Published in final edited form as:
Mol Ecol. 2013 June ; 22(11): 2864–2883. doi:10.1111/mec.12330.

The population structure and recent colonization history of Oregon threespine stickleback determined using RAD-seq

Julian Catchen[1,2], **Susan Bassham**[1,2], **Taylor Wilson**[1], **Mark Currey**[1], **Conor O'Brien**[1], **Quick Yeates**[1], and **William A. Cresko**[1,*]

[1]Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA 97403

Abstract

Understanding how genetic variation is partitioned across genomes within and among populations is a fundamental problem in ecological and evolutionary genetics. To address this problem we are studied the threespine stickleback fish, which has repeatedly undergone parallel phenotypic and genetic differentiation when oceanic fish have invaded freshwater habitats. While significant evolutionary genetic research has been performed using stickleback from geographic regions that have been de-glaciated in the last 20,000 years, less research has focused on freshwater populations that predate the last glacial maximum. We performed RAD-seq based population genomic analyses on stickleback from across Oregon, which was not glaciated during the last maximum. We sampled stickleback from coastal, Willamette Basin, and central Oregon sites, analyzed their genetic diversity using RAD-seq, performed *STRUCTURE* analyses, reconstructed their phylogeographic history, and tested the hypothesis of recent stickleback introduction into central Oregon, where incidence of this species was only recently documented. Our results showed a clear phylogeographic break between coastal and inland populations, with oceanic populations exhibiting the lowest levels of divergence from one another. Willamette Basin and central Oregon populations formed a clade of closely related populations, a finding consistent with a recent introduction of stickleback into central Oregon. Finally, genome wide analysis of genetic diversity (π) and correlations of alleles within individuals in subpopulations (F_{IS}) supported a role for introgressive hybridization in coastal populations and a recent expansion in central Oregon. Our results exhibit the power of next generation sequencing genomic approaches such as RAD-seq to identify both historical population structure and recent colonization history.

Keywords

population structure; isolation by distance; adaptation; human introduction; introgression; gene flow

1. Introduction

Biologists have long been interested in understanding the processes by which genetic variation is partitioned within and among populations (Wright 1931; Dobzhansky 1937; Fisher 1958; Wright 1978; Nielsen *et al*. 2007; Holsinger & Weir 2009). Historically,

*wcresko@uoregon.edu.
[2]co-first authors

Data Accessibility
Sampling locations: Table 1.
Raw RAD-seq reads NCBI SRA070979
Aligned RAD-seq reads: Available in Dryad entry doi:10.5061/dryad.62hb0
VCF file of genotype calls, and summary statistics file: DRYAD entry doi:10.5061/dryad.62hb0

empirical studies of these processes have been limited to just a handful of genetic markers (Holsinger & Weir 2009), sometimes leading to difficulties in inferring very recent phylogeographic events because of insufficient genetic information (Emerson *et al.* 2010; McCormack *et al.* 2011). In addition, genetic variation can exhibit significant heterogeneity across genomes due to demographic processes such as recent colonization events and range expansion, or because of diversifying and stabilizing natural selection that is distinct from the background level of divergence caused by neutral processes such as genetic drift (Lewontin & Krakauer 1973; Maynard Smith & Haigh 1974; Baer 1999; Beaumont & Balding 2004; Storz 2005; Stinchcombe & Hoekstra 2008; Akey 2009; Nosil *et al.* 2009; Pritchard *et al.* 2010; Renaut *et al.* 2011).

The recent development of next generation sequencing technologies has lowered this methodological barrier, allowing detailed population-level genomic studies in ways that were not possible even a few years ago (Asmann *et al.* 2008; Mardis 2008a; Mardis 2008b; Marguerat *et al.* 2008; Shendure & Ji 2008; Pool *et al.* 2010; Glenn 2011; McCormack *et al.* 2011; Fan *et al.* 2012; Hohenlohe *et al.* 2012b). A useful model for population genomic studies is the threespine stickleback fish, *Gasterosteus aculeatus* (Wootton 1976; McPhail 1984; Peichel *et al.* 2001; Peichel 2005; Cresko *et al.* 2007; Jones et al. 2012). This species diversified into three life history forms: marine, anadromous, and freshwater (Bell 1984; Bell & Foster 1994). Strictly marine stickleback fish spend their entire lives in the ocean, whereas anadromous stickleback fish spend adult life in the ocean and breed in fresh water (Foster & Baker 2004; Cresko *et al.* 2006). Freshwater stickleback fish inhabit water bodies that are often partially or completely isolated from the sea.

Repeated, rapid divergence in new freshwater habitats has resulted in a *G. aculeatus* species complex as individuals from the phylogenetically stable marine ancestor have invaded new freshwater habitats (Bell 1994; Bell & Andrew 1997; Bell *et al.* 2001). In many regions of the Northern Hemisphere, the majority of the present-day freshwater populations of stickleback did not exist before the last glacial maximum approximately 15,000 years ago (Cresko *et al.* 2007). In addition, recent work has shown that in some cases new freshwater populations have evolved in decades (Bell *et al.* 2004; Gelmond *et al.* 2009; Furin *et al.* 2012). For example, the Great Alaska Earthquake of 1964 uplifted Middleton Island in the Gulf of Alaska by 3.4 m and created new freshwater sites, which anadromous stickleback subsequently colonized. Gelmond *et al.* (2009) found that even in the presence of marine stickleback, the freshwater environment provided sufficient selective pressure to drive rapid divergence of stickleback in the new ponds.

The repeated, parallel evolution of stickleback populations in fresh water provides a set of repeated natural experiments that are useful for testing evolutionary hypotheses at different scales of time and geographic distances (McKinnon & Rundle 2002; Cresko *et al.* 2004; Albert & Schluter 2005; Boughman *et al.* 2005; Colosimo *et al.* 2005; Kimmel *et al.* 2005; Hohenlohe *et al.* 2010b; Kimmel *et al.* 2012a, 2012b). The stickleback populations in hundreds of lakes across Alaska could only have been established within the last 15,000 years after the Laurentide and Cordilleran ice sheets receded (Fig. 1). A surprising finding from studies of these relatively recently formed populations is that the same genomic regions are involved in adaptation to freshwater habitats in independently evolved populations (Colosimo *et al.* 2004; Cresko *et al.* 2004; Shapiro *et al.* 2004; Colosimo *et al.* 2005; Kimmel *et al.* 2005; Knecht *et al.* 2007; Miller *et al.* 2007; Chan *et al.* 2010; Hohenlohe *et al.* 2010b; Hohenlohe *et al.* 2012a; Kimmel *et al.* 2012a; Kimmel *et al.* 2012b, Jones *et al.* 2012).

Much less evolutionary genetic research has been performed examining the genomes of freshwater stickleback populations that may be significantly older than the last glacial

maximum. Contemporary stickleback habitat in Oregon (USA) was not ice covered during the last glacial maximum (Fig. 1), so we hypothesized that extant stickleback populations in Oregon may be older, providing an important opportunity to test the generality of previous population genomic results from recently de-glaciated regions. We used Restriction-site Associated DNA sequencing (RAD-seq; Miller *et al.* 2007; Baird *et al.* 2008; Davey & Blaxter 2010; Davey *et al.* 2011; Etter *et al.* 2011a; Etter *et al.* 2011b) and *Stacks* software (Catchen *et al.* 2011; Catchen et al. 2012, this volume) to determine the fine scale relationship of nine populations of threespine stickleback fish distributed across three different geographic regions in Oregon (Fig. 2 and Table 1). Oregon's unique geographic history allows us to establish the relationship of these populations to one another, document patterns of genetic variation and genomic architecture, and address a hypothesis of recent (< 50 years) human introduction of stickleback into central Oregon populations. These results provide a foundation for subsequent studies of the partitioning of genetic variation across these Oregon populations, and comparisons with previous findings from other evolutionary genetic studies of stickleback.

Methods

Sample Collection and DNA extraction

The stickleback samples used in this study were collected using minnow traps deployed overnight in stream, river, lake and estuarine slough shallow-water habitats (see Tables 1 & 2 and Figure 2 for complete set of collections). Stickleback were collected from both marine and freshwater sites along the coast, and from freshwater habitats in the Willamette Basin and central Oregon following previously published protocols (Cresko *et al.* 2004). All collections were made according to approved University of Oregon IACUC protocols, and under the auspices appropriate state and federal collecting permits.

The caudal fin and both pectoral fins from ethanol-preserved fish were removed for DNA extraction. The soma and fin clips of each fish were labeled with a unique identification number designating the collection and individual. Fin clips were stored at −80°C, and soma were fixed in 10% formalin solution overnight. Soma were cleared and stained in order to visualize mineralized skeletal traits according to previously published protocols (Cresko *et al.* 2004).

RAD Library Preparation and sequence analysis

Extracted genomic DNA was normalized to a concentration of 25 ng/μl in 96 well plates and processed into RAD libraries according to (Etter *et al.* 2011a), using the restriction enzyme SbfI-HF (NEB). This same enzyme has been used in previous population genomic studies of Alaskan threespine stickleback (Hohenlohe et al. 2010b; Hohenlohe et al. 2012a), making direct comparison of results possible. Libraries were sequenced in four lanes on an Illumina HiSeq2000 and six lanes on an Illumina GAIIx according to manufacturer's instructions.

Raw sequence reads were demultiplexed using *Stacks* (Catchen *et al.* 2011). Only those reads of sufficiently high sequencing quality, and that had the correct barcode and an unambiguous RAD site, were retained (Table 2). Demultiplexed reads were aligned to the threespine stickleback reference genome (version BROADs1, Ensembl release 64) using GSnap (Wu & Watanabe 2005). We required unique alignments, allowing for a maximum of 5 mismatches, the presence of up to 2 indels, and no terminal alignments. This last criterion prevented large fractions of either end of the sequence read from being soft-masked. Aligned reads were analyzed in S*tacks*, which derived each locus from overlapping GSnap alignments to produce a consensus sequence. SNPs were determined and genotypes called using a maximum likelihood statistical model implemented in *Stacks* (Hohenlohe *et al.*

2010b; Catchen *et al.* 2011; Hohenlohe *et al.* 2012b). To include a locus in the analysis we required it to be present in all nine populations and genotyped in at least 75% of the samples of each population; nearly all of the loci fit these criteria (Supp. Fig. 1). The small number of loci that did not meet these thresholds were removed from further analysis. For more details on the use of the *Stacks* software package for studies of model and non-model organisms see (Catchen *et al.* 2011; Catchen *et al.* 2012, this volume).

Population genetic statistics (major allele frequency, percent polymorphic loci, nucleotide diversity (π) and Wright's F statistics F_{IS} and F_{ST}) were calculated for every SNP using the `populations` program in *Stacks*. For bi-allelic SNP markers, π is a measure of expected heterozygosity and therefore a useful overall measure of genetic diversity in a population. F_{IS} measures the reduction in observed heterozygosity as compared to expected heterozygosity for an allele in a population, and positive values indicate nonrandom mating or cryptic population structure (Nei 1975; Nei 1987; Nei and Kumar 2000; Hartl & Clark 2006; Holsinger & Weir 2009). Several core population genomics statistics and analyses are now included in the *Stacks* software package (see Catchen et al. 2012, this volume), including kernel-smoothed values of π, F_{IS} and F_{ST}. Values for these statistics are part of the standard output of *Stacks*. All plots of π, F_{IS} and allele frequency spectrum presented in this paper were produced by analyzing these output data in the software package R (R Core Team, 2012). The neighbor-joining method of clustering of populations was implemented using the program *Neighbor* from the *Phylip* suite of programs (http://evolution.genetics.washington.edu/phylip.html) using average pairwise F_{ST} values as input, and the resultant tree was visualized using *FigTree* (http://tree.bio.edu.ac.uk/software/figtree/).

In order to analyze the organization of the populations using multilocus genotypic information, we used the `populations` program in *Stacks* to output SNP data from across all RAD sites into a *STRUCTURE*-format file (Pritchard *et al.* 2000; Falush *et al.* 2003, 2007; Hubisz *et al.* 2009). Due to computational limitations of handling many more than this number of loci in the current *STRUCTURE* application, we implemented a custom Perl script to randomly choose 1000 of these SNPs. A *STRUCTURE* output option, along with several other output options, is included in *Stacks* version 1.0 (see Catchen *et al.* 2012, this volume).

STRUCTURE analyses were performed on the complete set of data comprising all nine populations, and then separately on each population to more precisely determine if, within populations, fine structure exists but was obscured by the major axes of structure in the overall dataset analysis. For all analyses 20,000 burn-in steps and 20,000 replicates were used, with at least 10 replicates for each of several values of K, where K is the number of genotypic groups. For the entire population set, K ranged from 1 to 10, whereas for the analysis of each separate population the values of K ranged from 1 to 5. The optimal K for each analysis was chosen using the deltaK method of Evanno *et al.* (2005) and visual inspection of the change in the *Ln P(D)* of each model.

We also imported this same set of 1000 SNPs into the program *GenoDive* (Meirmans and Van Tienderen, 2004) to perform various tests of population differentiation. We first calculated pairwise F_{ST} values for all population pairs, accompanied by 80,000 randomization tests in order to test the hypothesis that each F_{ST} value is different from zero. We used a strict Bonferonni correction due to the multiple comparisons (Rice 1989). We also tested a higher level of population structure (populations nested within geographic region) using an AMOVA based approach (Excoffier *et al.* 1992) implemented in GenoDive. Finally, to fully visualize the major axes of genetic variation, we performed a principle component analysis (PCA) with an associated permutation test among populations

(10,000 replicates, Goudet et al. 1996) to identify axes that contribute to population structure, and plotted population means for the first two major axes of genetic variation. We used the same set of 1000 loci in *Spagedi* (Hardy and Vekemans, 2002) and *Arlequin* (Excoffier and Lischer, 2010), and obtained qualitatively similar results across all packages as well as in *Stacks* Finally, we examined the consistency across multiple independent random samples of 1000 loci and found the results to be very robust.

Results

Sequence data quality and processing

Ten lanes of sequencing produced a total of more than 821 million reads derived from 578 individuals, of which greater than 599 million reads passed our stringent quality threshold (Table 2, Supp. Fig. 1). After requiring loci to be present in all nine populations and in at least 75% of individuals from each population, 25,679 RAD loci were retained that contained zero, one, or more SNPs. All populations were well represented in the data set (Table 2 and Supp. Fig. 1). An average of 84% of the reads were aligned to the genome. Certain reads did not uniquely align primarily because they were from RAD sites that existed in regions of highly repetitive sequences, or because they fell into gaps in the present reference stickleback genome assembly. This second scenario is clearly identified because many of the unaligned reads form perfectly good stacks *de novo* (i.e., without genome alignment). Of the reads that were aligned to the reference genome, nearly 99% of them were used in subsequent analyses (Table 2). In all cases significant sample sizes were obtained to address our research goals and provide precise point estimates for population genomic statistics.

Genetic diversity within Oregon threespine stickleback populations

For all loci that were polymorphic in at least one population in the entire data set, the average major allele frequency (P) ranged from 0.948 to 0.983, and the average observed heterozygosity ranged from 0.0229 to 0.0714 (Table 3). When considering all nucleotide positions, including those not polymorphic anywhere in the dataset, as expected the values increase to 0.998–0.999 for the major allele frequency P, and the observed average heterozygosity decreases to a range of 0.0011 to 0.0034. The central Oregon populations demonstrated markedly reduced levels of genetic diversity as compared to the other populations (Table 3). This reduction in genetic variation of the central Oregon populations was particularly evident in the percentage of loci that are polymorphic. For sites that were polymorphic in at least one of our nine populations (Table 3, “Variant positions”), the percentage of polymorphic loci in coastal Oregon and Willamette Basin populations ranged from approximately 26% to 62%. For all three central Oregon populations, however, the percentage of polymorphic loci was between only 6% and 10%. Therefore, the coastal oceanic populations have approximately ten times the number of polymorphic loci present in the central Oregon populations.

In a previous study using similar RAD-seq approaches (Hohenlohe et al. 2010b) the overall nucleotide diversity (π) in stickleback populations in post-glacial Alaska was found to be 0.00336 across populations and between 0.0020–0.0027 within each population (Hohenlohe *et al.* 2010b). In concordance with these previous findings, we found similar levels of average nucleotide diversity in Oregon populations (0.0011–0.0037). However, when considering the geographic distribution of the nucleotide diversity averaged across all sites, it is clear that the Oregon coastal and Willamette Basin locations demonstrate markedly higher values of nucleotide diversity (0.0028 to 0.0037) as compared to the Alaskan populations (0.0020–0.0027). In sharp contrast, we found notably lower genetic diversity in the central Oregon populations (Table 3) than in Alaska and other populations from Oregon,

with π values ranging from 0.0011 to 0.0014, and with the lowest values from the high altitude Paulina Lake.

Related to these diversity measures, we found the allele frequency spectrum of major alleles across loci to be markedly variable across the populations (Figure 3). Whereas the coastal and Willamette Basin populations all exhibited a distribution of major allele frequencies consistent with older, genetically diverse populations at or near evolutionary equilibrium, all three central Oregon populations exhibited spectra of allele frequencies that were skewed towards 1.0. Skewed distributions of this sort are indicative of young populations that were formed by a small set of founder individuals, and that have yet to reach evolutionary equilibrium through the acquisition of new mutations or migrant alleles (Fisher 1958; Lewontin & Krakauer 1973; Wright 1978; Nei 1987; Nei and Kumar 2000; Hartl & Clark 2006).

Genetic and geographic relatedness of Oregon stickleback populations

To assess the genetic relatedness of Oregon stickleback populations, we calculated the average F_{ST} for pairwise comparisons of all sampled populations in the present study (Table 4 and Figure 4). Using a numerical randomization approach of 1000 randomly chosen loci, we found that all pairwise F_{ST} values were statistically significantly different from 0 at the alpha = 0.05 level when using a strict Bonferroni correction for multiple comparisons ($p < 0.00012$ for each comparison). Because of the very large amount of data included in this study, however, it should be noted that although all values are statistically significant, some values are very close to zero and therefore may not be biologically significant. The coastal populations clearly grouped together, with the three oceanic populations only slightly differentiated from one another (F_{ST} values of 0.003, 0.009 and 0.010 for the three pairwise comparisons). The coastal freshwater populations grouped together, and as expected under the hypothesis of independent derivation of each from the oceanic population, were approximately twice as differentiated from one another ($F_{ST} \sim 0.06$) than either was from the oceanic populations ($F_{ST} \sim 0.03$).

In contrast, the Riverbend Willamette Basin freshwater population was notably more divergent from the coastal oceanic and freshwater populations (F_{ST} values of 0.11 to 0.16 with respect to the oceanic populations, and 0.20 to 0.28 when compared to the coastal freshwater populations). The central Oregon populations of Paulina Lake, South Twin Lake and Crooked River all showed close genetic relatedness to one another, with pairwise F_{ST} values ranging from 0.022 to 0.058. Most surprising, these central Oregon populations also clustered together closely with Riverbend despite the large geographic separation between the Willamette Basin and central Oregon. In fact, the Riverbend and South Twin Lake populations were indistinguishable from one another (F_{ST} of ~ 0).

We also used an AMOVA approach to more precisely partition the genetic variation across individuals within populations, populations within regions (coastal, Willamette Valley and central Oregon), and across regions (Table 5). These results confirmed the pairwise phylogeographic and F_{ST} results. We found that the majority (~40%) of the genetic variation that is not attributable to variation among individuals (54%) was partitioned across regions, whereas only about 6% of the variation was accounted for by differences among populations nested within the geographic regions. A principle components analysis (PCA) of the entire genetic data set supported this final point. The first PC explained nearly 90% of the total variation in the data set (Table 6), and was the only statistically significant PC as determined by permutation. A bivariate plot of population means along PC1 and PC2 axes clearly shows that this first major axis defines the differences between coastal populations and those in the Willamette Valley and central Oregon (Figure 5).

Among and within population structure of Oregon stickleback

To test for any indication of cryptic assortative mating or hidden population structure, we examined Wright's inbreeding coefficient, $F_{IS,}$ which measures reductions in observed heterozygosity with respect to that expected under Hardy-Weinberg Equilibrium (HWE; Wright 1931; Wright 1978; Slatkin 1991; Charlesworth 1998). An usually large average value of F_{IS} can indicate significant assortative mating or cryptic population genetic structure and recent hybridization. When considering all loci that are polymorphic across Oregon populations, the average values of F_{IS} did not indicate significant cryptic population structure or assortative mating (Table 3). Although the distribution of F_{IS} values supports this primary point, a small but marked collection of outlier loci with significant F_{IS} could clearly be seen when comparing the global distribution of values across all loci in all populations (Supp. Fig. 2). Similarly, within each population the majority of loci had an F_{IS} value of zero or nearly so, indicating a lack of overall cryptic population structure. However, for the coastal oceanic populations (top row of Fig. 6), a noticeable fraction of loci exhibited F_{IS} values greater than one. These higher values were present, but to a lesser degree, in the coastal freshwater populations (Pony Creek and Winchester Creek) and the Willamette Basin population, and noticeably absent in the central Oregon populations (bottom row of Fig. 6). When only the loci polymorphic locally within each focal population were examined, however, the average F_{IS} values increased markedly within all but the central Oregon populations (Table 3), indicating the possibility of cryptic population structure in coastal and central Oregon populations.

As a further test of potential population structure, we analyzed 1000 randomly chosen SNPs using the software package *STRUCTURE* (Pritchard et al. 2000; Falush et al. 2003, 2007; Hubisz et al. 2009). Because loci in tight linkage should be avoided in *STRUCTURE* analyses (Pritchard et al. 2000), and multiple SNPs originating from a single RAD site are assumed to be in perfect linkage, only one SNP was chosen from each RAD site. By examining the change in *Ln P(D)*, and using the deltaK approach of Evanno et al. (2005), we found that a model with $K = 5$ best fits the data (Supp. Fig. 3). An examination of the plot of posterior probabilities (Fig. 7) clearly shows these five groupings. At the highest level of structure, the Oregon coastal populations are separated from the Willamette Basin and central Oregon populations. The Oregon coast populations were further separated into genotypes that represent a likely combination of oceanic alleles (Fig. 7, red), and two separate sets of freshwater combinations of alleles (Fig. 7, blue and green). These genotypic combinations appear to be present to varying degrees in individuals in each of the populations, supporting the possibility of cryptic population structure due to admixture between locally adapted marine and freshwater populations.

A more focused analysis using *STRUCTURE* within each population supported the hypothesis of introgressive hybridization between differentiated populations in Millport Slough, and potentially in Winchester Creek and South Jetty, although additional data will be required to fully test this hypothesis. In particular, Millport Slough exhibited equal proportions of two different genotypic classes in most individuals in the overall analysis, and a focused *STRUCTURE* analysis supported the presence of two genotypic clusters (Supp. Fig. 3).

In contrast to the patterns of mixed classes of genotypes in the coastal populations, the Willamette Basin and central Oregon populations tell a simpler story. All central Oregon populations clearly fell into a single group, and the dominant genotypes of these populations were present at low frequency in the Riverbend Willamette Basin population. Little evidence existed for cryptic population structure in Riverbend when examined individually using *STRUCTURE*, and the presence of some central Oregon genotypes (yellow in Fig. 7) in the

overall analysis was likely due to the sampling of a subset of alleles from this or a related valley population to form the central Oregon populations.

Genomic architecture of nucleotide diversity and inbreeding coefficients

We performed sliding window analyses of both nucleotide diversity (π) and Wright's inbreeding co-efficient F_{IS} (Figures 8 and 9), and found significant variation in values of π across the stickleback genome in the various populations. As expected, we found that overall levels of diversity across the stickleback genome were highest in the coastal populations. This result is expected for populations experiencing significant levels of gene flow. Supporting this inference was the observation that regions of higher and lower genetic diversity were localized to similar regions of the stickleback genome in each of these populations (Fig. 8). For example, much of linkage group IV (LGIV) exhibited elevated levels of genetic diversity in all four populations.

The Riverbend population exhibited slightly lower overall levels of genetic diversity and had distinct patterns of genomically localized high or low values of π as compared to the coastal populations. Examination of the π values for South Twin Lake, Paulina Lake and Crooked River revealed a strikingly different pattern, however. In addition to a clear global average reduction of genetic diversity, patterns of increased diversity were still evident in some genomic regions. However, these genomic regions often did not match those present in the coastal populations. For example, although the distal end of LGIV exhibited high levels of genetic diversity, most of the rest of the linkage group showed very low levels of diversity, in clear contrast to the results from the coastal and Willamette Basin populations.

The genome-wide pattern of F_{IS} values across populations was even more telling. As expected from the structure and average F_{IS} analyses, the Cushman Slough and South Jetty populations showed low average F_{IS} and variation in values across the genome that were episodically punctuated by high positive values. A notable exception was LGXIX, which consistently exhibited outlier F_{IS} values in all populations due to the presence of a non-recombining sex determining region on this linkage group (Peichel et al. 2004; Ross and Peichel 2008; Ross et al. 2009). A pattern of low average F_{IS} and variance, as seen in these populations, is indicative of older, stable populations. Another pattern emerged when examining Pony Creek Reservoir and Riverbend, and particularly Winchester Creek and Millport Slough. Many more genomic regions exhibited higher F_{IS} values and more significant variation, as one might expect due to cryptic population structure stemming from processes such as recurrent introgressive hybridization between differentiated populations. For example, regions of LGVII in Winchester Creek exhibited large regions of increased F_{IS}. Finally, the genomic F_{IS} patterns in the three central Oregon populations (South Twin Lake, Paulina Lake and Crooked River) presented strikingly different patterns from the rest of the populations, with much more significant variation in F_{IS} values. Numerous locations of the genome showed strong positive values of F_{IS}. Remarkably, many of these values were strongly negative, which is difficult to explain in stable populations because it represents an excess of heterozygosity over that expected by HWE.

Discussion

Genetic diversity within Oregon coastal and Willamette Basin threespine stickleback populations is higher than previously documented in Alaska

We found significant amounts of genetic diversity in each of the Oregon threespine stickleback populations, the levels of which were in most cases demonstrably higher than previously documented in RAD-seq studies of Alaskan stickleback populations (Hohenlohe *et al.* 2010b). Whereas the average values of nucleotide diversity (π) in Alaskan oceanic and

freshwater populations were between 0.0020–0.0027 within each population, we document here that the average within population π levels ranged from 0.0028 to 0.0037 in the coastal and Willamette Basin locations, an approximately 50% increase in genetic diversity over the values seen in Alaska. In contrast, the values of π were much lower in the central Oregon populations (0.0011 to 0.0014) than in other Oregon populations, or in Alaskan oceanic and freshwater populations. These patterns of genetic diversity extended across the entire genome as can be seen in the genome-wide plots of π (Figure 8). In particular, the central Oregon populations showed a genome-wide reduction in nucleotide diversity, although some discrete segments of increased genetic diversity were retained. These findings indicate that coastal and Willamette Basin populations maybe older, and perhaps larger, than those in Alaska, or are experiencing more recurrent gene flow in the case of the coastal populations, allowing for greater accumulation of genetic variation.

Genetic and geographic relatedness of Oregon threespine stickleback populations

Despite the greater levels of overall genetic diversity, we found similar levels of divergence between coastal oceanic and freshwater populations in Oregon (F_{ST} values from ~0.02 to 0.05) as had previously been reported for Alaskan populations (Hohenlohe *et al.* 2010b). However, we discovered significantly more genetic partitioning between the inland and the oceanic populations (F_{ST} from ~ 0.07 to 0.17), and even more divergence between the freshwater coastal and inland stickleback populations (F_{ST} ~ 0.16 to 0.31). This increased level of genetic diversity may reflect an expanded amount of time since the founding of the inland Oregon populations as compared to the relatively young Alaskan populations, and continual gene flow between the Oregon coastal oceanic and coastal freshwater populations that is absent with the inland populations.

Although the divergence between the coastal oceanic and freshwater populations was similar to that previously documented for Alaskan freshwater-ocean pairs, the Willamette Basin and central Oregon populations were very highly differentiated from the coastal populations, with F_{ST} values much higher than we found for any population pairs in Alaska (Hohenlohe et al. 2010b), and among the highest ever documented for this species. Supporting this point, the majority of genetic variation was partitioned between coastal as compared Willamette Valley and central Oregon populations as determined by AMOVA and principle component analyses (Tables 5 and 6, Figure 5). The population genetic data presented here, in combination with the relatively stable geological history of the Willamette Valley and absence of glaciation during the last maximum, supports the hypothesis that inland, freshwater stickleback in Oregon may be much older (millions of years) than those in recently deglaciated regions (thousands of years). If true these populations will be a valuable resource for understanding the global history of ancestral genetic variation that is apparently re-used repeatedly during adaptation to freshwater and oceanic habitats (Colosimo *et al.* 2005; Hohenlohe *et al.* 2010, 2012a; Jones et al. 2012).

Evidence for cryptic population structure in coastal and Willamette Valley populations

The very high levels of genetic diversity seen in near shore oceanic populations, such as Millport Slough and South Jetty, may indicate that these locations are zones of contact of different stickleback adapted to alternative environments increasing the rate of introgressive hybridization. If the individuals we included in our analyses from a single location were from different subpopulations, we would expect significant increases in the percent polymorphic loci and other values of genetic diversity. These observations led us to hypothesize that cryptic structure might exist in some populations. To test this hypothesis we analyzed the genome-wide average values of Wright's inbreeding coefficient F_{IS} for each population, as well as the distributions of this statistic across the stickleback genome in each population. Although we found that, when all loci are examined, the F_{IS} values are

close to zero, the average F_{IS} values are appreciable when examining only the loci that are polymorphic in the focal population (Table 3). Using only these loci is appropriate because the average value could be biased towards zero simply because a large number of monomorphic loci exist within each population.

This pattern of Oregon stickleback population structure can be clearly seen in the distribution of Bayesian posterior probabilities of group assignment for each individual with respect to their collecting location (Fig. 6). The *STRUCTURE* analysis supported five clusters of populations that can be most easily understood in terms of geographic isolation from coastal Oregon to central Oregon. All coastal populations contained a number of individuals with a significant multilocus signature of oceanic alleles. This was particularly true of the Cushman Slough and South Jetty populations, which comprise individuals that primarily belonged to one group (Fig. 6, red). In addition, the freshwater Pony Creek Reservoir population contain individuals that predominately have a multilocus genotype signature that is indicative of a freshwater genotype, and that are also at low frequencies in the Cushman and South Jetty populations. Interestingly, the Millport Slough and Winchester Creek populations appeared to have a significant number of individuals with a range of mixtures of both the oceanic and freshwater genotypes, possibly the result of ongoing hybridization between fish from the two habitats.

Another genotypic combination was only found in Millport Slough (Fig. 6, blue). This novel genotypic combination may be representative of introgressive hybridization with a local freshwater population that was not included in our sample. Subsequent genome analyses of freshwater populations in this region are needed to test this hypothesis. In contrast to the coastal populations, the Riverbend Willamette Basin and central Oregon populations showed a simpler pattern of population structure, with Riverbend comprising genotypic combinations that were largely unique to this sampling location (Fig. 6, purple), while all of the central Oregon populations were designated as members of a separate unit. Interestingly, the central Oregon specific genotypic combinations were present at low frequency in the Riverbend population (Fig. 6, yellow).

Furthermore, when examining the global distribution of F_{IS} values, the majority of loci were close to zero, but an appreciable number were above this value and approaching 1. This may indicate regions of the genome that show population structure due to local adaptation and barriers to introgressive hybridization. Anadromous Alaskan stickleback populations exhibit localized breeding that is offset from local resident freshwater populations (Gelmond *et al.* 2009). In contrast, stickleback seem to breed for more of the year in Oregon, and many of the freshwater and oceanic populations in Oregon have the opportunity for consistent introgressive hybridization (Yeates-Burghart *et al.* 2009).

Genomic localization of diversity signatures

An important benefit of the density of loci that can be assayed via RAD-seq is the possibility of discovering genomic regions of increased population structure that are recalcitrant to gene flow (Via & Hawthorne 2002; Via & West 2008; Feder & Nosil 2009; Via 2009; Nosil & Feder 2012a; Nosil & Feder 2012b). These genomic regions are particularly exciting because they may represent signatures of local adaptation or islands of divergence with gene flow, and may even be islands of speciation (Coyne 1992; Rieseberg 2001; Orr 2005; Hoffmann & Rieseberg 2008; Nosil 2008; Via & West 2008; Strasburg *et al.* 2009; Turner & Hahn 2010; Fan *et al.* 2012; Nosil & Feder 2012a; Nosil & Feder 2012b; Renaut *et al.* 2012; Strasburg *et al.* 2012). As a first look at the genomic localization of these regions, we found more variation in F_{IS} values across the genome in populations that appear to exhibit cryptic population structure (Millport Slough and Winchester Creek in Figure 9). Although the average F_{IS} value for these populations was still near zero, as it was for Cushman Slough

and South Jetty, particular locations exhibited increased levels of F_{IS} (for example, LGVII in Winchester Creek). These genomically localized signals of assortative mating provide a starting point for more detailed studies of potential genomic islands of divergence, particularly in the Millport Slough and Winchester Creek populations.

The origin of central Oregon stickleback populations

The diversity, AMOVA, *STRUCTURE*, F_{IS}, and Principle Component analyses all allowed us to address a specific hypothesis regarding the origins of the unusual central Oregon populations of stickleback, which are some of the most inland and highest altitude populations of stickleback in Western North America. The South Twin Lake and Crooked River are greater than 1000 meters above sea level, and at almost 3000 meters above sea level the population in Paulina Lake is to our knowledge the highest elevation stickleback population. The origin of these central Oregon stickleback has been an open question since significant increases in population densities of these fish were documented in the 1980‘s after not having been previously documented in these watersheds (Oregon Department of Fish and Wildlife, pers. comm.). These stickleback may have been resident for an extensive period of time, and increasing population densities may have been caused by environmental changes due to local (e.g. increasing nutrification) or global (e.g. climate change) anthropogenic environmental changes. Alternatively, these stickleback may have been introduced in the 1970s or early 1980s from a different location in Oregon, and then multiplied dramatically within the first few years after their introduction.

Our data clearly support the latter hypothesis, that stickleback were recently introduced into central Oregon. The percent polymorphic loci, average nucleotide diversities and heterozygosities are lower in the central Oregon populations, and the allele frequency spectra of these populations are biased towards values of 1.0, as would be expected for recently formed populations. The genome-wide patterns of reduced nucleotide diversity and increased variation in F_{IS} also support the recent introduction hypothesis. Finally, F_{ST} values are low between all pairs of Paulina Lake, South Twin Lake, and Crooked River, indicating their close genetic relatedness (Figure 4 and Table 4), despite being physically isolated from one another. Although the pairwise F_{ST} values are very low between Riverbend and the central Oregon populations (effectively zero between RB and STL), the STRUCTURE analysis still placed Riverbend into a different group. This result would occur if, as expected, the central Oregon populations are recently founded from a Willamette population and have similar allele frequencies across loci (influencing F_{ST}), but different multilocus genotypes, which would more significantly affect the Bayesian posterior probabilities in *STRUCTURE*.

Paulina Lake and South Twin Lake formed in isolated volcanic calderas (Sytsma 1985). Paulina Lake is a popular fishing location in the Newberry Crater that likely formed following the collapse of the Newberry volcanic peak 20–25,000 years ago. Paulina Lake has no connectivity from any downstream water bodies, as its perennial outflow to the Deschutes River descends an impassible, vertical waterfall (Sytsma 1985). South Twin Lake has a similar geologic history and also lacks connectivity with other water bodies. More than 700 km upstream and two mountain ranges away from the ocean, Crooked River also has no realistic connectivity with an ancestral stickleback habitat. Despite these biogeographic barriers, the divergence among these central Oregon populations is very low. The strongest piece of evidence for a recent introduction into central Oregon is the observation that the divergence between these populations and the Willamette Basin population are similarly very low despite being separated by hundreds of kilometers. The central Oregon and Willamette Basin populations are as geographically isolated from one another as the Willamette Basin populations are from the coastal populations, and yet the former are much less differentiated than the latter.

In addition, the *STRUCTURE* analyses of these populations provide evidence of similar multilocus genotypes present in the Riverbend population as exist in the central Oregon populations, a scenario expected if the latter populations were founded by a small number of colonists transplanted from the Willamette Basin and representing a subset of the genetic variation present in this region. Finally, the genome-wide pattern of F_{IS} also provides evidence of a recent population expansion. For each of the three central Oregon populations, much more variation in F_{IS} values is seen across the genome than in any other population in our study (Fig. 8). Remarkably, many of these values are strongly negative, which is difficult to explain in stable populations because negative F_{IS} values represent an excess of heterozygosity over that expected by HWE. However, if, as we hypothesize, these populations are the product of recent and rapid expansion, accompanied by significant bouts of ongoing natural selection, then this pattern may not be surprising. Patterns of high variance in F_{IS}, with significant negative values, can be attained in non-equilibrium populations due to founder effects and rapid expansion, associative overdominance due to heterogeneous selection, and dynamic patterns of gene flow in an unstable metapopulation.

Our data therefore support what is already a strong hypothesis for the origin of stickleback populations east of the Cascade Mountains; they are the outcome of a recent introduction by humans. Nonetheless, the fish that inhabit these central Oregon locales are phenotypically different from the Willamette Basin population to which they have genetic affinity (data not shown), indicating that selection may have rapidly moved them toward a different freshwater phenotype after introduction. A fruitful area of future research will be to examine the genomic regions that have lead to the local adaptation to the new habitats in central Oregon in the last four decades since their introduction, particularly to the high altitude population in Paulina Lake.

Conclusions

Here we report the first population genomic analyses of Oregon populations of threespine stickleback that live in habitats never glaciated during the last glacial maximum. We find higher levels of genetic diversity in these populations as compared to those in Alaska previously studied using similar RAD-seq methods. Also similar to what we previously documented in Alaska stickleback, we find little overall population structure exists among the oceanic stickleback populations, and small but significant differentiation of the coastal freshwater populations from the oceanic stickleback. In addition, we find evidence of cryptic population genetic structure within local geographic regions, particularly in the coastal oceanic and coastal freshwater populations. These data are consistent with persistent introgressive hybridization between oceanic and freshwater populations along the coast. In contrast, the inland populations of stickleback in both the Willamette Basin and central Oregon are significantly more differentiated from the coastal populations than any pairs of populations we previously examined in Alaska. Finally, the genetic and genomic affinity between Willamette Basin and central Oregon populations clearly support a recent human introduction of stickleback into habitats east of the Cascade Range, providing a novel opportunity to examine the genomic basis of rapid adaptation to these high altitude environments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of the Cresko and Phillips Labs at the University of Oregon for comments and insights into these data. We are grateful for detailed comments and suggestions from Emily Lescak and three anonymous reviewers, who improved the quality of our manuscript. This work has been generously supported by NSF grants IOS-1027283 and DEB-0949053 (WAC), NIH grant 1R24GM079486-01A1 (WAC), the M. J. Murdock Charitable Trust (WAC), and NIH NRSA Ruth L. Kirschstein fellowship F32GM095213-01 (JMC).

References

Akey JM. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. 2009; 19:711–722. [PubMed: 19411596]

Albert AYK, Schluter D. Selection and the origin of species. Curr Biol. 2005; 15:R283–R288. [PubMed: 15854890]

Asmann YW, Wallace MB, Thompson EA. Transcriptome profiling using next-generation sequencing. Gastroenterology. 2008; 135:1466–1468. [PubMed: 18848555]

Baer CF. Among-locus variation in F_{ST}: fish, allozymes and the Lewontin-Krakauer test revisited. Genetics. 1999; 152:653–659. [PubMed: 10353907]

Baird NA, Etter PD, Atwood TS, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE. 2008; 3:e3376. [PubMed: 18852878]

Beaumont MA, Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol. 2004; 13:969–980. [PubMed: 15012769]

Begun DJ, Holloway AK, Stevens K, et al. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans* . Plos Biol. 2007; 5:e310. [PubMed: 17988176]

Bell MA. Turner BJ. Evolutionary phenetics and genetics. The threespine stickleback, *Gasterosteus aculeatus*, and related species. Evolutionary Genetics of Fishes. 1984:431–527.

Bell, MA.; Andrew, CA. Evolutionary consequences of postglacial colonization of freshwater by primitively anadromous fishes. In: Streit, B.; Stadler, T.; Lively, CM., editors. Evolutionary Ecology of Freshwater Animals. Basel: Birkhauser Verlag; 1997.

Bell, MA. Palaeobiology and evolution of threespine stickleback. In: Bell, MA.; Foster, SA., editors. The Evolutionary Biology of the Threespine Stickleback. Oxford: Oxford University Press; 1994. p. 1-27.

Bell MA. Lateral plate evolution in the threespine stickleback: getting nowhere fast. Genetica. 2001; 112–113:445–461.

Bell MA, Aguirre WE, Buck NJ. Twelve years of contemporary armor evolution in a threespine stickleback population. Evolution. 2004; 58:814–824. [PubMed: 15154557]

Bell, MA.; Foster, SA. The evolutionary biology of the threespine stickleback. Vol. xii. Oxford science publications; 1994. 571 p.

Bonin A. Population genomics: a new generation of genome scans to bridge the gap with functional genomics. Mol Ecol. 2008; 17:3583–3584. [PubMed: 18662224]

Boughman JW, Rundle HD, Schluter D. Parallel evolution of sexual isolation in sticklebacks. Evolution. 2005; 59:361–373. [PubMed: 15807421]

Butlin RK. Population genomics and speciation. Genetica. 2010; 138:409–418. [PubMed: 18777099]

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. *Stacks*: building and genotyping Loci de novo from short-read sequences. G3 (Bethesda). 2011; 1:171–182. [PubMed: 22384329]

Catchen J, Hohenlohe P, Bassham S, Amores A, Cresko W. Stacks: an analysis tool set for population genomcis. Molecular Ecology, this volume. 2012

Chan YF, Marks ME, Jones FC, et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science. 2010; 327:302–305. [PubMed: 20007865]

Charlesworth B. Measures of divergence between populations and the effect of forces that reduce variability. Mol Biol Evol. 1998; 15:538–543. [PubMed: 9580982]

Charlesworth B, Nordborg M, Charlesworth D. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genetical Research. 1997; 70:155–174. [PubMed: 9449192]

Colosimo PF, Hosemann KE, Balabhadra S, et al. Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science. 2005; 307:1928–1933. [PubMed: 15790847]

Colosimo PF, Peichel CL, Nereng K, et al. The genetic architecture of parallel armor plate reduction in threespine sticklebacks. PLoS Biol. 2004; 2:E109. [PubMed: 15069472]

Coyne JA. Genetics and speciation. Nature. 1992; 355:511–515. [PubMed: 1741030]

Cresko WA, Amores A, Wilson C, et al. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. Proc Natl Acad Sci U S A. 2004; 101:6050–6055. [PubMed: 15069186]

Cresko WA, McGuigan KL, Phillips PC, Postlethwait JH. Studies of threespine stickleback developmental evolution: progress and promise. Genetica. 2007; 129:105–126. [PubMed: 16897450]

Davey JL, Blaxter MW. RADSeq: next-generation population genetics. Brief Funct Genomics. 2010; 9:416–423. [PubMed: 21266344]

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011; 12:499–510. [PubMed: 21681211]

Dobzhansky, T. Genetics and the Origin of Species. New York, NY: Columbia University Press; 1937.

Emerson KJ, Merz CR, Catchen JM, et al. Resolving postglacial phylogeography using high-throughput sequencing. Proc Natl Acad Sci U S A. 2010; 107:16196–16200. [PubMed: 20798348]

Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes - application to human mitochondrial-DNA restriction data. Genetics. 1992; 131:479–491. [PubMed: 1644282]

Excoffier L, Lischer HEL. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 2010; 10:564–567. [PubMed: 21565059]

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods Mol Biol. 2011a; 772:157–178. [PubMed: 22065437]

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA. Local de novo assembly of RAD paired-end contigs using short sequencing reads. PLoS One. 2011b; 6:e18561. [PubMed: 21541009]

Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005; 14:2611–2620. [PubMed: 15969739]

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003; 164:1567–1587. [PubMed: 12930761]

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes. 2007; 7:574–578. [PubMed: 18784791]

Fan S, Elmer KR, Meyer A. Genomics of adaptation and speciation in cichlid fishes: recent advances and analyses in African and Neotropical lineages. Philos Trans R Soc Lond B Biol Sci. 2012; 367:385–394. [PubMed: 22201168]

Feder JL, Nosil P. Chromosomal inversions and species differences: when are genes affecting adaptive divergence and reproductive isolation expected to reside within inversions? Evolution. 2009; 63:3061–3075. [PubMed: 19656182]

Fisher, RA. The Genetical Theory of Natural Selection. New York, NY: Dover; 1958.

Foster SA, Baker JA. Evolution in parallel: new insights from a classic system. Trends Ecol. Evol. (Amst.). 2004; 19:456–459. [PubMed: 16701305]

Furin CG, von Hippel FA, Bell MA. Partial reproductive isolation of a recently derived resident-freshwater population of threespine stickleback (gasterosteus aculeatus) from its putative anadromous ancestor. Evolution. 2012; 66:3277–3286. [PubMed: 23025615]

Gelmond O, von Hippel FA, Christy MS. Rapid ecological speciation in three-spined stickleback Gasterosteus aculeatus from Middleton Island, Alaska: the roles of selection and geographic isolation. J Fish Biol. 2009; 75:2037–2051. [PubMed: 20738670]

Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resources. 2011; 11:759–769.

Goudet J, Raymond M, De Meeus T, Rousset F. Testing differentiation in diploid populations. Genetics. 1996; 144:1933–1940. [PubMed: 8978076]

Hardy OJ, Vekemans X. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes. 2002; 2:618–620.

Hartl, DL.; Clark, AG. Principles of population genetics. 4th ed. Sunderland, MA: Sinauer Associates; 2006.

Hoffmann AA, Rieseberg LH. Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? Annu Rev Ecol Evol Syst. 2008; 39:21–42. [PubMed: 20419035]

Hohenlohe PA, Bassham S, Currey M, Cresko WA. Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. Philos Trans R Soc Lond B Biol Sci. 2012a; 367:395–408. [PubMed: 22201169]

Hohenlohe PA, Catchen J, Cresko WA. Population genomic analysis of model and nonmodel organisms using sequenced RAD tags. Methods Mol Biol. 2012b; 888:235–260. [PubMed: 22665285]

Hohenlohe PA, Phillips PC, Cresko WA. Using population genomics to detect selection in natural populations: Key concepts and methodological considerations. Int J Plant Sci. 2010a; 171:1059–1071. [PubMed: 21218185]

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 2010b; 6:e1000862. [PubMed: 20195501]

Holsinger KE, Weir BS. Genetics in geographically structure populations: defining, estimating and interpreting F_{ST} . Nature Reviews Genetics. 2009; 10:639–650.

Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour. 2009; 9:1322–1332. [PubMed: 21564903]

Jones FC, Grabherr MG, Chan YF, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012; 484:55–61. [PubMed: 22481358]

Kimmel CB, Cresko WA, Phillips PC, et al. Independent axes of genetic variation and parallel evolutionary divergence of opercle bone shape in threespine stickleback. Evolution. 2012a; 66:419–434. [PubMed: 22276538]

Kimmel CB, Hohenlohe PA, Ullmann B, Currey M, Cresko WA. Developmental dissociation in morphological evolution of the stickleback opercle. Evol Dev. 2012b; 14:326–337. [PubMed: 22765204]

Kimmel CB, Ullmann B, Walker C, et al. Evolution and development of facial bone morphology in threespine sticklebacks. P Natl Acad Sci Usa. 2005; 102:5791–5796.

Knecht AK, Hosemann KE, Kingsley DM. Constraints on utilization of the EDA-signaling pathway in threespine stickleback evolution. Evol Dev. 2007; 9:141–154. [PubMed: 17371397]

Lewontin RC, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics. 1973; 74:175–195. [PubMed: 4711903]

Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE. On the origin and spread of an adaptive allele in deer mice. Science. 2009; 325:1095–1098. [PubMed: 19713521]

Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008a; 9:387–402. [PubMed: 18576944]

Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008b; 24:133–141. [PubMed: 18262675]

Marguerat S, Wilhelm BT, Bahler J. Next-generation sequencing: applications beyond genomes. Biochem Soc Trans. 2008; 36:1091–1096. [PubMed: 18793195]

Maynard, Smith J.; Haigh, J. The hitch-hiking effect of a favourable gene. Genetical Research. 1974; 23:23–35. [PubMed: 4407212]

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol. 2011

McKinnon JS, Rundle HD. Speciation in nature: the threespine stickleback model systems. Trends in ecology & evolution. 2002; 17:480–488.

McPhail JD. Ecology and evolution of sympatric sticklebacks (Gasterosteus): Morphological and genetic evidence for a species pair in Enos Lake, British Columbia. Can. J. Zool. 1984; 62:1402–1408.

Meirmans PG, Van Tienderen PH. GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. Molecular Ecology Notes. 2004; 4:792–794.

Miller CT, Beleza S, Pollen AA, et al. cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. Cell. 2007; 131:1179–1189. [PubMed: 18083106]

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 2007; 17:240–248. [PubMed: 17189378]

Nei, M. Molecular Population Genetics and Evolution. New York, N.Y.: Elsevier; 1975.

Nei, M. Molecular Evolutionary Genetics. New York, N.Y.: Columbia University Press; 1987.

Nei, M.; Kumar, S. Molecular Evolution and Phylogenetics. New York, N.Y.: Oxford University Press; 2000.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. Nat Rev Genet. 2007; 8:857–868. [PubMed: 17943193]

Nosil P. Speciation with gene flow could be common. Mol Ecol. 2008; 17:2103–2106. [PubMed: 18410295]

Nosil P, Feder JL. Genomic divergence during speciation: causes and consequences. Philos Trans R Soc Lond B Biol Sci. 2012a; 367:332–342. [PubMed: 22201163]

Nosil P, Feder JL. Widespread yet heterogeneous genomic divergence. Mol Ecol. 2012b; 21:2829–2832. [PubMed: 22676072]

Nosil P, Funk DJ, Ortiz-Barrientos D. Divergent selection and heterogeneous genomic divergence. Mol. Ecol. 2009; 18:375–402.

Orr HA. The genetic basis of reproductive isolation: insights from Drosophila. P. Natl Acad Sci Usa. 2005; 102(Suppl 1):6522–6526.

Peichel CL, Nereng KS, Ohgi KA, et al. The genetic architecture of divergence between threespine stickleback species. Nature. 2001; 414:901–905. [PubMed: 11780061]

Peichel C, Ross J, Matson C, et al. The Master Sex-Determination Locus in Threespine Sticklebacks Is on a Nascent Y Chromosome. Current Biology. 2004; 14:1416–1424. [PubMed: 15324658]

Peichel CL. Fishing for the secrets of vertebrate evolution in threespine sticklebacks. Dev Dyn. 2005; 234:815–823. [PubMed: 16252286]

Pool JE, Hellmann I, Jensen JD, Nielsen R. Population genetic inference from genomic sequence variation. Genome Res. 2010; 20:291–300. [PubMed: 20067940]

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155:945–959. [PubMed: 10835412]

Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol. 2010; 20:R208–R215. [PubMed: 20178769]

Renaut S, Maillet N, Normandeau E, et al. Genome-wide patterns of divergence during speciation: the lake whitefish case study. Philos Trans R Soc Lond B Biol Sci. 2012; 367:354–363. [PubMed: 22201165]

Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L. SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (Coregonus spp.). Mol Ecol. 2011; 20:545–559. [PubMed: 21143332]

Rice WR. Analyzing tables of statistical tests. Evolution. 1989; 43:223–225.

Rieseberg LH. Chromosomal rearrangements and speciation. Trends Ecol Evol. 2001; 16:351–358. [PubMed: 11403867]

Ross JA, Peichel CL. Molecular cytogenetic evidence of rearrangements on the Y chromosome of the threespine stickleback fish. Genetics. 2008; 179:2173–2182. [PubMed: 18689886]

Ross JA, Urton JR, Boland J, Shapiro MD, Peichel CL. Turnover of sex chromosomes in the stickleback fishes (gasterosteidae). PLoS Genet. 2009; 5:e1000391. [PubMed: 19229325]

Core, R. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012. Team. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Shapiro MD, Marks ME, Peichel CL, et al. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature. 2004; 428:717–723. [PubMed: 15085123]

Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotech. 2008; 26:1135–1145.

Slatkin M. Inbreeding coefficients and coalescence times. Genet Res. 1991; 58:167–175. [PubMed: 1765264]

Stapley J, Reger J, Feulner PG, et al. Adaptation genomics: the next generation. Trends Ecol Evol. 2010; 25:705–712. [PubMed: 20952088]

Stephan W, Song YS, Langley CH. The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics. 2006; 172:2647–2663. [PubMed: 16452153]

Stinchcombe JR, Hoekstra HE. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity. 2008; 100:158–170. [PubMed: 17314923]

Storz JF. Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol. Ecol. 2005; 14:671–688. [PubMed: 15723660]

Strasburg JL, Scotti-Saintagne C, Scotti I, Lai Z, Rieseberg LH. Genomic patterns of adaptive divergence between chromosomally differentiated sunflower species. Mol Biol Evol. 2009; 26:1341–1355. [PubMed: 19276154]

Strasburg JL, Sherman NA, Wright KM, Moyle LC, Willis JH, Rieseberg LH. What can patterns of differentiation across plant genomes tell us about adaptation and speciation? Philos Trans R Soc Lond B Biol Sci. 2012; 367:364–373. [PubMed: 22201166]

Sytsma, JD., et al. Atlas of Oregon Lakes. Portland State University and Oregon Department of Environmental Quality; 2012 May 12. http://aol.research.pdx.edu/,

Turner TL, Hahn MW. Genomic islands of speciation or genomic islands and speciation? Mol. Ecol. 2010; 19:848–850. [PubMed: 20456221]

Via S. Natural selection in action during speciation. Proc Natl Acad Sci USA. 2009; 106(Suppl 1): 9939–9946. [PubMed: 19528641]

Via S, Hawthorne DJ. The genetic architecture of ecological specialization: correlated gene effects on host use and habitat choice in pea aphids. Am Nat. 2002; 159(Suppl 3):S76–S88. [PubMed: 18707371]

Via S, West J. The genetic mosaic suggests a new role for hitchhiking in ecological speciation. Mol. Ecol. 2008; 17:4334–4345. [PubMed: 18986504]

Wootton, R. The biology of the sticklebacks. Academic Press; 1976.

Wright S. Evolution in Mendelian Populations. Genetics. 1931; 16:97–159. [PubMed: 17246615]

Wright, S. Evolution and the genetics of populations. Vol. vol. 4. Chicago, IL: University of Chicago Press; 1978.

Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005; 21:1859–1875. [PubMed: 15728110]

Yeates-Burghart QS, O'Brien C, Cresko WA, Holzapfel CM, Bradshaw WE. Latitudinal variation in photoperiodic response of the three-spined stickleback Gasterosteus aculeatus in western North America. Journal of Fish Biology. 2009; 75:2075–2081. [PubMed: 20738673]

Figure 1.
Extent of the last glacial maximum approximately 18,000 years ago, showing how the present day State of Oregon outlined in yellow was not glaciated during this period. Image was modified from http://cosmographicresearch.org/Images/glacial_maxium_map2.jpg with permission.

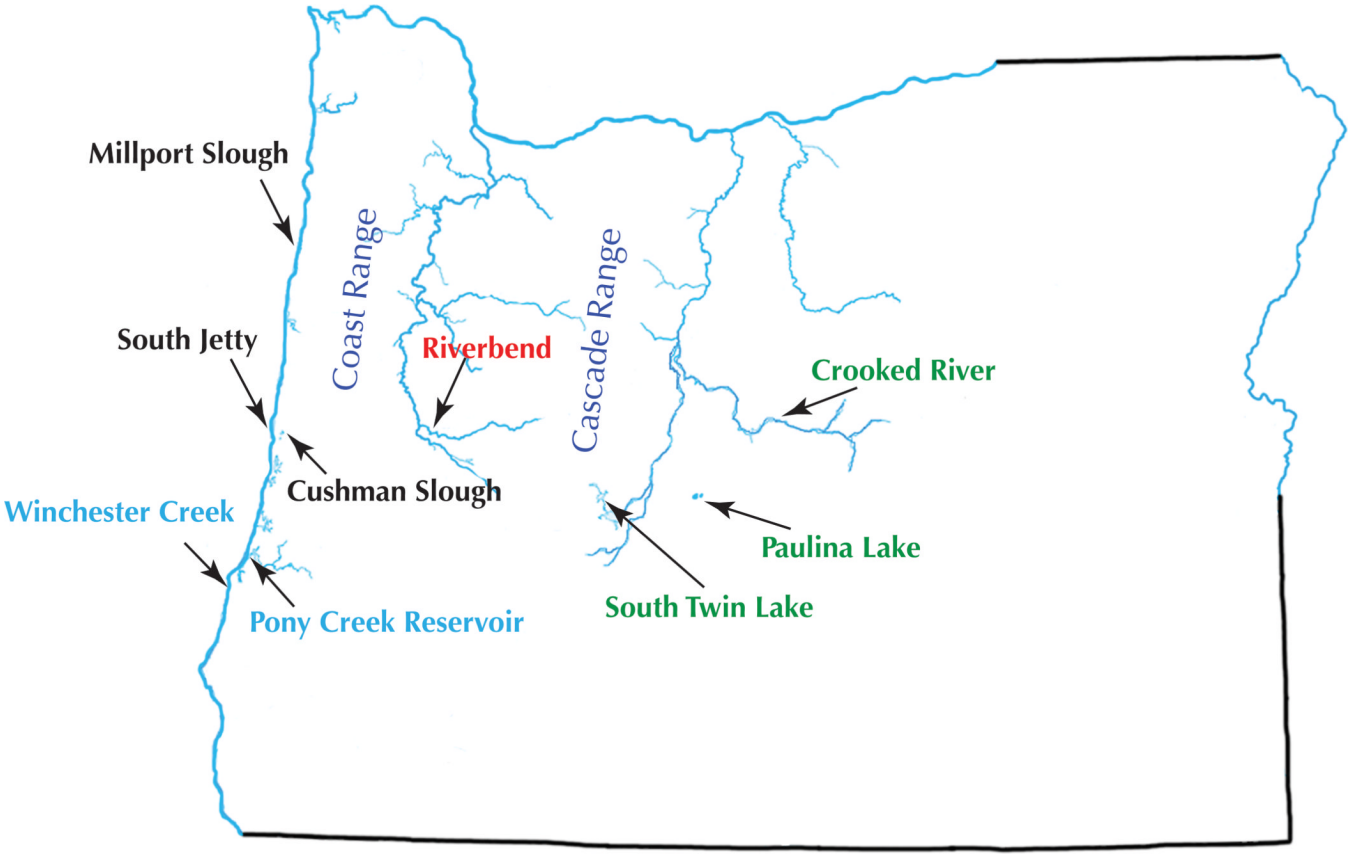

Figure 2.

Map of Oregon sample locations. Populations analyzed in this study come from the Pacific coast (Millport Slough, South Jetty, Cushman Slough, Winchester Creek, Pony Creek Reservoir), the Willamette Basin (Riverbend in the McKenzie River), and central Oregon (South Twin Lake, Paulina Lake, and Crooked River).

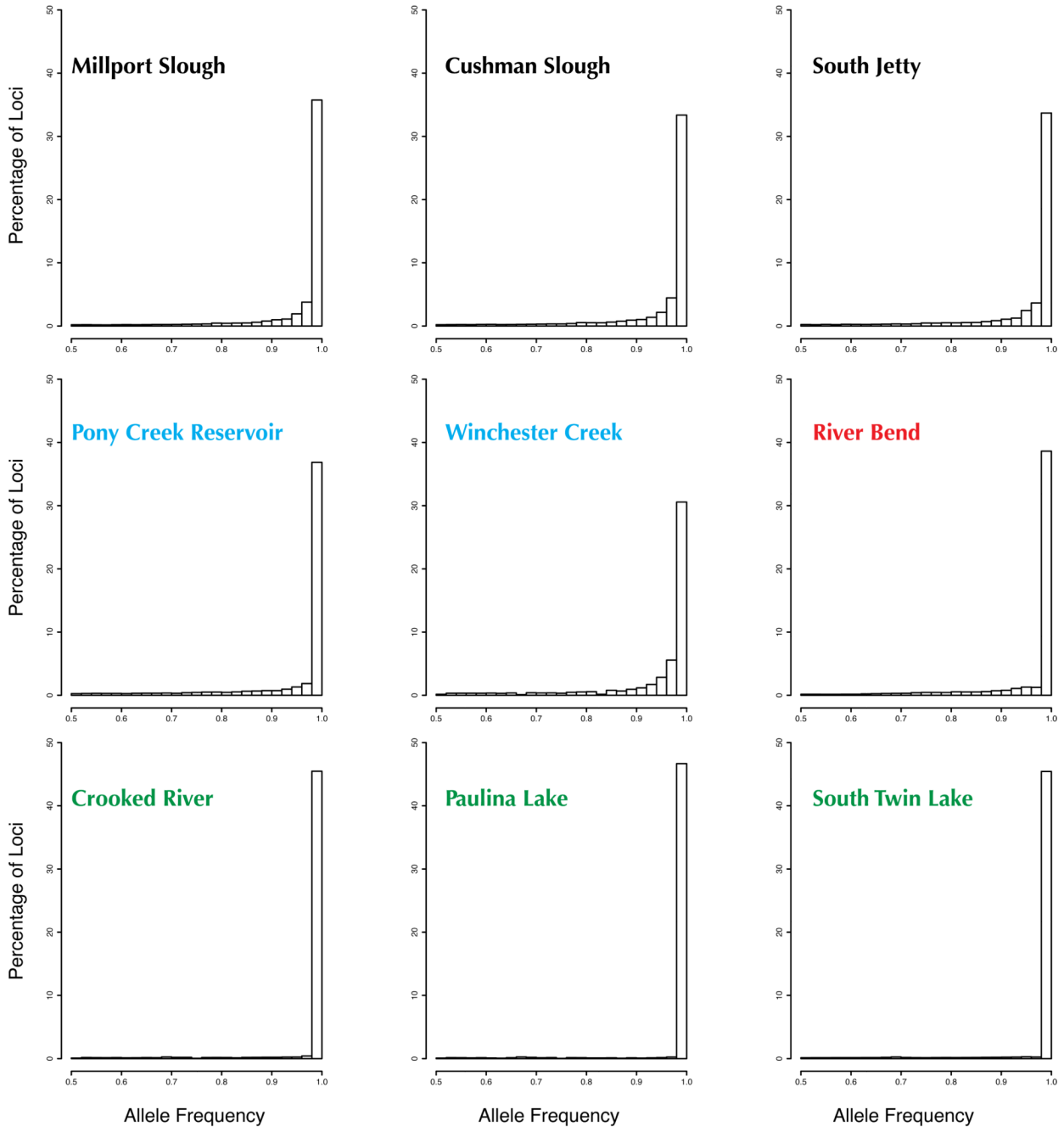

Figure 3.

Allele frequency spectrum distribution for loci in each population that are SNPs in the Oregon system. The x-axis represents categories of the major allele frequency (*p* in each population) with the frequency distribution of the loci in each category on the y-axis. The majority of loci that are polymorphic across populations are fixed within each population (right-most category equal to 1.0). For coastal and Willamette Basin populations a similar range of allele frequencies is seen, from 0.5 to 0.99 as expected for older populations at equilibrium. However, in central Oregon populations (bottom row), the distribution is clearly shifted to the right, as expected if these populations were recently founded by a small

number of individuals. (Coastal oceanic, black; coastal freshwater, blue; Willamette Basin, red; central Oregon, green).

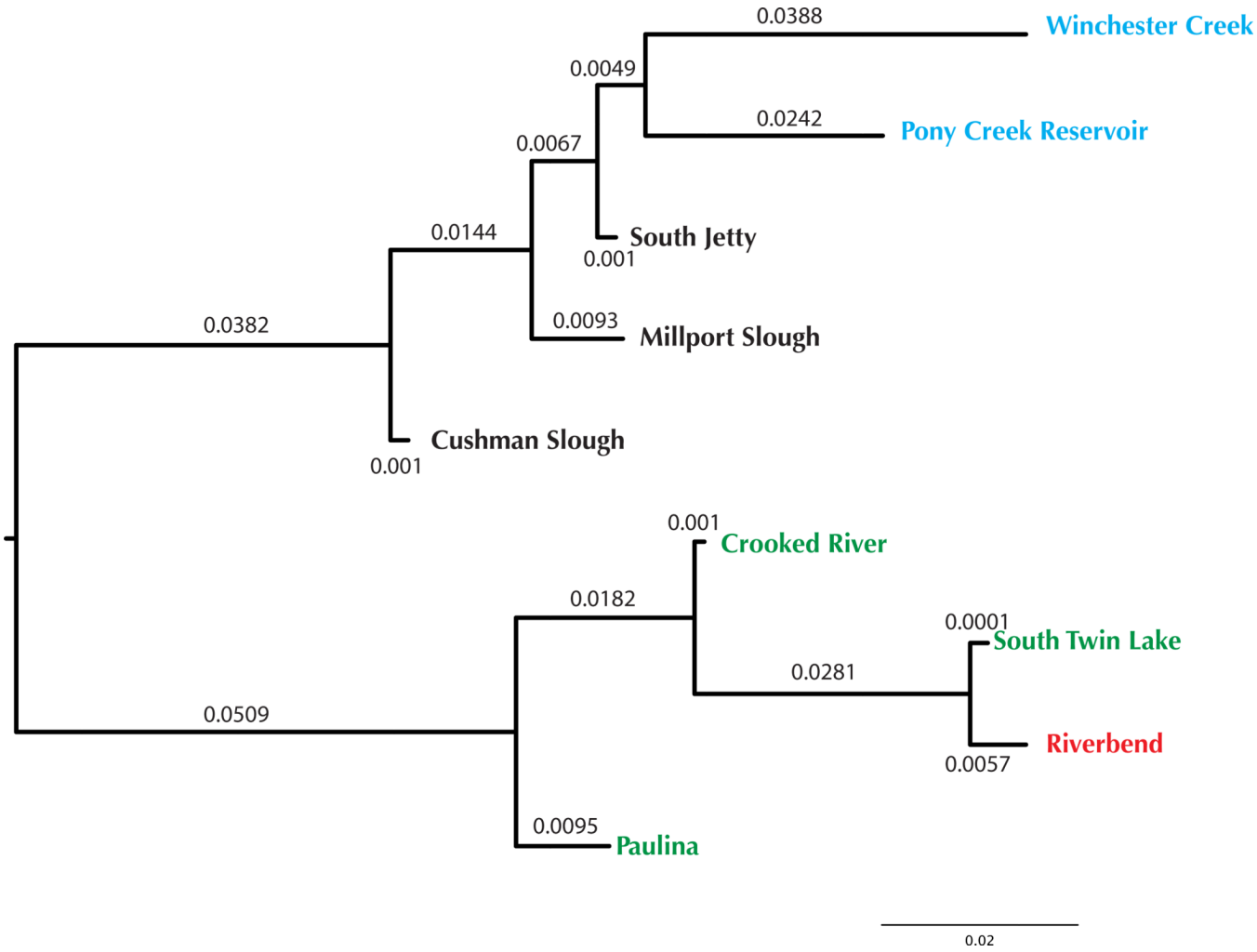

Figure 4.

Neighbor joining tree created using the pairwise F_{ST} values as a distance metric. The tree shows a clear split between the coastal populations (top; oceanic in black and freshwater in blue), and the inland populations (bottom). The central Oregon populations (green) clearly cluster together with the Willamette Basin population from the McKenzie River site at Riverbend (red), supporting the hypothesis of an introduction into central Oregon of stickleback from the Willamette Basin.

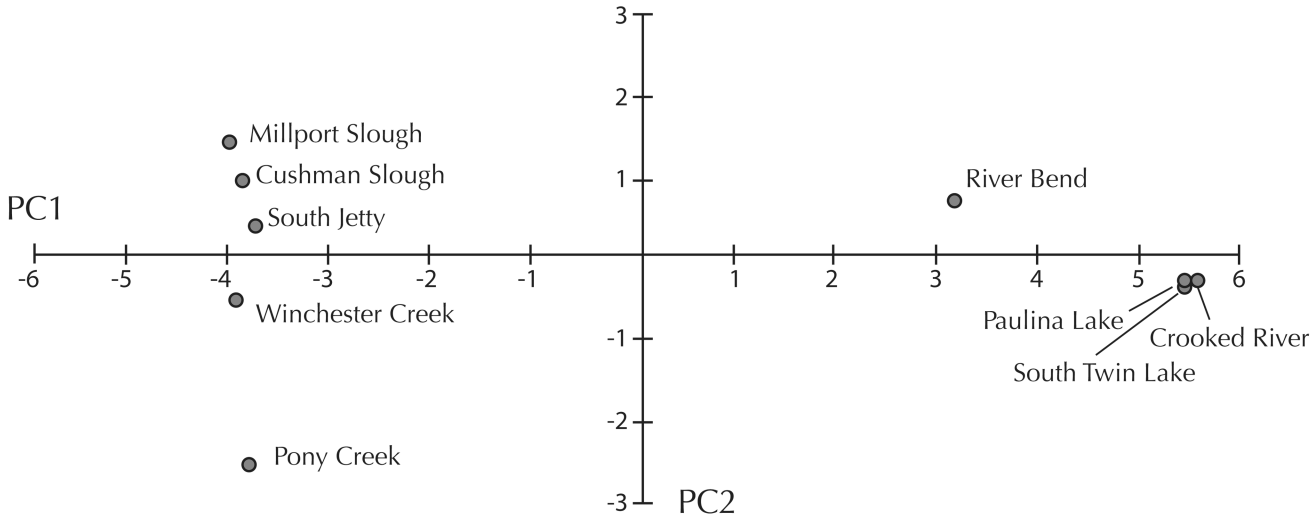

Figure 5.

Distribution of average scores along PC 1 and PC 2 axes of genetic variation for each population. PC1 encompasses the vast majority of the genetic variation (~90%), and clearly differentiates the coastal Oregon stickleback populations from those in the Willamette Valley and central Oregon. Although PC2 explains less of the overall genetic variation, within each region it still has some explanatory power for partitioning populations.

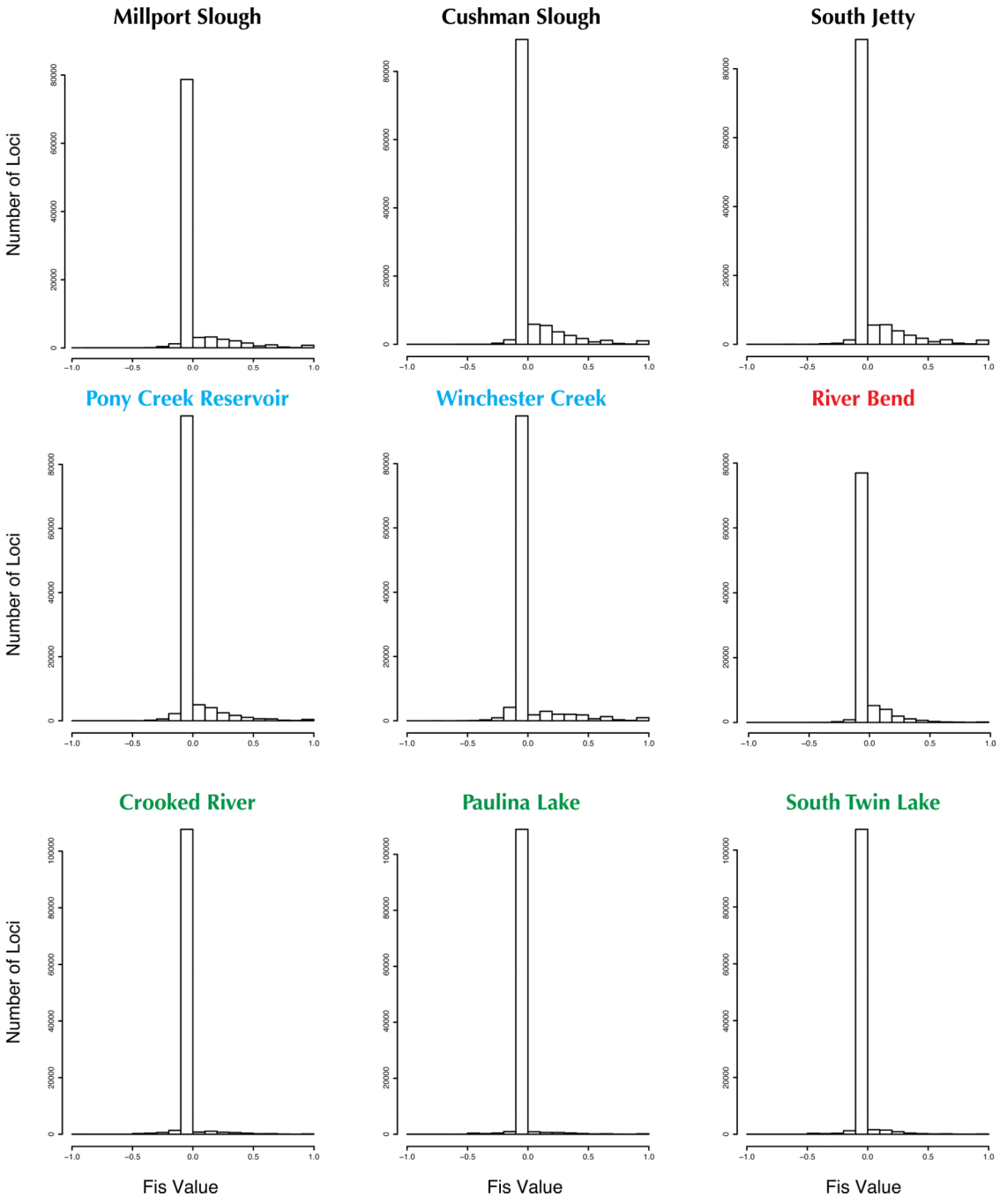

Figure 6.

Frequency distribution of F_{IS} values across loci within each population. The majority of loci within each population are zero or nearly so, indicating a lack of pervasive overall structure within each population. However, for several populations, including the coastal oceanic populations (top row), a noticeable fraction of loci exhibit values greater than one, including an appreciable number that are 1.0. These higher values are present in the coastal freshwater populations (Pony Creek and Winchester Creek), less in Riverbend, and noticeably absent in the central Oregon populations (bottom row). (Coastal oceanic, black; coastal freshwater, blue; Willamette Basin, red; central Oregon, green).

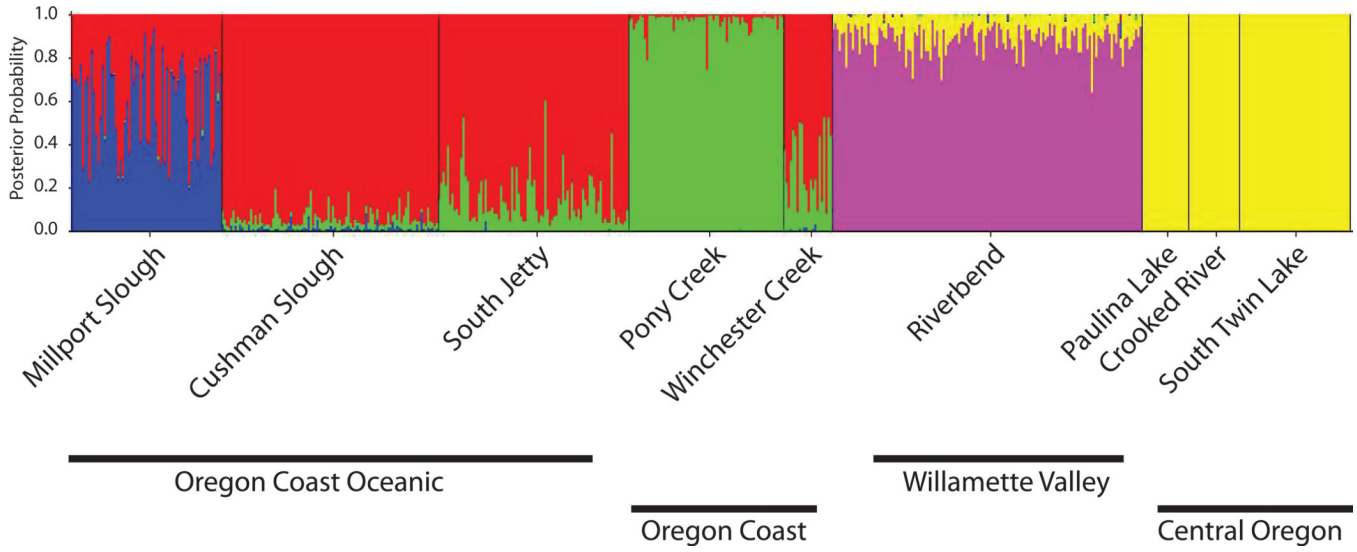

Figure 7.

Plots of posterior probabilities of group assignment of each individual into five clusters based upon the results of a *STRUCTURE* analysis. The results are grouped by population of origin for each individual. Each vertical bar represents a different individual from one of nine populations. The color proportion for each bar represents the posterior probability of assignment of each individual to one of five clusters of genetic similarity. The populations listed from left to right roughly correspond to populations from the Oregon coast eastward to the Willamette Basin and finally to the high desert of central Oregon.

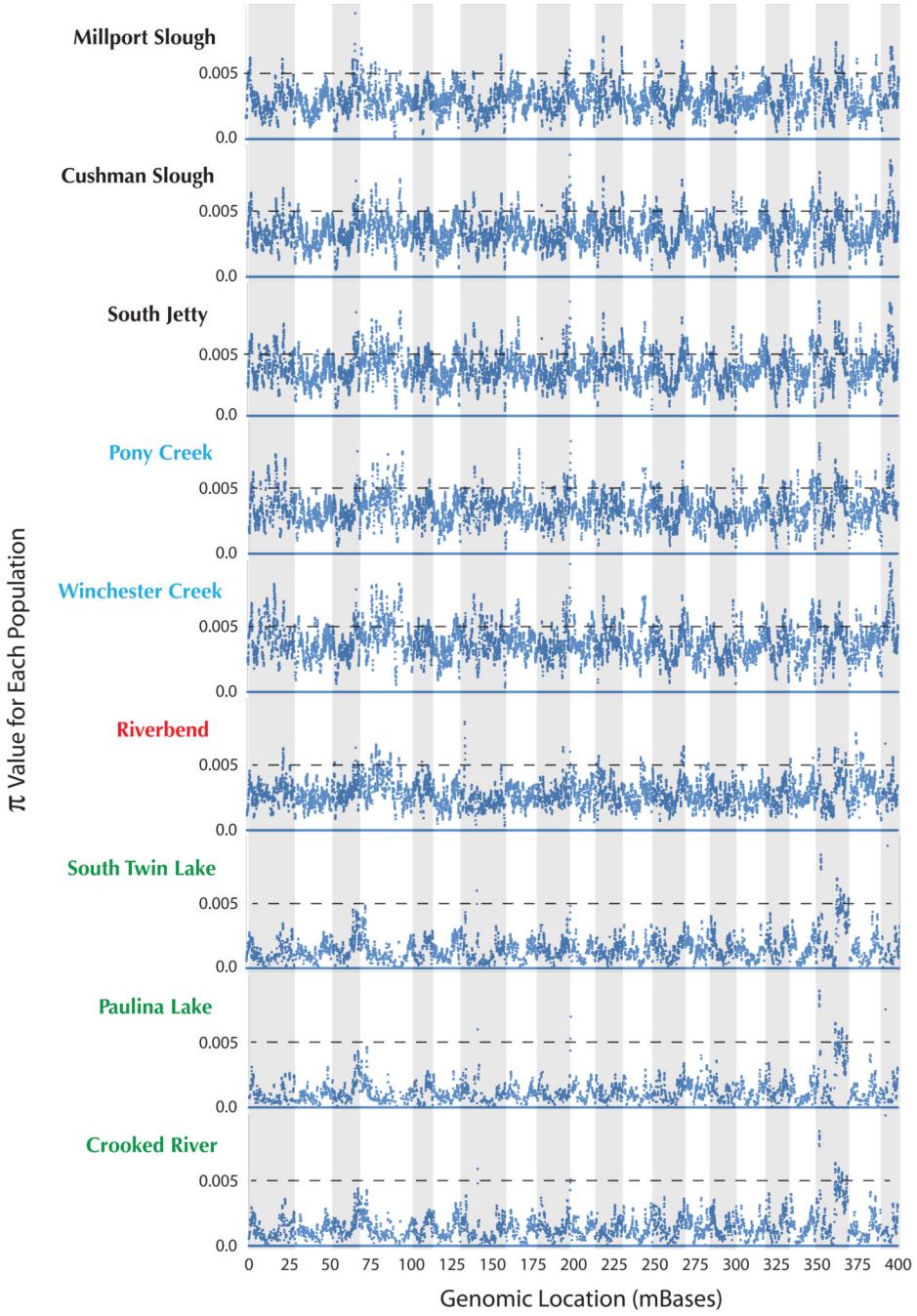

Figure 8.

Genome-wide distribution of smoothed π values across the stickleback genome. The x-axis is the entire stickleback genome in megabase pairs of sequence (mBases), with alternating gray and white representing the stickleback linkage groups. The populations are presented from top to bottom roughly from the coast eastward to central Oregon (coastal oceanic, black; coastal freshwater, blue; Willamette Basin, red; central Oregon, green). Significant variation exists across the genomes of each population, but a general reduction in genetic diversity is clearly seen in the central Oregon populations (South Twin Lake, Paulina Lake and Crooked River).

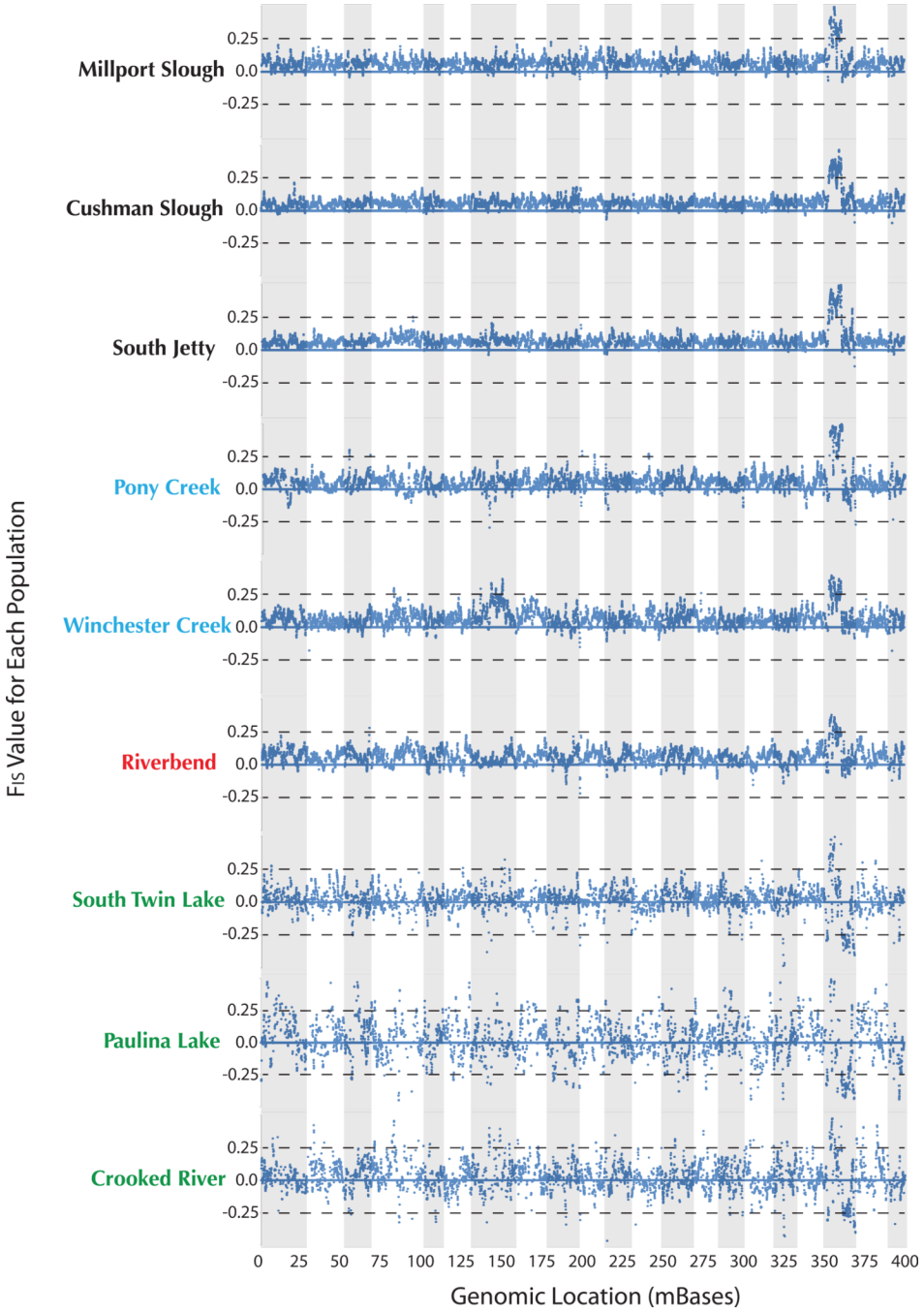

Figure 9.

Genome-wide distribution of smoothed F_{IS} values across the stickleback genome. The populations are presented from top to bottom roughly from the coast eastward to central Oregon. (Coastal oceanic, black; coastal freshwater, blue; Willamette Basin, red; central Oregon, green) For most of the coastal and Willamette Basin populations the values are close to zero, with exceptions of positive F_{IS} values in some genomic locations. Positive values appear to be more frequent in Millport Slough and Winchester Creek, which may be subject to more introgressive hybridization. In contrast, the central Oregon populations show extensive variation in F_{IS} values with nearly as many negative as positive values, as would

be expected for non-equilibrium populations that have experienced recent population growth.

Table 1

Sample collection data, including population name, geographic region, habitat type (FW = freshwater, OC = oceanic), latitude and longitude coordinates, and date of collection.

Population	Geographic Region	Habitat	Collection Site Coordinates	Collection Date
Millport Slough	Oregon Coast	OC	44 53' 14.68"N 123 59'46.20W	April 2010
Cushman Slough	Oregon Coast	OC	43°59'224"N 124°W	2007 Mar 2009
South Jetty	Oregon Coast	OC	44°00'030"N 124°W	Apr, May 2008 Mar, Apr 2009
Pony Creek Reservoir	Oregon Coast	FW	43°42'42.93"N 121°15'25.65"W	Mar 2007
Winchester Creek	Oregon Coast	FW	43°14'47.37"N 124°19'02.68"W	2007
Riverbend, McKenzie River	Willamette Basin	FW	44 04' 38.05"N 123 01'18.11"W	various
South Twin Lake	Central Oregon	FW	43°43'02.93"N 124°07'58.42"W	June 2007
Paulina Lake	Central Oregon	FW	43°42'736"N 121°W	Aug 2008
Crooked River	Central Oregon	FW	44°17'23.17"N 120°50'44.09"W	Aug 2008

Table 2

Summary statistics for all populations of RAD-seq data processing, including the number of individuals included in the final analyses (N), and the read length of Illumina data used (Read Length). For each population the per individual average of raw read counts (Raw Read Count), the number of high quality reads that were successfully aligned to the stickleback genome (GSNAP Aligned), and the number of the aligned reads subsequently fed into *Stacks* (Used by *Stacks*) are presented. These counts are also presented as percentages, with the final column presenting the summary of the overall percentage of raw read counts that were included in the final analysis. The bottom row presents the overall total count of individuals used in the analysis, as well as the total means presented for each of the counts and percentages. These data clearly show that the large majority of raw RAD-seq reads were utilized by *Stacks* for the population genomic analyses.

	N	Read Length	Raw Read Count	GSNAP Aligned	Used by *Stacks*	% raw reads aligned	% aligned reads used by *Stacks*	% overall of raw reads used
Millport Slough	68	74	762,465	649,709	643,188	85.2	98.9	84.3
Cushman Slough	98	95	1,042,323	878,531	872,539	84.3	99.2	83.7
South Jetty	86	95	1,380,400	1,157,152	1,152,357	83.8	99.5	83.4
Pony Creek Reservoir	70	95	1,241,193	1,037,801	1,031,021	83.6	99.2	82.9
Winchester Creek	22	95	1,352,115	1,136,290	1,126,940	84.1	99.2	83.4
Riverbend	140	74	689,025	578,664	562,634	84.0	97.2	81.6
Paulina Lake	21	95	1,185,899	993,219	987,604	83.8	99.4	83.2
Crooked River	23	95	1,123,269	939,813	933,684	83.7	99.3	83.1
South Twin Lake	50	95	1,223,269	1,012,726	1,009,027	82.8	99.6	82.5
TOTAL	**578**		**1,034,691**	**868,248**	**859,930**	**84.0**	**98.8**	**83.0**

Table 3

Summary genetic statistics for all populations split into those calculated for only nucleotide positions that are polymorphic in at least one Oregon population (top, "Variant positions"), as well as all nucleotide positions across all RAD sites regardless of whether they are polymorphic or fixed (bottom, "All positions"). These statistics include the average number of individuals genotyped at each locus (N), the number of variable sites unique to each population (Private), the number of polymorphic (top) or total (bottom) nucleotide sites across the data set (Sites), percentage of polymorphic loci (% poly), the average frequency of the major allele (P), the average observed heterozygosity per locus (H_{obs}), the average nucleotide diversity (π), and the average Wright's inbreeding coefficient (F_{IS}).

	N	Private	Sites	% poly	P	H_{obs}	π	F_{IS}
	Variant positions							
Millport Slough	65.70	7195	91466	45.62	0.958	0.0567	0.0631	0.0341
Cushman Slough	95.61	13110	114056	60.24	0.952	0.0667	0.0734	0.0395
South Jetty	85.40	13813	114036	61.68	0.949	0.0690	0.0766	0.0435
Pony Creek Reservoir	67.66	2139	114651	32.06	0.951	0.0664	0.0710	0.0190
Winchester Creek	21.78	3420	114553	38.70	0.948	0.0714	0.0786	0.0258
Riverbend	134.12	3412	91726	26.33	0.961	0.0540	0.0584	0.0176
South Twin Lake	49.63	271	114742	10.31	0.978	0.0294	0.0296	0.0015
Paulina Lake	20.55	392	114727	6.55	0.983	0.0229	0.0230	0.0011
Crooked River	22.63	1409	114743	8.91	0.980	0.0282	0.0284	0.0013
	All positions							
Millport Slough	66.61	7195	1897050	2.20	0.998	0.0027	0.0030	0.0016
Cushman Slough	96.36	13110	2433350	2.82	0.998	0.0031	0.0034	0.0018
South Jetty	86.01	13813	2433310	2.89	0.998	0.0032	0.0036	0.0020
Pony Creek Reservoir	68.06	2139	2434000	1.51	0.998	0.0031	0.0033	0.0009
Winchester Creek	21.88	3420	2433900	1.82	0.998	0.0034	0.0037	0.0012
Riverbend	136.32	3412	1897210	1.27	0.998	0.0026	0.0028	0.0009
Paulina Lake	20.62	271	2434030	0.31	0.999	0.0011	0.0011	0.0001
Crooked River	22.71	392	2434140	0.42	0.999	0.0013	0.0013	0.0001
South Twin Lake	49.81	1409	2434170	0.49	0.999	0.0014	0.0014	0.0001

Table 4

Pairwise comparison of genetic distance (F_{ST}) and geographic distance (km) among Oregon stickleback populations. Above diagonal; genetic divergence among populations as measured by F_{ST}. Below diagonal; geographic distance between populations (km), along drainage course.

	MPS	CS	SJ	PCR	WC	RB	STL	PL	CR
Millport S.		0.010	0.009	0.043	0.055	0.167	0.124	0.121	0.125
Cushman S.	118		0.003	0.022	0.038	0.111	0.078	0.088	0.093
South Jetty	110	10		0.017	0.028	0.121	0.074	0.083	0.088
Pony Creek R.	199	90	90		0.063	0.202	0.167	0.158	0.162
Winchester C.	194	105	95	20		0.277	0.307	0.170	0.171
Riverbend	617	724	711	804	799		–0.008	0.024	0.028
South Twin L.	854	770^{∞}	760^{∞}	835^{∞}	890^{∞}	1161		0.022	0.022
Paulina L.	833	815^{∞}	805^{∞}	880^{∞}	885^{∞}	1139	75^{∞}		0.058
Crooked R.	747	745	735	765	820	1054	175^{∞}	170^{∞}	

A ∞ symbol denotes no present water connectivity between the two populations.

Table 5

Results from an AMOVA analysis testing the partitioning of genetic variation across populations and geographic regions (coastal, Willamette Valley and central Oregon) that clearly show that a majority of the genetic variation that is not partitioned among individuals, is partitioned across regions. Little of the variation is partitioned across populations nested within region.

Source of Variation	Nested in	%var	F-stat	P-value
Within Individual	--	0.538	F_{IT}	--
Among Individual	Population	0.020	F_{IS}	0.00009999
Among Population	Region	0.038	F_{SC}	0.00009999
Among Region	--	0.404	F_{CT}	0.02169783

Table 6

Eigenvalues of PCs of genetic variation of 1000 randomly chosen loci. The first PC accounts for the vast majority of the genetic variation, and is the only axis that is statistically significantly associated with population differentiation as determined by randomization tests.

	Eigenvalue	%variance	cumulative	p-value
1	22.40	88.80	88.80	0.00099
2	1.13	4.48	93.28	1
3	0.67	2.64	95.93	1
4	0.33	1.30	97.23	1
5	0.31	1.25	98.47	1
6	0.22	0.85	99.33	1
7	0.09	0.34	99.67	1
8	0.08	0.33	100.00	1