Rapid genetic adaptation to a novel ecosystem despite a massive bottleneck

Morgan Sparks[1], Claire Schraidt[1], Xiaoshen Yin[2], Lisa Seeb[3], and Mark Christie[1]

[1]Purdue University
[2]Ocean University of China
[3]University of Washington

March 15, 2023

Abstract

Introduced and invasive species make excellent natural experiments for investigating rapid evolution. Here, we describe the effects of genetic drift and rapid genetic adaptation in pink salmon (Oncorhynchus gorbuscha) that were accidentally introduced to the Great Lakes via a single introduction event 31-generations ago. Using whole-genome resequencing for 134 fish spanning five sample groups across the native and introduced range, we estimate that the progenitor population's effective population size was 146,886 at the time of introduction, whereas the founding population's effective population size was just 72—a 2040-fold decrease. As expected with a severe bottleneck, we show reductions in genome-wide measures of genetic diversity, specifically a 37.7% reduction in the number of SNPs and an 8.2% reduction in observed heterozygosity. Despite this decline in genetic diversity, we provide evidence for putative selection at 47 loci across multiple chromosomes in the introduced populations, including missense variants in genes associated with circadian rhythm, immunological response, and maturation, which match expected or known phenotypic changes in the Great Lakes. For one of these genes, we use a species-specific agent-based model to rule out genetic drift and conclude that a strong response to selection occurred in a period gene (per2) that plays a predominant role in determining an organism's daily clock, matching large day length differences experienced by introduced salmon during important phenological periods. Together, these results inform how populations might evolve rapidly to new environments, even with a small pool of standing genetic variation.

Introduction

Rapid evolution—or evolution on the scale of tens of generations—is now a well-established cause of genetic change in populations (Colautti & Lau, 2015; Ellner et al., 2011; Hendry & Kinnison, 1999). Numerous studies have described the results of rapid evolution across diverse organisms (Campbell-Staton et al., 2017, 2021; Lee, 2002; Yin et al., 2021), as well as how rapid genetic change has been driven by different evolutionary processes (Barrett et al., 2019; Hof et al., 2016; Kardos et al., 2018; Lamichhaney et al., 2018). While rapid genetic adaptation is often the focus of these studies, both theoretical and empirical methods have demonstrated the strong effect of genetic drift on allele frequencies, even over a handful of generations, and is particularly apparent in bottlenecked populations with significant founder effects (Allendorf, 1986; Kardos et al., 2018). With this study, we show the genome-wide outcomes of a documented population bottleneck originating from a single introduction event. We also show a subsequent response to selection imposed by the introduction into a radically different environment.

Introduced and invasive species make for excellent model systems to study rapid evolution given the nature of most introductions—e.g., small founding populations, novel environments, or multiple progenitor sources (Kolbe et al., 2004; North et al., 2021; Stern & Lee, 2020). Because of these features, species introductions provide useful templates to answer questions about the evolutionary mechanisms that determine if

organisms are successful in responding to new environmental conditions. Species introductions can similarly serve as model systems for studying the potential for rapid evolution in threatened and endangered species, or less tractable populations, such as those that are long-lived (Lee, 2002; North et al., 2021).

Populations introduced into novel environments may experience rapid evolution because of different agents of genetic change. On one hand, the often-novel nature of recipient ecosystems—e.g., new pathogens, prey, and predators or differences in the availability of fundamental resources—is expected to result in rapid genetic adaptation (North et al., 2021). However, introduced populations are most often founded from far fewer individuals than the progenitor population and these introductions constitute a simultaneous demographic and genetic bottleneck even with multiple founder events (Allendorf et al., 2022). While a reduction in founding individuals relative to the source population will always result in demographic bottleneck, the effects of a genetic bottleneck are determined by the effective population size of the progenitor population (Waples, 2022). That is, other factors being the same, introduced populations that are founded from large, genetically diverse progenitor populations are expected to have a less extreme genetic bottleneck relative to populations founded from progenitors that are less genetically diverse and have smaller effective population sizes. Moreover, the genetic consequences (e.g., the magnitude of the reduction of genetic diversity and the likelihood of subsequent inbreeding) of the bottleneck will influence the likelihood of persistence as a function of rapid genetic adaptation.

In introduced populations, both genetic drift and natural selection can result in reduced genetic diversity relative to their progenitors. While it remains challenging, separating the not always mutually exclusive effects of genetic drift and genetic adaptation has become more feasible with whole-genome sequencing. The ability to contrast local with genome-wide patterns of allelic change provides a useful template with which to disentangle genetic adaption and genetic drift since the effects of drift are expected to occur more evenly across the genome while the effects of selection are expected to manifest within local genomic regions.

Salmon, trout, and their allies (Family *Salmonidae*) are among the most commonly introduced species globally because of their desirability as a commercial commodity and recreational species (Lowe et al., 2000). One of the iconic introductions of salmonids globally is the purposeful introduction of Pacific salmon and trout (*Oncorhynchus spp.*) to the Laurentian Great Lakes (hereafter Great Lakes) where they now make up the most popular targets of a $7 billion per year fishery (Melstrom & Lupi, 2013; Mills et al., 1994; Parsons, 1973). While the majority of salmon species introduced to the Great Lakes were planted with intent of new economic and recreational enterprises—with numerous stocking efforts made to drive those introductions—pink salmon (*O. gorbuscha*) were introduced in a single event in 1956 as the result of a single accidental introduction (Kwain & Lawrie, 1981; Schumacher & Eddy, 1960). The progeny of an odd-year spawning stock (1955, discussed below), collected from the Lakelse River in British Columbia, Canada, were intended for introduction into the Canadian Arctic (Baffin Bay, CA). However, thousands of juveniles were accidentally released in Lake Superior as hatched fry (*N [?]* 21,000, a subset of more than 750,000 progeny from the Artic introduction program made from ˜500 dame and sire pairs) in the Current River in Thunder Bay, Ontario (Gharrett & Thomason, 1987). Pink salmon spread rapidly throughout the Great Lakes with individuals being observed in every Great Lake, and populations growing to the 10s of thousands and potentially 100s of thousands, within 12 generations of the introduction (Kwain & Lawrie, 1981). This spread was despite what is presumed to be a strong population bottleneck at introduction—a founding population estimated on the order of hundreds of individuals—due to the very low survival of fry to adults (˜2%) and the likelihood of a significant portion of the population straying to new habitats (˜10%) (Gharrett & Thomason, 1987; Kwain & Lawrie, 1981; Quinn, 2018).

The introduction of pink salmon to the Great Lakes is an excellent system to study the combined effects of genetic drift and natural selection on rapid evolution because of their small founding population size, the lack of additional propagule pressure (no subsequent gene flow), the radically different environment they were introduced to in the Great Lakes, and the rise of novel age-at-maturity life histories. Pink salmon are obligate anadromous fish in their native range—meaning they must spend most of their pre-spawning life, about one-and-half of their two years, in a marine environment. This life history strategy is largely unique

in the genus *Oncorhynchus*, where pink salmon and their sister species, chum salmon (*O. keta*), migrate almost immediately to marine environments while the other anadromous species all maintain non-migratory, freshwater ecotypes or can have long periods of freshwater residency. Pink salmon are also unique within the genus in that they have a fixed two-year life cycle in their native range (i.e., time from hatching to death is 2 years for all individuals and other age-at-maturity phenotypes are exceedingly rare (Anas, 1959; Quinn, 2018; Turner & Bilton, 1968)). As a result of this feature, lineages of odd- and even- year fish have diverged, such that a river may have distinct populations spawning in the same habitat in both even and odd numbered years. Because of this lineage divergence, genetic analyses indicate pink salmon are more related to one another based on allochrony, rather than geography—e.g., all odd-year fish are more related to one another than they are to an even year fish in the exact same habitat—whereas geography is the dominant feature structuring populations in most other salmonids (Christensen et al., 2021; Fraser et al., 2011; Seeb et al., 2014; Tarpey et al., 2017). However, this fixed life history in pink salmon has eroded in the Great Lakes, with fish maturing at 1-, 2, and 3-years old, leading to the establishment of odd- and even-year spawning groups in the Great Lakes (despite only an odd-year progenitor) that appear to maintain gene flow across years (Bagdovitz et al., 1986; Gharrett & Thomason, 1987; Kwain & Chappel, 1978; Wagner & Stauffer, 1980). Given the presumed bottleneck at introduction, the novel recipient environment, and the rise of novel age-at-maturity life histories, pink salmon in the Great Lakes are expected to have rapidly evolved via multiple evolutionary forces.

Our study aimed to characterize and disentangle genetic drift and genetic adaptation of pink salmon introduced to the Great Lakes. We sequenced the genomes of pink salmon from the Great Lakes and their progenitor population to answer the following questions: i) what were the magnitude of the population and genetic bottlenecks at introduction and their corresponding reduction of genetic diversity?, and ii) what loci are associated with putative rapid genetic adaptation to the Great Lakes? We employed medium coverage (˜14x) whole genome-resequencing and found evidence for a severe genetic bottleneck at introduction along with associated reductions in genetic diversity, as well as multiple genomic regions that may have responded to selection in the novel ecosystem.

Methods

Sample sites and sample collection

Pink salmon were collected from the Lakelse River, British Columbia, Canada (Figure 1a) in the fall of 2006 and 2007 as part of an earlier population genetics study (Beacham et al., 2012). The odd-year spawning population from the Lakelse River was the sole progenitor population for pink salmon introduced to the Great Lakes (Gharrett & Thomason, 1987; Schumacher & Eddy, 1960). Lake Superior tributaries were sampled in September and October of 2018 and 2019 for even- and odd-year spawning fish, respectively. Fish collected in Lake Superior were from the Steel River on the north shore of Lake Superior (Figure 1b,c), one of the largest pink salmon spawning groups in the Great Lakes. The Steel River is 170 km east of the original site of introduction, the Current River (Kwain, 1982; Kwain & Lawrie, 1981). For Great Lakes fish, biometric data (length, weight, phenotypic sex) and a portion of the dorsal fin were collected for subsequent aging. Caudal fin tissue was collected for DNA extraction. In their native range, pink salmon from the Lakelse River form two distinct odd- and even-year populations without gene flow, however in the Great Lakes, fish spawning in odd- and even-years are maintaining gene flow across year classes (Figure S1). As such, we refer to all the groups in our study as sample groups, despite the strong population divergence between the even- and odd-year samples from the same location in the native range. Hereafter, the native range sample groups will be referred to by their year of spawning: British Columbia odd (progenitor) and British Columbia even, which we shorten to BC odd and BC even. We refer to the introduced two-year old sample groups as Great Lakes odd and Great Lakes even, which we shorten to GL odd and GL even. We also sampled 3-yr old pink salmon (a life history strategy unique to the Great Lakes), which we shorten to GL odd 3 (Table 1). It is important to note that a 3-yr old pink salmon that spawns in an odd-year would have parents that spawned in an even-year. These 3-yr old fish are the primary way that gene flow occurs between odd and even years in the Great Lakes and 3-yr old fish are present in both odd and even years though we did not sample any

in an even year.

Pink salmon aging

Dorsal fins from all sampled individuals were dried, separated, set in epoxy, and read according to the protocols outlined in Koch & Quist (Koch & Quist, 2007) and Little et al. (Little et al., 2012). To age fish, cross sections of dorsal fin rays set in epoxy and were read at 4.5x power on an Olympus SZ61 digital microscope. Samples were aged by counting annuli in fin rays. Photos were taken of each cross section using the digital microscope and associated IFINITY CAPTURE (v. 6.3.2) software. These photos were also sent to an outside reviewer to confirm accuracy of age determination. Samples from the native range did not include aging structures but were all safely assumed to be 2-years old due to their fixed age-at-maturity, which is the convention for all pink salmon populations in the native range (Gharrett & Thomason, 1987; Quinn, 2018).

DNA extraction, library preparation, and sequencing

Tissue samples collected from British Columbia were extracted using Promega Wizard extraction kits and protocols. Sample aliquots were dried down and stored frozen as pellets in population-specific 96-well plates. These aliquots were sent to Purdue University in 2019, where they were brought into solution with 10 mg of ultrapure water. Tissue samples collected from Great Lakes fish were extracted using Qiagen DNeasy blood and tissue kits using the standard protocol for tissue. Sample concentrations for all samples were determined using a Qubit 4 and the high-sensitivity dsDNA quantification assay.

For genomic sequencing, thirty samples each from BC odd and BC even were randomly chosen without respect to sex as it was unknown for those samples. For the verified 2-year old fish from the Great Lakes odd and even sample groups (2018 and 2019), thirty fish from each year were randomly chosen with an even sex ratio of 15 males to 15 females. Additionally, 15 fish from the GL odd 3yr sample group were included for sequencing (Table 1). Libraries were prepared at the Purdue Genomics core using Illumina DNA Prep kit and protocols with a 1/5 reaction dilution and a set of Illumina DNA Prep compatible amplification primers to add standard Illumina 384-well-design unique dual indexes. One sample failed after library preparation such that only 29 BC odd samples were sequenced. All 134 samples were transported to the Indiana University School of Medicine Genomics Core where they were sequenced using 2x150 base pair paired-end chemistry on an Illumina NovaSeq 6000 S4 for 300 cycles in two separate runs; all samples were included in each of the two runs.

Bioinformatics, variant calling, and filtering

Bioinformatic analyses were conducted using the Purdue Rosen Center for Advanced Computing (McCartney et al., 2014). After sequencing, there was an average of 285.43 million reads per sample (˜14.12× coverage of the 2.7 Gb genome, Table S1). Raw reads were trimmed of adapters, and low-quality reads removed using Trimmomatic v. 0.39 with the clipping command LEADING:20 TRAILING:20 MINLEN:30 (Bolger et al., 2014). For SNP calling, we used the Genome Analysis Toolkit v. 4.2.2 (GATK) germline short variant discovery pipeline with best practices unless otherwise noted (Van der Auwera & O'Connor, 2020). First, cleaned reads were aligned to the OgorEven_v1.0 even year pink salmon assembly (GCF_021184085.1) using bwa mem v. 07.17 with the -M flag to mark shorter splits as secondary to avoid aligning to multiple sites as paralogs are common in salmonid genomes (Christensen et al., 2021; Li, 2013). While chromosomal-level assemblies exist for both odd and even-year lineages, the even-year assembly for pink salmon is the reference assembly for the species; it is a much more complete assembly, was built using higher quality sequencing resources (i.e., Hi-C), and is the only assembly that contains annotation information (Christensen et al., 2021). The BAM files were merged into single files using Picard v. 2.20.8 AddOrReplaceReadGroups and MergeSamFiles programs (Piccard Tools, 2020) because reads were generated from two different sequencing runs (though all samples were included in each run). We then marked duplicate reads using GATK's MarkDuplicatesSpark to generate appropriate BAM files for variant calling.

Variants were called using HaplotypeCaller (in GVCF mode), consolidated with GenomicsDBImport,

and then jointly genotyped using GenotypeGVCFs in GATK. All steps were preformed jointly for all sample groups listed above, and the resulting pre-filtering file contained 24,071,073 SNPs. To filter SNPs, we first used a GATK recommended hard filter thresholds (see SI Methods) and all filtering was conducted using VCFtools v. 0.1.16 (Danecek et al., 2011). After the hard filtered loci were removed, samples were filtered in the following order: remove non-biallelic sites, remove individuals with >20% missingness (one individual from the GL even group was removed), remove sites with >20% missingness, remove sites with a minor allele frequency <0.05. Filtering for minor allele frequency was executed with both a joint vcf file with for all individuals (across all sample groups) (5,289,687 total SNPs remaining), as well as for sample group-specific vcf files. Hereafter, these vcfs are referred to as the joint vcf or sample group-specific vcfs, respectively (Table S2 includes the full array of differently filtered vcf files used in this study and the corresponding analyses for which they were used). For downstream analyses, our default approach was to use the joint vcf unless otherwise noted (e.g certain estimators can be biased if low frequency sites are removed, so we used the sample group-specific vcfs for those analyses). All filtered sample groups included even sample sizes (29 or 30 individuals after filtering and assessing for relatedness, Table 1)—one sample was removed because of relatedness (Figure S1, SI Methods) and one sample, described above, for high amounts of missing data. The Great Lakes odd 3-yr old group contained 15 individuals and this group was removed from analyses unless explicitly noted due to the imbalance in sample sizes.

Population genetic analysis

Variant Call Format (vcf) files were loaded into R v. 4.2.1. (R Core Team, 2022) using the gaston v1.5.7 package (Perdry et al., 2020). Using sample group-specific vcfs that were not filtered for minor allele frequency (to maintain rare alleles; Table S2), we calculated the number of SNPs both genome-wide and in 2.5 Mbp windows. Genome-wide estimates of observed heterozygosity were calculated using custom scripts in 2.5 Mbp windows. Tajima's D was calculated in VCFtools using the same window sizes but using the joint vcf before it was filtered on minor allele frequency (Table S2); Tajima's D is a sequence-based estimator and can be upwardly biased if invariant or low-frequency sites are not included. We used the R package SNPRelate v.1.32.2 (Zheng et al., 2012) to calculate mean pairwise F_{ST} (Weir & Cockerham, 1984), which used a linkage disequilibrium pruned (R^2 £ 0.2) subset of 123,924 SNPs. As an additional measure of population structure, we used the program fastSTRUCTURE v. 1.0-py27 (Raj et al., 2014) with our joint vcf to generate population assignments using a subset of 10,000 randomly selected SNPs using K values ranging from 1-5.

Effective population size estimates

We used two methods to understand the likely bottleneck and relative population sizes at introduction and in our contemporary samples. These two methods were employed based on their respective strengths and weaknesses: 1) the site frequency spectrum-based method (Stairway Plot 2) performs well at detecting historical changes in population size (Liu & Fu, 2020), and 2) the linkage-disequilibrium based method (GONE) (Santiago et al., 2020) performs well at estimating recent demographic shifts and contemporary (<100 generations) estimates of effective population size (Reid & Pinsky, 2022; Santiago et al., 2020). We used default input parameters in GONE which samples 50,000 random SNPs per chromosome and ran the program 100 times. We next calculated the average of those runs and computed the 95% quantiles. To compare with estimates of effective population sizes from farther back in time, we used Stairway Plot 2 (see SI Methods, Results).

SNP variant analysis and annotation

For outlier SNPs under putative selection, we first calculated (Weir & Cockerham, 1984) F_{ST} using VCFtools between the BC odd and GL odd samples. To identify outlier regions, we calculated sliding window outliers using a 100 kbp window with a 50 kbp step, for which we calculated ZF_{ST} relative to the chromosome a window was on and considered any window $^{3}5$ ZF_{ST} as a putative outlier region. We chose the 5 ZF_{ST} threshold as an especially conservative first metric for regions putatively under selection and then used a marginally less conservative SNP cutoff for individual locus annotation using snpEff (discussed below). For SNPs within putative outlier regions, we similarly transformed per SNP F_{ST} with a Z-transformation

normalized across the chromosome; any SNP with $^{3}4$ ZF_{ST} was retained for variant annotation. We visually compared our outlier windows with Tajima's D values calculated using VCFtools in 10 kbp steps. To validate these results, we conducted the same analyses using the same BC odd and the closely related GL even sample groups to ensure the same outlier windows were maintained. We also used eigenGWAS, an eigenvector decomposition-based method using a genome-wide association study approach (GWAS), but where genotypes are treated as the phenotypes in the GWAS (Chen et al., 2016). This method was similarly used to discover highly differentiated windows that would be consistent with putative selection at a locus. Variants from eigenGWAS were plotted against ZF_{ST} windows to further verify our best candidate outliers. We considered any window with a mean log transformed p-value ^{3} 6.64 (Bonferroni corrected p-value = 0.01) as significant.

We next used the programs snpEff and snpSift (v. 5.1) to annotate and describe the effect of outlier SNPs in our data set (Cingolani, Patel, et al., 2012; Cingolani, Platts, et al., 2012). We first filtered the joint vcf for individual SNPs as discussed above ($^{3}4$ ZF_{ST}) so that we only used SNPs highly differentiated between sample groups. The program snpEff uses the reference genome assembly and associated annotation report to determine the functional effect (e.g., missense, synonymous) of a given SNP relative to the assembly. Individual SNPs could have multiple effects, for example they may occur in the intronic region of one gene, in addition to occurring upstream of a separate gene.

Simulation of genetic drift for candidate loci

We leveraged estimated pink salmon effective population sizes through time (GONE results) to determine whether the empirically observed changes in allele frequencies were due to genetic drift alone at our best candidate locus (*per2,* which had a total of 38 SNPs; see Results). To do this, we first created an agent-based model that simulated the unique life-history characteristics of pink salmon (see SI Methods for full details). The carrying capacity of the adults was set equal to the mean effective population size calculated from GONE in each generation (averaged across the 100 GONE replicates) and was varied each year in the model. This approach was conservative as the ratio of N_e to N_c often equal 10-20% (Frankham, 1995; Waples, 2002); the smaller

used in our model means that, if anything, we are overestimating the effects of genetic drift.

We first wanted to estimate the allele frequencies in *per2* at the time of the introduction event. Because the effective population size in the native range has declined since the introduction event (see Results) and because we only sampled 30 individuals from the native range odd year, the estimates of allele frequencies from our sample may not reflect the range of allele frequencies present in BC odd at the time of introduction. After parameterizing the model with native range effective population sizes estimated for each generation and coupling that with sampling events at the end of the simulations (see SI Methods), we found that genetic drift alone could only explain an average change in allele frequency of 0.0000405 in the native range. Sampling error alone, however, was somewhat larger resulting in an average change in allele frequency of 0.000365. Our process allowed us to construct 95% confidence intervals around our estimates of allele frequencies from the native range that included both the effects of genetic drift and sampling error. Because the effects of both genetic drift and sampling were relatively small, we parametrized the starting allele frequencies of the Great Lakes simulations from our empirical estimates calculated from the progenitor population (BC odd).

We next wanted to determine whether the allele frequency at the time of the introduction could then drift to a frequency equal to that found in each of the 38 SNPs found within *per2*. We again used the forward-time agent-based model to simulate changes in *per2* allele frequencies, this time using the N_e values for BC odd for the first 75 generations (modeling from generations 100 to 26, where generation 26 reflected the population in 1956, the year of the introduction and then switching to the GL odd N_e estimates for the next 52 generations (generations 52 to 1) using the BC odd empirical estimates as our starting allele frequencies (See Figure S2 for modeled population sizes through time and Figure 3 for empirical estimates). We selected these generations to capture the most likely effective population sizes from the progenitor populations at the time of introduction and the smallest effective population sizes during the introduction event. At the end of each simulation, we again sampled 30 individuals and calculated the final allele frequency. This process was

replicated 1,000 times for each of the 38 SNPs. We then calculated how many simulated samples had allele frequencies as or more extreme than that found in our empirical GL sample to calculate the probability of the empirically observed allele frequencies occurring by genetic drift alone. We also plotted the changes in allele frequencies through time for one example SNP and calculated 95% CI's for all SNPs.

Results

Population structure

Population-level analyses indicated three distinct populations: BC even, BC odd, and the Great Lakes. Pairwise F_{ST} between BC even and BC odd was 0.091, and 0.088 between BC odd and GL odd suggesting nearly as much divergence between the progenitor population and the GL population as between the native-range odd and even lineages (Table 2). Among the Great Lakes sample groups F_{ST} ranged between 0.0006 (GL odd vs. GL odd 3) and 0.0046 (GL even vs. GL odd) (Table 2), indicating little to no population structure among Great Lakes sample groups. These results were supported by fastSTRUCTURE, which indicated a K of 3 (of K=1-5) was the best supported by maximized marginal likelihood where all samples had >97% assignment probability for their sample group (BC even, BC odd, Great Lakes) (Figure S3).

Measures of genetic diversity

Bottlenecks are theoretically predicted to cause a larger drop in allelic diversity than heterozygosity (Allendorf, 1986), which matches our empirical results where there was a 37.7% loss of segregating SNPs per 2.5 Mbp window (BC odd = 12,188 vs. GL odd=7,659), a 30.8% loss of SNPs genome-wide (BC odd = 7,337,647 and GL odd= 5,080,474), but just an 8.2% loss of heterozygosity (BC odd = 0.279 and GL odd=0.256) (Figure 2a-c). BC odd had the highest heterozygosity at 0.279, but all sample groups were similar with GL odd having 0.256, BC even with 0.250, GL odd 3-year old fish with 0.244, and GL even with 0.239 (Figure 2b). The distribution of the variance of these estimates did vary by group (Figure 2b), where native range groups have a greater standard deviation around their heterozygosity estimates. Two samples in the BC even sample group had much lower heterozygosity than the rest of the sample group (0.113) and when removed, the estimated heterozygosity for the group increased to 0.26. Counts of SNPs were highest in native-range sample groups, containing approximately 2.0-2.3 million more SNPs relative to the introduced Great Lakes sample groups (Figure 2c). For the native range populations, the odd-year population had greater heterozygosity than the even-year population (0.279 versus 0.256 or an 8.2% difference), but a similar number of SNPs (a difference of 170,365). This result provides further support for the idea that native range even-year lineages were likely founded from an odd-year lineage (Gordeeva & Salmenkova, 2011), the bottleneck of which still leaves a mark of reduced relative genetic diversity.

Demographic history of native and introduced pink salmon

Because recent bottlenecks may reduce low-frequency polymorphisms, positive Tajima's D can be indicative of recent demographic declines in populations (Nei et al., 1975). Both Great Lakes groups had elevated values (mean Tajima's D, GL odd = 0.41 GL even = 0.36) (Figure 3a, Figure S4). GONE results indicated that contemporary effective population sizes for the introduced sample groups have grown from a substantial bottleneck at introduction (calculated at the time of introduction—generation 31, GL odd = 71.67 and 95% quantiles 46.82-100.30, GL even = 68.23 and 95% quantiles 42.20-108.1) to larger contemporary estimates (GL odd = 604.62 and 95% quantiles 424.74-888.86, GL even = 833.71 and 95% quantiles 503.44-1265.41) (Figure 3b-d). Over the same period, the native range sample groups have experienced large declines. At the time of introduction, the progenitor's (BC odd) effective population size was estimated at 146,886.30 (95% quantiles 68,579.82-219,895.7) but reduced to 2,964.19 (95% quantiles 2,365.21-5,623.89) at the time of collection. During that same period the progenitor's even year complement (BC even) reduced from 50,212.30 (95% quantiles = 35,029.77 60,743.44) at introduction to 11,812.53 (95% quantiles = 8,804.14-14,990.69) at the time of collection. GONE results also suggest even smaller effective population sizes immediately preceding introduction for the Great Lakes sample groups, just 15.7 five and six generations before the introduction for GL odd and 27.7 four generations before introduction in GL even, which strongly suggests that the GONE's generation time estimate is off by a small factor. Furthermore, the observation that GONE

does not reproduce the large historic N_e values estimated from the native range populations in the ancestral Great Lakes sample groups suggests the extreme population change in the Great Lakes samples impedes the program from estimating their effective population sizes prior to introduction. StairwayPlot2 failed to reproduce the strong bottleneck indicated by GONE, Tajima's D, and our genomic diversity data, suggesting the program may fail to properly describe demography because of the strong bottleneck (SI Results, Figure S5,6).

Candidate genes responding to selection

Using sliding windows, we found 47 windows with ^{3} 5 ZF_{ST} across 19 of 26 chromosomes when comparing BC odd versus GL odd. Some of these windows overlapped and, when combined, resulted in 34 distinct windows (Figure S7, see SI Data for full annotation information). For individual variant sites within those windows, we found 1,320 SNPs ^{3} 4 ZF_{ST}. Using eigenGWAS, we found 4173 SNPs in those same windows. We plotted windows from both methods to verify that the same windows were found using multiple methods (Figure 4a,b). We used the significant SNPs (^{3} 4 ZF_{ST}) contained in the 34 ^{3} 5 ZF_{ST} 100 kbp windows for annotation using snpEff. Of the 1,320 SNPs queried, 5 resulted in synonymous changes, 9 resulted in nonsynonymous changes, and 1306 were found in non-coding regions. Of all significant variant effects, 57.83% (1,526) occurred in introns, 22.23% (613) in intergenic regions, and only 0.76% (20) in exons.

We selected genes that had putatively responded to selection in the novel environment as those that: 1.) had outlier windows identified by ZF_{ST} and eigenGWAS, 2) that were highly divergent in both comparisons of BC odd and GL odd and BC odd and GL even, and 3.) that resulted in nonsynonymous variants. For these regions, we plotted the outlier window in which they occurred with the genes and coding regions visualized (Figure 4c,d). Within each window, we illustrated F_{ST}, observed heterozygosity, and Tajima's D (10 kbp) as complementary indices of a selective sweep (Figure S7). Of our 9 missense variants, three occurred in the same gene, *LOC124013183* (CD209 antigen-like protein A), on chromosome 24. While we considered this strong evidence for selection, the flanking regions near the gene contained very few SNPs and heterozygosity and Tajima's D were not able to be calculated for most of the window. The remaining missense variants were each located on 6 separate genes, which were *gnrhr4* (gonadotropin releasing hormone receptor 4) on chromosome 2, *LOC124043346* (collagen alpha-2(V) chain-like) on chromosome 9, *LOC124000748* (B-cell CLL/lymphoma 7 protein family member A-like) on chromosome 16, *LOC124004644* (cytochrome P450 11B, mitochondrial-like) on chromosome 19, and *LOC124009961* (period circadian protein homolog 2-like) and *lpar5a* (lysophosphatidic acid receptor 5a) in the same window on chromosome 22. When further validated against the BC odd and GL even comparison, only gonadotropin releasing hormone receptor 4, period circadian protein homolog 2-like, lysophosphatidic acid receptor 5a, and CD209 antigen-like protein A were shared and thus presented in our analyses (Figure 4c,d, Figure S8). Of genes with missense variants, *LOC124009961* (period circadian protein homolog 2-like, hereafter *per2*) had a particularly strong signal of elevated F_{ST} (mean within gene = 0.72, mean within window = 0.23), as well as large local decreases in heterozygosity and Tajima's D (Figure 4d). There were 38 total SNPs located within the gene, with the majority being close to fixation in the progenitor sample group (36/38 SNPs in BC odd) but switching to near fixation for the alternate allele in the introduced sample group (34/38 SNPs in GL odd). However, there were two genes, one upstream (*LOC124009266*, an uncharacterized long non-coding RNA) and one downstream (*LOC124009839*, transcription cofactor HES-6-like) that overlapped the same window of elevated F_{ST} occupier by *per2* (Figure 4d). Both loci had lower overall mean F_{ST}, (0.50 at 177 SNPs) and 0.55 (1 SNP), respectively, when all SNPs in each locus were considered.

Lastly, we used an agent-based model to determine whether the allele frequency differences between BC odd and GL odd for the SNPs located within *per2* could be explained by genetic drift alone. As an illustrative example, we first used a *per2* SNP with an allele frequency of 0.033 in the BC odd sample group. The empirically observed allele frequency at this locus in GL odd was 0.933. After simulating the demographic effects of the population bottleneck 1000 times (parameterized by empirically estimated *Ne* estimates through time; see Methods) we found that 26.2% of simulations fixed at 0, while the upper 95% quantile occurred at an allele frequency 0.298 (Figure 5a). None of the 1000 simulations resulted in allele

frequency [3] 0.933, which was the observed allele frequency in the Great Lakes. Because we parameterized our agent-based model with conservatively low estimates of effective population sizes, these results strongly suggest that a response to selection imposed by the new environment drove the observed increase in allele frequency. We next calculated 95% confidence intervals around our empirical estimates of allele frequencies from both BC odd and GL odd that included both the effects of genetic drift and sampling error (Figure S9, Figure 5b) for all 38 SNPS found in *per2*. This approach also allowed us to calculate the probability that the change in allele frequency for each *per2* SNP was driven by drift alone (Figure S9). A total of 32 out of 38 SNPs had a p-value less than 0.05 (Figure 5b, Figure S9).

To assess the driver of putative selection at *per2*, which is closely linked to determining on organism's day length, we plotted daylength change using phenological data for the Great Lakes from (Bagdovitz et al., 1986) and daylight hours calculated using the R package geosphere v.1.5-18 (Hijmans et al., 2022). We found that spawning corresponded to a period with very similar day lengths, whereas emergence and outmigration occurred in one of the more extreme windows of deviation (0.75 – 1 hours less daylight) in the introduced range (Figure 5c).

Discussion

The mechanisms that allow populations to rapidly adapt from small pools of standing genetic variation has broad evolutionary and conservation implications. Here, we present results showing a severe bottleneck—representing a more than 2040-fold reduction in effective population size and loss of 37.2% of SNPs—during the introduction of a non-native fish into the Great Lakes. We also present evidence for numerous regions across the genome that potentially aided in the rapid genetic adaptation of pink salmon to the Great Lakes. For loci under putative selection, we find seven genes with one or more missense variants. One of those genes, *per2*, which is a period family gene with a strong effect on daily circadian patterns, displayed a near total switch from the reference alleles in British Columbia to the alternate alleles at 34/38 SNPs in the Great Lakes. Moreover, this response to selection matches large day length changes between the native and non-native habitat that overlap with important phenological periods in the pink salmon life history (Figure 5c). These results provide evidence for how pink salmon could have rapidly adapted to the Great Lakes, despite the large effects of genetic drift.

Our results suggest a 2040-fold reduction in effective population size, which resulted in 8.2% decrease in genome-wide heterozygosity (BC odd = 0.279,GL odd = 0.256) and a 37.2% loss of SNPs. This pattern matches a similar species, rainbow trout (*O. mykiss*), which was also introduced to the Great Lakes in the last century (Willoughby et al., 2018). That work described a 9.49% reduction in genome-wide genetic diversity (using pooled heterozygosity) approximately 25 generations after introduction, a similar time scale to this study. Another study, which used allozymes from pink salmon in the Great Lakes, found a loss of polymorphisms in 11 loci relative to the progenitor 12 generations after introduction (Gharrett & Thomason, 1987). Indeed, populations that are successful in non-native environments are expected to have lower genetic diversity than their founders (Allendorf et al., 2022; Barrett & Kohn, 1991), except in cases where multiple introduction events contribute to a more diverse admixture, potentially leading to successful adaptation on the back of this admixed genetic diversity (Gomez-Uchida et al., 2018; Kolbe et al., 2004, 2014).

The census population size for pink salmon in the Lakelse system (BC odd and even) was as high as 1.5 million individuals as recently as the 1980s, the ratio of N_e/N_c for our GONE estimates at introduction would be ˜10% using 146,886 as the effective population size for the progenitor at the time of introduction. Frankham (1995) suggested that a typical ratio between the two metrics in wild populations was around 10%, but that was later amended by Waples (2002) to be closer to 20%. Using either benchmark, the results from GONE closely match these predicted estimates. This conclusion is also supported by work in Chinook salmon (*O. tshawytscha*), which found ratios that were generally 5% or greater (Shrimpton & Heath, 2003) where the smallest ratio estimates were in populations with the largest census population size.

While we observed substantial decreases in genetic diversity, our data likely do not reflect the full pool of genetic diversity from the founding population, and we hypothesize that more genetic diversity

may have been lost than our results indicate. Historically, both the odd and even spawning populations in the Lakelse River (BC odd and even) were quite large, reaching as high as 1.5 million adults returning in a single year (Skeena Fisheries Commission, 2003). Additionally, while the founding population may have only been around 21,000 juveniles, they were part of larger introduction effort in Arctic Canada that produced nearly 750,000 juveniles from ˜500 dam and sire pairs (Gharrett & Thomason, 1987). Furthermore, observed heterozygosity is not an ideal metric for the loss of genetic diversity resulting from a bottleneck, as rare alleles are much more prone to be lost than heterozygosity (Allendorf, 1986, 2017). As such, we also show a loss of 37.2% of SNPs. Our GONE results indicate large declines in effective population size of the founding population, which might indicate the genetic diversity in our contemporary progenitor population does not fully reflect the standing genetic variation present at the time of introduction. The effect of the genetic diversity in founder populations remains understudied as a mechanism promoting persistence in non-native species and could be an area of focus in future work forecasting invasion risk (North et al., 2021).

We identified 4 significantly differentiated genes that were shared across the multiple sample group comparisons and contained missense variants. One gene, *LOC124013183* (CD209 antigen-like protein A), contained three missense variants, two of which were only 11 base pairs apart. This gene is predicted to play a role in immune response, specifically upstream or within regulation of T cell proliferation. Numerous studies suggest *CD209a* interacts with immunological defenses against macrophages (Lu et al., 2012, p. 209) and schistosomiasis (Kalantari et al., 2018; Ponichtera et al., 2014; Ponichtera & Stadecker, 2015) in mouse models (*Mus musculus*), suggesting it may play a role in combating novel pathogens in the introduced environment. Another gene with a single missense variant, *gnrhr4* (gonadotropin releasing hormone receptor 4), enables the release of the hormone gonadotropin and may contribute to gonadal maturation (Corchuelo et al., 2017). Pink salmon in the Great Lakes have broken the obligate two-year life cycle displayed by fish in their native range (Kwain & Chappel, 1978; Kwain & Lawrie, 1981). While our results are too preliminary to suggest if this gene plays a role in maturity timing of pink salmon in the Great Lakes, it is worth further investigation if this locus might contribute to these differences. Meaningful changes in the allele frequencies of these loci match the rapid, large changes documented by Gharrett & Thomason (1987), who found rare alleles in allozymes in the progenitor population transitioned to very common in populations in the Great Lakes after 12 generations.

The locus with the strongest evidence for putative selection was *per2* on chromosome 22. In addition to containing a missense variant, nearly all SNPs within the gene transitioned from near fixation for the reference allele in the native range, to near fixation for the alternate allele in the introduced range (Figure 5) and had much higher mean F_{ST} than genes in the same window. Similarly, this corresponded with local negative Tajima's D values as well as reduced local heterozygosity, consistent with selection in both the native and introduced range. Research into *per2* in other organisms indicates the gene as a core member of period genes regulating the circadian clock (Albrecht et al., 2007). In particular, the differential phosphorylation of the *per2* gene by casein kinase 1 (CK1) alters daily circadian period length in other organisms (Albrecht, 2007; Narasimamurthy et al., 2018), however allele frequencies within and around the homolog for CK1 were largely unchanged between BC odd and GL odd (Figure S10). A study in Alaskan sockeye salmon (*O. nerka*) suggests local water temperature has a large effect on emergence date and may vary substantially from year-to-year or occur over many months indicating these differences may be even more compounded from one year to the next (Sparks et al., 2019). It is worth noting the Steel River fish in the Great Lakes (our samples) spawn in essentially the northern-most habitat in the entire Great Lakes system available to pink salmon, and their entire lake phase is expected to occur south of that location. By contrast, pink salmon in the native British Columbian region of the Lakelse River would typically spend their ocean phase in the Gulf of Alaska, which may be many hundreds of kilometers farther north than their spawning location (Quinn, 2018). The spawning habitat of sample groups from this study are over 550 kilometers south of their native range spawning location. Research into the effect of *per2* in salmonids indicates the expression of *per2* became arrhythmic under unnaturally shorter or longer day periods in Atlantic salmon (*Salmo salar*) (Davie et al., 2009; McStay, 2012). Given the known effects of this gene in another salmonid, as well as in better described model systems, we hypothesize that putative selection in this gene is related to necessary circadian

changes to persist in the novel environment given the large day length changes experienced by salmon in the Great Lakes. While the exact mechanism driving this selection is unknown, ripe areas for future research might be changes around important phenological events, such as hatching or emergence, or daily behavioral trends of juveniles or subadults treating the lakes as surrogate oceans.

Conclusion

In this study, we show the genome-wide effects of rapid evolution resulting from both genetic drift and genetic adaptation in an introduced fish. Our results indicate rapid genetic adaptation is possible despite a more than 2040-fold reduction in effective population size and associated loss of genetic diversity. We provide evidence for at least 47 regions under putative selection, with 7 genes in those regions containing missense variants. Of those genes, we show simulated and observed data indicative of a strong response to selection in *per2* (period circadian protein homolog 2-like) on chromosome 22, which research in other organisms suggests plays a significant role in an individual's daily clock and matches considerable changes in day length between the progenitor and introduced populations. Finally, the combined result of genetic adaptation despite a significant bottleneck has important implications not only for how and when introduced species might colonize new habitat, especially invasive European pink salmon (Diaz Pauli et al., 2022; Sandlund et al., 2018), but also for populations or species at conservation risk that might experience similar demographic declines.

Acknowledgements

We thank J. Markovitz for her help with field sampling and the Upper Great Lakes Management Unit of the Ontario Ministry of Natural Resources and Forestry, especially F. Fischer, K. Rogers, and P. Drombolis, for their help in permitting, locating, and sampling salmon in Lake Superior tributaries. C. Pascal assisted with the preparation and transportation of native range DNA samples. Great Lakes DNA samples were prepared at the Purdue Genomics Core by P. San Miguel and V. Krasnyanskaya and sequenced by the staff at the Center for Medical Genomics at the Indiana School of Medicine. We also thank T. Beacham and Fisheries and Oceans Canada for the previous collection efforts of salmon in their native range. K. Christensen provided useful discussion of the pink salmon assembly and genomic features. Finally, we thank T. Kennedy for discussion of aging salmon using fin rays and his help as independent reviewer for our samples. M. Sparks was supported in part by the Purdue Ecology and Evolutionary Biology Waser Fellowship. This project was funded by the Department of Biological Sciences and National Science Foundation grant (DEB-1856710) to M. Christie.

References:

Albrecht, U. (2007). Per2 has time on its side. *Nature Chemical Biology*, *3*(3), https://doi.org/10.1038/nchembio0307-139

Albrecht, U., Bordon, A., Schmutz, I., & Ripperger, J. (2007). The multiple facets of per2. *Cold Spring Harbor Symposia on Quantitative Biology*, *72*, 95–104. https://doi.org/10.1101/sqb.2007.72.001

Allendorf, F. W. (1986). Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology*, *5*(2), 181–190.

Allendorf, F. W. (2017). Genetics and the conservation of natural populations: Allozymes to genomes. *Molecular Ecology*, *26*(2), 420–430. https://doi.org/10.1111/mec.13948

Allendorf, F. W., Funk, W. C., Aitken, S. N., Byrne, M., & Luikart, G. (2022). *Conservation and the genomics of populations*. Oxford University Press, USA.

Anas, R. E. (1959). Three-year-old pink salmon. *Journal of the Fisheries Research Board of Canada*, *16*(1), 91–94. https://doi.org/10.1139/f59-010

Bagdovitz, M. S., Taylor, W. W., Wagner, W. C., Nicolette, J. P., & Spangler, G. R. (1986). Pink salmon populations in the U.S. waters of Lake Superior, 1981–1984. *Journal of Great Lakes Research*, *12*(1), 72–81.

https://doi.org/10.1016/S0380-1330(86)71701-2

Barrett, R. D. H., Laurent, S., Mallarino, R., Pfeifer, S. P., Xu, C. C. Y., Foll, M., Wakamatsu, K., Duke-Cohan, J. S., Jensen, J. D., & Hoekstra, H. E. (2019). Linking a mutation to survival in wild mice. *Science*, *363*(6426), 499–504. https://doi.org/10.1126/science.aav3824

Barrett, S. C., & Kohn, J. R. (1991). Genetic and evolutionary consequences of small population size in plants: Implications for conservation. *Genetics and Conservation of Rare Plants*, 3–30. Oxford University Press, USA.

Beacham, T. D., McIntosh, B., MacConnachie, C., Spilsted, B., & White, B. A. (2012). Population structure of pink salmon *(Oncorhynchus gorbuscha)* in British Columbia and Washington, determined with microsatellites. Fishery Bulletin, *110*(2), 242-256.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Campbell-Staton, S. C., Arnold, B. J., Goncalves, D., Granli, P., Poole, J., Long, R. A., & Pringle, R. M. (2021). Ivory poaching and the rapid evolution of tusklessness in African elephants. *Science*, *374*(6566), 483–487.

Campbell-Staton, S. C., Cheviron, Z. A., Rochette, N., Catchen, J., Losos, J. B., & Edwards, S. V. (2017). Winter storms drive rapid phenotypic, regulatory, and genomic shifts in the green anole lizard. *Science*, *357*(6350), 495–498. https://doi.org/10.1126/science.aam5512

Chen, G.-B., Lee, S. H., Zhu, Z.-X., Benyamin, B., & Robinson, M. R. (2016). EigenGWAS: Finding loci under selection through genome-wide association studies of eigenvectors in structured populations. *Heredity*, *117*(1), https://doi.org/10.1038/hdy.2016.25

Christensen, K. A., Rondeau, E. B., Sakhrani, D., Biagi, C. A., Johnson, H., Joshi, J., Flores, A.-M., Leelakumari, S., Moore, R., Pandoh, P. K., Withler, R. E., Beacham, T. D., Leggatt, R. A., Tarpey, C. M., Seeb, L. W., Seeb, J. E., Jones, S. J. M., Devlin, R. H., & Koop, B. F. (2021). The pink salmon genome: Uncovering the genomic consequences of a two-year life cycle. *PLOS ONE*, *16*(12), e0255752. https://doi.org/10.1371/journal.pone.0255752

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics*, *3*, 35.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92.

Colautti, R. I., & Lau, J. A. (2015). Contemporary evolution during invasion: Evidence for differentiation, natural selection, and local adaptation. *Molecular Ecology*, *24*(9), 1999–2017. https://doi.org/10.1111/mec.13162

Conserving Lakelse fish and their habitat. (2003). 1-65, Skeena Fisheries Commission.

Corchuelo, S., Martinez, E. R. M., Butzge, A. J., Doretto, L. B., Ricci, J. M. B., Valentin, F. N., Nakaghi, L. S. O., Somoza, G. M., & Nobrega, R. H. (2017). Characterization of Gnrh/Gnih elements in the olfacto-retinal system and ovary during zebrafish ovarian maturation. *Molecular and Cellular Endocrinology*, *450*, 1–13. https://doi.org/10.1016/j.mce.2017.04.002

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. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

Davie, A., Minghetti, M., & Migaud, H. (2009). Seasonal variations in clock-gene expression in Atlantic Salmon (*Salmo salar*). *Chronobiology International*, *26*(3), 379–395. https://doi.org/10.1080/07420520902820947

Diaz Pauli, B., Berntsen, H. H., Thorstad, E. B., Homrum, E. i, Lusseau, S. M., Wennevik, V., & Utne, K. R. (2022). Geographic distribution, abundance, diet, and body size of invasive pink salmon (*Oncorhynchus gorbuscha*) in the Norwegian and Barents Seas, and in Norwegian rivers. *ICES Journal of Marine Science*, fsac224. https://doi.org/10.1093/icesjms/fsac224

Ellner, S. P., Geber, M. A., & Hairston Jr, N. G. (2011). Does rapid evolution matter? Measuring the rate of contemporary evolution and its impacts on ecological dynamics. *Ecology Letters*, *14*(6), 603–614.

Frankham, R. (1995). Effective population size/adult population size ratios in wildlife: A review. *Genetics Research*, *66*(2), 95–107. https://doi.org/10.1017/S0016672300034455

Fraser, D. J., Weir, L. K., Bernatchez, L., Hansen, M. M., & Taylor, E. B. (2011). Extent and scale of local adaptation in salmonid fishes: Review and meta-analysis. *Heredity*, *106*(3), 404–420.

Gharrett, A. J., & Thomason, M. A. (1987). Genetic changes in pink salmon (*Oncorhynchus gorbuscha*) following their introduction Into the Great Lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, *44*(4), 787–792. https://doi.org/10.1139/f87-096

Gordeeva, N. V., & Salmenkova, E. A. (2011). Experimental microevolution: Transplantation of pink salmon into the European North. *Evolutionary Ecology*, *25*(3), 657–679. https://doi.org/10.1007/s10682-011-9466-x

Hendry, A. P., & Kinnison, M. T. (1999). The pace of modern life: Measuring rates of contemporary microevolution. *Evolution*, *53*(6), 1637–1653. https://doi.org/10.2307/2640428

Hijmans, R. J., Karney (GeographicLib), C., Williams, E., & Vennes, C. (2022). *geosphere: Spherical Trigonometry* (1.5-18). https://CRAN.R-project.org/package=geosphere

Hof, A. E. van't, Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C., & Saccheri, I. J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, *534*(7605), 102–105. https://doi.org/10.1038/nature17951

Kalantari, P., Morales, Y., Miller, E. A., Jaramillo, L. D., Ponichtera, H. E., Wuethrich, M. A., Cheong, C., Seminario, M. C., Russo, J. M., Bunnell, S. C., & Stadecker, M. J. (2018). CD209a synergizes with dectin-2 and mincle to drive severe Th17 cell-mediated schistosome egg-induced immunopathology. *Cell Reports*, *22*(5), 1288–1300. https://doi.org/10.1016/j.celrep.2018.01.001

Kardos, M., Akesson, M., Fountain, T., Flagstad, O., Liberg, O., Olason, P., Sand, H., Wabakken, P., Wikenros, C., & Ellegren, H. (2018). Genomic consequences of intensive inbreeding in an isolated wolf population. *Nature Ecology & Evolution*, *2*(1), Article 1. https://doi.org/10.1038/s41559-017-0375-4

Koch, J. D., & Quist, M. C. (2007). A technique for preparing fin rays and spines for age and growth analysis. *North American Journal of Fisheries Management*, *27*(3), 782–784. https://doi.org/10.1577/M06-224.1

Kolbe, J. J., Glor, R. E., Rodriguez Schettino, L., Lara, A. C., Larson, A., & Losos, J. B. (2004). Genetic variation increases during biological invasion by a Cuban lizard. *Nature*, *431*(7005), Article 7005. https://doi.org/10.1038/nature02807

Kwain, W. (1982). Spawning behavior and early life history of pink salmon (*Oncorhynchus gorbuscha*) in the Great Lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, *39*(10), 1353–1360. https://doi.org/10.1139/f82-182

Kwain, W., & Chappel, J. A. (1978). First evidence for even-year spawning pink salmon, *Oncorhynchus gorbuscha*, in Lake Superior. *Journal of the Fisheries Research Board of Canada*, *35*, 1373–1376.

Kwain, W., & Lawrie, A. H. (1981). Pink salmon in the Great Lakes. *Fisheries*, *6*(2), 2–6. https://doi.org/10.1577/1548-8446(1981)006<0002:PSITGL>2.0.CO;2

Lamichhaney, S., Han, F., Webster, M. T., Andersson, L., Grant, B. R., & Grant, P. R. (2018). Rapid hybrid speciation in Darwin's finches. *Science*, *359*(6372), 224–228.

Lee, C. E. (2002). Evolutionary genetics of invasive species. *Trends in Ecology & Evolution*, *17*(8), 386–391. https://doi.org/10.1016/S0169-5347(02)02554-5

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997, https://arxiv.org/abs/1303.3997v2

Little, D., MacLellan, S. E., & Charles, K. (2012). *A guide to processing fin-rays for age determination* (*No. 3002; Canadian Technical Report of Fisheries and Aquatic Sciences*). Department of Fisheries and Oceans.

Liu, X., & Fu, Y.-X. (2020). Stairway Plot 2: Demographic history inference with folded SNP frequency spectra. *Genome Biology*, *21*(1), 1–9.

Lowe, S., Browne, M., Boudjelas, S., & De Poorter, M. (2000). 100 of the world's worst invasive alien species: A selection from the global invasive species database (Vol. 12). Invasive Species Specialist Group Auckland.

Lu, X.-J., Chen, J., Yu, C.-H., Shi, Y.-H., He, Y.-Q., Zhang, R.-C., Huang, Z.-A., Lv, J.-N., Zhang, S., & Xu, L. (2012). LECT2 protects mice against bacterial sepsis by activating macrophages via the CD209a receptor. *Journal of Experimental Medicine*, *210*(1), 5–13. https://doi.org/10.1084/jem.20121466

McCartney, G., Hacker, T., & Yang, B. (2014). Empowering faculty: A campus cyberinfrastructure strategy for research communities. *Educause Review*. https://er.educause.edu/articles/2014/7/empowering-faculty-a-campus-cyberinfrastructure-strategy-for-research-communities

McStay, E. (2012). Photoperiod regulation of molecular clocks and seasonal physiology in the Atlantic salmon (*Salmo salar*). http://hdl.handle.net/1893/11012

Melstrom, R. T., & Lupi, F. (2013). Valuing recreational fishing in the Great Lakes. *North American Journal of Fisheries Management*, *33*(6), 1184–1193.

Mills, E. L., Leach, J. H., Carlton, J. T., & Secor, C. L. (1994). Exotic species and the integrity of the Great Lakes. *BioScience*, *44*(10), 666–676.

Narasimamurthy, R., Hunt, S. R., Lu, Y., Fustin, J.-M., Okamura, H., Partch, C. L., Forger, D. B., Kim, J. K., & Virshup, D. M. (2018). CK1δ/ε protein kinase primes the PER2 circadian phosphoswitch. *Proceedings of the National Academy of Sciences*, *115*(23), 5986–5991. https://doi.org/10.1073/pnas.1721076115

Nei, M., Maruyama, T., & Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations. *Evolution*, 1–10.

North, H. L., McGaughran, A., & Jiggins, C. D. (2021). Insights into invasive species from whole-genome resequencing. *Molecular Ecology*, *30*(23), 6289–6308. https://doi.org/10.1111/mec.15999

Parsons, J. W. (1973). *History of salmon in the Great Lakes, 1850-1970* (Vol. 68). US Bureau of Sport Fisheries and Wildlife.

Perdry, H., Dandine-Roulland, C., Bandyopadhyay, D., & Kettner, L. (2020). gaston: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models (1.5.7). https://CRAN.R-project.org/package=gaston

Ponichtera, H. E., Shainheit, M. G., Liu, B. C., Raychowdhury, R., Larkin, B. M., Russo, J. M., Salantes, D. B., Lai, C.-Q., Parnell, L. D., Yun, T. J., Cheong, C., Bunnell, S. C., Hacohen, N., & Stadecker, M. J. (2014). CD209a expression on dendritic cells is critical for the development of pathogenic Th17 cell responses in murine schistosomiasis. *The Journal of Immunology*, *192*(10), 4655–4665. https://doi.org/10.4049/jimmunol.1400121

Ponichtera, H. E., & Stadecker, M. J. (2015). Dendritic cell expression of the C-type lectin receptor CD209a: A novel innate parasite-sensing mechanism inducing Th17 cells that drive severe immunopathology in murine schistosome infection. *Experimental Parasitology*, *158*, 42–47. https://doi.org/10.1016/j.exppara.2015.04.006

Quinn, T. P. (2018). *The Behavior and Ecology of Pacific Salmon and Trout.* University of Washington Press. Seattle, Washington.

R Core Team. (2022). *R: A language and environment for statistical computing.* https://www.R-project.org/.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Reid, B. N., & Pinsky, M. L. (2022). Simulation-based evaluation of methods, data types, and temporal sampling schemes for detecting recent population declines. *Integrative and Comparative Biology. 62*(6), 1849-1863.

Sandlund, O. T., Berntsen, H. H., Fiske, P., Kuusela, J., Muladal, R., Niemelä, E., Uglem, I., Forseth, T., Mo, T. A., Thorstad, E. B., Veselov, A. E., Vollset, K. W., & Zubchenko, A. V. (2018). Pink salmon in Norway: The reluctant invader. *Biological Invasions*. https://doi.org/10.1007/s10530-018-1904-z

Santiago, E., Novo, I., Pardiñas, A. F., Saura, M., Wang, J., & Caballero, A. (2020). Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. *Molecular Biology and Evolution*, *37*(12), 3642–3653.

Schumacher, R. E., & Eddy, S. (1960). The appearance of pink salmon, *Oncorhynchus gorbuscha* (Walbaum), in Lake Superior. *Transactions of the American Fisheries Society*, *89*(4), 371–373. https://doi.org/10.1577/1548-8659(1960)89[371:TAOPSO]2.0.CO;2

Seeb, L. W., Waples, R. K., Limborg, M. T., Warheit, K. I., Pascal, C. E., & Seeb, J. E. (2014). Parallel signatures of selection in temporally isolated lineages of pink salmon. *Molecular Ecology*, *23*(10), 2473–2485. https://doi.org/10.1111/mec.12769

Sparks, M. M., Falke, J. A., Quinn, T. P., Adkison, M. D., Schindler, D. E., Bartz, K., Young, D., & Westley, P. A. H. (2019). Influences of spawning timing, water temperature, and climatic warming on early life history phenology in western Alaska sockeye salmon. *Canadian Journal of Fisheries and Aquatic Sciences*, *76*(1), 123–135. https://doi.org/10.1139/cjfas-2017-0468

Stern, D. B., & Lee, C. E. (2020). Evolutionary origins of genomic adaptations in an invasive copepod. *Nature Ecology & Evolution*, *4*(8), https://doi.org/10.1038/s41559-020-1201-y

Tarpey, C. M., Seeb, J. E., McKinney, G. J., Templin, W. D., Bugaev, A., Sato, S., & Seeb, L. W. (2017). Single-nucleotide polymorphism data describe contemporary population structure and diversity in allochronic lineages of pink salmon (*Oncorhynchus gorbuscha*). *Canadian Journal of Fisheries and Aquatic Sciences*, *75*(6), 987–997. https://doi.org/10.1139/cjfas-2017-0023

Picard Tools. (2020). http://broadinstitute.github.io/picard/ *Broad Institute.*

Turner, C. E., & Bilton, H. T. (1968). Another pink salmon (*Oncorhynchus gorbuscha*) in its third year. *Journal of the Fisheries Research Board of Canada*, *25*(9), 1993–1996. https://doi.org/10.1139/f68-176

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

Wagner, W. C., & Stauffer, T. M. (1980). Three-year-old pink salmon in Lake Superior tributaries. *Transactions of the American Fisheries Society*, *109*(4), 458–460. https://doi.org/10.1577/1548-8659(1980)109<458:TPSILS>2.0.CO;2

Waples, R. S. (2002). Definition and estimation of effective population size in the conservation of endangered species. *Population Viability Analysis*, 147–168.

Waples, R. S. (2022). What Is Ne, Anyway? *Journal of Heredity*, *113*(4), 371–379. https://doi.org/10.1093/jhered/esac023

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.2307/2408641

Willoughby, J. R., Harder, A. M., Tennessen, J. A., Scribner, K. T., & Christie, M. R. (2018). Rapid genetic adaptation to a novel environment despite a genome-wide reduction in genetic diversity. *Molecular Ecology*, *27*(20), 4041–4051. https://doi.org/10.1111/mec.14726

Yin, X., Martinez, A. S., Perkins, A., Sparks, M. M., Harder, A. M., Willoughby, J. R., Sepulveda, M. S., & Christie, M. R. (2021). Incipient resistance to an effective pesticide results from genetic adaptation and the canalization of gene expression. *Evolutionary Applications*, *14*(3), 847–859.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Data Accessibility

Code used for this project will be released as a DOI enabled archive on Zenodo and raw sequence data will be made available on SRA at the acceptance of the manuscript.

Tables

Table 1. Year, habitat range, sample size after bioinformatic filtering, and location (latitude, longitude) of sample groups used in this study.

Sample group	Range	Year	Sample size	Location
BC odd	Native	2007	29	54.367763, -128.59778
BC even	Native	2006	29	54.367763, -128.59778
GL odd	Introduced	2019	30	48.777236, -86.886525
GL odd 3	Introduced	2019	15	48.777236, -86.886525
GL even	Introduced	2018	29	48.777236, -86.886525

Table 2. Pairwise mean estimates of Weir and Cockheram's F_{ST} for sample groups used in this study using a subset of 123,924 linkage-disequilibrium pruned SNPs.

	BC even	BC odd	GL even	GL odd
GL odd 3	0.1638	0.0850	0.0039	0.0006
GL odd	0.1620	0.0882	0.0046	

GL even	0.1610	0.0885
BC odd	0.0911	

Figure Captions

Figure 1. Map of study sites. Inset a) shows British Columbia, Canada, where the progenitor and its even year complement were collected from the Lakelse River (blue), a tributary of the Skeena River (pale blue). Inset b) shows Lake Superior in Canada (dark grey) and the United States (light grey). Pink salmon were introduced into the Current River, Ontario, Canada (pale blue) on the western edge of Lake Superior. Fish for this project were sampled in the Steel River (blue) on the north shore of Lake Superior. Inset c) shows a male pink salmon captured in the Steel River for this project (Photo M. Sparks).

Figure 2. Measures of genetic diversity for introduced pink salmon in the Great Lakes, as well as the odd-year progenitor and its complementary even-year sample group. Panel a) shows the total number of SNPs in BC odd (blue) and GL odd (orange) in 2.5 Mbp windows across chromosomes. Panel b) shows observed heterozygosity for individuals from the five sample groups used in this study as raw data (points), density distributions, and box and whisker plots (middle line indicates the mean, outer boxes the 25% and 75% quartiles, and the whiskers 1.5*inter quartile range). Panel c) shows the number of SNPs present in each sample group's vcf. For this figure, data colored blue indicates the progenitor samples (BC odd), green it's even year complement (BC even), and orange the introduced, Great Lakes groups (orange = 2-yr odd year group, dark orange = 3-yr odd group, light orange = 2-yr even group). Notice that genetic diversity, particularly when measured as the number of SNPs, is substantially reduced in the introduced Great Lakes sample groups relative to British Columbia.

Figure 3. Measures of demography for pink salmon in the Great Lakes and their progenitor. Panel a) shows Tajima's D in 2.5 Mbp windows across chromosomes for pink salmon introduced to the Great Lakes (2-yr odd year group) and the progenitor stock. Positive genome-wide Tajima's D is indicative of a sudden population contraction (colors match those in legend in panel, b). Panels b-d show effective population size for those groups as calculated from GONE, a linkage disequilibrium method. Two panels (b,c) show the same data but differentially subset on the y-axis to better visualize the results for the respective groups. The dashed lines in b and c show the approximate timing of the introduction event. Panel d shows the mean and 95% quantile for samples in the year the samples were collected (i.e., contemporary estimates). For this figure, data colored blue indicates the progenitor samples (BC odd), green it's even year complement (BC even), and orange the introduced Great Lakes groups (orange = 2-yr odd year group, light orange = 2-yr even group).

Figure 4. Indices of putative selection for pink salmon introduced into the Great Lakes; analyses are based on a comparison of BC odd to GL odd. Panel a), and b) show 100 kbp outlier windows calculated using ZF_{ST} and eigenGWAS with outliers indicated as red points. Shaded rectangles either represent windows with missense variants (grey) represented in panel c) or the period circadian protein homolog 2-like gene (*per2;* blue) represented in panel d). Panel c shows raw F_{ST} for the windows containing two genes with missense variants shared between comparisons between BC odd and GL odd a). Genes are represented in the top of each figure as rectangles with black (coding sequence) or transparent (non-coding sequence) segments. The gene affected by the missense variant is represented by the grey rectangle, and the missense variant is shown with a red line. Panel d) shows a similar representation of the window containing the *per2* locus (and

lpar5, which contains our other missense variant at the locus*)* along with Tajima's D and heterozygosity calculated in 10 kbp steps across the window, with lines representing the introduced (orange) and progenitor (blue) sample groups used in the methods for panels a) and b).

Figure 5. Simulated and observed data indicating genetic drift, and a possible response to selection, in the period circadian protein homolog 2-like (*per2*-like) gene of chromosome 22. Panel a) shows the N_e values on a log scale used in our model paired with the results of 1000 simulations exploring the effect of genetic drift for a single SNP (starting allele frequency = 0.033). Values above the dashed line indicate the threshold the upper 95$^{\text{th}}$ percentile of final allele frequencies, with the frequency distribution of all values indicated to the right. Notice that no simulations ever exceeded the observed value, suggesting that the change in allele frequency was unlikely to occur by drift alone. The observed values for the progenitor (blue) and introduced sample group (orange) are shown with colored points. Panel b) shows simulated and observed allele frequencies for all SNPs within the *per2*-like gene. For the simulated portion, lines connect the SNPs observed empirically in the BC odd sample group with the maximum 95$^{\text{th}}$ percentile change observed in allele frequencies after simulating drift associated with the bottleneck at introduction (see panel a as an example). For the observed data, the lines simply connect the observed allele frequencies between BC and GL odd. In both, SNPs are connected by a line and the missense variant is shown in red. Solid lines in the observed values represent observed changes that were greater than the 95$^{\text{th}}$ percentiles of the simulated values suggesting changes were the result of selection (see Figure S9 for 95% CIs for every SNP. Notice that 32 of 38 SNPs, including the missense SNP, show a greater change in allele frequencies than is conservatively estimated by drift alone. Panel c) shows the daylight change experienced by the sample groups in the Great Lakes relative to the progenitor sample group over the course of a hypothetical lifecycle for 2-year-old fish in the Great Lakes. Important phenological events are indicated with colored polygons using approximate timing from Bagdovitz et al. (1986).

Hosted file

`PinkSalmonAdaptation_sub1.docx` available at https://authorea.com/users/594602/articles/628973-rapid-genetic-adaptation-to-a-novel-ecosystem-despite-a-massive-bottleneck

Figure 1

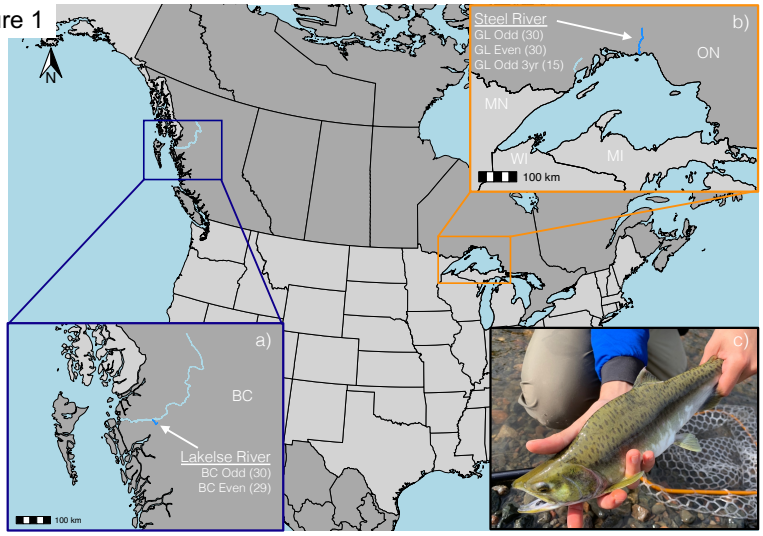

Figure 2

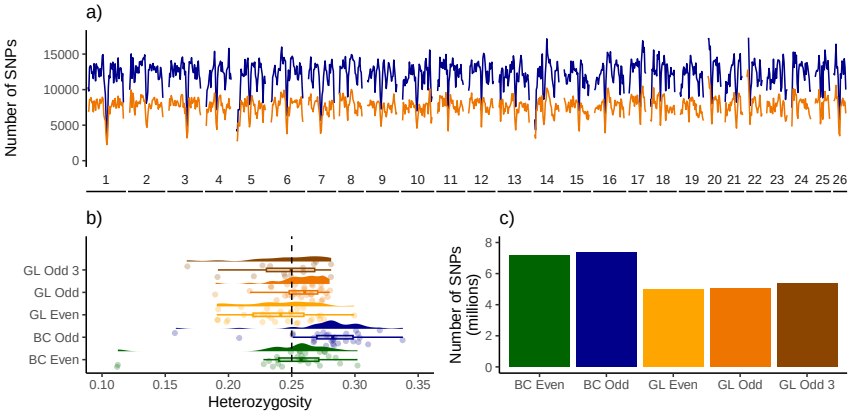

Figure 3

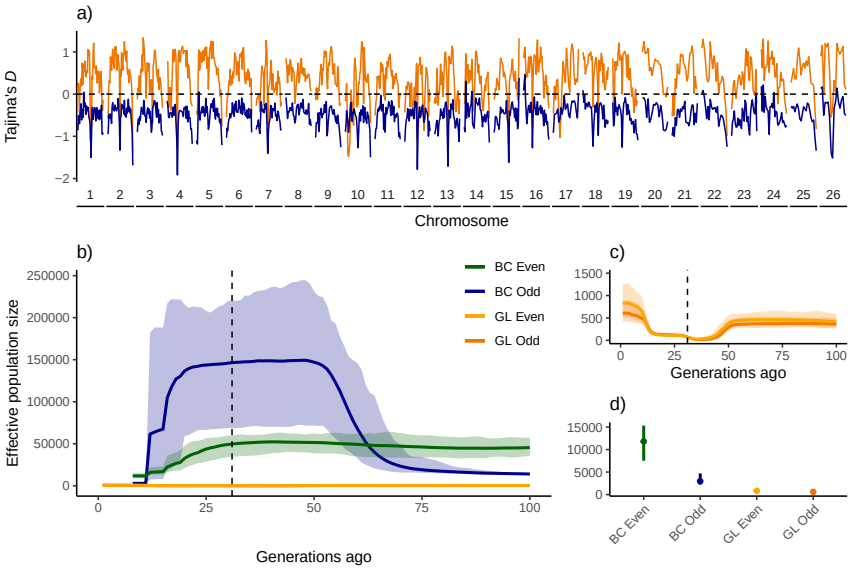

Figure 4

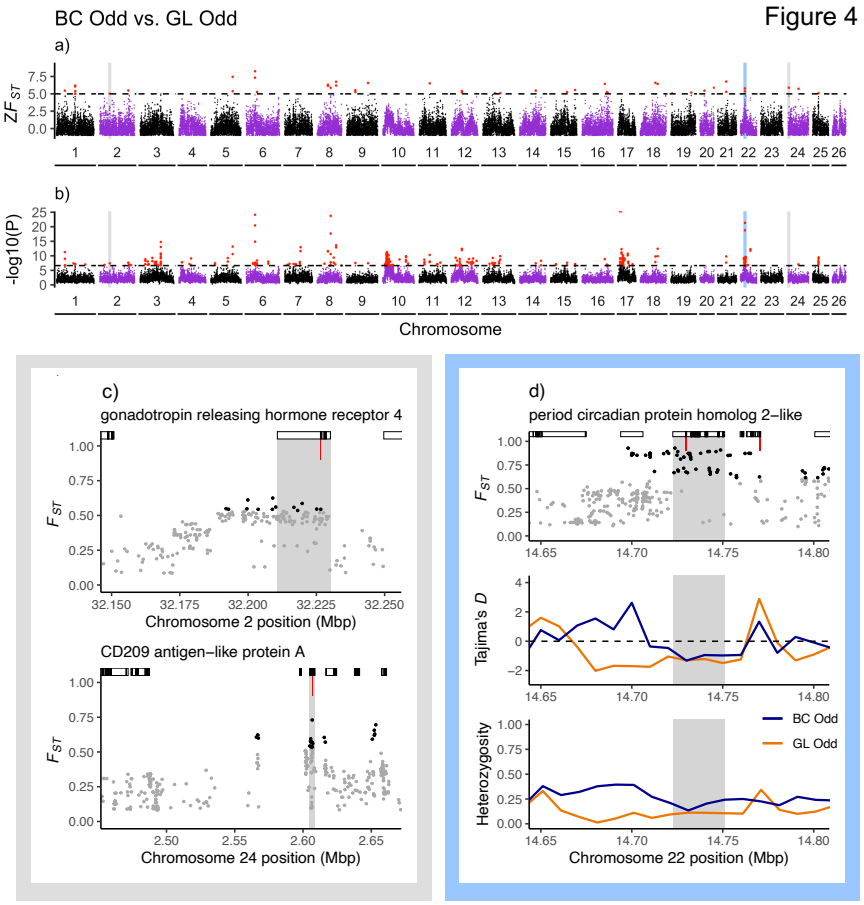

Figure 5