitat (Merkel & Associates, 2023), lacking significant physical barriers other than an industrial shipping channel, and our genetic estimates lie well within the genetic neighborhood size previously estimated for the species (de Graaf, 2006). These findings suggest that gene flow is ongoing at this spatial scale and that bay pipefish disperse effectively within continuous estuarine habitats, aligning with broader trends observed across the Syngnathidae and supporting the hypothesis proposed by Fritzsche (1980) that bay pipefish populations are genetically homogeneous within bays but may be differentiated across larger spatial scales. While our study is limited to two nearby locations within a single estuary, the patterns of panmixia, high diversity, and minimal kinship structure are consistent with this hypothesis. Future population genomic studies from other bays near Coos Bay can extend tests of Fritzsche’s predictions.

Previous population genetic studies in syngnathids have revealed a diversity of patterns, with some congeners of bay pipefish (the dusky pipefish, *S. floridae* and the gulf pipefish, *S. scovelli*), maintaining high genetic diversity and fine-scale population structuring (Mobley et al., 2010; Flanagan et al., 2016) in a pattern of site fidelity and low dispersal, while the broadnosed pipefish, *S. typhle*, and the straight-nosed pipefish, *N. ophidion*, both associated with eelgrass habitats, show evidence of adult movement across eelgrass patches and weak site fidelity consistent with greater gene flow potential (Vincent et al., 1995). These contrasting patterns have been linked to differences in dispersal ability, larval duration, habitat fragmentation, and species-specific behavior—all of which can influence the spatial scale of gene flow and connectivity. In bay pipefish, panmixia might be aided by its large adult body size relative to many syngnathids, its large reproductive brood sizes, and the robust continuity of eelgrass habitat in estuaries like Coos Bay.

To more fully resolve the spatial genetic dynamics of bay pipefish, future studies should extend sampling to additional sites throughout Coos Bay and across other Pacific Coast estuaries. Broader spatial and temporal sampling, combined with focused studies of site fidelity and larval dispersal, will be essential for understanding the mechanisms that generate and maintain connectivity in this species. Our study helps establish a baseline for understanding genetic variation in bay pipefish in Coos Bay and provides a useful reference for future monitoring and comparisons across West Coast estuaries. More broadly, it contributes to our understanding of how spatial scale, habitat continuity, and life-history traits shape genetic structure in syngnathid fishes and offers guidance for future conservation and sampling efforts in estuarine systems.

## ACKNOWLEDGMENTS

Fish were collected under Oregon Department of Fish and Wildlife scientific taking permit number 26987. All animal research was conducted under University of Oregon Institutional Animal Care and Use Committee protocol AUP-20-23. Sequencing was performed at the Genomics and Cell Characterization Core Facility at the University of Oregon. This research was supported by the National Science Foundation grant IOS-2015301 to W.A.C. and by research funding from the University of Oregon.

**Author Contributions:** M.C.C., M.A.W., V.G., and W.A.C. contributed to conceptualization and methodology. M.C.C., M.A.W., V.G., and W.A.C. conducted field investigations and specimen collection. M.C.C., M.A.W., V.G., and S.B. performed DNA extractions. M.C.C. constructed RAD-seq libraries, conducted formal analysis, created visualizations, and curated data. M.C.C. and M.A.W. wrote the original draft. All authors contributed to review and editing. W.A.C. acquired funding, provided resources, and supervised the project.

## DATA ACCESSIBILITY

Raw RAD-seq data generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA851781. The *Syngnathus leptorhynchus* reference genome has been submitted to NCBI GenBank (accession GCA\_XXXXXXXXXX, to be assigned upon release). Custom R scripts used for data analysis and figure generation are available on GitHub at <https://github.com/wcresko/bay-pipefish-popgen>.

## REFERENCES

- Adams, S. M.** 1976. The ecology of eelgrass, *Zostera marina* (L.), fish communities. I. Structural analysis. *Journal of Experimental Marine Biology and Ecology* 22:273–291.
- Barrett, T., M. Dowle, A. Srinivasan, J. Gorecki, M. Chirico, T. Hocking, B. Schwendinger, and I. Krylov.** 2025. data.table: Extension of data.frame. <https://CRAN.R-project.org/package=data.table>
- Bertola, L. D., J. T. Boehm, N. F. Putman, A. T. Xue, J. D. Robinson, S. Harris, C. C. Baldwin, I. Overcast, and M. J. Hickerson.** 2020. Asymmetrical gene flow in five co-distributed syngnathids explained by ocean currents and rafting propensity. *Proceedings of the Royal Society B: Biological Sciences* 287:20200657.
- Braga Goncalves, I., L. Cornetti, A. S. Couperus, C. J. G. Van Damme, and K. B. Mobley.** 2017. Phylogeography of the snake pipefish, *Entelurus aequoreus* (Family: Syngnathidae) in the northeastern Atlantic Ocean. *Biological Journal of the Linnean Society* 122:787–800.
- Catchen, J., S. Bassham, T. Wilson, M. Currey, C. O'Brien, Q. Yeates, and W. A. Cresko.** 2013a. The population structure and recent colonization history of Oregon threespine stickleback

499 determined using restriction-site associated DNA-sequencing. *Molecular Ecology* 22:2864–  
500 2883.

501 **Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko.** 2013b. Stacks: An  
502 analysis tool set for population genomics. *Molecular Ecology* 22:3124–3140.

503 **Chang, C. C., C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee.**  
504 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets.  
505 *GigaScience* 4:7.

506 **Cowen, R. K., and S. Sponaugle.** 2009. Larval dispersal and marine population connectivity.  
507 *Annual Review of Marine Science* 1:443–466.

508 **Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E.**  
509 **Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin.** 2011. The  
510 variant call format and VCFtools. *Bioinformatics* 27:2156–2158.

511 **Dawson, C. E.** 1985. Indo-Pacific pipefishes (Red Sea to the Americas). Gulf Coast Research  
512 Laboratory, Ocean Springs, Mississippi.

513 **de Graaf, R. C.** 2006. Fine-scale population genetic structure of the eastern Pacific bay pipefish,  
514 *Syngnathus leptorhynchus*. Unpubl. M.S. thesis, University of British Columbia, Vancouver,  
515 British Columbia, Canada.

516 **Do, C., R. S. Waples, D. Peel, G. M. Macbeth, B. J. Tillett, and J. R. Ovenden.** 2014.  
517 NeEstimator v2: Re-implementation of software for the estimation of contemporary effective  
518 population size ( $N_e$ ) from genetic data. *Molecular Ecology Resources* 14:209–214.

519 **Flanagan, S. P., E. Rose, and A. G. Jones.** 2016. Population genomics reveals multiple drivers  
520 of population differentiation in a sex-role-reversed pipefish. *Molecular Ecology* 25:5043–5072.

521 **Foster, S. J., and A. C. J. Vincent.** 2004. Life history and ecology of seahorses: Implications  
522 for conservation and management. *Journal of Fish Biology* 65:1–61.

523 **Fritzsche, R. A.** 1980. Revision of the Eastern Pacific Syngnathidae (Pisces: Syngnathiformes),  
524 including both recent and fossil forms. *Proceedings of the California Academy of Sciences*  
525 42:181–227.

526 **Kim, S.** 2015. ppcor: Partial and semi-partial (part) correlation. R package. [https://CRAN.R-](https://CRAN.R-project.org/package=ppcor)  
527 [project.org/package=ppcor](https://CRAN.R-project.org/package=ppcor)

528 **Kolde, R.** 2025. Package "pheatmap": Pretty heatmaps. R package. [https://CRAN.R-](https://CRAN.R-project.org/package=pheatmap)  
529 [project.org/package=pheatmap](https://CRAN.R-project.org/package=pheatmap)

530 **Kopelman, N. M., J. Mayzel, M. Jakobsson, N. A. Rosenberg, and I. Mayrose.** 2015.  
 531 Clumpak: A program for identifying clustering modes and packaging population structure  
 532 inferences across K. *Molecular Ecology Resources* 15:1179–1191.

533 **Louie, K.** 2003. Genetic structure and phylogeography of four eastern Pacific estuarine species,  
 534 with an emphasis on eelgrass dependent taxa. Unpubl. Ph.D. diss., University of California, Los  
 535 Angeles, California.

536 **Lourie, S. A., D. M. Green, and A. C. J. Vincent.** 2005. Dispersal, habitat differences, and  
 537 comparative phylogeography of Southeast Asian seahorses (Syngnathidae: *Hippocampus*).  
 538 *Molecular Ecology* 14:1073–1094.

539 **Lowe, W. H., and F. W. Allendorf.** 2010. What can genetics tell us about population  
 540 connectivity? *Molecular Ecology* 19:3038–3051.

541 **McKinney, G., M. V. McPhee, C. Pascal, J. E. Seeb, and L. W. Seeb.** 2020. Network analysis  
 542 of linkage disequilibrium reveals genome architecture in chum salmon. *G3: Genes, Genomes,*  
 543 *Genetics* 10:1553–1561.

544 **Meirmans, P. G.** 2020. genodive version 3.0: Easy-to-use software for the analysis of genetic  
 545 data of diploids and polyploids. *Molecular Ecology Resources* 20:1126–1131.

546 **Merkel & Associates, Inc.** 2023. Oregon International Port of Coos Bay WRDA Section 204(f)  
 547 proposed channel modification, Coos County, Oregon: baseline eelgrass survey. Merkel &  
 548 Associates, Inc., Sacramento, California.

549 **Mobley, K. B., C. M. Small, and A. G. Jones.** 2011. The genetics and genomics of  
 550 Syngnathidae: pipefishes, seahorses and seadragons. *Journal of Fish Biology* 78:1624–1646.

551 **Mobley, K. B., C. M. Small, N. K. Jue, and A. G. Jones.** 2010. Population structure of the  
 552 dusky pipefish (*Syngnathus floridae*) from the Atlantic and Gulf of Mexico, as revealed by  
 553 mitochondrial DNA and microsatellite analyses. *Journal of Biogeography* 37:1363–1377.

554 **Nei, M., and F. Tajima.** 1981. Genetic drift and estimation of effective population size.  
 555 *Genetics* 98:625–640.

556 **Orsi, J. A., R. K. Gish, and B. L. Wing.** 1991. Northern range extensions of four nearshore  
 557 marine fishes in Alaska. *The Canadian Field-Naturalist* 105:82–86.

558 **Partridge, C., A. Boettcher, and A. G. Jones.** 2012. Population structure of the Gulf pipefish in  
 559 and around Mobile Bay and the northern Gulf of Mexico. *Journal of Heredity* 103:821–830.

560 **Pritchard, J. K., M. Stephens, and P. Donnelly.** 2000. Inference of population structure using  
 561 multilocus genotype data. *Genetics* 155:945–959.

562 **R Core Team.** 2024. R: The R Project for Statistical Computing. <https://www.r-project.org/>  
 pg. 23

563 **Ramesh, B., C. M. Small, H. Healey, B. Johnson, E. Barker, M. Currey, S. Bassham, M.**  
564 **Myers, W. A. Cresko, and A. G. Jones.** 2023. Improvements to the Gulf pipefish *Syngnathus*  
565 *scovelli* genome. *GigaByte* 2023:76.

566 **Scott-Somme, K., S. McTierney, R. Brittain, F. Perry, M. Brenen, Marine Biological**  
567 **Association Genome Acquisition Lab, and Darwin Tree of Life Consortium.** 2023. The  
568 genome sequence of the greater pipefish, *Syngnathus acus* (Linnaeus, 1758). *Wellcome Open*  
569 *Research* 8:274.

570 **Stiller, J., N. G. Wilson, S. Donnellan, and G. W. Rouse.** 2017. The leafy seadragon,  
571 *Phycodurus eques*, a flagship species with low but structured genetic variability. *Journal of*  
572 *Heredity* 108:152–162.

573 **Teske, P. R., M. I. Cherry, and C. A. Matthee.** 2003. Population genetics of the endangered  
574 Knysna seahorse, *Hippocampus capensis*. *Molecular Ecology* 12:1703–1715.

575 **Vincent, A. C. J., A. Berglund, and I. Ahnesjö.** 1995. Reproductive ecology of five pipefish  
576 species in one eelgrass meadow. *Environmental Biology of Fishes* 44:347–361.

577 **Vincent, A. C. J., S. J. Foster, and H. J. Koldewey.** 2011. Conservation and management of  
578 seahorses and other Syngnathidae. *Journal of Fish Biology* 78:1681–1724.

579 **Wickham, H.** 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York,  
580 New York.

581 **Wickham, H., R. François, L. Henry, K. Müller, and D. Vaughan.** 2023. dplyr: A Grammar  
582 of Data Manipulation. R package. <https://CRAN.R-project.org/package=dplyr>

583 **Wilson, A.** 2006. Genetic signature of recent glaciation on populations of a near-shore marine  
584 fish species (*Syngnathus leptorhynchus*). *Molecular Ecology* 15:1857–1871.

585 **Wolf, M., B. L. da Silva Ferrette, R. T. F. Coimbra, M. de Jong, M. Nebenführ, D.**  
586 **Prochotta, Y. Schöneberg, K. Zapf, J. Rosenbaum, H. A. Mc Intyre, J. Maier, C. C. S. de**  
587 **Souza, L. M. Gehlhaar, M. J. Werner . . . S. Winter.** 2024. Near chromosome-level and highly  
588 repetitive genome assembly of the snake pipefish *Entelurus aequoreus* (Syngnathiformes:  
589 Syngnathidae). *GigaByte* 2024:1–13.

590 **Woodall, L. C., F. Otero-Ferrer, M. Correia, J. M. R. Curtis, N. Garrick-Maidment, P. W.**  
591 **Shaw, and H. J. Koldewey.** 2018. A synthesis of European seahorse taxonomy, population  
592 structure, and habitat use as a basis for assessment, monitoring and conservation. *Marine Biology*  
593 165:19.

594 **Wu, T. D., and C. K. Watanabe.** 2005. GMAP: A genomic mapping and alignment program  
595 for mRNA and EST sequences. *Bioinformatics* 21:1859–1870.

**Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir.** 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28:3326–3328.

**FIGURE CAPTIONS**

**Figure 1. Sampling locations for bay pipefish in Coos Bay, Oregon.** The blue point indicates Valino Island, while the orange point marks the Causeway site, positioned adjacent to the Trans Pacific Pkwy.

**Figure 2. Principal Component Analysis of genetic variation between two populations.** PCA plot showing genetic differentiation between Causeway (dark gray circles) and Valino Island (light gray triangles) individuals. PC1 (2.0%) and PC2 (1.6%) explain 3.6% of the total genetic variance. Each point represents an individual, with color and shape distinguishing populations.

**Figure 3. Results of STRUCTURE analysis, a Bayesian clustering method used to infer population structure.** Each vertical bar represents an individual, with colors showing the proportion of that individual’s genome assigned to each genetic cluster for K = 1 and K = 2.

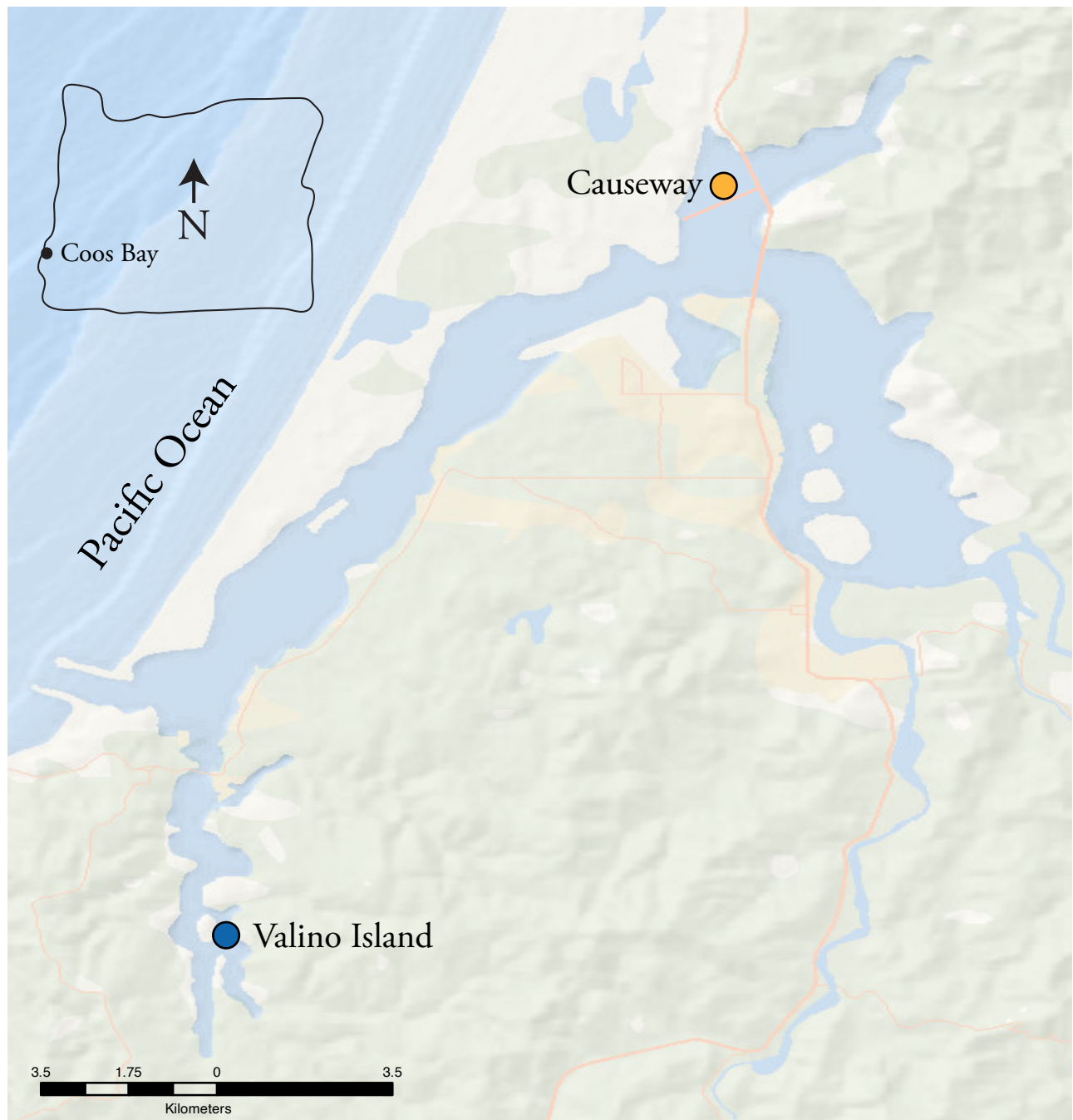
**Figure 4. Genome-wide  $F_{ST}$  across chromosomes.** Alternating shaded blocks represent successive chromosomes (1-22). Each point represents the  $F_{ST}$  estimate for an individual SNP, while the solid line indicates a smoothed trend across the genome. The black dot on the y-axis marks the genome-wide  $F_{ST}$  average of 0.0074.

**Figure 5. Genome-wide linkage disequilibrium (LD) patterns across 22 chromosomes.** (A) Heatmap of maximum  $r^2$  values and (B) heatmap of mean  $D'$  values between 1 Mb genomic bins, highlighting intra- and inter-chromosomal LD. Black lines mark chromosome boundaries; the diagonal indicates perfect self-linkage ( $LD = 1$ ). Color scales range from 0 (blue, no LD) to 1 (red, complete LD).

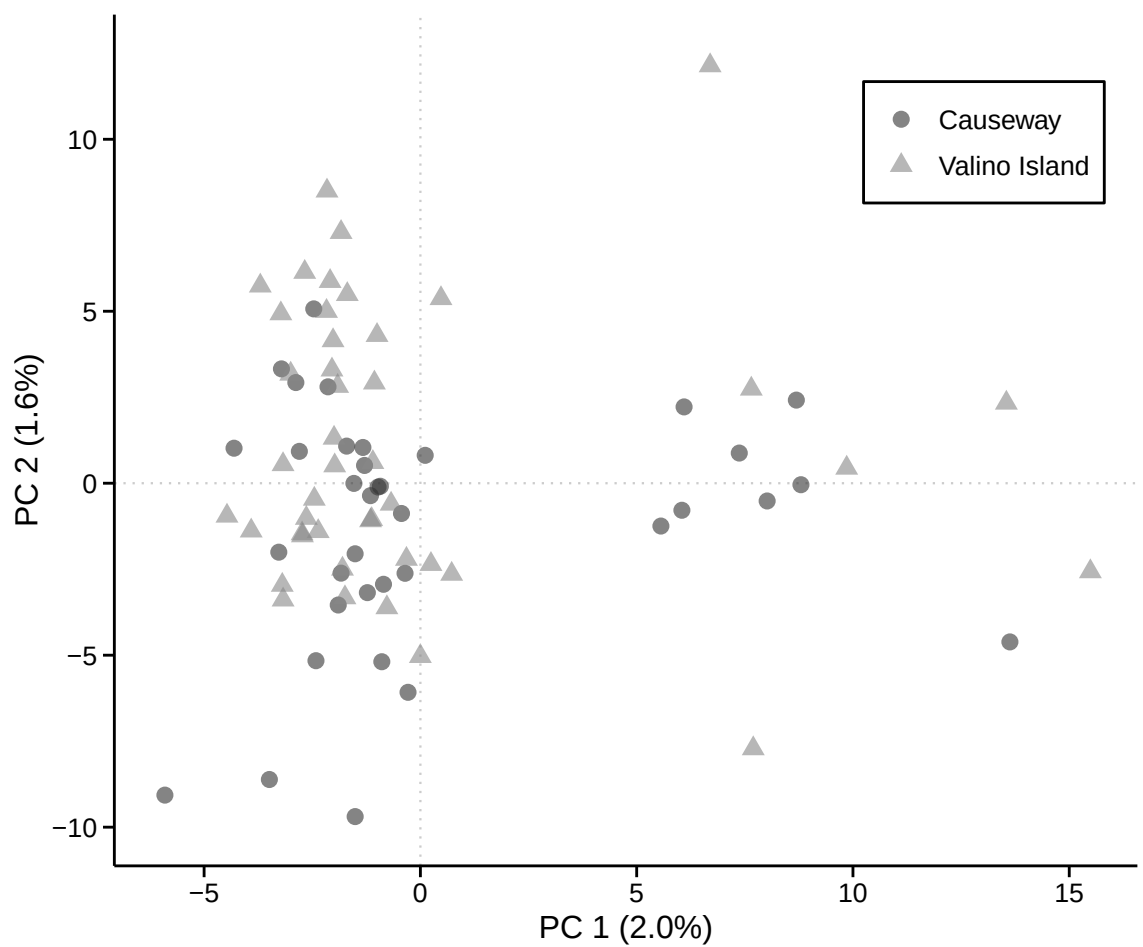
**Table 1. Sampling site information.**

This table presents collection details for two sampling sites. Columns show sample site name, GPS coordinates, elevation in meters and feet, distance to ocean in kilometers and miles, collection dates, and total number of samples collected at each site.

| Sample Site   | Sample site GPS coordinates | Elevation (m/ft) | Distance to ocean (km/miles) | Collection dates   | Total samples |
|---------------|-----------------------------|------------------|------------------------------|--------------------|---------------|
| Causeway      | 43°26'46"N<br>124°13'23"W   | 0.52/2           | 16.24/10.09                  | September 14, 2023 | 50            |
| Valino Island | 43°18'45"N<br>124°18'55"W   | 1.96/6           | 7.64/4.75                    | September 14, 2023 | 66            |



Esri, Garmin, GEBCO, NOAA, NGDC, and others contributors



K = 1

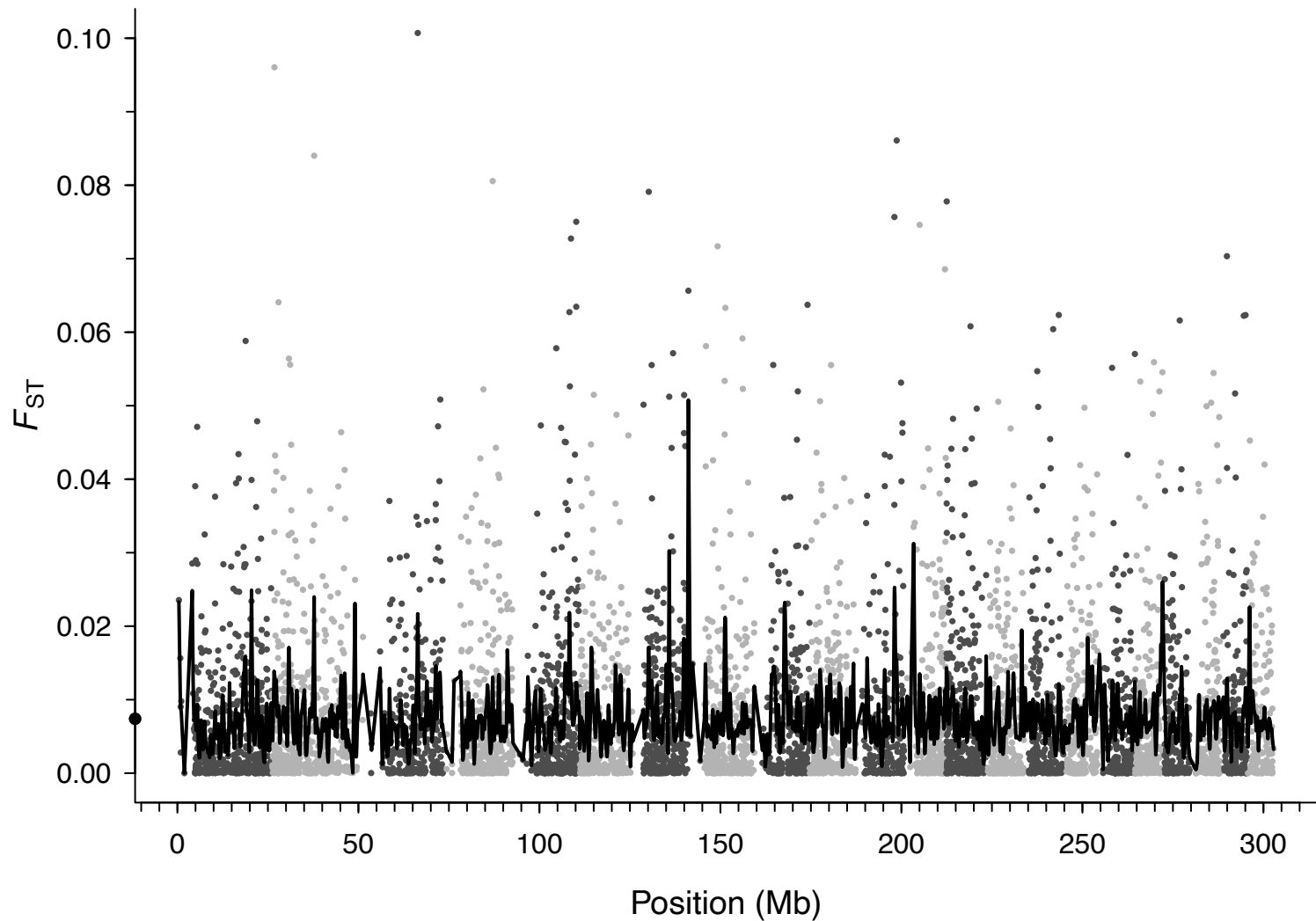


K = 2

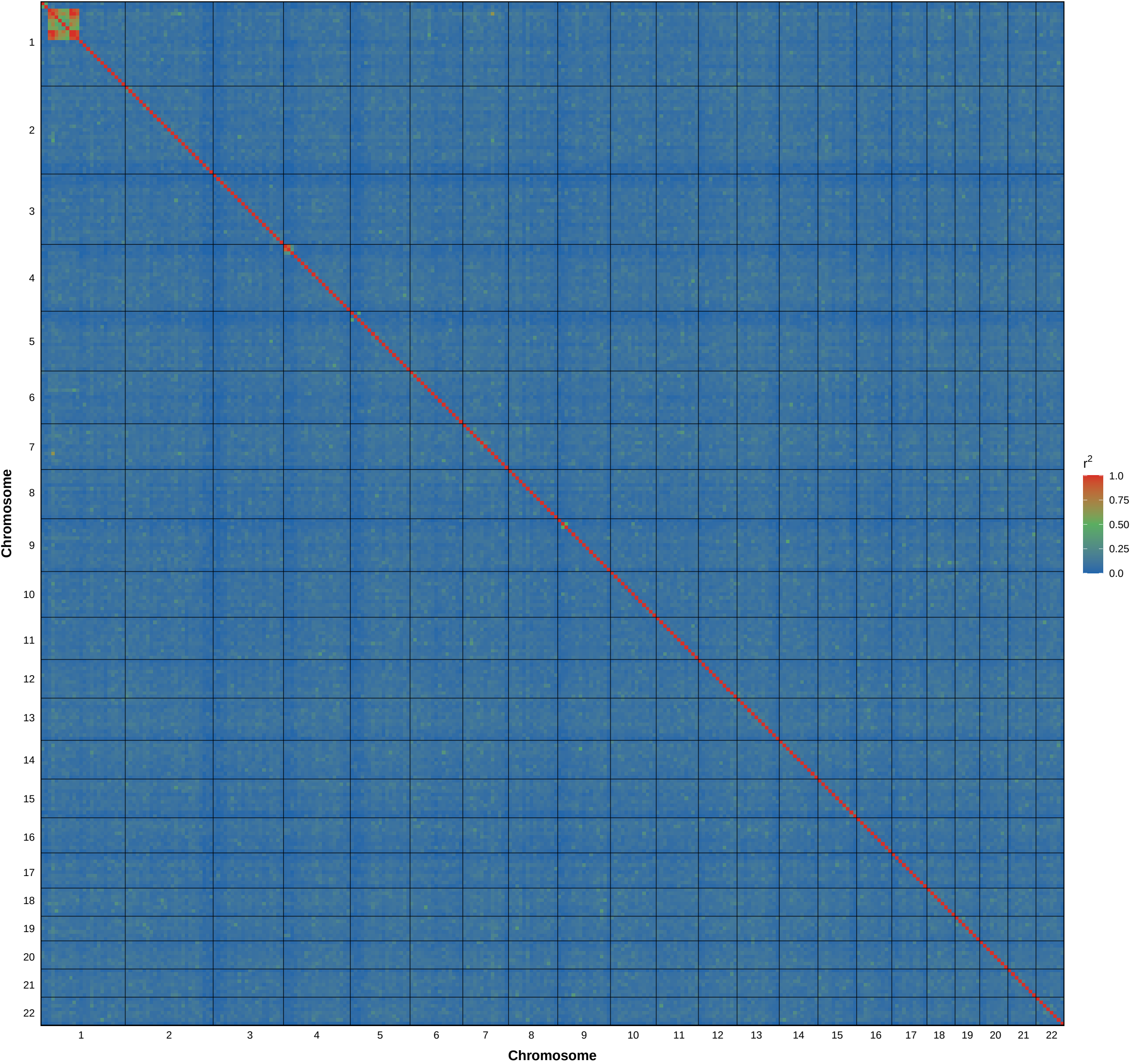


Causeway

Valino Island



**A**  $r^2$  linkage disequilibrium



**B**  $D'$  linkage disequilibrium

