Genomic islands of divergence infer a phenotypic landscape in Pacific lamprey

Jon Hess[1], Jeramiah Smith[2], Nataliya Timoshevskaya[2], Cyndi Baker[3], Christopher Caudill[4], David Graves[5], Matthew Keefer[4], Andrew Kinziger[6], Mary Moser[7], Laurie Porter[1], Greg Silver[1], Steven Whitlock[8], and Shawn Narum[1]

[1]Columbia River Inter-Tribal Fish Commission
[2]University of Kentucky
[3]Oregon Department of Fish and Wildlife
[4]University of Idaho
[5]The Columbia River InterTribal Fish Commission
[6]Humboldt State University
[7]Northwest Fisheries Science Center, NOAA Fisheries
[8]University of Washington

April 28, 2020

Abstract

Pacific lamprey (Entosphenus tridentatus) is a culturally important and imperiled anadromous fish with a parasitic ocean phase. Biological uncertainties challenge restoration efforts and life-history research is needed to explain observed trait variation and inform management actions. Using two new whole genome assemblies and genotypes from 7,716 single nucleotide polymorphism (SNP) loci in 518 individuals from across the species range, we identified four large regions of high genomic divergence (on chromosomes 01, 02, 04, and 22). We genotyped a subset of 302 broadly distributed SNPs in 2,145 individuals for genotype-by-phenotype trait associations for adult body size, sexual maturity, migration distance and timing, adult swimming ability, and larval growth. Body size traits were strongly associated with SNPs on chromosomes 02 and 04. Moderate associations also implicated SNPs on chromosome 01 as being associated with variation in female maturity. Using genotypic frequencies of candidate SNPs for female maturity and body size, we extrapolated a heterogeneous spatiotemporal distribution of these traits based on independent datasets of larval and adult collections. These maturity and body size results guide future studies to validate these predicted phenotypic distributions across the geographic range and elucidate factors driving regional optimization of these traits for fitness.

Introduction

Highly dispersive species like Pacific lamprey (*Entosphenus tridentatus*) present an evolutionary conundrum for adaptation. Adaptation is facilitated when particular combinations of gene variants that confer optimal fitness in an environment can be passed on to the next generation. However, high rates of gene flow can impede inheritance of these optimal combinations of gene variants via the action of recombination. Yet there is evidence from Pacific lamprey and other dispersive species that local adaptation may occur despite these high rates of gene flow. For example, Pacific lamprey body size is correlated with upstream migration distance in the Columbia River (Keefer *et al* . 2009; Hess et al. 2014) and traits in other dispersive species appear to be optimized for specific environments within their broader range (Asaduzzaman et al. 2019, Miller *et al.* 2019, Phair*et al.* 2019).

Genomic architecture appears to be one factor that can influence local adaptation in highly dispersive species. In general, the closer two genes occur in the genome the smaller the chance for recombination events that may separate an optimal combination of variants (Yeaman and Whitlock 2011). Inversions resist recombination between inverted haplotypes and can effectively lock an optimal combination of variants together over longer distances within the inverted segment (sometimes referred to as a supergene); the fitness conferred by these inversions can help maintain them as a polymorphism in a population through both forces of balancing and divergent selection (Wellenreuther and Bernatchez 2018, Faria *et al.* 2019). In Pacific lamprey, if there are particular phenotypes that have a polygenic basis and confer differential fitness across environments, we might expect to identify long polymorphic intervals of DNA sequence.

Several traits in Pacific lamprey have been found to have a genetic basis. These include body size, reproductive migration-timing (Hess*et al.* 2014, 2015), and advanced maturity of females at onset of freshwater migration (i.e. ocean-maturing versus river-maturing ecotypes, Parker *et al.* 2019). There also appears to be evidence for statistical linkage of multiple loci that show high divergence in the species' range (Hess *et al.* 2013). One thing that is unclear is whether range-wide divergence that has been observed can be explained by phenotype-by-genotype associations reported thus far. Phenotypic traits are often interrelated, which can obscure the true target of selection (Powell and MacGregor 2011). Testing a large variety of phenotypic trait associations with genotypes at different sites in the species' range can help to disentangle these correlations and help elucidate the true target of selection. Once phenotype-by-genotype associations are confirmed across geographic sites, these relationships can be exploited to extrapolate a phenotype across large geographic areas in which only genotypes have been measured. This genetic tool then becomes a powerful predictor and can generate hypothesis testing frameworks to guide future studies aimed to validate these predicted phenotypic distributions across the range and elucidate factors driving regional optimization of these traits.

In this study we addressed four major objectives: 1) *Divergence mapping* : Test whether previously observed genomic divergence across the species' range is either concentrated or diffusely organized in the genome, 2) *Association testing* : Test phenotypic trait associations with genotypes across geographic sites to identify robust phenotype-by-genotype relationships, 3) *Association mapping* : Test whether phenotypic-by-genotype associations mapped to the genome can explain genomic divergence across the species' range, 4) *Extrapolation of spatiotemporal phenotypic distributions* : Use candidate SNP genotypic distributions across time and space to characterize the ecological niche of life history traits. Our findings supported a high concentration of genomic divergence to regions within four chromosomes, referred to as genomic islands. Two of these four genomic islands showed robust correlation with maturity and body-size traits and could be used to predict their spatiotemporal distributions across the species' range.

Methods

Divergence mapping

Two new Pacific lamprey genome assemblies were constructed using the whole genome sequence from the milt and blood from a male (representing the gametic and somatic genomes; Genbank Assession #: PRJNA613923) and the blood of a female (Genbank Assession #:XXXXX), and using a high density linkage map (Smith et al. 2018) to validate and extend higher order scaffolding of chromosomes (Supplemental Materials).

For characterization of SNP densities and F_{ST}statistics, we used a set of 7,716 unique SNP loci from previously published RAD-seq datasets (Hess *et al.* 2013; Smith *et al.*2018), which passed a set of population genetic quality control filters (Supplemental Materials). This set of 7,716 unique SNPs was a combination of overlapping groups of SNPs from a previous dataset (Hess*et al.* 2013; SNPs N = 8,772) and a *de novo* linkage mapping dataset (Smith *et al.* 2018; SNPs N = 7,977). BOWTIE2 (Langmead and Salzberg 2012) was used to align these two datasets to the male reference assembly to define homologous loci. For the 7,716 total SNPs passing the QC filters, 4,046 loci were unique to Hess *et al.*2013, 1,418 loci were unique to Smith *et al.* 2018, and 2,252 SNPs were shared across datasets. Marker positions based on BOWTIE2 alignments

were compared between Pacific lamprey male and female genomes and the Pacific lamprey male and sea lamprey male gametic genome (GenBank assembly accession: GCA_002833325.1) to characterize synteny.

Using these 7,716 SNPs genotyped for the same individuals from Hess et al. (2013; i.e., 16 collections with >20 individuals which totaled 482 individuals; Table S1), LOSITAN (Antao *et al.* 2008) was run using parameter settings of 50,000 simulations, confidence interval of 0.99, false discovery rate set to 0.1, subsample size of 20, simulated F_{ST} of 0.019 and an attemptedF_{ST} of 0.021. We considered loci candidates for positive selection above a probability level of 0.995, and neutral loci were defined as falling between the 10th and 90th quantiles of theF_{ST} distribution. Any remaining SNPs were conservatively considered undetermined (neither candidates nor neutral).

Genes located within adaptive regions were identified using published sea lamprey gene annotations that were found in the homologous regions corresponding to the following Pacific lamprey male genome positions 1) chromosome 01 positions: 8939466...14772759 (sea lamprey scaf_00003: 6777250...13554086), 2) chromosome 02 positions: 3351206...18794404 (sea lamprey scaf_00006:1198871-13859281), 3) chromosome 04 positions: 6408032...19202839 (sea lamprey scaf_00005: 2591251...16864119), and 4) chromosome 22 positions: 617460...11364740) (sea lamprey scaf_00012: 1160196...12993068). We used the website Enrichr (https://amp.pharm.mssm.edu/Enrichr/) (Kuleshov et al. 2016) to gain insights into the potential function of these genes via both the manifested phenotypes in mammals (i.e., MGI Mammalian Phenotype Level 4 2019) and in fishes (FishEnrichr; Phenotype AutoRIF Predicted Z score).

Association testing

Genotyping-in-thousands by sequencing (GT-seq, Campbell *et al.*2015) was employed to genotype 308 genetic markers for the association testing analyses. The GT-seq 308 loci were a subset of markers developed from the paired end consensus reads from the Hess *et al.* (2013) RAD-seq dataset. The selection of loci and steps in development are described in detail in Supplemental Materials. Locus selection began with a group of 457 total SNP loci considered in round 1, which included 120 that had been already designed for TaqMan assays (Hess *et al.*2015). Final optimization left 308 loci that worked best in GT-seq genotyping. For all samples used below in the association testing we filtered out individuals missing >10% of genotypes at the 308 loci. Excluding the four species diagnostic loci and two duplicated loci provided 302 unique loci for association tests.

There were six samples, five comprised of adults (JDD, S_BON, T_BON, WFA, and WFA) and one comprised of larvae (GAR), with which we performed association testing (Table S1). Adult samples were from the following three locations: males (WFA, N=136) and females (WFA, N=133) from Willamette Falls collected in 2016 (Willamette River, Oregon City, OR; 205.6 Rkm upstream from the Columbia River mouth), two samples (S_BON, N=295 and T_BON, N=883) from Bonneville Dam in 2014 (235.1 Rkm upstream from the Columbia River mouth), and one sample (JDD, N=656) from John Day Dam in 2014 and 2015 (346.9 Rkm upstream from the Columbia River mouth). The following five adult traits were measured on all adult samples: ordinal "day" of collection (timing of migration to the sample point), girth (mm), total "length" (mm), weight (g), and distance between dorsal fins ("interdorsal", mm). Interdorsal measurements have been suggested to serve as an indicator of maturation status in Pacific lamprey because the distance tends to decrease with maturation (Clemens *et al.* 2009). We measured an additional migration trait for three adult samples (S_BON, T_BON, and JDD) via a combination of passive integrated transponder (PIT) and radio tagging of individual fish and observing their furthest upstream detection from the release location ("Rkm"). Further, since the males and females collected at Willamette Falls (WFA and WFA) were being harvested, we were able to measure gonad weight as a proxy for maturity in those samples. Finally, a subset of the adult sample from Bonneville Dam (S_BON) was used in a swim trial experiment within a flume (Kirk*et al.* 2016), in which the following three swimming behavioral traits were measured: "approached" experiment, passed challenge ("pass"), and passed challenge without fallback ("passrep"). Details of these swimming performance experiments can be found in Kirk et al. (2016) and Supplemental Materials.

A single group of larvae were artificially propagated using adults captured at Bonneville Dam. These larvae

were reared in a common garden experiment to generate early larval growth ("GAR") rate data (N=337). All larvae were spawned in the spring of 2015 and allowed to rear from 30 to 163 days after hatching. Growth rate was measured as length / time ("growth"), and also corrected growth rate ["growth rate_b"; (length – 4 mm) / time] to correct for length at hatch (~4 mm).

Intercorrelation among all measured traits in these six samples (i.e. JDD, S_BON, T_BON, WFA, WFA, and GAR) was examined (based on Pearson's r) to avoid excessive redundancy of predictor variables ($|r| > 0.95$), and P -values were calculated (SAS Institute, Inc. 2000). We performed univariate analyses using a general linear model (GLM) and a mixed linear model (MLM) with TASSEL v. 5.1.0 (Bradbury *et al.* 2007). The GLM is a fixed effects linear model that is used in TASSEL to identify significant associations between phenotypes and genotypes. TASSEL takes population structure into account by using genetic principal coordinate axes as covariates in the model. The MLM is similar to GLM but includes both fixed effects (e.g. population structure, and genetic marker) and random effects (i.e., relationships among individuals) and can thus account for both population structure and kinship to reduce false positive associations (Yu *et al.* 2006). Details on the covariates and ways in which loci were used taking population structure and relatedness into account in the GLM and MLM tests are provided in the Supplemental Materials. To account for multiple tests, only those associations with P -values less than the critical value as determined using the false discovery rate procedure described by Benjamini and Hochberg (1995) were considered significant. The Benjamini and Hochberg (1995) false discovery rate approach has more power to detect significant differences than sequential Bonferroni correction (Narum 2006). Critical values were calculated using the function p.adjust within the R package stats (RDC Team 2019).

Association mapping

The 308 SNP loci on the GT-seq panel were aligned to reference genomes using BOWTIE2. There were 306 that were assigned to a single location on the Pacific lamprey male genome (99.4%), covering 70 different chromosomes with an average of 4.4 loci per chromosome (range 1 – 22). Marker locations were based on the alignments of marker sequences to the Pacific lamprey male and female genomes, homologous scaffolds of the sea lamprey genome, and positions on the previously published Pacific lamprey linkage map (Smith *et al.* 2018).

Adjusted P -values from the association testing described above were log transformed (-LOG10) and plotted by consensus genome position on the Pacific lamprey male genome. We tested correlation of association tests -LOG10(P) with F_{ST} from the rangewide divergence to understand whether trait associations may explain the high divergence observed at the rangewide scale for the subset of markers shared between datasets. Among the 308 SNPs, there were 230 neutral SNPs, 41 adaptive markers SNPs, and a set of 31 "intermediate" SNPs that did not fit definitions of putatively neutral and putatively adaptive (divergence mapping). Finally, four loci were species diagnostic (Hess *et al.* 2015), and 2 loci were duplicated. Therefore, there were 302 unique markers available for these association analyses. These markers included 38 SNPs that were mostly adaptive loci that were categorized into the following 4 groups of statistically linked loci: A (N=10), B (N=13), C (N=7), and D (N=8, Hess *et al.*2013).

Extrapolation of spatiotemporal phenotypic distributions

We characterized candidate SNP genotypic distributions across time and space to better understand the ecological niche of life history traits. These spatiotemporal distributions were characterized using the candidate SNPs with the most robust associations with body size (chromosome 02) and sexual maturity (chromosome 01). Distributions of representative SNPs of the other adaptive chromosomes (Chromosomes 04 and 22) were also characterized (Fig. S1, S2).

We used two independent datasets to characterize spatial and temporal distributions of genetic variation (Table S1). These datasets were independent of each other and separate from the association testing samples, and they were optimally suited for these characterizations. For the spatial dataset, we primarily used collections of larvae and juveniles (95% of dataset of N=3,435) but included some adult collections that were distributed widely across the species' range. Larvae and juveniles were the ideal life stage to represent

genotypic distributions of individuals that successfully spawned at discrete locations throughout the range. Adult collections were used to fill in portions of the range where larval samples were not available. Genotyping was partially conducted with a TaqMan assay panel (Hess*et al.* 2015), which overlapped the GT-seq panel by 85 SNPs they had in common. COLONY v. 2.0.6.5 (Jones and Wang 2010) was used to reconstruct full-sibling families (Wang 2004) using the 85 shared SNPs on each of the 70 collections. We analyzed all collections together as one using the following parameter settings: polygamous mating for males and females without inbreeding, full-likelihood, medium length of run, no allele updating, and no sibship priors. Only 1 collection out of the 7 adult collections had full siblings (N=13, Stamp River, B.C.) which were maintained to accurately represent this small spawning segment. We excluded duplicate genotypes, 797 full siblings, and collections with fewer than 5 individuals, resulting in a final set of 57 collections consisting of a total of 2,581 individuals each representing a unique family (Table S2). This dataset was then used to calculate allele frequencies across collections for the representative candidate SNPs Etr_464 and Etr_5713 within the adaptive regions on chromosomes 01 and 02, respectively.

For the temporal dataset, we used individuals collected from two successive spawning runs at Willamette Falls (2014 – 2015; N of 868 and 581, respectively) over which it was possible to randomly sample the majority of the annual adult migration of Pacific lamprey (typically Feb – August) in weekly strata. A daily abundance estimate (Whitlock*et al.* 2019) was used to expand candidate SNP allelic proportions in the weekly strata. One biological complexity was that a portion of the adults encountered before May probably overwintered and experienced shrinkage in body size due to advanced maturation (Beamish 1980). Therefore, in addition to characterizing allele frequencies of candidate SNPs Etr_464 and Etr_5713, we categorized fish by body length to provide insight into the transition between overwintered fish and newly-arrived migrants.

Results

Divergence mapping

Outlier analyses identified 311 (4.0%) SNPs as candidates for positive selection (out of a total of 7,716 SNPs; $P > 0.995$). LOSITAN was also used to identify neutral loci, which we defined using a conservative threshold range of probabilities between 0.10 and 0.90. There were 350 (4.5%) and 4 ($< 0.1\%$) SNPs below and above this range, respectively (i.e. candidates for balancing and positive selection, respectively), and 7,051 neutral loci (91.4% of 7716 loci) that fell within these probability levels.

A total of 7385 out of 7716 loci (95.7%) and 7366 out of 7716 loci (95.4%) aligned to the Pacific Lamprey female and male genome assemblies, respectively, and 4916 out of 7716 loci (63.7%) aligned to the male gametic sea lamprey genome. The alignment to the Pacific lamprey male genome was used to order the loci by scaffold position, and in cases in which only alignments to the other assemblies were available we interpolated values to estimate relative positions. Manhattan plot was used to visualize the distribution of the outlier SNPs in both the Pacific lamprey and sea lamprey male genome assemblies (Fig. 1), and alignments were generated between Pacific lamprey male and female genomes (Fig. S3). These results illustrated that 65% of the outlier loci are localized to each of the following four chromosomes:01, 02, 04, and 22; which share homology with sea lamprey chromosomes 03, 06, 05, and 12, respectively (Fig. 1). The patterns of synteny within these four chromosomes indicated large regions of inversions that overlapped with concentrations of outlier SNPs (e.g., chromosomes 01 and 02, Fig. S3) and may be polymorphic within Pacific lamprey given the differences between male and female genome assemblies. The same inversion patterns on chromosomes 01 and 02 were present between species (male assemblies, Fig. S3).

Association testing

Examination of the intercorrelation of predictor variables indicated that many of the morphological variables related to body size attributes were highly correlated, however, none had a significant correlation above 0.95 (Table S3 – S8), and therefore all were retained for association analysis. Significant intercorrelations among traits were consistent across samples and using these correlations we categorized traits into the following four main groups: 1) body size, 2) female sexual maturity, 3) larval growth, and 4) swimming ability. The "Body size" category included the body metrics of length, weight, girth, and interdorsal, which were all significantly

positively correlated across datasets (Table S3 – S8). This trait category also included migration distance which was positively correlated with increasing body size metrics, and migration timing which was negatively correlated with these body size metrics. The male gonad mass was intended to serve as a proxy for male sexual maturity, but it also was significantly positively correlated with other body size metrics. However, the female gonad metric was not correlated with the other body size metrics and was potentially an accurate proxy for "female sexual maturity". The "larval growth" category was populated by the only two measures of growth in the common garden experiment. The "swimming ability" category contained the three metrics of swimming performance which were all significantly positively correlated, and included migration day, which was positively correlated.

We examined the relative strength of associations of the total 302 SNPs with the traits within each of the four trait categories (Table S9), but were primarily interested in associations of SNPs on the four chromosomes chr01, chr02, chr04 and chr22, where evidence for range-wide adaptive divergence was concentrated.

Body size traits

Highly significant associations (adjusted $P < 0.001$) were observed for body size traits and SNPs on the four adaptive regions across datasets (WFA, WFA, T_BON, S_BON, and JDD; Table 1). The strongest associations between SNPs and body size traits (-LOG(10P) > 30) were with length (Fig. 2), weight, and girth and SNPs on chromosome 02 for the T_BON dataset; however, the S_BON and JDD datasets also showed significant associations for the same SNPs and traits (-LOG(10P) > 2). This result was similar to the findings of Hess *et al.* (2014) where they show that a SNP (Etr_5317), herein mapped to chromosome 02, had strongest association to body size traits at Bonneville Dam. In contrast, the WFA and WFA datasets had fewer total significant associations with body size traits across the four adaptive linkage groups. The WFA and WFA datasets significant associations to body size traits were concentrated on chromosome 02 and chromosome 04, and of these two linkage groups, chromosome 04 appeared to have the strongest associations with body size traits, and primarily with length and weight. For WFA, similar to the length and weight traits, male gonad size was also associated with chromosome 04 and chromosome 02 SNPs, which likely owes to the high intercorrelation observed among these traits. Overall, the results support strong association of SNPs with body size traits, primarily length and weight, compared to other intercorrelated traits (e.g., migration timing and migration distance). The chromosome that showed the highest association with body size was chromosome 02 at the BON and JDD sites. This chromosome 02 association with body size was consistent among sites both within and outside the Columbia River basin (Parker*et al.* 2019).

The genotypes in the chromosome 02 adaptive region (SNP Etr_5317) in the T_BON sample were also predictive of average lengths, such that the average size of homozygotes for large body size alleles "AA", heterozygotes "AC", and homozygotes for small body size alleles "CC" were 677 mm, 627 mm, and 592 mm, respectively (Fig. 3b). Although the average body sizes differed across sites and sexes (WFA were larger on average than WFA), the trends were consistent (Fig. 3b). Further, similar genotype and average length associations have been detected in Pacific lamprey collected from the Klamath River (634 mm, 602 mm, and 557 mm for the AA, AC, and CC genotypes at Etr_5317, respectively; Parker et al. 2019).

Female sexual maturity

Female gonad size was significantly associated with chromosome 01 and none of the other three adaptive regions in the WFA collection (Table 1, Fig. 2). This finding is concordant with Parker et al. (2019) who also resolved a significant association between female gonad size and chromosome 01. The average gonad sizes associated with genotypes at candidate SNP Etr_464 (chromosome 01) were 25 g, 20g, and 18g for the AA, AC, and CC genotypes, respectively (Fig. 3a). While gonad mass was less in the Klamath River collection, categorizations by genotype were consistent (average egg mass of 13 g, 7 g, and 6 g for the AA, AC, and CC genotypes at Etr_464, respectively; Parker et al. 2019).

Swimming ability and Larval Growth rate

No significant associations were observed on the four adaptive linkage groups or any of the other chromosomes

for the short-term swimming performance trials or the larval growth rates (Table 1).

Genotypic prediction of phenotypic traits and potential gene-interaction effects

In most of the trait associations, the representative candidate SNPs chosen on the four adaptive chromosomes (Etr_464, Etr_5317, Etr_1806, and Etr_4281 on chromosomes 01, 02, 04, and 22, respectively) represented above average genotype-by-phenotype associations of all 34 of the significant SNPs on these four chromosomes. In many cases, these four SNPs lie at the extreme end of the range of observed P values (Table S9).

Parker *et al.* (2019) found evidence for epistatic interactions that involved loci on chromosomes 01 and 04 (referred to previously as linkage groups D and B, respectively), which were found to represent the model with highest predictive ability for the female maturity trait (or ocean- and river-maturing ecotypes). We used single SNP locus representatives for each chromosome and conducted gene interaction tests for the maturity and body-size candidate loci following Parker et al. (2019, Supplemental Materials). For both traits, the model with highest support was a single locus model such that Etr_464 (chromosome 01) and Etr_5713 (chromosome 02) were the loci with highest predictive ability for the female maturity and adult total length traits, respectively (Table S10).

Overlap of genomic divergence and association mapping

The range-wide F_{ST} values that were mapped to the male Pacific lamprey genome were plotted by genomic position with the subset of SNPs used in the association testing for the two traits with consistent strong associations (i.e. adult total body length as measured at Bonneville Dam "T_BON" and female gonad size as measured at Willamette Falls "WFA"; Fig. 2a,b). For chromosomes 01 and 02, the adjusted -LOG10(P) values from the association tests were highly correlated with the genomic divergence as measured by F_{ST} for the female gonad size and adult body size traits, respectively (Fig. 2a,b). We quantified the overlap of genomic divergence and trait association by regressing range-wide F_{ST} and the adjusted -LOG10(P) values for the 302 SNPs in the GT-seq panel. These 302 SNPs show positive linear trends for both traits, but the linear trends with highest slope and R^2 were observed for SNPs on chromosomes 01 and 02 for the gonad size and body length traits, respectively (Fig. S4). These results suggest that the high geographic divergence exhibited by SNPs on chromosomes 01 and 02 may be related to selection on the traits with which these same SNPs are highly associated.

Functions enriched within adaptive genomic regions

The mammalian phenotype terms showed some significant tests for enrichment based on the overlap of annotated genes from the four chromosome regions (Table S11). The top three terms with lowest P-value from our list of 98 candidate genes on chromosome 01 (which is associated with maturity) were abnormal social investigation, abnormal blood urea nitrogen level, and induced hyperactivity. The three top terms output for our list of 260 genes on chromosome 02 (which is associated with body size traits) were short tibia, decreased brown fat cell lipid droplet size, and Purkinje cell degeneration. The phenotypes in fishes did not show significant Fisher's exact tests for enrichment of terms, although for chromosome 01, the top three ranked phenotypic terms (based on combined score, Chen et al. 2013) may be relevant to the associations we observed in this study. These terms were reproductive behavior, response to absence of light, and entrainment of circadian clock by photoperiod (Table S12).

Extrapolation of spatial and temporal distribution of phenotypic traits based on candidate SNPs

One general pattern observed consistently across candidate SNPs (e.g. Etr_464 on chromosome 01 and Etr_5713 on chromosome 02) was a divergence between coastal and interior collections (Fig. 4). This pattern was most evident within the Columbia River basin, where both candidate SNPs increased in one allelic variant with increasing distance from the river mouth (Fig. S5). The most dramatic increase was with the frequency of the allelic variant of Etr_5713 associated with large adult body-size ("A" allele) from ~10% at the river mouth to near fixation (~98%) upstream of river kilometer 644. A more moderate increase was observed for the allelic variation of Etr_464 associated with small gonad size (40% to 95% shift in "C"

allele from the river mouth to river kilometer 644). Similar clines were observed within the Willamette River subbasin of the Columbia River (Fig. S5), such that strong linear trends were observed with a change in frequencies 53% to 77% (Etr_464, R^2=0.42) and 22% to 61% (Etr_5713, R^2=0.87) over the span of 180 river kilometers. We classified all collections of the spatial dataset into putative "Mature" and "Premature" forms based on the whether the Etr_464 mature allele frequency was [?]50% or <50%, respectively (Table 6S). We also classified all collections of the spatial dataset into putative "Small" and "Large" body-size forms based on the whether the Etr_5713 large body-size allele frequency was <50% or [?]50%, respectively (Table 6S).

Intra-annual temporal heterogeneity was observed at Willamette Falls among the adult Pacific lamprey returning in run years 2014 and 2015. The abundance of the AA genotype of Etr_464 (chromosome 01) associated with large gonad size arrived earlier than the CC genotype associated with small gonad size (Fig. 5a,b). When the abundance of each run year was divided into equal halves, we estimated that the AA genotype decreased by 3X and 2X between the first and second halves of the run in 2014 and 2015, respectively (Fig. 5a,b).

For the genotypes at Etr_5317 (chromosome 02) associated with adult body-size we did not observe consistent intra-annual trends across years (Fig S6). Genotype proportions at Etr_5317 were similar for both halves of the runs in 2014 and 2015. However, when we paired phenotypic body-size with the genotypes at Etr_5317, we observed a relatively large and consistent trend of a decrease in proportions of AA and AC genotypes that exhibited phenotypic small body-sizes across the runs in 2014 and 2015 (Fig. 5c,d). We estimated that the AA and AC genotypes with phenotypic small body-size decreased by >2.5X between the first and second halves of the run (Fig. 5a,b). On the basis of multiyear observations of migration patterns, we infer that this category of AA and AC Etr_5317 genotype with phenotypic small body-size is a proxy for fish that exhibit advanced maturity. Association testing conducted on pre-mature adults with genotypes AC and AA at Etr_5317 (e.g. at Bonneville Dam) demonstrated strong association with intermediate to large adult body-size, respectively. However, Willamette Falls samples contained mixtures of fish in varying states of maturity. Since these fish tend to shrink in body size as they reach maturity, these AC and AA genotypes can also be found at Willamette Falls in adults with relatively small-body size (i.e., total length in the lower 50% of the length distribution). Our proxy for fish with advanced maturity (AA and AC Etr_5317 genotype with phenotypic small body-size) made it possible to demonstrate that these fish arrive shortly before spawning as compared to premature fish that spend several months in freshwater prior to spawning (Fig. 5c,d).

Discussion

Phenotypic trait associations explain existence of genomic islands of divergence on chromosomes

Trait associations with adult body-size metrics and female gonad size (a proxy for maturity) appear to explain the presence of high levels of genomic divergence on two of the four major adaptive chromosomes in Pacific lamprey (i.e., chromosomes 01 and 02). Genotype-phenotype association testing across multiple data sets from the Columbia and Klamath River basins consistently had strong association of body size with chromosome 02. Using samples from the Klamath River in California, Parker *et al* . (2019) associated the maturity trait ("ocean" and "river-maturing" ecotypes) with markers we have now mapped on chromosome 01. In this study, the association has been extended geographically to include Willamette Falls, in Oregon City, OR. We have evidence, particularly on chromosome 01, that the divergent alleles on these chromosomes are tightly linked across extensive genomic regions because they are captured within inversions that are polymorphic in the species. This concentrated genomic architecture could be key to the landscape genetics and apparent local adaptation for this highly dispersive and near panmictic species. The genotype-by-phenotype associations of candidate markers were exploited for their predictive ability to extrapolate putative distributions of the phenotypes across the species' range. These predicted phenotypic distributions provide insight into how these traits may be heterogeneously distributed in space: ocean-mature and small-bodied lamprey appear concentrated in coastal streams, whereas stream-mature and large-bodied lamprey are concentrated in interior streams. These phenotypes also appear temporally heterogeneous based on their

arrival at Willamette Falls where stream-mature fish return to freshwater long before spawning in contrast to ocean-mature fish that arrive shortly before spawning. This predicted heterogeneity of the spatiotemporal distribution of maturity and body size traits provides a basis for understanding what combinations of traits may be optimally suited for particular freshwater habitats across the species' range.

Genetic architecture of Pacific lamprey body forms

The genetic architecture underlying these traits related to body size and maturity is highly concentrated. However, this result is somewhat expected given the high degree of gene flow exhibited in Pacific lamprey (Spice *et al* . 2012). When natural selection is strongly acting on a particular trait in the face of high gene-flow, the genes involved in the trait tend to become highly concentrated and physically linked within the genome. Concentrated genetic architecture (i.e., few quantitative trait loci, QTL, of large effect) has been predicted to evolve under a set of conditions that include, among other factors, higher rates of gene flow between diverging populations compared to conditions leading to more diffuse genetic architecture (i.e., many QTL of small effect, Yeaman and Whitlock 2011). We have previously found that the adaptive genetic markers were statistically linked (i.e. exhibited linkage disequilibrium within populations) and that allowed categorization of these markers into four groups of linked loci (groups A, B, C, and D; Hess *et al.* 2013). Now we can confirm that groups previously characterized as A, B, C, and D loci (Hess et al. 2015; Parker et al. 2019) localize to chromosomes 02, 04, 22, and 01, respectively, out of a total of 83 chromosomes characterized. Further, we observed that several of these chromosomes have one or more inversion alleles that distinguish the species from a non-inverted state (based on sea lamprey), and appear polymorphic within the species (based on alignments between male and female Pacific lamprey genome assemblies). These inversions appear to coincide with the adaptive SNPs identified asF_{ST} outliers, particularly for the cases of chromosome 01 and 04, which suggests that polymorphic inversions may play an important role in the adaptation of Pacific lamprey to local environments. Since recombination is highly reduced, the fitness conferred by these inverted haplotypes can help maintain them as a polymorphism in a population through both forces of balancing and divergent selection (Wellenreuther and Bernatchez 2018; Faria et al. 2019; Pearse et al. 2019).

Targets of selection

Although the phenotypes that we measured and observed in association with candidate SNPs cannot be concluded to be the causal variants that are the actual targets of selection, we can use the genotype-by-phenotype associations to guide future research aimed to identify these targets. Further, we can also narrow down some traits and life stages that do not appear relevant to any observed adaptive variation. For example, the adaptive genetic variation had no predictive ability for the short-term swimming ability of migration-phase adult lamprey at Bonneville Dam or the growth differences among young-of-year larvae. These two cases of failure to reject a null hypothesis help narrow down the search for a mechanism that manifests in large body size adults that tend to travel further upstream to spawn (a trait highly associated with genes on chromosome 2). These genes apparently do not confer adult swimming endurances, at least for the short timeframe that could be tested in swim trials at Bonneville Dam. Further, these adult body size differences do not translate to faster growth in young of year larvae. However, these adult body size differences could be influenced by differential growth at older life stages; prey selection, length of time in the ocean, or ocean distribution likely affect growth (Clemens et al. 2019). It will require further investigation of multiple life stages in both freshwater and the ocean to understand Pacific lamprey life history strategies. For the maturity trait, gene ontology could suggest other traits to examine as potential targets of selection including circadian rhythm.

Trait comparisons with other anadromous fishes

Despite strong differences in philopatry and population genetic structure, there are similarities between the Pacific lamprey ecotypes and those of steelhead trout (anadromous *Oncorhynchus mykiss*) and Chinook salmon (*O. tshwaytscha*), as were described by Parker*et al.* (2019). In this study, we observed even greater similarities with steelhead of these ecotypes in the Willamette River than were apparent in the Klamath River Pacific lamprey. Notably, like steelhead (Hess *et al.* 2016), the Pacific lamprey ocean- and

river-maturing ecotypes exhibit seasonal separation where premature fish return to freshwater long before spawning in contrast to mature fish that arrive shortly before spawning. Also similar to steelhead (Micheletti *et al.* 2018), the Pacific lamprey ocean-maturing form is only distributed in coastal regions and the river-maturing ecotype is distributed further inland. However, it is unknown whether inland migrating Pacific lamprey exhibit both early and late arrival to spawning grounds as observed for inland migrating steelhead (Micheletti et al. 2018) and Chinook salmon (Narum et al. 2018). Finally, although we found no evidence that the homologous genes were conserved with salmonids, Pacific lamprey ecotypes were associated with a single locus of major effect as shown in steelhead (Hess *et al.* 2016; Micheletti et al. 2018) and Chinook salmon (Prince et al. 2017; Narum et al. 2018).

The adult body size trait that was associated with chromosome 02 genes in Pacific lamprey may share similarities with the age-at-maturity trait described in Pacific salmon (McKinney *et al.* 2019), steelhead (Copeland *et al.* 2017), and Atlantic salmon (*Salmo salar* , Barson *et al.* 2015). In salmonids, the number of consecutive years spent in the ocean before returning to freshwater as a mature adult is highly correlated to body size (e.g., Chinook salmon, Lewis*et al.* 2015). It has also been shown that larger, older Chinook and Sockeye salmon (*O. nerka*) tend to arrive earlier at Bonneville Dam compared to the smaller 1-ocean-age adults (Anderson and Beer 2009). Similarly, larger Pacific lamprey arrive earlier at Bonneville Dam than the smaller bodied forms (Keefer *et al.* 2009, 2013), which may be related to life history decisions in seasonal environments. The primary growth of Pacific lamprey occurs during the ocean phase of the lamprey's parasitic life cycle and so bigger lamprey may also be older in ocean age. There is not yet an accurate way to measure the total age or ocean age of lampreys since they lack bony structures, but this hypothesis could be tested once an aging method is developed (e.g., statolith microstructure). Collectively, the convergence of traits in salmonids and lamprey suggest strong tradeoffs between allocation of resources in capital breeding fishes, whereby long distance migration constrains maturation schedules and in at least some cases (e.g., lamprey) body size.

Hypothesis testing framework

Pacific lamprey genetic traits and their associated phenotypes appear to be inherited independently and may occur in combinations that manifest as different life history strategies to fit unique ecological niches throughout the species' range. For example, in the Willamette River basin there is relatively high diversity of traits and nearly equal portions of genetic variants associated with alternate forms of small and large-bodied adults and ocean- versus river-maturing ecotypes. The following multiple strategies appear to be represented: 1) stream-maturing small- and large-bodied fish that have overwintered below the Falls, 2) ocean-maturing small- and large-bodied fish that arrive shortly before spawning above the Falls, and 3) stream-maturing small- and large-bodied fish that arrive after June and ascend the Falls. The fact that there are four separate chromosomes with important adaptive genes (some with undetermined trait associations) provides for the possibility that various combinations of adaptive variants at these four chromosomes could underpin a multitude of life history strategies. Patterns in the occurrence of the two phenotypic traits we emphasized in this study suggest that Pacific lamprey life history traits may exhibit differential fitness across the range. For example, extrapolation predicts a predominance of large-bodied, stream-maturing forms in northern B.C. and the interior Columbia River, small-bodied ocean- and stream-maturing forms in Puget Sound, intermediate-bodied ocean- and stream-maturing forms in the lower Columbia, and large-bodied ocean- and stream-maturing forms in the southern coastal range. This provides a hypothesis testing framework to examine the incidence likelihoods of life history traits across the range, and understand factors driving optimization of these traits.

Acknowledgements

This work was funded by Bonneville Power Administration, and we appreciate the contributed sampling efforts by Jeffery Jolley, Luke Schultz, Andrew Wildbill, Matt Fox, Ralph Lampman, and the Army Corp of Engineers, as well as the laboratory and coordination efforts by Travis Jacobson, Stephanie Harmon, Brian McIlraith, and Jeff Stephenson.

References

Anderson, J. J., & Beer, W. N. (2009). Oceanic, riverine, and genetic influences on spring chinook salmon migration timing. Ecological Applications, 19(8), 1989-2003.

Antao, T., Lopes, A., Lopes, R. J., Beja-Pereira, A., & Luikart, G. (2008). LOSITAN: a workbench to detect molecular adaptation based on a F st-outlier method. BMC bioinformatics, 9(1), 323.

Asaduzzaman, M., Wahab, M. A., Rahman, M. J., Nahiduzzzaman, M., Dickson, M. W., Igarashi, Y., ... and Wong, L. L. (2019). Fine-scale population structure and ecotypes of anadromous Hilsa shad (Tenualosa ilisha) across complex aquatic ecosystems revealed by NextRAD genotyping. Scientific reports, 9(1), 1-14.

Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., ... & Kent, M. (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. Nature, 528(7582), 405-408.

Beamish, R.J. 1980. Adult biology of the river lamprey (*Lampetra ayresi*) and the Pacific lamprey (*Lampetra tridentata*) from the Pacific coast of Canada. Canadian Journal of Fisheries and Aquatic Sciences 37: 1906–1923.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics, 23(19), 2633-2635.

Campbell, N. R., Harmon, S. A., & Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Molecular ecology resources, 15(4), 855-867.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., ... & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics, 14(1), 128.

Clemens, B.J., van de Wetering, S.J., Kaufman, J., Holt, R.A., and Schreck, C.B. 2009. Do summer temperatures trigger spring maturation in adult Pacific lamprey, Entosphenus tridentatus? Ecol. Freshw. Fish, 18: 418–426.

Clemens, B. J., Weitkamp, L., Siwicke, K., Wade, J., Harris, J., Hess, J., ... & Orlov, A. M. (2019). Marine biology of the pacific lamprey Entosphenus tridentatus. Reviews in Fish Biology and Fisheries, 1-22.

Copeland, T., Ackerman, M. W., Wright, K. K., and Byrne, A. (2017). Life history diversity of Snake River steelhead populations between and within management categories. North American Journal of Fisheries Management, 37(2), 395-404.

Endelman, J. B., & Jannink, J. L. (2012). Shrinkage estimation of the realized relationship matrix. G3: Genes, Genomes, Genetics, 2(11), 1405-1413.

Faria, R., Johannesson, K., Butlin, R. K., and Westram, A. M. (2019). Evolving inversions. Trends in ecology and evolution.

Hess, J. E., Campbell, N. R., Close, D. A., Docker, M. F., and Narum, S. R. (2013). Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. Molecular Ecology, 22(11), 2898-2916.

Hess, J. E., Caudill, C. C., Keefer, M. L., McIlraith, B. J., Moser, M. L., and Narum, S. R. (2014). Genes predict long distance migration and large body size in a migratory fish, Pacific lamprey. Evolutionary applications, 7(10), 1192-1208.

Hess, J. E., Campbell, N. R., Docker, M. F., Baker, C., Jackson, A., Lampman, R., ... and Wildbill, A. J. (2015). Use of genotyping by sequencing data to develop a high-throughput and multifunctional SNP panel for conservation applications in Pacific lamprey. Molecular Ecology Resources, 15(1), 187-202.

Hess, J. E., Zendt, J. S., Matala, A. R., and Narum, S. R. (2016). Genetic basis of adult migration timing in anadromous steelhead discovered through multivariate association testing. Proceedings of the Royal Society B: Biological Sciences, 283, 20153064.

Jones, O. and Wang, J. (2010) COLONY: a program for parentage and sibship inference from multilocus genotype data. Molecular Ecology Resources 10: 551–555.

Keefer, M. L., Moser, M. L., Boggs, C. T., Daigle, W. R., and Peery, C. A. (2009). Effects of body size and river environment on the upstream migration of adult Pacific lampreys. North American Journal of Fisheries Management, 29(5), 1214-1224.

Keefer, M. L., C. C. Caudill, T. S. Clabough, M. A. Jepson, E. L. Johnson, C. A. Peery, M. D. Higgs et al. 2013. Fishway passage bottleneck identification and prioritization: a case study of Pacific lamprey at Bonneville Dam. Canadian Journal of Fisheries and Aquatic Sciences 70:1551–1565.

Kirk, M. A., Caudill, C. C., Tonina, D., and Syms, J. C. (2016). Effects of water velocity, turbulence and obstacle length on the swimming capabilities of adult Pacific lamprey. Fisheries Management and Ecology, 23(5), 356-366.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., ... & McDermott, M. G. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research, 44(W1), W90-W97.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357.

Lewis, B., Grant, W. S., Brenner, R. E., and Hamazaki, T. (2015). Changes in size and age of Chinook Salmon Oncorhynchus tshawytscha returning to Alaska. PLoS One, 10(6), e0130184.

McKinney, G. J., Seeb, J. E., Pascal, C. E., Schindler, D. E., Gilk-Baumer, S. E., & Seeb, L. W. (2019). Y-chromosome haplotypes drive variation in size and age at maturity in male Chinook salmon. bioRxiv, 691063.

Micheletti, S., Hess, J. E., Zendt, J. S., and Narum, S. R. (2018). Selection at a genomic region of major effect is responsible for complex life histories in anadromous steelhead. BMC Evolutionary Biology, 18, 140.

Miller, A. D., Hoffmann, A. A., Tan, M. H., Young, M., Ahrens, C., Cocomazzo, M., ... and Sherman, C. D. (2019). Local and regional scale habitat heterogeneity contribute to genetic adaptation in a commercially important marine mollusc (Haliotis rubra) from southeastern Australia. Molecular ecology.

Narum, S. R. (2006). Beyond Bonferroni: less conservative analyses for conservation genetics. Conservation genetics, 7(5), 783-787.

Parker, K. A., Hess, J. E., Narum, S. R., and Kinziger, A. P. (2019). Evidence for the genetic basis and epistatic interactions underlying ocean-and river-maturing ecotypes of Pacific Lamprey (Entosphenus tridentatus) returning to the Klamath River, California. Molecular ecology.

Phair, N. L., Toonen, R. J., Knapp, I., and von der Heyden, S. (2019). Shared genomic outliers across two divergent population clusters of a highly threatened seagrass. PeerJ, 7, e6806.

Powell, Matthew and MacGregor, Johnryan. (2011). A geographic test of species selection using planktonic foraminifera during the Cretaceous/Paleogene mass extinction. Paleobiology. 37. 426-437. 10.2307/23014731.

Smith, J. J., Timoshevskaya, N., Ye, C., Holt, C., Keinath, M. C., Parker, H. J., ... and Kaessmann, H. (2018). The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. Nature genetics, 50(2), 270.

SAS Institute, Inc. 2000. SAS/STAT User's Guide, version 8. SAS Institute,

Cary, NC.

Spice, E. K., Goodman, D. H., Reid, S. B., & Docker, M. F. (2012). Neither philopatric nor panmictic: microsatellite and mtDNA evidence suggests lack of natal homing but limits to dispersal in Pacific lamprey. Molecular Ecology, 21(12), 2916-2930.

Team, R. D. C. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011. URL https://www. R-project. org.

Wang, J. (2004). Sibship reconstruction from genetic data with typing errors. Genetics, 166(4), 1963-1979.

Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. Trends in ecology & evolution, 33(6), 427-440.

Whitlock, S. L., Deweber, J. T., and Peterson, J. T. 2019. Assessment of Pacific Lamprey Monitoring Techniques at Willamette Falls, OR. Draft Report to the Confederated Tribes of Warm Springs Reservation of Oregon.

Yeaman, Sam, and Whitlock, Michael C. (2011). The genetic architecture of adaptation under migration-selection balance. Evolution : International Journal of Organic Evolution., 65(7), 1897–1911. https://doi.org/10.1111/j.1558-5646.2011.01269.x

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., ... & Kresovich, S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics, 38(2), 203-208.

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., ... & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. Nature genetics, 42(4), 355.

Data Accessibility

The population genomic divergence dataset of 7716 quality-filtered SNP loci (Dryad repository)

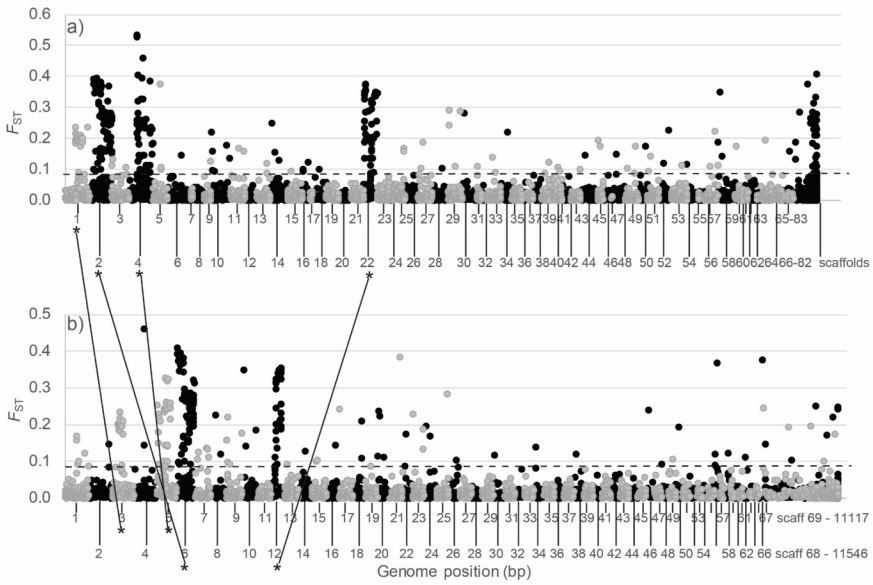
The six datasets used for association mapping with 302 unique SNP loci (Dryad repository)

The temporal dataset of Willamette Falls (2014-2015) adults with 302 unique SNP loci (Dryad repository)

The spatial dataset of larvae/juveniles rangewide with 85 SNP loci (Dryad repository)

Author Contributions

JEH designed the study, analyzed the data, wrote the manuscript; JS & NT analyzed and compiled genome assemblies; CB, CC, MK, MM, LP, GS directly sampled, contributed analysis, and coordinated data collections; DG performed spatial interpolations; SW performed daily abundance analysis; SRN advised analysis and study design; all authors contributed to improving drafts of the manuscript.



Figures

Figure 1. Manhattan plot of the Pacific lamprey SNPs aligned to the (a) Pacific lamprey male gametic (N=7,366; 95.5% of total loci) and the (b) sea lamprey (N=4,916; 63.7% of total loci) genome .F_{ST} values were generated in LOSITAN using 7,716 SNPs and the 16 collections with >20 individuals (N=518 individuals). The critical F_{ST} value of 0.07 is indicated by the dashed line (SNPs above this line were considered outliers P > 0.995). SNPs are shown by alternating odd (black) and even (gray) linkage groups. The asterisks indicate the synteny between Pacific lamprey and sea lamprey for concentrations of outlier loci.

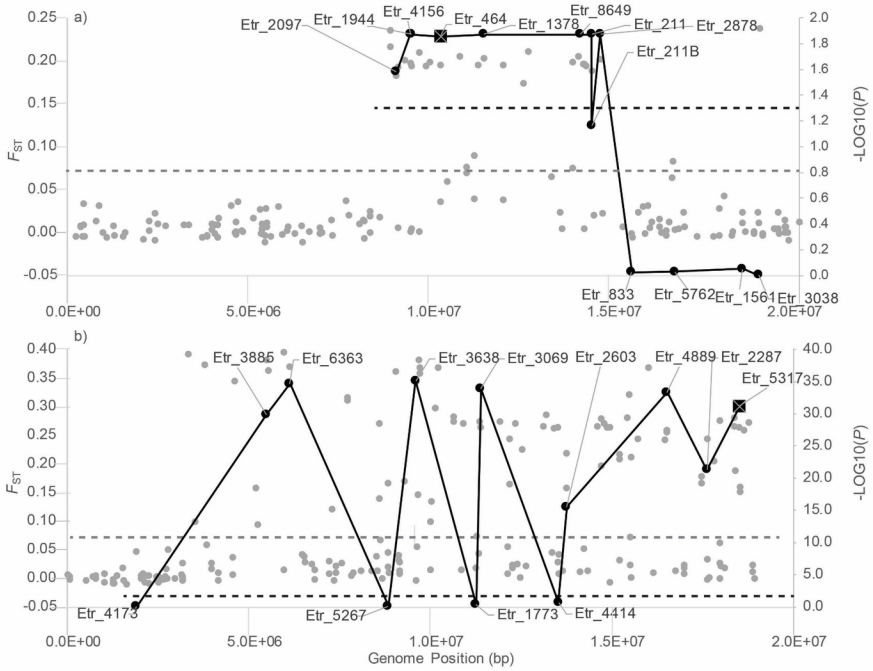


Figure 2. Manhattan plots of SNP positions on the male Pacific lamprey gametic genome of a) chromosome 1 and b) chromosome 2. F_{ST} values among range-wide collections generated in LOSITAN are indicated on the primary y-axis (gray). The critical F_{ST} value of 0.07 is indicated by the gray dashed line (SNPs above this line were considered outliers P > 0.995). The -LOG10(P) values from association testing are indicated on the secondary y-axis (black), and show values from a) testing gonad size in females for the WFA sample and b) testing adult body size for the T_BON sample. The critical value of 1.3 -LOG10(P) indicates the adjusted P -values using the Benjamini and Hochberg (1995) false discovery rate for alpha = 0.05.

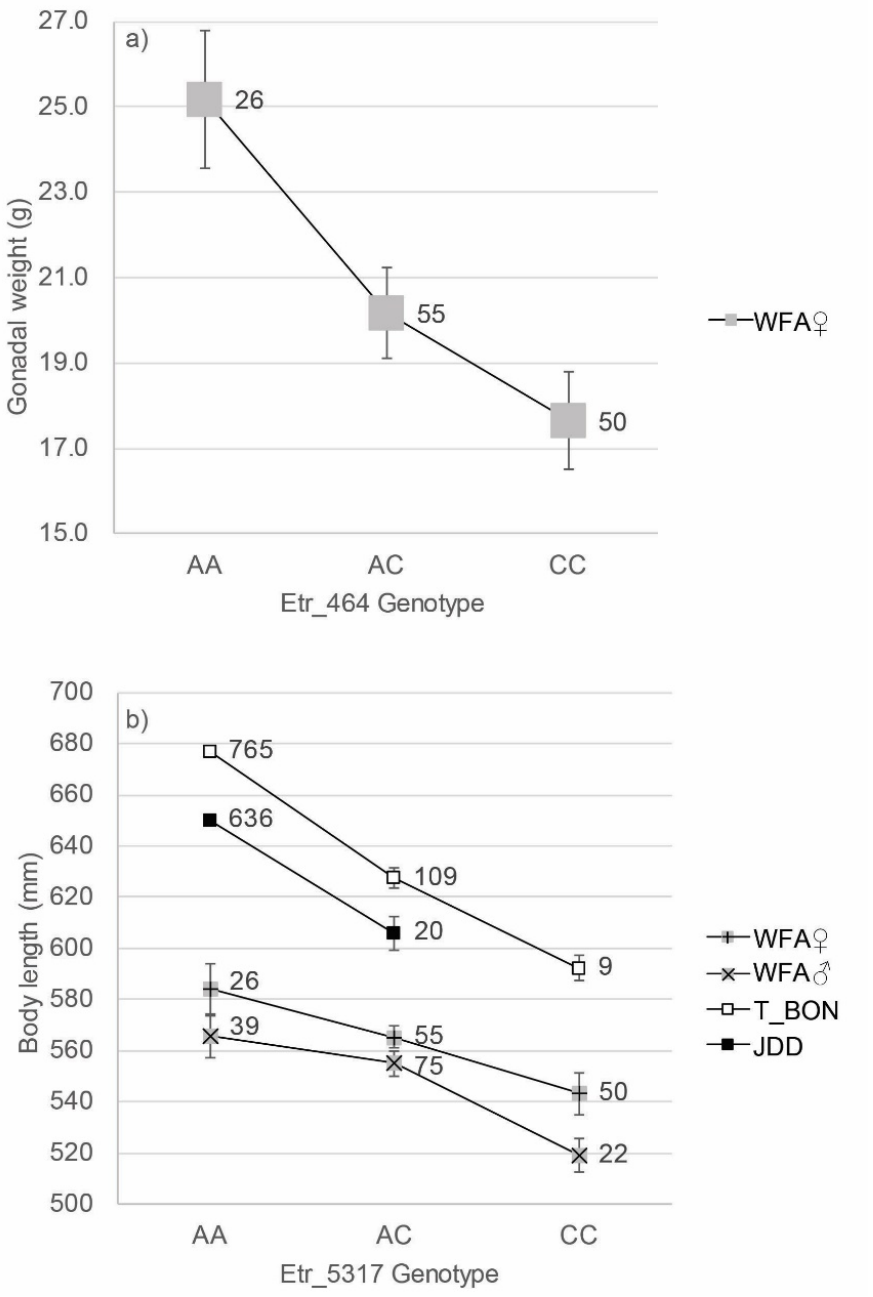


Figure 3. Average gonad weight (a) and average total body length (b) for genotypes at the representative candidate SNPs on chr01 and chr02, respectively . For each point, the sample size and standard error bars are shown. Dataset abbreviations as in Table 1.

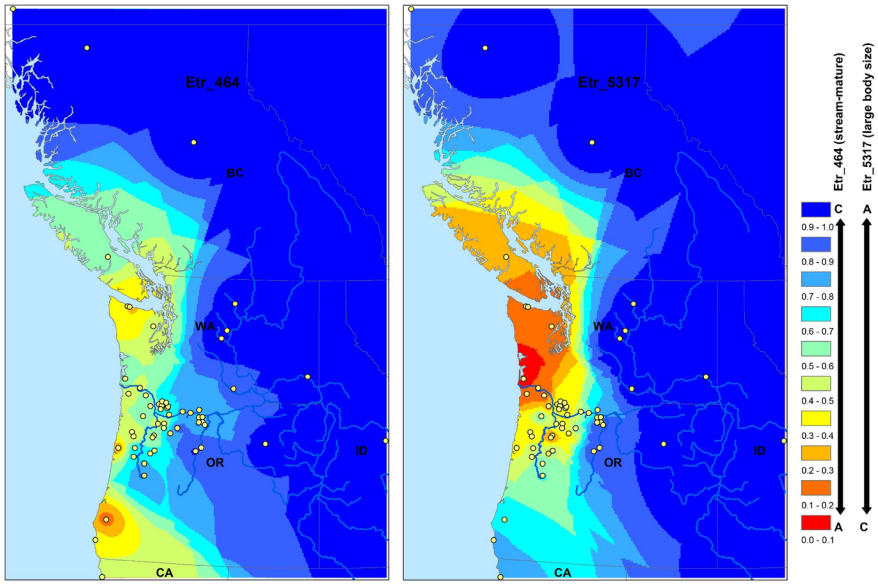


Figure 4. Interpolated candidate gene distribution map for prediction of maturity (Etr_464, left) and body size (Etr_5317, right) phenotypic trait distributions. Allele proportions are color coded from low to high proportions of the C allele for Etr_464 and the A allele for Etr_5317 which are associated with premature ("stream-mature") female gonad and large adult body size, respectively.

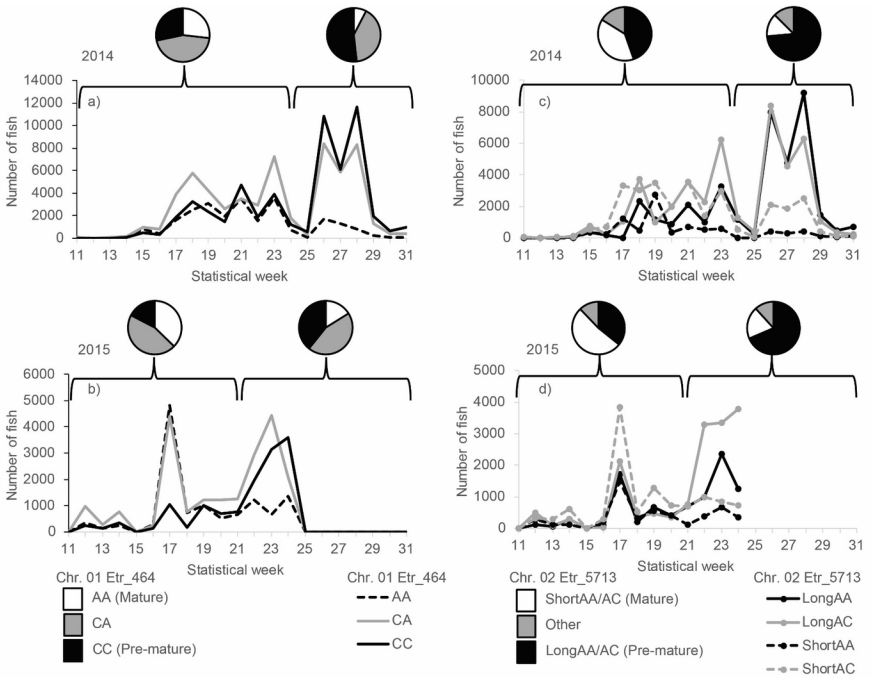


Figure 5. Temporal distributions of candidate SNPs for characterizing phenotypes of the adult Pacific lamprey migration at Willamette Falls. The timing of AA versus CC genotypes at Etr_464 demonstrates relative migration of the "mature" and "premature" ecotype for both a)2014 and b)2015 runs. The line charts (a – b) indicate the relative abundance of genotypes at Etr_464 across statistical weeks and the pies indicate the relative proportions of those genotypes for first and second halves of the run. The migration timing of the AA and AC genotypes at Etr_5713 of fish that were phenotypically small-bodied "ShortAA/AC" versus phenotypically large-bodied "LongAA/AC" was relatively early and late, respectively, for both c)2014 and d)2015 runs. Late stages of maturation causes body size shrinkage (Clemens*et al* . 2009) and so even genotypes associated with large and intermediate body size (i.e. genotypes AA and AC at Etr_5317) can exhibit relatively short body lengths in the lower 50% distribution of length ("ShortAA/AC") when they mature. The line charts (c – d) indicate relative abundance of the "Short" and "Long" phenotypes of the Etr_5317 genotypes and the pies indicate the relative proportions of those phenotype-genotype groups for the first and second halves of the run

Table 1. The number of significant association tests on four evolutionarily important chromosomes for traits and datasets analyzed in this study.

			Chromosome 1	Chromosome 1	Chromosome 1	Chromosome 1	Chromoson
Category	**Trait**	**Dataset**	**N**	**P_N**	**Avg.**	**Range**	**Etr_464**
Body size	Length	JDD	13	0			0.3
		S_BON	13	0			1.0
		T_BON	13	9	6.6	6.4 – 6.8	6.6
		WFA	13	0			0.1
		WFA	13	0			0.4
	Weight	JDD	13	0			0.1
		S_BON	13	0			0.7
		T_BON	13	9	5.9	5.7 – 6.2	6.0
		WFA	13	0			0.5
		WFA	13	0			0.1
	Girth	JDD	13	0			0.2
		S_BON	13	0			0.4
		T_BON	13	9	4.4	4.0 – 4.5	4.5
		WFA	13	0			0.1
		WFA	13	0			0.0
	Dorsal	JDD	13	0			0.3
		S_BON	13	9	1.5	1.4 – 1.6	1.5
		T_BON	13	9	4.1	3.9 – 4.3	4.2
		WFA	13	0			0.2
		WFA	13	0			0.5
	Day	JDD	13	0			0.0
		S_BON	13	0			0.1
		T_BON	13	9	1.8	1.4 – 2.2	1.7
		WFA	13	0			0.0
		WFA	13	0			0.0
	Rkm	JDD	13	0			0.0
		S_BON	13	0			1.3
		T_BON	13	9	5.1	4.8 – 5.3	5.1
	Gonad	WFA	13	0			0.0
Larval Growth	Growth	GAR	13	0			0.0
	Growth_b	GAR	13	0			0.0
Maturity	Gonad	WFA	13	8	1.8	1.6 – 1.9	1.9
Swimming	Approach	S_BON	13	0			0.1

		Chromosome 1	Chromosome 1	Chromosome 1	Chromosome 1	Chromoson
Pass	S_BON	13	0			0.0
Passrep	S_BON	13	0			0.0

Note: P-values were adjusted for multiple testing using the Benjamini and Hochberg (1995) false discovery rate and then transformed with -Log10(P). The -Log10(P) values were shaded light to dark to indicate critical values of 1.3, 2.0, and 3.0 corresponding to alpha levels of 0.050, 0.010, and 0.001, respectively. Included in this table are the total number of loci (N) genotyped on each chromosome, the number of loci with significant adjusted P-values (P N), and the "average" and "range" of -Log10(P) values across loci with significant adjusted P-values. A more detailed Table is available in the Supplemental Materials.

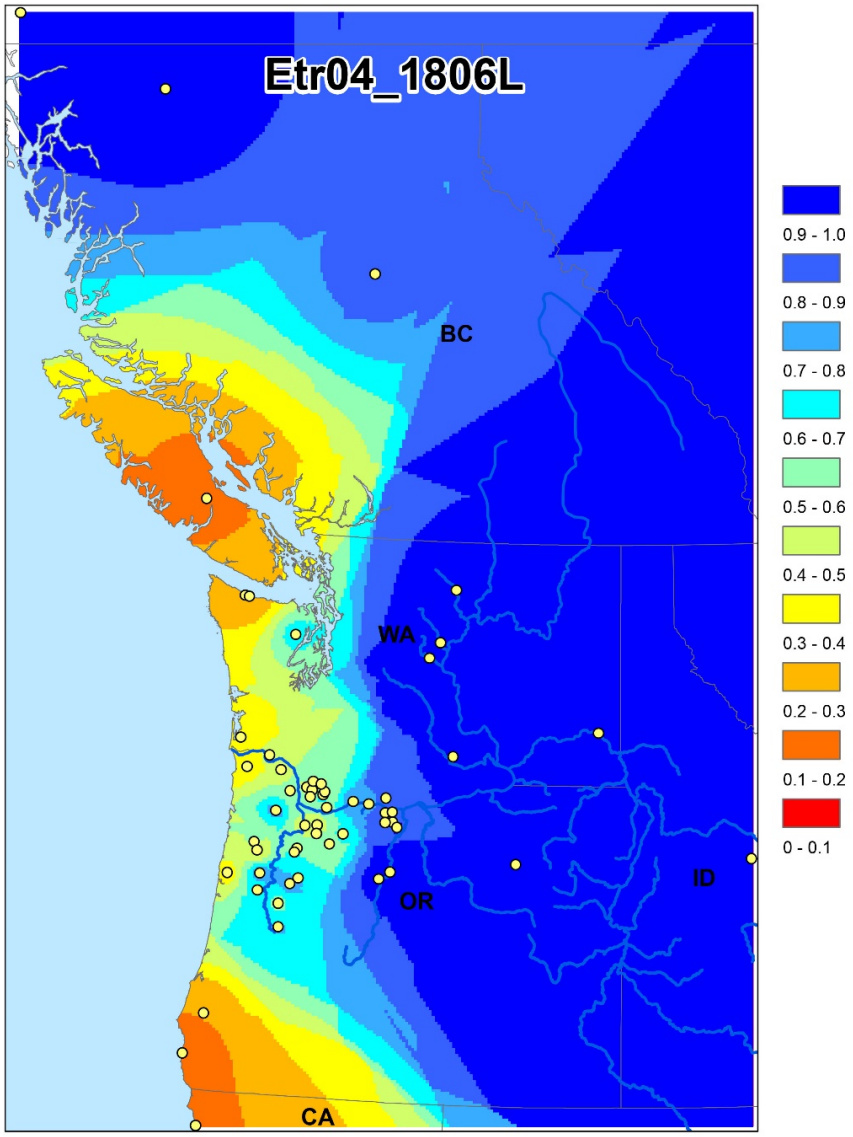


Figure S1. Interpolated candidate gene distribution map for the candidate SNP Etr_1806 on chromosome 04. Allele proportions are color coded from low to high proportions of the A allele for Etr_-

1806 which was associated with large adult body size in the Columbia River. Despite relatively high numbers of large adults in the Klamath River, the alternative allele G was found in higher frequencies in California, which made association with body size inconsistent for this locus and could indicate an unmeasured trait is the true target of selection.

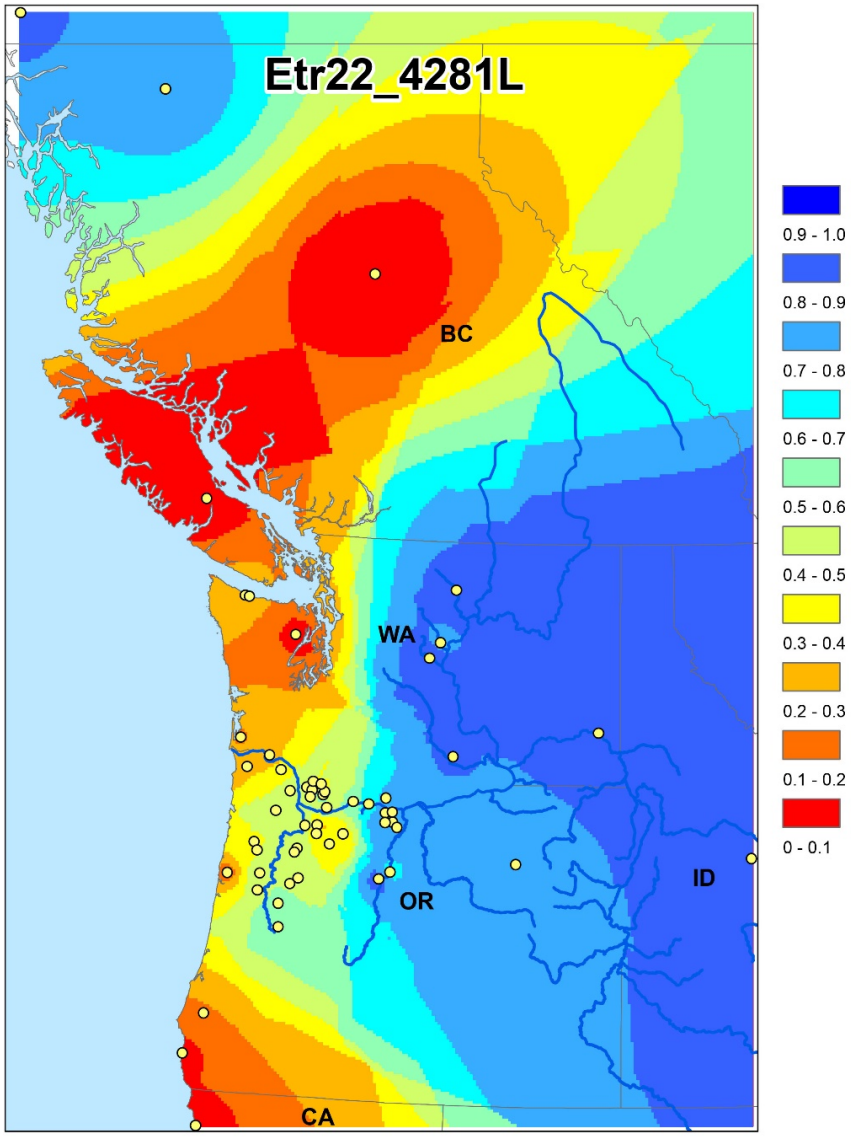


Figure S2. Interpolated candidate gene distribution map for the candidate SNP Etr_4281 on chromosome 22. Allele proportions are color coded from low to high proportions of the A allele for Etr_4281 which was associated with large adult body size in the Columbia River. Body size associations with this locus were weak or inconsistent in other parts of the range. Further, the alternative alleles A and T were found to be fixed between Skeena and Fraser Rivers, B.C. and may indicate an unmeasured trait that is divergent between these river basins.

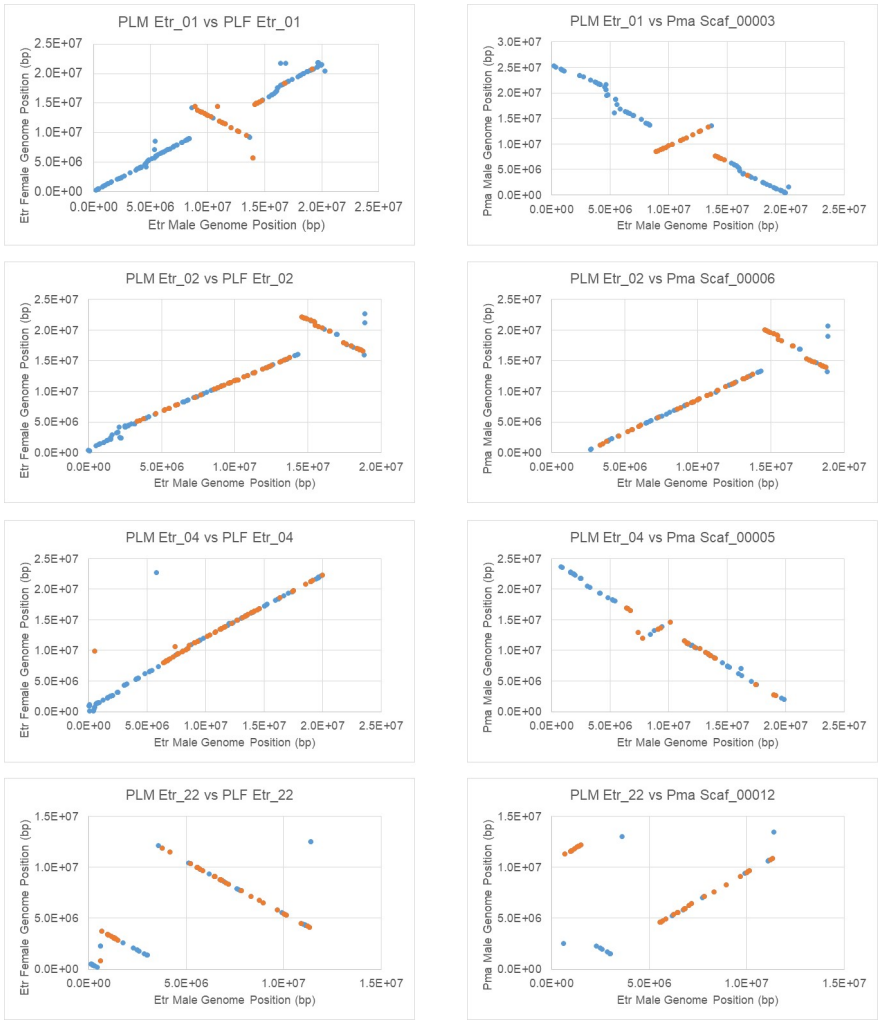


Figure S3. Comparison of Pacific lamprey male and female genome assemblies and Sea Lamprey genome assembly positions (bp) of the total (blue) and outlier (orange) SNPs surveyed across rangewide Pacific lamprey collections on the following four pairs of homologous chromosomes that represent islands of divergence: a) Etr01 versus Pma Scaffold_00003, b) Etr02 versus Pma Scaffold_00006, c) Etr04 versus Pma Scaffold_00005, and d) Etr22 versus Pma Scaffold_00012 .

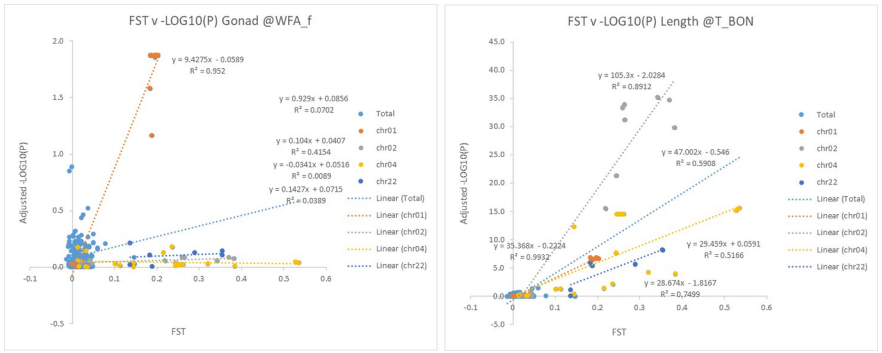


Figure S4. Correlation of rangewide FST and the adjusted -LOG10(P) from association testing with a) female gonad size and b) adult body length. The female maturity and body length traits showed strongest relationship with chr01 and chr02, respectively. The critical value of 1.3 -LOG10(P) indicates the adjusted P -values using the Benjamini and Hochberg (1995) false discovery rate for alpha = 0.05

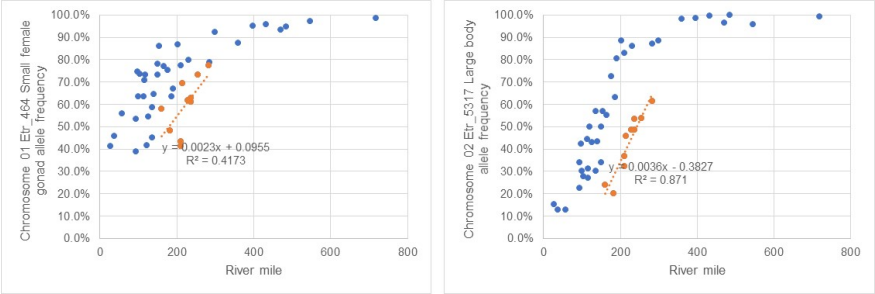


Figure S5. Candidate SNP allele frequencies by river mile location of collections of Pacific lamprey distributed throughout the range (blue) and within the Willamette River basin (orange). The candidate SNP Etr_464 represents the adaptive region on chromosome 1 that is associated with female maturity and the allele associated with small gonad size is plotted (left). The candidate SNP Etr_5713 represents the adaptive region on chromosome 2 that is associated with adult body size and the allele associated with large body size is plotted (right).

[CHART][CHART][CHART][CHART]

Figure S6. **Temporal distributions of candidate SNP Etr_5317 using the adult Pacific lamprey migration at Willamette Falls.** The timing of AA, AC, and CC genotypes at Etr_5317 for both a)2014 and b)2015 runs. The pies indicate the relative proportions of those genotypes for first and second halves of the run. The migration timing of the AA and CC genotypes are generally associated with large versus small body sizes among premature adults migrants. However, late stages of maturation causes body size shrinkage (Clemens *et al* . 2009) and so even genotypes associated with large and intermediate body size (i.e. genotypes AA and AC at Etr_5317) can exhibit relatively short body lengths in the lower 50% distribution of length when they mature. Genotypes at Etr_5317 did not show consistent trends in the first and second halves of the runs in 2014 and 2015, however when body size phenotypes and Etr_5317 genotypes were considered together there were consistent patterns (see Figure 5).

Table S1. Summary of datasets analyzed in this study

Dataset	Description	Life stage	Collection
Range-wide F_{ST}	16 N>20 collections (Hess et al. 2013)	Adult/ Larvae	1995 – 2011
JDD	John Day Dam	Adult	2014 – 2015
S_BON	Bonneville Dam Flume Experiment	Adult	2014
T_BON	Bonneville Dam Adult Fish Facility	Adult	2014
WFA	Willamette Falls Harvest	Adult	2016
WFA	Willamette Falls Harvest	Adult	2016
GAR	Common garden experiment	Larvae	2015
WFA2014	Willamette Falls fish ladder	Adult	2014
WFA2015	Willamette Falls fish ladder	Adult	2015
Larval/juvenile Range-wide	Primarily larvae and juveniles from unique families (Table S2)	Larvae	2011 – 2015

Note: The measured traits were categorized into the following groups: 1)Body size- Length, weight, girth,

and interdorsal distance; 2)migration timing (day of arrival); 3)migration distance (distance in river kilometer traveled upstream from collection site); 4)growth following hatch; 5)maturity (measured by gonad weight); 6) swimming performance.

Table S2. Summary information on the spatial dataset used to extrapolate phenotypes using candidate SNP allele frequencies.

								01	02	04	2
#	Name	N	Siblings	duplicate	Total	Lat	Long	Etr_464 mature	Etr_5317 large	Etr_1806 large	E la
1	Nass	26			26	54.9780	-129.8890	0.00	0.81	1.00	0.
2	Skeena	11			11	54.3000	-126.6300	0.00	1.00	1.00	0.
3	Frasier/Thompson	10			10	52.2680	-121.9880	0.00	1.00	0.85	0.
4	Stamp	30	13*		30	49.3310	-124.9162	0.42	0.30	0.10	0.
5	Deep	65	2	1	68	48.1729	-124.0263	0.59	0.15	0.28	0.
6	Twin	10			10	48.1657	-123.9470	0.80	0.10	0.20	0.
7	Hood Canal	7			7	47.7398	-123.0347	0.57	0.14	0.64	0.
8	EllsworthCr	8	16		24	46.4111	-123.8868	0.43	0.00	0.31	0.
9	Klaskanine	23	18		41	46.0520	-123.7258	0.59	0.15	0.33	0.
10	Elochoman	35	1		36	46.2207	-123.3427	0.54	0.13	0.41	0.
11	Clatskanie	35			35	46.0489	-123.1201	0.44	0.13	0.44	0.
12	NFScappoose	29	8		37	45.7973	-122.9247	0.47	0.22	0.53	0.
13	LockwoodCr	22	8		30	45.8551	-122.6369	0.61	0.34	0.36	0.
14	EFLewisLP	45			45	45.8227	-122.5346	0.26	0.42	0.66	0.
15	LewisRockCr	13	29		42	45.7754	-122.3382	0.27	0.27	0.69	0.
16	UpperEFLewis	29	8		37	45.8143	-122.3113	0.29	0.31	0.47	0.
17	CedarCr	36			36	45.9312	-122.5253	0.26	0.28	0.65	0.
18	UpperCedarCr	26	14		40	45.9068	-122.3817	0.37	0.44	0.42	0.
19	DeepCreek	20	3		23	45.3926	-122.4071	0.55	0.30	0.43	0.
21	ClearCreek	24			24	45.2856	-122.4080	0.35	0.43	0.52	0.
24	Clackamas	34	4		38	45.1670	-122.1695	0.22	0.34	0.56	0.
29	Tualitin	29	2		31	45.5367	-123.1544	0.41	0.57	0.74	0.
32	Willamette	12			12	45.3809	-122.6207	0.58	0.50	0.46	0.
34	Butte	25			25	45.0859	-122.7327	0.42	0.24	0.52	0.
36	Abiqua	25			25	45.0334	-122.7730	0.52	0.20	0.64	0.
38	Willamina	29	6		35	45.1299	-123.4914	0.59	0.32	0.53	0.
40	Mill	15	9		24	45.0242	-123.4333	0.57	0.37	0.33	0.
42	Luckiamute	59	338		397	44.7419	-123.3467	0.31	0.46	0.79	0.
45	Thomas	30			30	44.7161	-122.6799	0.39	0.53	0.70	0.
48	Crabtree	34			34	44.6401	-122.8092	0.38	0.49	0.75	0.
49	MarysRiver	35			35	44.5321	-123.3704	0.37	0.49	0.63	0.
50	Calapooia	26	5		31	44.3899	-122.9890	0.27	0.54	0.56	0.
51	Mohawk	31	1		32	44.0926	-122.9577	0.23	0.61	0.71	0.
52	SalmonCr	30	11		41	45.7379	-122.5578	0.37	0.30	0.57	0.
53	Washougal	35			35	45.6178	-122.2591	0.46	0.43	0.57	0.
54	SalmonR	36	1		37	45.3066	-121.9435	0.14	0.57	0.64	0.
55	Wind	28			28	45.7133	-121.7952	0.27	0.50	0.80	0.
57	Hood	77	2		79	45.6942	-121.5130	0.23	0.55	0.75	0.
62	Klickitat	171	8		179	45.7791	-121.2140	0.25	0.72	0.82	0.
65	MillCr	38	2		40	45.5935	-121.2099	0.36	0.63	0.84	0.
66	8_15Confluence	56	34		90	45.6061	-121.0867	0.33	0.80	0.86	0.
67	FifteenmileCr	67	8		75	45.4845	-121.0730	0.23	0.83	0.91	0.

								01	02	04	2
71	EightmileCr	56	65		121	45.4765	-121.2101	0.13	0.88	0.87	0.
73	Deschutes	96			96	45.4191	-120.9918	0.20	0.86	0.91	0.
75	WarmSprings	35	5		40	44.8623	-121.0754	0.21	0.87	0.96	0.
76	ShitikeCr	39	1		40	44.7688	-121.2680	0.08	0.88	0.96	0.
77	NFJohnDay	31	4		35	45.0118	-118.8780	0.05	0.98	0.98	0.
78	Yakima	28	39	1	68	46.3304	-120.0448	0.13	0.98	1.00	0.
81	Wenatchee	110	27		137	47.5443	-120.5502	0.07	0.96	0.99	0.
82	Entiat	37	3		40	47.7445	-120.3584	0.05	1.00	1.00	0.
85	Methow	34	22		56	48.4021	-120.1056	0.03	0.96	0.99	0.
87	Snake	451	39	2	492	46.6616	-117.4329	0.04	1.00	0.99	0.
89	MFSalmonR	127	54		181	45.0823	-114.7282	0.02	0.99	0.99	0.
91	Siletz	37			37	44.7197	-123.9167	0.78	0.34	0.32	0.
92	Coquille	10			10	42.9500	-124.1100	0.83	0.60	0.20	0.
93	Rogue	30			30	42.4300	-124.4066	0.67	0.75	0.12	0.
94	Klamath	34			34	41.5470	-124.0840	0.50	0.76	0.18	0.
		2581	797	4	3382						

Note: The number of individuals removed due to full sibship are indicated and only 1 adult collection (*) was found to have full siblings, all of which were maintained in the dataset. The "Maturity" extrapolated phenotypic category is based on whether a collection had $<$ or [?] 50% of the mature allele at Etr_464, i.e. "PREMATURE" or "MATURE", respectively. The "Size" extrapolated phenotypic category is based on whether a collection had $<$ or [?] 50% of the large body-size allele at Etr_5713, i.e. "SMALL" or "LARGE", respectively. The other candidate SNP loci (Etr_1806 and Etr_4281) reflect the frequency of the allele associated with large body-size in association tests at Bonneville Dam. However, these loci were not consistently associated with body-size traits at other sites in which association testing was conducted and therefore should be treated as adaptive markers with undetermined traits associations.

Table S3-S8. Pearson r correlations among traits measured for the following six datasets used for association testing: JDD, S_BON, T_BON, WFA, WFA, and GAR.

Table S3	JDD	Body size	Body size	Body size	Body size	Larval Growth	Larval Growth	Mat
Category	**Trait**	Length	Weight	Girth	Dorsal	Growth	Growth_b	Gon
Body size	Length	-	**<0.0001**	**<0.0001**	**<0.0001**			
	Weight	0.8843	-	**<0.0001**	**<0.0001**			
	Girth	0.7488	0.8382	-	0.5940			
	Dorsal	0.2573	0.3228	-0.0209	-			
Larval Growth	Growth					-		
	Growth_b						-	
Maturity	Gonad							-
Migration timing	Day	-0.2629	-0.2805	-0.1083	-0.2186			
Migration Distance	Rkm	0.0719	0.1147	0.0727	0.0913			
Swimming	Approach							
	Pass							
	Passrep							
Table S4	S_BON	Body size	Body size	Body size	Body size	Larval Growth	Larval Growth	Mat
Category	**Trait**	Length	Weight	Girth	Dorsal	Growth	Growth_b	Gon
Body size	Length	-	**<0.0001**	**<0.0001**	**<0.0001**			
	Weight	0.8971	-	**<0.0001**	**<0.0001**			
	Girth	0.8070	0.9328	-	**<0.0001**			

Table S3	JDD	Body size	Body size	Body size	Body size	Larval Growth	Larval Growth	Mat
	Dorsal	0.5146	0.4251	0.3957	-			
Larval Growth	Growth					-		
	Growth_b						-	
Maturity	Gonad							-
Migration timing	Day	-0.0480	-0.0748	-0.0183	0.0059			
Migration Distance	Rkm	0.2657	0.2707	0.2890	0.1951			
Swimming	Approach	-0.0779	-0.101	-0.0891	-0.0395			
	Pass	-0.0415	-0.0656	-0.0547	0.061			
	Passrep	0.0108	-0.0264	-0.0205	0.0963			
Table S5	T_BON	Body size	Body size	Body size	Body size	Larval Growth	Larval Growth	Mat
Category	**Trait**	Length	Weight	Girth	Dorsal	Growth	Growth_b	Gon
Body size	Length	-	**<0.0001**	**<0.0001**	**<0.0001**			
	Weight	0.8865	-	**<0.0001**	**<0.0001**			
	Girth	0.7783	0.9344	-	**<0.0001**			
	Dorsal	0.5268	0.4582	0.4161	-			
Larval Growth	Growth					-		
	Growth_b						-	
Maturity	Gonad							-
Migration timing	Day	-0.2496	-0.2720	-0.2681	-0.2976			
Migration Distance	Rkm	0.2720	0.2607	0.2687	0.2079			
Swimming	Approach							
	Pass							
	Passrep							
Table S6	WFA	Body size	Body size	Body size	Body size	Larval Growth	Larval Growth	Mat
Category	**Trait**	Length	Weight	Girth	Dorsal	Growth	Growth_b	Gon
Body size	Length	-	**<0.0001**	**<0.0001**	**<0.0001**			0.07
	Weight	0.9074	-	**<0.0001**	**<0.0001**			0.96
	Girth	0.6458	0.7670	-	**<0.0001**			0.48
	Dorsal	0.6773	0.6635	0.5441	-			0.55
Larval Growth	Growth					-		
	Growth_b						-	
Maturity	Gonad	-0.1563	-0.0037	0.0618	-0.0512			-
Migration timing	Day	-0.2616	-0.3579	-0.7903	-0.3183			-0.15
Migration Distance	Rkm							
Swimming	Approach							
	Pass							
	Passrep							
Table S7	WFA	Body size	Body size	Body size	Body size	Larval Growth	Larval Growth	Mat
Category	**Trait**	Length	Weight	Girth	Dorsal	Growth	Growth_b	Gon
Body size	Length	-	**<0.0001**	**<0.0001**	**<0.0001**			**<0.**
	Weight	0.8911	-	**<0.0001**	**<0.0001**			**<0.**
	Girth	0.7342	0.8091	-	**<0.0001**			**<0.**
	Dorsal	0.6628	0.6762	0.6199	-			**<0.**
Larval Growth	Growth					-		
	Growth_b						-	
Maturity	Gonad	0.6942	0.8082	0.6686	0.5866			-
Migration timing	Day	-0.3689	-0.4171	-0.7939	-0.4517			-0.38
Migration Distance	Rkm							
Swimming	Approach							
	Pass							

Table S3	JDD	Body size	Body size	Body size	Body size	Larval Growth	Larval Growth	Mat
	Passrep							
Table S8	GAR	Body size	Body size	Body size	Body size	Larval Growth	Larval Growth	Mat
Category	**Trait**	Length	Weight	Girth	Dorsal	Growth	Growth_b	Gon
Body size	Length	-						
	Weight		-					
	Girth			-				
	Dorsal				-			
Larval Growth	Growth					-	**<0.0001**	
	Growth_b					0.5817	-	
Maturity	Gonad							-
Migration timing	Day							
Migration Distance	Rkm							
Swimming	Approach							
	Pass							
	Passrep							

Note: Only the correlations for the group of traits measured for each dataset are provided in each table. P-values and Pearson's r values are shown in the upper and lower triangles, respectively. B-Y FDR corrections for multiple testing at the alpha level 0.05 are indicated by bolded p-values (Narum 2006). See Supplemental Materials methods for association testing for details on traits and datasets.

Table S9a. The number and values of significant association tests on four evolutionarily important chromosomes and their representative candidate SNPs for traits and datasets analyzed in this study.

			Chromosome 1	Chromosome 1	Chromosome 1	Chromosome 1	Chromoson
Category	**Trait**	**Dataset**	**N**	**P_N**	**Avg.**	**Range**	**Etr_464**
Body size	Length	JDD	13	0			0.3
		S_BON	13	0			1.0
		T_BON	13	9	6.6	6.4 – 6.8	6.6
		WFA	13	0			0.1
		WFA	13	0			0.4
	Weight	JDD	13	0			0.1
		S_BON	13	0			0.7
		T_BON	13	9	5.9	5.7 – 6.2	6.0
		WFA	13	0			0.5
		WFA	13	0			0.1
	Girth	JDD	13	0			0.2
		S_BON	13	0			0.4
		T_BON	13	9	4.4	4.0 – 4.5	4.5
		WFA	13	0			0.1
		WFA	13	0			0.0
	Dorsal	JDD	13	0			0.3
		S_BON	13	9	1.5	1.4 – 1.6	1.5
		T_BON	13	9	4.1	3.9 – 4.3	4.2
		WFA	13	0			0.2
		WFA	13	0			0.5
	Day	JDD	13	0			0.0
		S_BON	13	0			0.1
		T_BON	13	9	1.8	1.4 – 2.2	1.7

			Chromosome 1	Chromosome 1	Chromosome 1	Chromosome 1	Chromoson
		WFA	13	0			0.0
		WFA	13	0			0.0
	Rkm	JDD	13	0			0.0
		S_BON	13	0			1.3
		T_BON	13	9	5.1	4.8 – 5.3	5.1
	Gonad	WFA	13	0			0.0
Larval Growth	Growth	GAR	13	0			0.0
	Growth_b	GAR	13	0			0.0
Maturity	Gonad	WFA	13	8	1.8	1.6 – 1.9	1.9
Swimming	Approach	S_BON	13	0			0.1
	Pass	S_BON	13	0			0.0
	Passrep	S_BON	13	0			0.0

Note: P-values were adjusted for multiple testing using the Benjamini and Hochberg (1995) false discovery rate and then transformed with -Log10(P). The -Log10(P) values were shaded light to dark to indicate critical values of 1.3, 2.0, and 3.0 corresponding to alpha levels of 0.050, 0.010, and 0.001, respectively. Included in this table are the total number of loci (N) genotyped on each chromosome, the number of loci with significant adjusted P-values (P N), and the "average" and "range" of -Log10(P) values across loci with significant adjusted P-values. The adjusted -Log10(P) values for representative candidate SNPs Etr_464, Etr_5317, and Etr_1806 are provided for chromosomes 1, 2, and 4, respectively.

Table S9b. The number and values of significant association tests on four evolutionarily important chromosomes and their representative candidate SNPs for traits and datasets analyzed in this study.

			Chromosome 22	Chromosome 22	Chromosome 22	Chromosome 22	Chrom
Category	**Trait**	**Dataset**	**N**	**P_N**	**Avg.**	**Range**	**Etr_4**
Body size	Length	JDD	7	2	2.3	1.9 – 2.6	2.6
		S_BON	7	5	4.4	2.1 – 5.3	5.3
		T_BON	7	5	6.6	5.4 – 8.2	8.1
		WFA	7	0			0.4
		WFA	7	0			0.6
	Weight	JDD	7	0			1.1
		S_BON	7	3	1.6	1.4 – 1.7	1.7
		T_BON	7	4	2.3	1.5 – 2.8	2.5
		WFA	7	0			1.3
		WFA	7	0			0.1
	Girth	JDD	7	0			0.4
		S_BON	7	4	1.9	1.4 – 2.3	2.3
		T_BON	7	4	2.1	1.6 – 2.8	1.8
		WFA	7	0			0.3
		WFA	7	0			0.0
	Dorsal	JDD	7	0			0.1
		S_BON	7	0			0.6
		T_BON	7	0			0.8
		WFA	7	0			0.3
		WFA	7	0			0.3
	Day	JDD	7	0			0.2
		S_BON	7	0			0.0
		T_BON	7	0			0.1

			Chromosome 22	Chromosome 22	Chromosome 22	Chromosome 22	Chrom
		WFA	7	0			0.0
		WFA	7	0			0.0
	Rkm	JDD	7	0			0.0
		S_BON	7	0			0.6
		T_BON	7	1	1.5	1.5 – 1.5	1.1
	Gonad	WFA	7	0			0.0
Larval Growth	Growth	GAR	7	0			0.0
	Growth_b	GAR	7	0			0.0
Maturity	Gonad	WFA	7	0			0.1
Swimming	Approach	S_BON	7	0			0.0
	Pass	S_BON	7	0			0.0
	Passrep	S_BON	7	0			0.0

Note: P-values were adjusted for multiple testing using the Benjamini and Hochberg (1995) false discovery rate and then transformed with -Log10(P). The -Log10(P) values were shaded light to dark to indicate critical values of 1.3, 2.0, and 3.0 corresponding to alpha levels of 0.050, 0.010, and 0.001, respectively. Included in this table are the total number of loci (N) genotyped on each chromosome, the number of loci with significant adjusted P-values (P N), and the "average" and "range" of -Log10(P) values across loci with significant adjusted P-values. The adjusted -Log10(P) values for representative candidate SNP Etr_4281 are provided for chromosomes 22. Any chromosome with significant adjusted P-values not on the four "major" adaptive chromosomes 1, 2, 4, and 22 are listed as "other" chromosomes.

Table S10. Generalized Multifactor Dimensionality Reduction for egg mass and total length, including the number of loci, SNPs in the model, cross-validation consistency, and testing balanced accuracy.

WFA Egg			
Number of Loci	Model	Testing Balanced Accuracy ++	Cross-Validation Co
1	Etr_464		6/10
2	Etr_464, Etr_190		4/10
3	Etr_4142, Etr_6389, Etr_8909		3/10
4	Etr_464, Etr_2517, Etr_4142, Etr_6389		2/10
WFA Total Length			
Number of Loci	Model	Testing Balanced Accuracy	Cross Validation Co
1	Etr_1806	0.7649	10/10
2	Etr_1806, Etr_951		6/10
3	Etr_1806, Etr_464, Etr_6369		1/10
4	Etr_1806, Etr_1210, Etr_485, Etr_6369		3/10
T_Bon Total Length			
Number of Loci	Model	Testing Balanced Accuracy	Cross Validation Co
1	Etr_5317	0.6650	10/10
2	Etr_5317, Etr_1806	0.7191	10/10
3	Etr_5317, Etr_1806, Etr_4281	0.7271	10/10
4	Etr_5317, Etr_1806, Etr_4281, Etr_1104		3/10

+ Cross-validation consistency is defined as the number of times the same model is identified in all 10 training data sets.

++ Testing balanced accuracy is ((TP/(TP+FN)) + (TN/(TN+FP))/2, where TP = True Positive, FP = False Positive, TN = True Negative, and FN = False Negative. Note, testing balanced accuracy is biased

unless the same locus model was identified in all 10 training data sets. Thus testing balanced is reported only when cross-validation consistency was 10/10.

Functions enriched within adaptive genomic regions

Table S11. Pacific lamprey male genome chromosome adaptive region annotations based on Enrichr results for the MGI Mammalian Phenotype Level 4 2019.

Genomic interval
Chr1: 8,939,400-14,772,759 Aig1;ANXA11;ARHGAP6;ARPC1A;AXIN1;CAD;ccg-8;CEP112;CNGA3;COG7;crf2;Ctgf;Cyfip2;DCHS1;DDX1;DDX47;Dhr
Chr2: 3,351,200-18,794,405 Abca1;ABCD3;ABHD17B;acer1;ACO1;Adal;ADAMTS17;Adamts7;ADAMTS7;Adamtsl1;ADAMTSL5;ADPGK;Aen;AGL;A
Chr4: 6,408,000-19,202,840 ABCB1;Abcb1a;ABI1;ACLY;ACTR3;ADCY1;Agap1;AGAP1;AGR3;AHR;AK7;AKAP9;Ambp;AMPH;ANKAR;ANKH;APE
Chr22: 617,450-11,364,750 73;Aak1;AAK1;ABCG2;Abcg3;ACSL6;Adam12;AKR1;ANO2;ANTXR1;Antxr2;ANXA7;ARHGAP24;ATOH7;Bicc1;bmp1;B

Note: Genomic intervals that bound the F_{ST} outlier loci on four chromsomes were used to compile a list

of genes from the corresponding homologous regions in the sea lamprey genome. For each chromosome's genomic interval, the annotated genes are listed and were used to generate the functional "terms" based on the Enrichr human phenotypes "MGI Mammalian Phenotype Level 4 2019" and Fisher's exact tests for enrichment. The table includes "overlap" of each term with the input gene list, "P-value" from Fisher's exact test, "Adjusted P-value" using the Benjamini-Hochberg method for correction for multiple hypotheses testing, "Odds ratio" deviation from expected rank, "Combined Score" is combination of p-value and odds ratio, and "Genes" that were identified as enriched for a particular term. The top ten terms are listed and sorted by lowest to highest "P-value".

Table S12. Pacific lamprey male genome chromosome adaptive region annotations based on FishEnrichr results for the Phenotype AutoRIF Predicted Z-score.

Genomic region
Chr1: 8,939,400-14,772,759 Aig1;ANXA11;ARHGAP6;ARPC1A;AXIN1;CAD;ccg-8;CEP112;CNGA3;COG7;crf2;Ctgf;Cyfip2;DCHS1;DDX1;DDX47;Dhr
Chr2: 3,351,200-18,794,405 Abca1;ABCD3;ABHD17B;acer1;ACO1;Adal;ADAMTS17;Adamts7;ADAMTS7;Adamtsl1;ADAMTSL5;ADPGK;Aen;AGL;A
Chr4: 6,408,000-19,202,840 ABCB1;Abcb1a;ABI1;ACLY;ACTR3;ADCY1;Agap1;AGAP1;AGR3;AHR;AK7;AKAP9;Ambp;AMPH;ANKAR;ANKH;APE
Chr22: 617,450-11,364,750 73;Aak1;AAK1;ABCG2;Abcg3;ACSL6;Adam12;AKR1;ANO2;ANTXR1;Antxr2;ANXA7;ARHGAP24;ATOH7;Bicc1;bmp1;B

Genomic region

Note: Genomic intervals that bound the F_{ST}outlier loci on four chromsomes were used to compile a list of genes from the corresponding homologous regions in the sea lamprey genome. For each chromosome's genomic interval, the annotated genes are listed and were used to generate the functional "terms" based on the FishEnrichr phenotypes "FishEnrichr Phenotype AutoRIF Predicted Z-score" and Fisher's exact tests for enrichment. The table includes "overlap" of each term with the input gene list, "P-value" from Fisher's exact test, "Adjusted P-value" using the Benjamini-Hochberg method for correction for multiple hypotheses testing, "Z-score" deviation from expected rank, "Combined Score" is combination of p-value and the z-score, and "Genes" from the enrichment tests that were identified as enriched for a particular term. The top ten terms are listed and sorted by lowest to highest "Z-score".

Supplemental Materials

Divergence mapping

Two new Pacific lamprey genome assemblies were constructed using the whole genome sequence from the milt and blood from a male (representing the gametic and somatic genomes) and the blood of a female, and using a high density linkage map (Smith et al. 2018) to validate and extend higher order scaffolding of chromosomes. High molecular weight DNA was extracted from these tissues by Amplicon Express (Pullman, WA, USA), and 10X sequencing was performed on an Illumina Nova-seq (University of Illinois Urbana-Champaign). NT-10X Genomics linked-reads from male milt and blood were first deduplicated with hts_SuperDeduper tool, that is part of HTStream pipeline (https://ibest.github.io/HTStream/), and combined together providing 54X effective read coverage and estimated mean molecule size of 57Kb. De novo assembly was performed by Supernova assembler v2.1 (Weisenfeld *et al* . 2017) and then ALLMAPS (Tang*et al* . 2015) was used for further scaffolding based on linkage map, placing 63% of assembled sequence to the 83 linkage groups. These steps resulted in the assembly of the Pacific lamprey male genome of 974 Mb in size with a scaffold N50 of 7.8 Mb and longest scaffold reaching 21Mb.

Linked-reads sequenced from female blood had longer mean molecule size (87Kb) and effective coverage of 42X. They were also assembled with Supernova v2.1 and then 69% of the assembled sequence was placed to the linkage groups by running ALLMAPS. These steps generated an assembly of the Pacific lamprey female genome that is 997 Mb in size with longest scaffold of 22 Mb and a scaffold N50 of 10 Mb.

For characterization of SNP densities and F_{ST}statistics, we used a set of 7,716 unique SNP loci from previously published RAD-seq datasets (Hess *et al.* 2013; Smith *et al.*2018), which passed the following a set of population genetic QC filters. The 518 individuals distributed among 21 samples and across the species' range (described in Table 1, Hess *et al.* 2013) had no more than 20% missing genotypes, and SNP loci had >1% minor allele frequency averaged across the subset of 16 samples with N > 20; and SNP loci had <3 Hardy-Weinberg deviations within 5 aggregated samples (following methods to minimize potential Wahlund effects by pooling individuals into the following five test populations as described in Hess et al. (2013)). This set of 7,716 SNPs was a combination of a group of SNPs from a previous dataset (Hess*et al.* 2013; SNPs N = 8,772 of which 6,295 passed these population genetic QC filters) and a group of SNPs discovered *de novo* for a linkage mapping dataset (Smith *et al.* 2018; SNPs N = 7,977 of which 3,670 passed these population genetic QC filters. BOWTIE2 (Langmead and Salzberg 2012) was used to align datasets of 8,772 (Hess*et al.* 2013) and 7,977 SNPs (Smith *et al.* 2018) to the male reference assembly to define homologous loci. For the 7,716 total SNPs passing the QC filters, 4,046 loci were unique to Hess *et al.* 2013, 1,418 loci were unique to Smith *et al.* 2018, and 2,252 SNPs were shared across datasets. Marker positions based on BOWTIE2 alignments were compared between Pacific lamprey male and female genomes and the Pacific lamprey male and sea lamprey male gametic genome (GenBank assembly accession: GCA_002833325.1) to characterize synteny.

The program minimap2.1 with parameters (–cs=long -cx asm20) was used for alignment between the Pacific lamprey male and Sea lamprey genomes. The function maf-convert (from LAST (Kiełbasa et al. 2011) was used to generate a chain file, that was used by CrossMap (Zhao et al. 2013) to lift over gene annotations from Sea lamprey to the Pacific lamprey male genome assembly.

Association testing

Genotyping-in-thousands by sequencing panel optimization:

Genotyping-in-thousands by sequencing (GT-seq, Campbell et al. 2015) was employed to genotype 308 genetic markers for the association testing analyses. The GT-seq 308 loci were a subset of markers developed from the paired end consensus reads from the Hess et al. (2013) RAD-seq dataset. The selection of loci and steps in development began with a group of 457 total SNP loci considered in round 1, which included 120 that had been already designed for TaqMan assays (Hess et al. 2015). We chose 337 SNPs that had not been designed previously, and we ensured that all SNP sites were located at base pair position 30 or higher to accommodate the assay primer site in flanking DNA. We established the following set of guidelines for choosing SNPs: 1) Pass QC filters for Rangewide dataset, 2) only align to 1 locus in Bowtie to itself test, 3) Overlapped with loci in the linkage map (Smith et al. 2018), 4) High concordance in alignments to the sea lamprey genome across overlapping markers in the Hess et al. (2013) and Smith et al. (2018) datasets, 5) Previously developed as Tagman 96 assays + some species ID loci, 6) Spaced 5cM or greater apart on a linkage group, 7) Mostly neutral and high MAF for parentage power. 8) Adaptive SNPs chosen to be equally representative across four groups of statistically linked loci. A PERL script was run to screen out loci that appeared to have too many heterozygotes and were likely duplicated regions. There were 401 loci that passed this filter. Although we already had 120 primers designed from previous work, we had to construct consensus sequence for the rest using paired-end sequence data from Hess et al. (2013) and were successful developing 266 primer pairs for the loci. A PERL script was used to identify 28 primer interactions which were resolved by dropping 26 primer pairs. This filter resulted in a remaining set of 360 loci (240 new + 120 original primer pairs). Final optimization left 308 markers that worked best in GT-seq genotyping. For all samples used in the association testing we filtered out individuals missing >10% of genotypes at the 308 loci. Excluding the four species diagnostic loci and two loci that were duplicates, provided 302 unique loci for association tests.

Details on the covariates and ways in which loci were used taking population structure and relatedness into account in the GLM and MLM tests:

Equations for the GLM and MLM are described in the TASSEL manual (Bradbury *et al.* 2007). A kinship matrix using the 'scaled IBS' method (Endelman *et al.* 2012) based on 76 vetted neutral SNPs (the neutral SNPs that overlap the 85 neutral SNPs characterized by Hess*et al.* 2015) was generated in TASSEL to represent cryptic familial relationships. The MLM was implemented using default options, i.e. 'P3D' (Zhang *et al.* 2010) parameter option and the 'optimum compression' option. The GLM effectively represents a 'maximum compression' option, and thus provides contrast to the MLM. Permutation tests (1000) were used to calculate P -values to determine significant associations of SNPs with traits. The association tests using a GLM were performed with covariates of population structure. For population structure, the first 3 Principal Coordinate (PC) axes of the 76 neutral SNPs were used. Datasets were also analyzed using an MLM, and the kinship matrix was included as an additional covariate. In all datasets except for GAR, the GLM was a better fit to the data based on the fact that most of the neutral loci aligned closer to the 1:1 line on the Q-Q plot; therefore, reported P-values were generated exclusively from the GLM.

Phenotypic traits measured for association testing

1. Willamette ("WFA") Falls 2016 adult Pacific lamprey were split into separate analyses for a group of males (N=136) and females (N=133) collected as part of the tribal harvest. The following six traits were measured: ordinal "day" of arrival, girth, total "length", weight, "gonad" weight, and distance between dorsal fins ("dorsal"). Willamette Falls is 205.6 Rkm upstream from the Columbia River mouth.

2. Total ("T_BON") sample of Bonneville Dam adult Pacific lamprey in 2014 was measured with six traits: Ordinal day, length, weight, dorsal, girth, and upstream migration distant ("Rkm") (N=883). Bonneville Dam is 235.1 Rkm upstream of the Columbia River mouth.
3. Swimming ("S_BON") trials of Bonneville Dam adult Pacific lamprey were a subset (N=295) of the fish included in the T_BON sample, and included 3 swimming behavioral trait categories, in addition to the 6 traits: Ordinal day, length, weight, dorsal, girth, and upstream migration distant ("Rkm"). Swimming trial traits were measured from previous work (Kirk et al. 2016), and included 3 swimming behavioral trait categories: approached experiment ("approached"), passed challenge ("pass"), and passed challenge without fallback ("passrep"). The "approached" category refers to those that approached *vs* . did not approach the flume, which was a possible indicator of motivation. The "passed" category refers to those that approached and passed the swimming challenge compared to those that approached and did not pass the challenge, which was the major descriptor of performance and success (Kirk *et al.* 2016). Finally, the "passrep" category refers to a subset of fish that approached and passed the challenge, and unlike others that passed the challenge, they did not fall back downstream of the challenge. In summary, "approach" indicates motivation, "pass" indicates motivation + performance and "passreq" indicates motivation, performance and perseverance.
4. Sample of John Day Dam ("JDD") adult Pacific lamprey with 6 traits: Ordinal day, length, weight, dorsal, girth, and upstream migration distant ("Rkm") (N=656). Year and Translocation status were used as covariates. John Day Dam is 346.9 Rkm upstream of the Columbia River mouth. Most of the fish were translocated upstream to Ice Harbor Dam (N = 571, 537.7 Rkm) in both years (2014, 2015), and a portion (N=85) were released at John Day Dam in 2014.
5. Common garden experiment using artificial propagation of larval Pacific lamprey with early larval growth ("GAR") rate data (N=334). Traits of growth rate were calculated as: (length / time), and a corrected growth rate ("growth rate_b") was calculated as: [(length - 4mm) / time] to correct for length at hatch (~4 mm). The MLM that used a kinship matrix as a covariate was a better fit to the data as compared to the GLM, and so the P-values generated by the MLM were used for all downstream analysis.

Potential gene-interaction effects

To identify potential gene-gene interactions across the four primary adaptive chromosomes related to body size and maturity we conducted Generalized Multifactor Dimensionality Reduction (GMDR). Analyses were conducted for maturity using adult females from the WFA (N=133) data set and for total length using the WFA and T_BON (N=883) data sets. We used the software GMDR version 0.9 (Lou et al. 2007; Chen 2011) to conduct an exhaustive search for all possible one to four locus models. The best model was defined as the model with the maximal cross-validation consistency. For additional details on GMDR and analysis methods see Parker et al. (2019).

Results

The gene-interaction analysis using GMDR for egg mass in the WFA collection identified Etr_464 (chromosome 1) as the best single-locus model (Table S10). However, this model was only identified in 6 of the 10 training data sets, indicating limited support. Additionally, cross-validation accuracy for higher order models (two-locus 4/10; three-locus 3/10; four-locus 2/10) indicated the lack of support for gene combinations associated with egg mass. This result contrasts with Parker et al. (2019) who found evidence for a two-locus interaction model including chromosomes 1 and 4 for egg mass in Klamath River collections of Pacific lamprey. The discrepancy may be explained by the differences in collections of Pacific lamprey that have recently initiated their freshwater migration (Parker et al. 2019) versus collection of individuals further upstream (herein). The latter data set likely contains a mixture of current year and hold-over individuals whereas the former contained only current year migrants.

For total length, the GMDR produced different results depending upon the data set (Table S10). The gene-interaction analysis for WFA collection produced support for a single-locus model including Etr_-1806 (Chromosome 4) with cross-validation accuracy (10/10) and testing balance accuracy (77%). Higher

order models with more loci were not supported. In contrast, for T_BON the model with maximal cross-validation accuracy (10/10) and highest testing balance accuracy (73%) was a three-locus interaction model (Table S10). The testing balance accuracy for the one-locus model (Etr_5317/Chromosome 2) was 67%. A 5% increase in testing balanced accuracy was realized in a two-locus interaction model (72%) that included Etr_5317/Chromosome 2 and Etr_1806/Chromosome 4. However, only a 1% increase was observed in the three-locus interaction model (73%), which included Etr_5317/Chromosome 2, Etr_1806/Chromosome 4, and Etr_4281/Chromosome 22. Models involving four loci had considerably lower cross validation accuracy (3/10) indicating lack of support. Under the best three-locus model, if Etr_5317 = AA and Etr_1806 = AA and Etr_4281 = AA, or if Etr_5317 = AA and Etr_1806 = AA and Etr_4281 = AT then individuals are classified as large body size whereas all other genotype combinations are classified as small body size. Classifying T_BON individuals using these methods produces a mean total length for large body size of 681 mm and for small body size of 632 mm. The analysis of Parker et al. (2019) also suggested support for a two-locus interaction model for total length involving chromosomes 2 and 4.

References

Chen, G. B., Xu, Y., Xu, H. M., Li, M. D., Zhu, J., & Lou, X. Y. (2011). Practical and theoretical considerations in study design for detecting gene-gene interactions using MDR and GMDR approaches. PLoS ONE, 6, e16981.

Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011; 21:487–93.

Lou, X-Y., Chen, G.B., Yan, L., Ma, J. Z., Zhu, J., Elston, R. C., & Li, M.D. (2007). A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *The American Journal of Human Genetics* , 80, 1125-1137.

Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable J, Schnable P, Lyons E, Lu J. (2015) ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biology 16(1):3

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. Genome research, 27(5), 757-767.

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., & Wang, L. (2013) CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics (Oxford, England), btt730.