Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags

Paul A. Hohenlohe¹*, Susan Bassham¹*, Paul D. Etter², Nicholas Stiffler³, Eric A. Johnson², William A. Cresko¹

1 Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, Oregon, United States of America, 2 Institute of Molecular Biology, University of Oregon, Eugene, Oregon, United States of America, 3 Genomics Core Facility, University of Oregon, Eugene, Oregon, United States of America

Abstract

Next-generation sequencing technology provides novel opportunities for gathering genome-scale sequence data in natural populations, laying the empirical foundation for the evolving field of population genomics. Here we conducted a genome scan of nucleotide diversity and differentiation in natural populations of threespine stickleback (Gasterosteus aculeatus). We used Illumina-sequenced RAD tags to identify and type over 45,000 single nucleotide polymorphisms (SNPs) in each of 100 individuals from two oceanic and three freshwater populations. Overall estimates of genetic diversity and differentiation among populations confirm the biogeographic hypothesis that large panmictic oceanic populations have repeatedly given rise to phenotypically divergent freshwater populations. Genomic regions exhibiting signatures of both balancing and divergent selection were remarkably consistent across multiple, independently derived populations, indicating that replicate parallel phenotypic evolution in stickleback may be occurring through extensive, parallel genetic evolution at a genome-wide scale. Some of these genomic regions co-localize with previously identified QTL for stickleback phenotypic variation identified using laboratory mapping crosses. In addition, we have identified several novel regions showing parallel differentiation across independent populations. Annotation of these regions revealed numerous genes that are candidates for stickleback phenotypic evolution and will form the basis of future genetic analyses in this and other organisms. This study represents the first high-density SNP–based genome scan of genetic diversity and differentiation for populations of threespine stickleback in the wild. These data illustrate the complementary nature of laboratory crosses and population genomic scans by confirming the adaptive significance of previously identified genomic regions, elucidating the particular evolutionary and demographic history of such regions in natural populations, and identifying new genomic regions and candidate genes of evolutionary significance.

Introduction

Population genetics provides a rich and mathematically rigorous framework for understanding evolutionary processes in natural populations. This theory was built over the last hundred years by modeling the processes of selection, genetic drift, mutation and migration in spatially distributed populations [1–6]. The field has concentrated primarily on the dynamics of one or a small number of genetic loci, largely because of methodological limitations. However, genes are not islands, but rather form part of a genomic community, integrated both by physical proximity on chromosomes and by various evolutionary processes [7–10]. With technological advances, such as Next Generation Sequencing (NGS) [11–13], the emerging field of population genomics now allows us to address evolutionary processes at a genomic scale in natural populations [14–20]. Population genetic measures like Wright’s $F$ statistics [2,21,22], traditionally viewed as point estimates, can now be examined as continuous distributions across a genome [23–29]. As a result, in addition to estimating genome-wide averages for such statistics, we can identify specific genomic regions that exhibit significantly increased or decreased differentiation among populations, indicating regions that have likely been under strong diversifying or stabilizing natural selection [9,30–41]. These signatures of selection can then be used to identify candidate pathways, genes and alleles for targeted functional analyses [42–47].

An excellent opportunity for this type of population genomics approach exists in the threespine stickleback, Gasterosteus aculeatus [48–50]. This small fish is distributed holarctically and inhabits a large number of marine, estuarine and freshwater habitats in Asia, Europe and North America. In many regions replicate extant freshwater stickleback populations have been independently derived from oceanic ancestors when stickleback became isolated postglacially in newly created freshwater habitats [49,51]. Population genetic data support this inference, and also indicate that present day oceanic populations can be used as surrogates for stock that gave rise to nearby derived freshwater populations [52–64]. Because of the varied selection regimes in novel habitats,
Author Summary

Oceanic threespine stickleback have invaded and adapted to freshwater habitats countless times across the northern hemisphere. These freshwater populations have often evolved in similar ways from the ancestral marine stock from which they independently derived. With the exception of a few identified genes, the genetic basis of this remarkable parallel adaptation is unclear. Here we show that the parallel phenotypic evolution is matched by parallel patterns of nucleotide diversity and population differentiation across the genome. We used a novel high-throughput sequence-based genotyping approach to produce the first high density genome-wide scans of threespine stickleback populations and identified several genomic regions indicative of both divergent and balancing selection. Some of these regions have been associated previously with traits important for freshwater adaptation, but others were previously unidentified. Within these genomic regions we identified candidate genes, laying the foundation for further genetic and functional study of key pathways. This research illustrates the complementary nature of laboratory mapping, functional genetics, and population genomics.

derived stickleback populations have quickly evolved along numerous phenotypic axes, leading to significant variation in behavior, life history, and morphology [55–75]. Importantly, despite little or no gene flow between them, populations in similar freshwater habitats often evolve in parallel along the same phenotypic trajectories at a variety of local, regional and global scales [59,76–80].

Because of their extreme diversification some stickleback populations are actually incipient [81–83] or completely differentiated species [84–88]. Diversification has happened very rapidly, on the order of just a few thousand years [50,58,60,84], or in a few rare instances in just a few decades [82,89]. Thus, the biogeography of stickleback offers an excellent opportunity to examine the developmental genetic and genomic basis of rapid adaptation by comparing ancestral oceanic and derived freshwater populations. Importantly, these population genomic analyses are greatly advanced by a first draft of the stickleback genome, generated from a line derived from one of the populations used in this study (Bear Paw Lake; Ensembl: http://www.ensembl.org/Gasterosteus_aculeatus/Info/Index).

Stickleback can be crossed in the laboratory to produce viable offspring and genetic mapping crosses [79,90,91] which have been used to successfully identify nearly two dozen quantitative trait loci (QTL; [78,79,91–97]). A surprising result of this work is that, at least in some cases, parallel phenotypic evolution is due to different types of parallel genetic changes. The parallel evolution appears to occur mostly through the fixation of alleles of the same genes from the standing genetic variation in oceanic populations [78–80,93,95], but these alleles may be the product of single [93] or multiple [96] mutational events. Despite these advances in our understanding of evolutionary genetics in natural populations, a fundamental question remains: Are these instances of parallel evolution at individual loci representative of genome-wide patterns of parallel evolution in independently derived freshwater populations?

To address this question we have performed the first analysis of genome-wide patterns of polymorphism and differentiation using densely spaced single-nucleotide polymorphism (SNP) markers in replicate derived freshwater and ancestral oceanic stickleback populations. We used a novel and efficient genotyping approach based on Illumina sequencing of libraries of Restriction-site Associated DNA (RAD) tags [90,99]. Using short sequence reads, this technique provides genotype information on a large number of SNP markers, although it does not provide gametic phase across SNPs in different tags or haplotype sequence information. We use a kernel-smoothing analysis of these SNP genotype data aligned to the reference genome sequence to assess genome-scale patterns. Here we present a population genomic analysis based on several thousand SNPs across the genomes of 100 individuals from five populations. We focus on three freshwater populations which previous evidence suggests are quite young (less than 10,000 years old) and are independently derived from oceanic ancestral populations, with little or no gene flow directly among them [55,57,79]. Because of this history, we expect most of the adaptive evolution in the freshwater habitats to be the result of selection on standing genetic variation present in the founding populations. Accordingly, we focus primarily on measures of nucleotide diversity and differentiation in allele frequencies between the derived freshwater populations and two replicate oceanic populations, quantified with the statistic FST [7,21,22,32,100,101]. We further support our inferences with genomic distributions of private allele density and Tajima’s D [102]. We have identified numerous genomic regions that are likely under diversifying selection, and a smaller number of regions that appear subject to balancing selection. We find that many of these regions are shared across the independently derived populations, confirming past results on the genetic basis of morphological evolution from laboratory crosses, and also implicating many other previously unidentified genomic regions as adaptively significant.

Results

RAD tag genome coverage and sequencing depth

RAD tag sequencing provided a genome-wide distribution of over 45,000 single nucleotide polymorphisms (SNPs) that were simultaneously identified, scored, and used in a genome-wide scan of 100 individuals, 20 each from two oceanic and three freshwater stickleback populations (Figure 1 and Figure 2; Table 1). The published stickleback genome contains 22,830 identifiable SmI restriction sites across the 21 linkage groups and unassembled scaffolds (Ensembl, assembly Broad S1). Each site is expected to produce at most two RAD tags (sequence reads in each direction from the restriction site), and our sequencing effort recovered a large proportion of the expected RAD tags (Table S1). The sites were spread evenly throughout the genome (Figure 3A), and on average each tag was sequenced approximately five to ten times in every individual (Figure 3B). This depth of coverage allowed the identification of SNPs and statistical estimation of the diploid genotype for each individual at most nucleotide sites; sites at which coverage was insufficient were not assigned a genotype (see Methods). The overall frequency of SNPs (Table 1) agrees well with previous estimates of nucleotide polymorphism in stickleback populations.

Genome-wide estimates of genetic diversity and population differentiation

From these SNP genotype data we identified significant genetic variation within and across populations, with average genetic diversity (π) equal to 0.00336 across all populations and 0.0020–0.0027 within each population (Table 2). These findings are in rough agreement with previous studies of genetic variation within and among stickleback populations [55,57,59,60], although they are somewhat reduced. This may be a consequence of the
patterns of genetic diversity distributed across the genome

To assess genome-wide patterns we examined mean nucleotide diversity \( \pi \) and heterozygosity \( H \) using a Gaussian kernel smoothing function across each linkage group (Figure 4 and Figure S1). Although the overall mean diversity and heterozygosity values are 0.00336 and 0.00187, respectively, values vary widely across the genome. Nucleotide diversity within genomic regions ranges from 0.00003 to over 0.01, whereas heterozygosity values range from 0.00001 to 0.00003. This variation in diversity across the genome provides important clues to the evolutionary processes that are maintaining genetic diversity. For example, while expected \( \pi \) and observed \( H \) heterozygosity largely correspond, they differ at a few genomic regions (e.g., on Linkage Group XI). Genomic regions that exhibit significantly (\( p < 10^{-5} \)) low levels of diversity and heterozygosity (e.g., on LG II and V, Figure 4 and Figure S1) may be the result of low mutation rate, low recombination rate, purifying or positive selection that is consistent across populations, or some combination of factors [9,36,105–107].

In contrast, other genomic regions, such as those on LG III and XIII (Figure 4), show very high levels of both diversity and heterozygosity. The most striking such region, found near the end of LG III, corresponds precisely with a region of reduced differentiation among populations (Figure 5). This suggests the presence of balancing selection maintaining a common pool of genetic variation at this genomic region within and among populations. To further investigate the pattern of increased genetic variation on LG III, we delineated a region from 14.8 to 16.1 Mb (Figure 5; see Methods). Within the corresponding 1.3-Mb interval in the published stickleback genome are several candidate targets of balancing selection, namely genes implicated in the first line of defense against pathogens [108]; ZEB1 (ENSGACG00000017648), and two adjacent APOL genes (ENSGACG00000017778, ENSGACG00000017779). Supporting the importance of this region in immune response, there are also orthologs of several inflammation pathway genes: LTB4R (ENSGACG00000017812), SHARPIN (ENSGACG00000017834), and CEBPD (ENSGACG00000017927) [109–111]. The region of significantly elevated nucleotide diversity on LG XIII (18.1–19.1 Mb) also contains candidate targets of balancing selection including a TRIM14 (ENSGACG00000014283) and three TRIM35 genes (ENSGACG00000014250, ENSGACG00000014251, ENSGACG00000014403). Many members of this large gene family have been implicated in innate immune response (reviewed in [112]), and one gene, TRIM5alpha, bears the signature of balancing selection in primates [113]. The stickleback TRIM cluster on LG XIII provides a second example of balancing selection acting at a TRIM locus.

Evidence for balancing selection on Major Histocompatibility (MHC) loci is somewhat weaker. An MHC Class II gene (ENSGACG00000017967) falls nearly 380 kb outside the interval...
of maximum nucleotide diversity on LG III, although both $\pi$ and $H$ are moderately elevated at this region as well ($\pi = 0.0046$, $p < 0.02$; $H = 0.0030$, $pH > 0.01$). In addition, a 250 kb unassembled genomic contig (scaffold 131) contains a block of six MHC class II genes (ENSGACG00000000330, ENSGACG00000000336, ENSGACG00000000344, ENSGACG00000000346, ENSGACG00000000348, ENSGACG00000000350).

Figure 2. Schematic diagram of population genomic data analysis using RAD sequencing. (A) Following Illumina sequencing of barcoded fragments, sequence reads (thin lines) are aligned to a reference genome sequence (thick line). Depth of coverage varies across tags. Reads that do not align to the genome, or align in multiple locations, are discarded. (B) Sample of reads at a single RAD site. The recognition site for the enzyme Sbf1 is indicated along the reference genome sequence (top), and sequence reads typically proceed in both directions from this point, at which they overlap. At each nucleotide site, reads showing each of the four possible nucleotides can be tallied (solid blue box). (C) Nucleotide counts at each site for each individual are used in a maximum likelihood framework to assign the diploid genotype at the site. In this example, G/T heterozygote is the most likely genotype; the method provides the log-likelihood for this genotype, a maximum-likelihood estimate for the sequencing error rate $e$, and a likelihood ratio test statistic comparing G/T to the second-most-likely genotype, G/G homozygote. (D) Each individual now has a diploid genotype at each nucleotide site sequenced, and single nucleotide polymorphisms (SNPs, shown in red) can be identified across populations. Note, however, that haplotype phase is still unknown across RAD tags. (E) SNPs (red ovals) are distributed across the genome (thick line), and population genetic measures (e.g. $F_{ST}$) are calculated for each SNP. (F) A kernel smoothing average across multiple nucleotide positions is used to produce genome-wide distributions of population genetic measures.

doi:10.1371/journal.pgen.1000862.g002
Patterns of population differentiation distributed across the genome

Profiles of population differentiation across each linkage group are generally consistent with the genome-wide average $F_{ST}$ values described above. In agreement with the genome-wide results of little genetic structuring among the oceanic populations, we found no genomic regions that exhibit either significantly elevated or reduced ($p<10^{-5}$) differentiation between the two oceanic populations (Figure 6A). In contrast, comparisons between the ancestral oceanic and individual derived freshwater populations (Figure 6B–6D) exhibit several genomic regions of significant differentiation, with $F_{ST}>0.35$, as do the overall freshwater-ocmc genomic regions responding to directional selection across freshwater populations. Second, some, but not all, of these peaks also appear in the overall oceanic-freshwater comparison (Figure 6E). A striking example of this situation is seen on LG XXXI (Figure 8D), where a remarkable consistency in both the levels of $F_{ST}$ and the location of peak margins across the three freshwater populations is matched by a large peak in the overall oceanic-freshwater comparison. Nucleotide diversity and heterozygosity are reduced in the freshwater populations in this region as well (at 5.7 Mb, $H=0.0030$, $p=0.0003$; $H=0.0006$, $p=0.0003$).

We delineated the nine most consistent and significant of these peaks (see specific criteria in Methods). These regions occur on six linkage groups (I, IV, VII, VIII, XI, XXI) and are highlighted in Figure 7 and Figure 8. Also plotted in Figure 7 and Figure 8 are all linkage groups illustrating the overall oceanic-freshwater comparison (Figure 6E). A striking example of this situation is seen on LG XXXI (Figure 8D), where a remarkable consistency in both the levels of $F_{ST}$ and the location of peak margins across the three freshwater populations is matched by a large peak in the overall oceanic-freshwater comparison. Nucleotide diversity and heterozygosity are reduced in the freshwater populations in this region as well (at 5.7 Mb, $p<0.001$, $p=0.0003$; $p=0.0003$).

We delineated the nine most consistent and significant of these peaks (see specific criteria in Methods). These regions occur on six linkage groups (I, IV, VII, VIII, XI, XXI) and are highlighted in Figure 7 and Figure 8. Also plotted in Figure 7 and Figure 8 are all $F_{ST}$ values at individual SNPs where population differentiation in the overall oceanic-freshwater comparison is significant at the $\alpha=10^{-20}$ level (equivalent to $p<6.58 \times 10^{-23}$) following false discovery rate correction of individual $G$-tests (see Methods). These highly significant SNPs largely correspond with the genomic

<table>
<thead>
<tr>
<th>Linkage group</th>
<th>Length</th>
<th>Sites</th>
<th>RS</th>
<th>RB</th>
<th>BP</th>
<th>BL</th>
<th>ML</th>
<th>OC</th>
<th>FW</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>28,185,914</td>
<td>125,496</td>
<td>994</td>
<td>1,316</td>
<td>688</td>
<td>812</td>
<td>1,025</td>
<td>1,694</td>
<td>1,549</td>
<td>2,417</td>
</tr>
<tr>
<td>II</td>
<td>23,295,652</td>
<td>100,502</td>
<td>764</td>
<td>1,074</td>
<td>566</td>
<td>620</td>
<td>893</td>
<td>1,336</td>
<td>1,329</td>
<td>1,979</td>
</tr>
<tr>
<td>III</td>
<td>16,798,506</td>
<td>88,770</td>
<td>840</td>
<td>1,191</td>
<td>697</td>
<td>763</td>
<td>1,035</td>
<td>1,499</td>
<td>1,574</td>
<td>2,257</td>
</tr>
<tr>
<td>IV</td>
<td>32,632,948</td>
<td>138,898</td>
<td>999</td>
<td>1,408</td>
<td>749</td>
<td>865</td>
<td>1,278</td>
<td>1,774</td>
<td>1,842</td>
<td>2,871</td>
</tr>
<tr>
<td>V</td>
<td>12,251,397</td>
<td>59,631</td>
<td>497</td>
<td>656</td>
<td>347</td>
<td>394</td>
<td>561</td>
<td>851</td>
<td>813</td>
<td>1,243</td>
</tr>
<tr>
<td>VI</td>
<td>17,083,675</td>
<td>77,914</td>
<td>688</td>
<td>907</td>
<td>440</td>
<td>512</td>
<td>799</td>
<td>1,140</td>
<td>1,082</td>
<td>1,615</td>
</tr>
<tr>
<td>VII</td>
<td>27,937,443</td>
<td>115,092</td>
<td>838</td>
<td>1,092</td>
<td>677</td>
<td>739</td>
<td>984</td>
<td>1,429</td>
<td>1,489</td>
<td>2,312</td>
</tr>
<tr>
<td>VIII</td>
<td>19,368,704</td>
<td>87,664</td>
<td>700</td>
<td>933</td>
<td>456</td>
<td>589</td>
<td>774</td>
<td>1,188</td>
<td>1,141</td>
<td>1,736</td>
</tr>
<tr>
<td>IX</td>
<td>20,249,479</td>
<td>91,100</td>
<td>731</td>
<td>971</td>
<td>511</td>
<td>560</td>
<td>787</td>
<td>1,250</td>
<td>1,171</td>
<td>1,798</td>
</tr>
<tr>
<td>X</td>
<td>15,657,440</td>
<td>69,574</td>
<td>602</td>
<td>827</td>
<td>427</td>
<td>477</td>
<td>661</td>
<td>1,040</td>
<td>979</td>
<td>1,490</td>
</tr>
<tr>
<td>XI</td>
<td>16,706,052</td>
<td>82,787</td>
<td>699</td>
<td>948</td>
<td>495</td>
<td>586</td>
<td>763</td>
<td>1,215</td>
<td>1,172</td>
<td>1,801</td>
</tr>
<tr>
<td>XII</td>
<td>18,401,067</td>
<td>74,887</td>
<td>634</td>
<td>806</td>
<td>473</td>
<td>535</td>
<td>703</td>
<td>1,055</td>
<td>1,063</td>
<td>1,630</td>
</tr>
<tr>
<td>XIII</td>
<td>20,083,130</td>
<td>91,333</td>
<td>794</td>
<td>998</td>
<td>538</td>
<td>634</td>
<td>847</td>
<td>1,307</td>
<td>1,255</td>
<td>1,897</td>
</tr>
<tr>
<td>XIV</td>
<td>15,246,461</td>
<td>73,639</td>
<td>611</td>
<td>874</td>
<td>462</td>
<td>505</td>
<td>773</td>
<td>1,072</td>
<td>1,084</td>
<td>1,560</td>
</tr>
<tr>
<td>XV</td>
<td>16,198,764</td>
<td>75,415</td>
<td>618</td>
<td>837</td>
<td>414</td>
<td>476</td>
<td>645</td>
<td>1,041</td>
<td>938</td>
<td>1,438</td>
</tr>
<tr>
<td>XVI</td>
<td>18,115,788</td>
<td>74,669</td>
<td>653</td>
<td>795</td>
<td>392</td>
<td>464</td>
<td>642</td>
<td>1,039</td>
<td>981</td>
<td>1,519</td>
</tr>
<tr>
<td>XVII</td>
<td>14,603,141</td>
<td>65,431</td>
<td>606</td>
<td>772</td>
<td>401</td>
<td>427</td>
<td>598</td>
<td>1,004</td>
<td>882</td>
<td>1,370</td>
</tr>
<tr>
<td>XVIII</td>
<td>16,282,716</td>
<td>80,526</td>
<td>678</td>
<td>923</td>
<td>484</td>
<td>544</td>
<td>799</td>
<td>1,170</td>
<td>1,156</td>
<td>1,709</td>
</tr>
<tr>
<td>XIX</td>
<td>20,240,660</td>
<td>89,505</td>
<td>582</td>
<td>919</td>
<td>594</td>
<td>664</td>
<td>814</td>
<td>1,118</td>
<td>1,180</td>
<td>1,689</td>
</tr>
<tr>
<td>XX</td>
<td>19,732,071</td>
<td>78,669</td>
<td>558</td>
<td>777</td>
<td>463</td>
<td>472</td>
<td>659</td>
<td>988</td>
<td>996</td>
<td>1,538</td>
</tr>
<tr>
<td>XXI</td>
<td>11,717,487</td>
<td>51,484</td>
<td>428</td>
<td>552</td>
<td>339</td>
<td>359</td>
<td>526</td>
<td>730</td>
<td>751</td>
<td>1,169</td>
</tr>
<tr>
<td>Other</td>
<td>60,744,953</td>
<td>303,308</td>
<td>2,536</td>
<td>3,891</td>
<td>2,940</td>
<td>4,507</td>
<td>4,767</td>
<td>6,618</td>
<td>6,768</td>
<td>8,751</td>
</tr>
<tr>
<td>TOTAL</td>
<td>461,533,448</td>
<td>2,092,294</td>
<td>16,870</td>
<td>23,467</td>
<td>13,305</td>
<td>14,937</td>
<td>21,073</td>
<td>29,707</td>
<td>31,045</td>
<td>45,789</td>
</tr>
</tbody>
</table>

1 Linkage group of the stickleback genome (Ensembl), where “Other” includes all unassembled scaffolds.
2 Total length (bp) of each linkage group.
3 The total number of nucleotide sites for which sequence information was generated in at least one individual, after trimming restriction enzyme recognition sequence.
4 The remaining columns give the number of single-nucleotide polymorphisms identified within each population. Oceanic populations are RS (Rabbit Slough) and RB (Resurrection Bay); freshwater populations are BP (Bear Paw Lake), BL (Boot Lake), and ML (Mud Lake); OC is both oceanic populations (RS + RB); FW is all freshwater populations (BP + BL + ML); ALL is all 5 populations combined.

doI: 10.1371/journal.pgen.1000862.t001

Table 1. Nucleotide sites and SNPs identified on each linkage group.
regions of elevated differentiation, indicating that the averaged results from the kernel smoothing analysis are not anomalous. Of the 44,841 SNPs in this comparison at which $F_{ST}$ and a $G$-statistic could be calculated, 307 were significant at this level. Of these 307, 227 occur on these six linkage groups, and 119 of these are within the boundaries of the nine peaks, despite the fact that these nine regions collectively account for just $\sim$2.5 percent of the entire genome.

In contrast, some of the genomic regions that show consistent differentiation in all of the individual freshwater populations do not exhibit a peak in the overall oceanic-freshwater comparison. An example of this situation is observed on LG II (Figure 7B), where substantial peaks in each of the individual freshwater comparisons cover the same genomic region but differ slightly in their precise location. Accordingly, we do not observe significant differentiation in the overall comparison, and the freshwater populations are substantially differentiated from each other in this region; in fact, the largest peak in the among-freshwater $F_{ST}$ ($F_{ST} = 0.5147$, $p < 10^{-7}$; Figure 6F) occurs at this region. Both of these patterns are observed together on LG IV. Of the three LG IV peaks highlighted in Figure 7C, the third is most consistent in its height, width, and location across the freshwater populations. It corresponds to the most substantial peak of the three in the overall oceanic-freshwater comparison ($F_{ST} = 0.4262$, $p < 10^{-7}$) and shows virtually no differentiation among the freshwater populations. In contrast, the second peak and neighboring region to 22.5 Mb shows more variation among the freshwater populations and is substantially lower in the overall oceanic-freshwater comparison ($F_{ST} = 0.3269$, $p < 10^{-7}$).

Finally, there are peaks of differentiation observed in one or two, but not all three, freshwater populations. One example of this is seen at 11.5–12 Mb on LG VIII (Figure 8B), where the Mud Lake population exhibits a peak in differentiation ($F_{ST} = 0.4262$, $p < 10^{-7}$) and shows virtually no differentiation among the freshwater populations. In contrast, the peak at $\sim$8.3 Mb on the same linkage group, which is consistent across the three populations and also observed in the overall oceanic-freshwater comparison ($F_{ST} = 0.3844$, $p < 10^{-7}$), but not present in the comparison among freshwater populations.

The interpretation of these peaks of population differentiation as foci of selection is further supported by the genome-wide distributions of other statistics (Figure 9). First, we estimated

| Table 2. Pairwise nucleotide diversity and population differentiation among five stickleback populations. |

<table>
<thead>
<tr>
<th></th>
<th>RS</th>
<th>RB</th>
<th>BP</th>
<th>BL</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS</td>
<td>0.00216</td>
<td>0.00267</td>
<td>0.00277</td>
<td>0.00290</td>
<td>0.00308</td>
</tr>
<tr>
<td>RB</td>
<td>0.0076</td>
<td>0.00250</td>
<td>0.00291</td>
<td>0.00296</td>
<td>0.00308</td>
</tr>
<tr>
<td>BP</td>
<td>0.1391</td>
<td>0.0650</td>
<td>0.00203</td>
<td>0.00269</td>
<td>0.00295</td>
</tr>
<tr>
<td>BL</td>
<td>0.1040</td>
<td>0.0462</td>
<td>0.1310</td>
<td>0.00227</td>
<td>0.00299</td>
</tr>
<tr>
<td>ML</td>
<td>0.1252</td>
<td>0.0849</td>
<td>0.0798</td>
<td>0.0868</td>
<td>0.00268</td>
</tr>
</tbody>
</table>

1 Above the diagonal is average nucleotide diversity ($\pi$) in each combined pair of populations; along the diagonal is $\pi$ within each single population; below the diagonal is $F_{ST}$ between the two populations. Population abbreviations are as in Table 1.
Tajima’s $D$ [102] across the genome in the oceanic populations (Figure 9A). (Because of their young age and expected non-equilibrium allele frequency distributions, we did not consider this statistic to be informative in the freshwater populations). $D$ is negative overall in the oceanic populations, perhaps as a result of demographic processes affecting the entire genome equally. However, regions of significantly negative $D$ correspond with peaks of freshwater-oceanic differentiation. In addition, we examined the genomic distribution of the density of private alleles—alleles that are found in only a single population or group of populations in a comparison. Overall, the private allele density ($r$) is higher in oceanic populations compared to freshwater than vice versa (Figure S2). This is consistent with the view that the genetic variation in the freshwater populations is largely a sample from the oceanic stock. However, peaks in private allele density in freshwater populations relative to the ocean (Figure 9B–9D) correspond well with $F_{ST}$ peaks in the freshwater-oceanic comparisons (with the exception of the peaks on LG I and XI). Thus the peaks in $F_{ST}$ are largely the result of alleles that we did not detect in the oceanic populations. The hypothesis that these are new mutations in the freshwater populations is rejected by the absence of corresponding peaks in private allele density among the freshwater populations (Figure 9E–9G). Instead, while selection in freshwater has acted on haplotypes that were rare (and not detected in our samples) in the oceanic stock, these haplotypes are nonetheless shared among the independently derived freshwater populations. Previous work has shown that freshwater-adapted alleles may persist at a very low frequency in the ocean, low enough that we would not expect to detect many of them in our sample of 40 individuals [74]. However, the maintenance of such low-frequency alleles in the ocean by gene flow from freshwater populations, combined with selection against them in the oceanic habitats, could also account for the significantly negative Tajima’s $D$ in the ocean at these genomic regions.

Figure 4. Genome-wide patterns of nucleotide diversity. Each plot shows a smoothed distribution of the statistical measure across the genome (black lines). Colored bars above and below the distributions indicate regions of significantly elevated ($p<10^{-5}$, blue; $p<10^{-7}$, red) and reduced ($p<10^{-5}$, green) values, assessed by bootstrap resampling. Vertical shading indicates the 21 linkage groups and the unassembled scaffolds greater than 1 Mb in length, and gold shading indicates two regions showing evidence of balancing selection as discussed in the text. (A) Nucleotide diversity ($\pi$) across all five stickleback populations sampled. (B) Heterozygosity ($H$) across all five populations.

doi:10.1371/journal.pgen.1000862.g004

Figure 5. Evidence for balancing selection on Linkage Group III. Population genetic measures plotted along Linkage Group III. (A) Nucleotide diversity ($\pi$) and (B) heterozygosity ($H$) across all five (blue), the three freshwater (red), and the two oceanic (green) populations. (C) Population differentiation ($F_{ST}$) between oceanic and freshwater (blue), among freshwater (red), and between oceanic (green) populations. Colored bars indicate significant ($p<10^{-5}$) regions of elevated (above the plots) or reduced (below the plots) values of each statistic for the corresponding set of populations. Vertical yellow shading indicates the region of putative balancing selection used for candidate gene annotation.

doi:10.1371/journal.pgen.1000862.g005
Figure 6. Genome-wide differentiation among populations. $F_{ST}$ across the genome, with colored bars indicating significantly elevated ($p \leq 10^{-5}$, blue; $p \leq 10^{-7}$, red) and reduced ($p \leq 10^{-5}$, green) values. Vertical gray shading indicates boundaries of the linkage groups and unassembled scaffolds, and gold shading indicates the nine peaks of substantial population differentiation discussed in the text. (A) $F_{ST}$ between the two oceanic populations (RS and RB; note that no regions of $F_{ST}$ are significantly elevated or reduced). (B,C,D) Differentiation of each single freshwater population from the two oceanic populations, shown as the mean of the two pairwise comparisons (with RS and RB): (B) BP, (C) BL, (D) ML. Colored bars in each plot represent regions where both pairwise comparisons exceeded the corresponding significance threshold. (E) Overall population differentiation between the oceanic and freshwater populations. (F) Differentiation among the three freshwater populations (BP, BL, ML).

doi:10.1371/journal.pgen.1000862.g006
Figure 7. Differentiation among oceanic and freshwater populations on Linkage Groups I, II, and IV. For each linkage group, the upper panel shows population differentiation ($F_{ST}$) of each freshwater population from the two oceanic populations, plotted as the mean of the two freshwater versus oceanic comparisons for each freshwater population: BP (blue), BL (red), ML (green). Colored bars indicate regions of bootstrap significance ($p \leq 10^{-5}$) for the corresponding population. The lower panel shows $F_{ST}$ for the overall oceanic-freshwater comparison (black), $F_{ST}$ among the three freshwater populations (orange), and corresponding regions of significance ($p \leq 10^{-5}$), along with $F_{ST}$ values (blue circles) at single nucleotide polymorphisms at which population differentiation is significant at the level of $\alpha = 10^{-20}$ in a G-test corrected for false discovery rate. Vertical shading indicates boundaries of the peaks used for candidate gene annotation. (A) LG I. (B) LG II. (C) LG IV.

doi:10.1371/journal.pgen.1000862.g007
Exceptions to the pattern described above are found at the $F_{ST}$ peaks on LG I and XI. Here, the private allele density in freshwater does not differ significantly from the genome-wide average (Figure 9B–9D), but private allele density in the ocean relative to freshwater is significantly higher (Figure S2B). In addition, $\pi$ is elevated in oceanic populations at the LG I region (Figure S1A, S1B, S1C). These data suggest the hypothesis that the oceanic environment may be permissive for multiple haplotypes at these genomic regions, of which only a subset have relatively high fitness in freshwater. In contrast, in the region centered at 13.3 Mb...

Figure 8. Differentiation among oceanic and freshwater populations on Linkage Groups VII, VIII, XI, and XXI. All panels show population differentiation as in Figure 7. (A) LG VII. (B) LG VIII. (C) LG XI. (D) LG XXI.

doi:10.1371/journal.pgen.1000862.g008
Figure 9. Genome-wide distributions of allele frequency spectrum and private allele density. (A) Tajima’s $D$, a measure of allele frequency spectrum, within the combined oceanic population (RS and RB). Colored bars above and below the distribution indicate regions of significantly elevated ($p < 10^{-2}$, green) or reduced ($p < 10^{-2}$, blue; $p < 10^{-4}$, red) values, assessed by bootstrap resampling. (B–G) Private allele density ($\rho$) in single freshwater populations. Colored bars indicate regions of significantly elevated ($p < 10^{-3}$, blue; $p < 10^{-5}$, red) or reduced ($p < 10^{-3}$) values. (B) Private allele density in BP relative to combined oceanic populations (OC). (C) BL relative to OC. (D) ML relative to OC. (E) Private allele density in BP relative to other freshwater populations (FW). (F) BL relative to FW. (G) ML relative to FW. Across all panels, vertical gray shading indicates Linkage Groups I-XXI and unassembled scaffolds, and gold shading indicates the nine peaks of population differentiation highlighted in Figure 7 and Figure 8. doi:10.1371/journal.pgen.1000862.g009
on LG II, the freshwater populations exhibit high densities of private alleles, both with respect to the oceanic populations and with respect to each other (Figure 9B-9G). These correspond with peaks in $F_{ST}$ both between oceanic and freshwater populations and among freshwater populations (Figure 7B). Here different haplotypes have evolved to high frequency among the different freshwater populations.

**Identification of genes of adaptive significance**

To set our results in the context of previous QTL mapping studies, and to explore a set of putative candidate genes associated with adaptation to freshwater, we focused on the nine peaks highlighted in Figure 6. Our results are complementary to previous QTL mapping of traits relevant to freshwater adaptation, although direct comparison with QTL results is complicated because many of those previous studies used microsatellite markers placed on a genetic linkage map. The order of those markers on the genetic map does not always correspond with the marker order on the physical map of the stickleback genome (Ensembl, database version 56.1j, assembly Broad S1), leading in some cases to quite large physical distances between QTL-associated markers. Also, some of the previously used microsatellite markers do not appear at all in the genome sequence. Nonetheless, of the nine peaks we identified, the three on LG IV co-occur with previously identified QTL and specific genes [78,79,93,97,99]. This includes the gene Ecotoxylaspin A (Eda), implicated in the loss of the lateral plate phenotype [99], which occurs within the first peak of population differentiation that we identified on LG IV. An additional three peaks show the possibility of an association with previous QTL: Shapiro et al. [95] identified very broad QTL that overlap large portions of LG IV and VII, including all five peaks we identified on those linkage groups, and Albert and colleagues [97] identified a QTL adjacent to our peak on LG XXI. In addition, evidence for directional selection based on microsatellite markers has been found just adjacent to two of our delineated peaks. One of these occurs at ~22.3 Mb on LG I [103] (but see reanalysis by [28]). The other lies at ~9.5–9.8 Mb on LG VIII [104], just outside the strict delineation of the peak in Figure 8C, but within the broader region in which we detected substantially elevated $F_{ST}$ values and highly significant SNPs. Other regions outside the nine most significant peaks also exhibit a correspondence with QTL studies. For example, the peak on LG XII (Figure 6E) contains many osteogenesis genes and overlaps a QTL peak for many skeletal characters [97]. In contrast, the region at the distal end of LG VII previously associated with the pelvic structure phenotype, specifically containing the *Pitx1* gene [79,95,99], did not correspond to elevated levels of divergence in any of our comparisons.

To evaluate potential candidate genes, we identified all loci overlapping the boundaries of the nine most consistent peaks (Table S2 provides the complete list). Many genes in these defined intervals are already annotated by name and orthology in the *Gasterosteus* genome database (Ensembl, database version 56.1j, assembly Broad S1); the orthology relationships of the remaining genes, those for which no gene name is yet listed, were further analyzed by a BLAST comparison of the predicted protein sequence for each of them against the NCBI protein database. We then assessed the ontological relationships of all protein coding genes in each interval with respect to skeletal biology and to osmoregulation, two axes of the phenotype known to change drastically as stickleback evolve in response to freshwater environments with very different ecological and chemical conditions than the ocean. Table 3 identifies genes for which a strong association with either of these two broad ontological classes is supported in the literature. From the nine annotated peaks, covering a total of 12.2 Mb, we list 31 candidate genes: 23 candidates for patterning and homeostasis of skeletal traits, 8 candidates for response to osmotic stress and development of osmoregulatory organs, and three candidates with pleiotropic roles in both skeletogenesis and osmoregulation. The total numbers of all protein-coding genes within each peak are also listed in Table 3.

The abundance of annotated genes within the nine consistent peaks of differentiation does not appear to be an artifact of the distribution of genes across the genome (Figure S3). Rather, gene density shows no apparent correlation with the regions of population differentiation that we identified here.

Although we focused on the nine significant peaks of differentiation that appear most consistent across freshwater populations, several other regions show strong evidence of selection in derived freshwater populations and contain candidate genes worthy of further study. In particular, large regions of LG IV and LG VII outside the delineated peaks appear to be important in differentiation of freshwater stickleback, and these two linkage groups have been the focus of much previous attention. Intriguingly, duplicate synteny groups containing six genes (CLINT1, EBF1, IL12B, ADRB2, ABLIM3 and AFAP1L1) lie just adjacent to Peak 1 of LG IV and partially overlapping Peak 2 of LG VII. Of these, EBF1, IL12B and ADRB2 are skeletal trait candidates [114–116]. As mentioned above, a region of LG XII previously implicated by QTL analysis also shows a signature of selection here. We provide a list of candidate genes in these additional genomic regions in Table S3.

**Discussion**

RAD sequencing is a useful tool for population genomic analysis

Population genomic studies depend on having a very high density of markers that can be scored across many individuals. Depending upon demographic factors such as population size and structure, and the strength and nature of selection [117,118], blocks of linkage disequilibrium (LD) can be as small as a few hundred base pairs (as in flies [105]) to several dozens of kilobases (kb) (as in dogs [119]). For most natural populations, the likely size is on the order of 1 to 100 kb, meaning that tens or hundreds of thousands of markers are required to adequately cover an average-sized genome. Furthermore, population genetic sampling variances occur for single point estimates at each marker, requiring numerous individuals to be analyzed from each group or subpopulation of a study. Illumina-sequenced RAD tags provide a powerful new tool to meet these needs, generating a dense battery of SNP markers that are likely to cover a large proportion of the LD blocks produced by stickleback adaptation, and which can be simultaneously identified and scored across entire genomes. The density of markers that can be scored across individuals using RAD-seq holds promise for association mapping of phenotypic traits in natural populations of other organisms.

Although we used the stickleback reference genome sequence for the alignment of RAD tags, this tool can be used for population genomic studies in organisms that do not yet have a sequenced genome. Instead of aligning against a genome, the sequence reads can instead be aligned to one another, with SNPs identified and zygositities scored for individuals in the same manner as we describe here (Hohenlohe and Cresko unpublished data). Although these identified RAD sites are initially unanchored with respect to one another, if scored in an $F_{ST}$ or backcross mapping family, they could be ordered to produce a high-density linkage map. This genetic map could then be used to perform genome scans, as well
Table 3. Candidate genes related to morphology and osmoregulation, identified within the nine major peaks of parallel differentiation.1

<table>
<thead>
<tr>
<th>Location</th>
<th>Gene</th>
<th>p-value</th>
<th>OD</th>
<th>BD</th>
<th>TO</th>
<th>CF</th>
<th>OS</th>
<th>KF</th>
<th>IG</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>LG 1: 1 Mb, 52 genes</td>
<td>TNS1</td>
<td>&lt;10^-7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[160]</td>
</tr>
<tr>
<td></td>
<td>IGFBP5</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[160,161]</td>
</tr>
<tr>
<td></td>
<td>IGFBP2</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[160,162]</td>
</tr>
<tr>
<td>LG IV Peak 1: 1 Mb, 43 genes</td>
<td>EDA</td>
<td>&lt;10^-7</td>
<td>Yes/T</td>
<td>Yes/T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[78,93,139,163]</td>
</tr>
<tr>
<td></td>
<td>FLT4</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[164]</td>
</tr>
<tr>
<td></td>
<td>PDLM7</td>
<td>2.6x10^-5</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[165]</td>
</tr>
<tr>
<td></td>
<td>ANX6</td>
<td>0.0043</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[166]</td>
</tr>
<tr>
<td>LG IV Peak 2: 1.1 Mb, 31 genes</td>
<td>WNT7B</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td></td>
<td></td>
<td>Yes/T</td>
<td></td>
<td></td>
<td></td>
<td>[163,167]</td>
</tr>
<tr>
<td></td>
<td>FBLN1</td>
<td>&lt;10^-7</td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td>[142]</td>
</tr>
<tr>
<td></td>
<td>LEMD3</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[144]</td>
</tr>
<tr>
<td></td>
<td>PRL</td>
<td>0.0073</td>
<td>Yes/T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[148]</td>
</tr>
<tr>
<td></td>
<td>SCUBE1</td>
<td>1.2x10^-6</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[138]</td>
</tr>
<tr>
<td></td>
<td>NFBY</td>
<td>0.0005</td>
<td>Yes/T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[137]</td>
</tr>
<tr>
<td></td>
<td>PDDXL</td>
<td>0.0006</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[168]</td>
</tr>
<tr>
<td></td>
<td>SLC26A3</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[169,170]</td>
</tr>
<tr>
<td></td>
<td>OSBPL8</td>
<td>10^-5</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[171]</td>
</tr>
<tr>
<td>LG VII Peak 1: 0.8 Mb, 42 genes</td>
<td>CAMKK1</td>
<td>2.1x10^-6</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[172]</td>
</tr>
<tr>
<td></td>
<td>CA4</td>
<td>9x10^-5</td>
<td>Yes</td>
<td></td>
<td></td>
<td>Yes/T</td>
<td></td>
<td></td>
<td></td>
<td>[146,152]</td>
</tr>
<tr>
<td></td>
<td>HRH2</td>
<td>1.6x10^-6</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[173]</td>
</tr>
<tr>
<td></td>
<td>AR</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[174]</td>
</tr>
<tr>
<td></td>
<td>ADRB2</td>
<td>0.044</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[115]</td>
</tr>
<tr>
<td></td>
<td>IL12B</td>
<td>0.044</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[114]</td>
</tr>
<tr>
<td>LG VIII: 1.1 Mb, 50 genes</td>
<td>LEPR</td>
<td>0.0012</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[175]</td>
</tr>
<tr>
<td></td>
<td>ADAMTS10</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[176]</td>
</tr>
<tr>
<td>LG XI: 1 Mb, 55 genes</td>
<td>FZD2</td>
<td>&lt;10^-7</td>
<td>Yes/T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[177]</td>
</tr>
<tr>
<td></td>
<td>STAT3</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[178]</td>
</tr>
<tr>
<td>LG XXI: 2.6 Mb, 119 genes</td>
<td>BM1</td>
<td>1.1x10^-6</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[179]</td>
</tr>
<tr>
<td></td>
<td>RDH10</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[141]</td>
</tr>
<tr>
<td></td>
<td>EYA1</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td>Yes/T</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[140,180,181]</td>
</tr>
<tr>
<td></td>
<td>SGK3</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[182]</td>
</tr>
<tr>
<td></td>
<td>CRH</td>
<td>&lt;10^-7</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[183]</td>
</tr>
<tr>
<td></td>
<td>FLT1</td>
<td>2x10^-7</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[149–151,184,185]</td>
</tr>
<tr>
<td></td>
<td>LNX2</td>
<td>0.0017</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[186]</td>
</tr>
<tr>
<td></td>
<td>ATP6V1A</td>
<td>0.076</td>
<td>Yes/T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[147]</td>
</tr>
</tbody>
</table>

1 Shown are possible skeletal and osmoregulatory targets of selection and their positions within nine peaks highlighted in Figure 7 and Figure 8. Also listed for each interval is the total number of protein coding genes annotated in the Gasterosteus aculeatus genome (Ensembl, version 56.1j). P-values represent bootstrap significance of FST in the overall oceanic-freshwater comparison in the region centered on the nearest 100 kb to the midpoint of each gene (see Methods). Genes are connected to one or more ontology categories of morphology (OD, osteoblast differentiation; BD, bone density and mineralization; TO, tooth organogenesis; CF, craniofacial development) or osmoregulation (OS, response to osmotic stress; KF, kidney function or development; IG, ion transport across gills or gut epithelia). Supporting information from teleost fish is indicated by “Yes/T”, while “Yes” denotes information from other vertebrates. For the complete list of protein-coding genes in each peak, see Table S2.

doi:10.1371/journal.pgen.1000862.t003
as to help order a physical map from subsequent genome sequencing projects. Such data may be useful even when a preliminary genome assembly already exists. For instance, our approach revealed that nearly 60 Mb - equivalent to two of the largest chromosomes - of the stickleback genome are segregating alleles and show significant signatures of selection, but have not been incorporated into the existing assembly of 21 linkage groups (Ensembl, Broad S1 assembly). A forthcoming RAD genetic map will help incorporate this nearly 10% of the genome into its proper locations. In sum, RAD sequencing has the potential to combine population genetic and genomic studies with genetic and association mapping in populations of both model and non-model organisms, and in addition can help quickly produce or enhance essential genomic resources for organisms that presently have few.

**Parallel genetic evolution in stickleback**

We produced genome-wide estimates of population diversity and differentiation for five stickleback populations that have been the focus of intense previous research. These data are largely in agreement with previous estimates of genetic diversity for stickleback, and support the view that oceanic stickleback populations have differentiated little from each other due to extensive gene flow over long distances. Each freshwater population exhibits a greater amount of divergence from the oceanic populations and from the other freshwater populations, but the overall amount is generally moderate and in line with previous estimates of population genetic divergence derived from microsatellite markers [55]. Taken together our data support the biogeographic hypothesis that large populations of oceanic stickleback have given rise repeatedly to freshwater populations, which have become phenotypically differentiated on a background of minor neutral population divergence [55,79].

Furthermore, we were able to determine the distribution across the genome of genetic diversity and differentiation among the replicate populations. Identifying genomic regions of significantly increased or decreased diversity and differentiation allows us to make inferences about evolutionary processes, and to generate hypotheses about the evolutionary role of specific loci. Overall, the genome-wide patterns showed remarkable consistency across replicate populations and across pairwise comparisons. For example, the region with the most substantially elevated nucleotide diversity, observed on LG III, was consistent across populations and also exhibited increased heterozygosity and greatly reduced diversity, observed on LG III, was consistent across populations and also exhibited increased heterozygosity and greatly reduced differentiation among populations. This pattern indicates balancing selection. This situation is best known for the vertebrate Major Histocompatibility (MHC) loci, which encode proteins responsible for tagging and presenting antigens to the immune system [120]. Greater levels of heterozygosity increase the range of antigens that can be identified by the immune system. Other genes that mediate a host's ability to repel or mitigate infection by parasites and other pathogens may also be the object of balancing selection [108]. Such loci can show strong signatures of balancing selection such as the persistence of old and highly polymorphic alleles (e.g., [121]). The region on stickleback LG III contains several candidates that fit this description. In mammals, ZEB1 helps maintain viral latency by binding the promoter of a virally encoded latency-to-lysogeny switch gene [122]. The direct interaction of ZEB1 with the viral genome makes it an attractive candidate target for host-pathogen co-evolution and balancing selection. The LG III peak contains a stickleback ZEB1 and two members of the APOL gene family, which encode proteins that may also directly interact with pathogens. APOL1 is a secreted protein that causes the lysis and death of trypanosome parasites in the blood, and variation at this locus affects resistance to trypanosome infection in humans [123]. Among primates, APOL genes show signs of rapid evolution and selective sweeps, possibly linked to their role in immunity [124]. Interestingly, the signature of balancing selection in the region of these host-pathogen-related loci was stronger than that in two regions with MHC orthologs: one MHC class IIB ortholog adjacent to the peak identified on LG III, and a cluster of six MHC class II loci on scaffold 131. Members of this latter group were found in a previous microsatellite analysis to show evidence of balancing selection in stickleback [125].

Similarly, the interval of increased nucleotide diversity on LG XIII overlaps a region rich in TRIM family genes, and includes a TRIM14 and three TRIM35 genes. Antiviral gene TRIM5alpha provides a rare example of balancing selection in primates [113]. It is possible that the increase in polymorphism on stickleback LG XIII has likewise been driven by selection on innate immunity genes, as has been suggested for clusters of other TRIM genes in teleost fish [126]. The patterns of nucleotide diversity and $F_{ST}$ across this LG XIII interval in stickleback provides a second example of balancing selection acting at a TRIM cluster locus and bolsters the hypothesis that the largely unstudied mammalian TRIM14 and TRIM35 genes may be involved in immune response [127]. The inference of balancing selection on these identified regions is clearly not conclusive, but can be used as the starting point for more focused, sequence-based or functional analyses.

We can draw further evolutionary inferences by focusing on the patterns of differentiation among replicate oceanic and freshwater stickleback populations, taking advantage of the rapid and often parallel phenotypic evolution coupled with little background population genetic structuring. In comparisons between freshwater and oceanic populations, we found numerous regions of the stickleback genome that exhibit significantly greater differentiation than observed in the rest of the genome, providing clear signatures of divergent selection distributed across numerous linkage groups. Although there were several instances in which a private signature could be observed in just one population, the strikingly common pattern is one of very similar regions being selected in all three independently derived populations. We can thus answer the question posed in the Introduction: the previously identified parallel genetic basis for the loss of armor traits in stickleback appears to be a general rule across the genome, in that much of the adaptation of stickleback populations to freshwater conditions likely involves the repeated use of the same repertoire of developmental and physiological systems, genes, and perhaps even alleles. However, the details of this parallel evolution – for example, whether it involves independent fixation of alleles that are identical by descent in multiple derived populations, or fixation of different alleles at the same locus – appear to differ in different parts of the genome. Population genomic scans of replicate derived populations in combination with laboratory mapping and sequence-based studies provide a powerful repertoire of tools for distinguishing among these hypotheses.

**Distinguishing among modes of adaptive evolution**

Other researchers [32,34,35,128,129] have distinguished between two types of selective sweeps. A hard sweep occurs when one or a small number of haplotypes present in standing genetic variation (in this case, in the ancestral oceanic pool) is selected to high frequency (in this case, in the newly established freshwater populations). Following such a hard sweep, a large proportion of the haplotypes at a given genomic region will be identical by descent. This is contrasted with a soft sweep, in which multiple alleles at a locus or genomic region are selected to high frequency.
Hard sweeps are expected to produce regions with reduced nucleotide diversity within populations, more significant differentiation between populations, and more extensive linkage disequilibrium (LD) [14,16,36,117,130,131]. Soft sweeps are expected to be more easily detected by changes in patterns of LD than by alterations of diversity or differentiation [24,32,34,35].

In the case of replicate freshwater stickleback populations, we can identify instances of parallel hard sweeps, in which the same or a few haplotypes present in the ancestral oceanic population were selected to high frequency independently in multiple freshwater populations. Alternatively, non-parallel sweeps are observed when different alleles from the oceanic standing variation are swept to high frequency in different derived freshwater populations, producing a hard sweep pattern within each freshwater population. The distinctions between these cases are apparent in the overall oceanic-freshwater comparison and in the comparison among freshwater populations. In fact, the ability to differentiate between parallel and non-parallel hard sweeps is a particular strength of natural systems with multiple replicate populations like stickleback. For example, the examination of parallel hard sweeps in several populations may help identify causative mutations if each sweep is only partially overlapping, narrowing the search to the region common in all populations.

The strongest example of a parallel hard sweep was observed here on LG XXI. Each of the three freshwater populations was strongly diverged from the oceanic ancestors, the overall oceanic-freshwater differentiation was similarly elevated, and there was no substantial differentiation among the freshwater populations (Figure 8D). In addition, nucleotide diversity within each population was substantially reduced in this region (Figure S1). Matching the FST results, private allele density was significantly elevated in freshwater relative to oceanic populations (Figure 9B-9D), but not in reciprocal comparisons among freshwater populations (Figure 9E-9G). These data suggest that the same haplotype, likely present at low frequency in the standing genetic variation in the ancestral oceanic stock, was selected to high frequency independently in all three freshwater populations. Despite their likely independent derivation from ancestral oceanic stocks, these three freshwater populations have evolved in a remarkably consistent manner at this genomic region. Alternative alleles at this region are favored in oceanic populations, leaving a signature of selection against the low-frequency freshwater alleles that are maintained by gene flow from freshwater back to the ocean.

In contrast, the region of LG II centered at 13.3 Mb provides an example of a non-parallel sweep, in which all three freshwater populations underwent substantial differentiation from the ancestor at the same region, but without exhibiting such consistency in the overall oceanic-freshwater comparison. Such a situation leads to several alternative hypotheses: the same allele at a particular locus was selected to high frequency in each population, but LD with surrounding variation was reduced in the oceanic pool. Alternatively, the same gene was under selection but different alleles were fixed in each freshwater population. Lastly, different genes in a genomic cluster may have responded to selection in each population. In this case, further data support the latter two hypotheses: private allele density is elevated in the freshwater populations, with respect to both the oceanic populations and the other freshwater populations. Additional peaks of population differentiation and private allele density in the broader genomic region, somewhat coincident across freshwater populations, also suggest that multiple loci in this section of LG II may have responded to selection in freshwater.

The examples highlighted above are the most striking of the general patterns observed, and many genomic regions are intermediate in their structure of population differentiation. In fact there is roughly continuous variation in the degree to which selective sweeps show a parallel genetic basis across replicate freshwater populations. Nonetheless, the large majority of genomic regions exhibiting substantial differentiation are shared across the freshwater populations. While the particular nature of allelic variation responding to selection appears to differ among these genomic regions, the adaptive significance of the regions themselves remains consistent. In this respect, genomic patterns of evolution are remarkably parallel among these populations.

Comparison of our results with previous microsatellite-based genome scans

Previous studies [103,104,133] used a set of microsatellite markers across the genome to identify selective sweeps in replicate stickleback populations in Finland, identifying a region of significant differentiation between oceanic and freshwater populations on LG VIII. That analysis focused on the region from ~9.3 Mb to 9.9 Mb on LG VIII [103,104], just adjacent to the peak delineated in Figure 8B. In fact, in this region of LG VIII we observed signatures of both a parallel hard sweep (from ~8.0 to ~9.0 Mb), in which differentiation among freshwater populations is reduced but the overall oceanic-freshwater comparison is elevated, and a non-parallel sweep (from ~9.3 to 10.0 Mb), in which differentiation among freshwater populations is elevated. Taken together, these results suggest the intriguing hypothesis that this region includes two adjacent genomic regions of importance for freshwater adaptation, at least one of which has undergone rapid evolution in both Alaskan and Fennoscandian populations, and which demonstrate two different modes of adaptive evolution in Alaskan populations.

Linking population genomics and QTL mapping

Comparisons between QTL mapping and population genomic studies can help discern the pattern of adaptation [see 42,43,45].
for a fine example of this approach). Laboratory mapping of
phenotypic variation in stickleback has been quite successful,
leading to the identification of numerous QTL for a variety of
different morphological and behavioral traits [50]. An open
question is whether these QTL-containing regions also exhibit
patterns of selective sweeps in natural populations. Our data
clearly show this to be the case for some QTL, but also provide
novel insights into the precise evolutionary trajectories. For
example, major loci for the loss of the bony lateral plates and
pelvic structures have been mapped previously to LG IV and LG
VII respectively, including in two of the three freshwater
populations used in this study [79,99].

On LG IV, the three regions of differentiation between oceanic
and freshwater populations that we observed (Figure 7C) were
previously associated with the lateral plate phenotype in QTL
studies of laboratory crosses. The first peak contains the gene
Ectodysplasin A (Eda, found at ~12.8 Mb), which has specifically
been implicated in the parallel loss of bony lateral plates in
freshwater populations [78]. Furthermore, previous mapping
studies using RAD genotyping in our laboratory have shown that
two additional regions of LG IV, corresponding to the second and
third peaks recovered here, also co-segregate with the lateral plate
phenotype [99]. Thus all three of these regions previously
identified in laboratory mapping studies show evidence of a hard
selective sweep within each of the freshwater populations and
varying degrees of parallel evolution across the populations. The
presence of three regions spread across nearly 20 Mb of a
chromosome associated with a single phenotype was difficult to
explain in the previous mapping cross. However, if loci in all three
regions interact epistatically then the entire region may be subject
to selection. If true, then although alleles along LG IV may be
recombined in the oceanic environment, selection acting in
isolated populations to favor haplotypes that contain the high
fitness multilocus genotype could manifest as a hard sweep across
the freshwater populations.

In contrast to the lateral plate QTL on LG IV, the major pelvic
structure reduction QTL exhibits a very different pattern with
respect to signatures of selection. The major locus for pelvic loss
was mapped to the very distal end of LG VII in two of these three
populations [79,95,134]. Additional work on other populations
pointed to Paxl as a likely candidate responsible for loss of the
pelvic structure [95]. Although we found significant signatures of
selection on LG VII (Figure 8A), none of them corresponds to the
region of the pelvic structure QTL mapped in laboratory crosses.
In fact, the distal 7.5 Mb of LG VII exhibits levels of
differentiation in all populations that is indistinguishable from
background levels. Furthermore, one of these populations, Mud
Lake, retains a full pelvic structure, whereas fish from both Bear
Paw and Boot Lakes exhibit pelvic reduction. Despite these
phenotypic differences, the three populations show very similar
levels of differentiation from each other and the oceanic
populations. This may be because selection has not occurred on
the locus despite the loss of pelvic structure in two of the three
populations. Alternatively, multiple different pelvic-loss alleles that
are not identical by descent may have been selected in each of the
pelvic reduced populations, leading to a soft sweep pattern. This
hypothesis is supported by results from previous laboratory
complementation results [79]. Although crosses between the
derived populations did not show evidence for complete
complementation, there was a statistically significant increase in
the size of the pelvic structure. We interpreted this quantitative
complementation result as likely due to different alleles at the same
major pelvic locus having the ability to partially complement one
another [79]. These new population genomic data fit this scenario.

In addition to these two major armor QTL, others have been
identified in stickleback crosses for a variety of traits. Previous
QTL mapping analyses, using crosses between oceanic and
freshwater stickleback populations or among freshwater ecotypes,
uncovered genomic regions co-segregating with various morpho-
logical traits, including the aforementioned presence or absence of
lateral plate or pelvic armor elements and aspects of head and
body geometry [91,135]. A few of these QTL overlap peaks
uncovered in our SNP marker genome scan. For example, Albert
and colleagues [97] found that changes in jaw and head
morphology are associated with regions on LG IV and XII; in
our analysis, peaks overlapping these regions contain orthologs of
SCUBE1, NFYB, and WNT5A, all known or suspected to impact
craniofacial development (Table 3, Table S3) [136–138]. Com-
plementary to the fruits of QTL mapping, our study highlights
new genomic regions that had not yet been recognized as
important in the evolution of freshwater phenotypes from oceanic,
namely significant peaks on Linkage Groups I, VII, VIII, XI, and
XXI.

These examples demonstrate the ways in which QTL mapping
and population genomic studies complement each other. While
QTL studies can implicate genomic regions and specific genes in
the evolution of particular phenotypes, population genomic results
such as those presented here can provide evidence for the adaptive
significance of these genomic regions in natural populations. A
population genomics approach covering multiple replicate popu-
lations provides further insight into the standing genetic variation,
types of selective sweeps, and extent of parallel evolution across
natural populations for genes previously linked to particular
phenotypes. A population genomics approach may also narrow a
region of interest previously identified in mapping studies,
especially when blocks of linkage disequilibrium in natural
populations are smaller than in laboratory crosses. Even situations
in which a population genomic approach does not implicate a
genomic region previously identified as a QTL, as here on LG
VII, are informative. The type of soft sweep postulated for the
pelvic structure locus may lead to a bias against detecting selection
on some previously identified loci with a genome scan. In addition,
the converse situation is also informative: population genomic
studies can identify putative regions of adaptive significance and
candidate genes that no previous mapping approach has
identified.

Candidate loci for adaptation to freshwater

We identified a list of candidate genes within peaks of parallel
divergence among stickleback populations that may be important
for adaptation to freshwater. Most work on adaptation to
freshwater in stickleback has focused on genes and pathways
associated with bone development and skeletal morphology.
Changes in teeth, jaw and gill elements correlate with feeding
mode in some lacustrine threespine stickleback populations
[91,135]. An assumption that differently shaped fish might be
adapted, for example, to capturing suspended zooplankton or to
foraging on benthic prey is reflected in the label “ecotypes” [83].
Likewise, derived states of loss or reduction in the number and
robustness of bony elements in freshwater stickleback populations
might be driven by predator regime or by the reduced mineral
availability of fresh water [73]. Differences between oceanic and
freshwater stickleback predict that selection acts on developmental
processes that shape the skeleton and on pathways that regulate
bone density and ion physiology.

Orthologs of many genes known to affect bone development by
modulating specification, differentiation, proliferation, migration
and patterning of skeletogenic tissues fall within genomic regions
associated with differentiation between oceanic and freshwater stickleback. In other vertebrates, profound effects on the developmental patterning of the teeth, jaw, and other branchial arches result from changes in expression of EDA, EYA1, FBLN1, NPYB, RDIH10, and Wnt5a genes [136,137,139–142]. Orthologs of these six genes fall within genomic intervals associated with differentiation between oceanic and freshwater sticklebacks (Table 3 and Table S3). Skeletal structure is continuously maintained and shaped throughout life by a balance between bone deposition and removal, carried out by osteoblasts and osteoclasts. Several osteogenic candidates in genomic regions differing between oceanic and lake stickleback are orthologs of genes that are also associated with human bone density variation, including imbalanced, disease states such as osteoporosis and osteopetrosis. These genes include LEMD3, LEPR, ARHGEF3 and RHOA (Table 3 and Table S3) [143–145].

Nonadromous fish such as salmon undergo smoltification, a set of morphological and physiological changes that prepare the juvenile fish for the demanding transition from freshwater to marine. Stickleback entrained in freshwater lakes have lost this portion of their life history, and are probably no longer under strong selection pressure to maintain tolerance and physiological adaptability to saline conditions. On the other hand, fish adapted to freshwater must contend with limited access to minerals (e.g., calcium) and with a steep gradient of internal to external ion concentration. Peaks of oceanic-freshwater differentiation on LG IV, VII and XXI in stickleback contain genes associated with acute physiological adaptation to hypo- or hyperosmotic conditions in other species of fish, namely PRL2, a hormone controlling osmoregulation, and CA4 and ATP6V1A, important for ion transport across the gill epithelium and skin (Table 3) [146–149].

Two genes, CA4 and FLT1, of which we found stickleback orthologs within peaks of differentiation on LG VII and XXI, have pleiotropic roles in both bone biology and osmoregulation [146,149–152], suggesting a possible pleiotropic basis for coordinated evolutionary responses to freshwater conditions in skeletal characters and ion physiology. Evolved responses to the host of physical and biological constraints that differ between freshwater and oceanic life histories are expected to be genetically complex. It is not surprising, therefore, that we find many genomic regions displaying strong patterns of differentiation between populations. What is surprising is the consistency of the regions of differentiation and the number of compelling candidate targets for selection they contain, suggesting the possible co-selection of functionally related, multilocus genotypes.

Conclusions

This work represents the first whole-genome analysis of threespine stickleback in which high-density SNP markers reveal signatures of selection in natural populations. The patterns we detected confirm findings from earlier studies that used QTL analysis in controlled crosses or research that used microsatellite markers in natural populations to scan the genome. However, because of the dense coverage of SNPs across the genome, and our ability to sample numerous individuals in multiple populations, our findings are a significant extension of previous work. The present investigation complements these prior efforts by exposing new genomic regions that had not yet been recognized as important in the transition from oceanic to freshwater life histories. In particular, we find remarkably similar patterns of conservation and differentiation between three independently derived freshwater populations as compared to a common oceanic ancestor. Our data support the view that these patterns are driven in part by alleles that are repeatedly selected for in freshwater populations, and maintained at low frequency in oceanic populations by a balance between gene flow from freshwater and selection against them in the ocean. Previous work supported the role of parallel genetic evolution associated with parallel phenotypic evolution in a small number of traits. Our data indicate that this pattern is not limited to these traits, and that parallel phenotypic evolution in stickleback may be underlain by extensive, genome-wide, parallel genetic evolution.

Methods

Collection of stickleback samples

Threespine stickleback were collected from five populations in Alaska: Rabbit Slough (oceanic), Resurrection Bay (oceanic), Bear Paw Lake (freshwater), Boot Lake (freshwater), and Mud Lake (freshwater) (Figure 1). Fish were collected by beach seine (Resurrection Bay) or by minnow trap (lakes and Rabbit Slough) from wild populations in the summers of 1997 and 1998. Bear Paw Lake (61°36’N, 149°45’W, elev. 38 m), Boot Lake (61°43’N, 150°07’W, elev. 55), and Mud Lake (61°56’N, 150°58’W, elev. 38 m) are all in different drainage systems, separated by geographic barriers of distance and elevation. Rabbit Slough (61°32’N, 149°15’W, elev. 5 m) and Resurrection Bay (60°07’N, 149°23’W, elev. 14 m) empty to opposite sides of the Kenai Peninsula. Fish were anaesthetized with a tricaine methane sulphonate solution (MS222), frozen on dry ice in the field, and later transferred to 100% ethanol. Genomic DNA was purified from fin tissue using the DNeasy Blood & Tissue Kit (Qiagen).

Creation of RAD tag libraries

Genomic DNA was purified from 20 individuals from each of the five populations. DNA from each fish was digested with high fidelity TaqI (New England Biolabs). RAD tag libraries were created as in Baird et al. [99] with the following modifications: only barcodes that differed by at least three nucleotides were used, a longer P2 adapter (with the following sequences: P2-2 top oligo 5’/5Phos/GATCGGAAGAGCGGTTCAGCAGGAATGCCGA-GACCGATCGAAACAAG3’; P2-2 bottom oligo 5’ CAAGCAGA-AGACGGCACATCGAGATCGGTTCGCGAATTCTCGTA-ACCGCTCTTCGGATCT3’) was used in the production of all libraries, libraries produced for the May 2009 run and thereafter used P1 and P2 adapters modified with a phosphorothioate bond between the last two 3’ nucleotides on both oligos of the P1 adapter and the bottom oligo of the P2 adapter. Adaptor ligated DNA was subjected to fewer rounds (14 or 16) of PCR amplification and PCR products were gel purified by excising a DNA fraction of 400–600 bp. Each Illumina sequencing lane contained a library representing approximately equal amounts of DNA from 16 individual fish (refer to Table S1). Sequences are available at the NCBI Short Read Archive [http://www.ncbi.nlm.nih.gov/Traces/sra; accession number SRA010788.9].

Infering genotypes

Sequence reads from the Illumina runs were filtered as follows: reads with a barcode that did not match one of the expected barcodes (i.e. a sequencing error in the barcode), and sequence reads of poor overall quality, were removed from the analysis. Sequence reads were then sorted by barcode and aligned to the stickleback genome using Bowtie [153] with a maximum of 2 mismatches within the first 28 bases and a sum of base quality for all mismatches in the read no greater than 70. Following alignment, the read counts of the four possible nucleotides at each nucleotide site were tallied for each individual (see Figure 2). Reads were further
truncated by removing the portion of the sequence within the restriction enzyme recognition site, since any nucleotide polymorphism in this area would result in the absence of RAD tags, and including these data would underestimate total nucleotide diversity.

Diploid genotypes at each nucleotide site for each individual were determined in a maximum likelihood statistical framework as follows. For a given site in an individual, let \( n \) be the total number of reads at that site. Let \( n = n_1 + n_2 + n_3 + n_4 \), where \( n_i \) is the read count for each possible nucleotide at the site (disregarding ambiguous reads). For a diploid individual, there are ten possible genotypes (four homozygous and six heterozygous genotypes). We calculate the likelihood of each possible genotype by using a multinomial sampling distribution, which gives the probability of observing a set of read counts \((n_1, n_2, n_3, n_4)\) given a particular genotype. For example, the likelihoods of a homozygote (genotype 1/1) or a heterozygote (1/2) are, respectively:

\[
L(1/1) = P(n_1, n_2, n_3, n_4 | 1/1) = \frac{n!}{n_1! n_2! n_3! n_4!} \left( 1 - \frac{3e}{4} \right)^{n_1} \left( \frac{e}{4} \right)^{n_2} \left( \frac{e}{4} \right)^{n_3} \left( \frac{e}{4} \right)^{n_4}
\]

and

\[
L(1/2) = P(n_1, n_2, n_3, n_4 | 1/2) = \frac{n!}{n_1! n_2! n_3! n_4!} \left( 0.5 - \frac{e}{4} \right)^{n_1} \left( \frac{e}{4} \right)^{n_2} \left( \frac{e}{4} \right)^{n_3} \left( \frac{e}{4} \right)^{n_4}
\]

where \( e \) is the sequencing error rate. If we let \( n_1 \) be the count of the most observed nucleotide, and \( n_2 \) be the count of the second-most observed nucleotide, then the two equations in (1) give the likelihood of the two most likely hypotheses out of the ten possible genotypes. For all the analyses that follow, we assigned a diploid genotype to each site based on a likelihood ratio test between these two most likely hypotheses with one degree of freedom. If this test was significant at the \( \alpha = 0.05 \) level, we assigned the most likely genotype at the site. If this test was not significant, we did not assign a genotype at the site for that individual. This effectively removes data for which there are too few sequence reads to determine a genotype, instead of establishing a constant threshold for sequencing coverage. We account for the resulting variance in sample size among sites in the analyses below.

This basic multinomial-based statistical framework has been proposed elsewhere [154]. Our approach differs from that of Lynch [154], however, in that we estimate the sequencing error rate \( e \) separately by maximum likelihood for each nucleotide site, rather than assuming or estimating a single global error rate. We have found empirical evidence that sequencing error varies among sites, and that this approach is more robust to other assumptions than using a single global error rate (Hohenlohe and Cresko, unpublished data). Note that equations (1) allow for a random sequencing error rate but do not account for any systematic biases in, for instance, the frequency of sequence reads for alternative alleles at a heterozygous site. The generation of likelihoods for each of the ten possible genotypes at each site also allows for more sophisticated methods than were used here of carrying error and uncertainty through the analysis to the final population genetic measures. We will address these and other aspects of this statistical genotyping method in a forthcoming paper (Hohenlohe and Cresko, in preparation).

Calculating population genomic statistics

We first calculated four population genetic measures at each nucleotide site for the population(s) under examination. To estimate nucleotide diversity, we calculated \( \pi \) (equivalent to expected heterozygosity) as

\[
\pi = 1 - \sum \left( \frac{n_i}{2} \right) / \left( \frac{n}{2} \right)
\]

where \( n_i \) is the count of allele of population \( j \) and \( \pi \) is the nucleotide diversity within population \( j \) from equation (2), and \( \pi_* \) is the total nucleotide diversity across the pooled populations. We compared this measure of \( F_{ST} \) to others, including the analysis of variance approach of [21], and found that it gave similar results but performed well with small sample sizes. In particular, the consistency and location of the peaks examined in detail here did not change with different methods of estimating \( F_{ST} \) (not shown).

For each population in a comparison we assessed whether each single nucleotide polymorphism (SNP) was the result of a private allele. Here \( \rho_j = 1 \) if an allele at the SNP is found only in population \( j \) and at least one individual was genotyped at that nucleotide site in each population, and \( \rho_j = 0 \) otherwise.

To generate smooth genome-wide distributions of these four population genetic measures, we used a kernel-smoothing moving average. For each genomic region centered on a nucleotide position \( c \), the contribution of the population genetic statistic at position \( j \) to the region average was weighted by the Gaussian function \( \exp \left( -\frac{(p-c)^2}{2\sigma^2} \right) \), where \( \sigma = 150 \) kb. For computational efficiency, we truncated this weighted average at \( 3\sigma \) in each direction (beyond which nucleotide sites have a relative weight less than \( \sim 0.01 \). We evaluated multiple choices for the width \( \sigma \) and found 150 kb to be large enough to overcome sampling variance but still small enough to detect relatively narrow genomic regions of differentiation, with a precision greater than many QTL studies (data not shown). For example, in the overall freshwater-oceanic comparison each 6 \( \sigma \) window contained a mean of 81.6 SNPs. We shifted the moving average by a step size of 100 kb. Because of the variance in sample size across sites (due to sampling variance in Illumina sequencing and sites where a genotype could not be assigned using the maximum likelihood technique above), we further weighted each statistic at each nucleotide position by \( n_k - 1 \), where \( n_k \) is the number of alleles sampled at site \( k \) [156]. As above, we explored different weighting formulas, as well as unweighted averages, and these did not appreciably change the consistency or location of major peaks in population differentiation (not shown). Nucleotide diversity \( \pi \) and heterozygosity \( H \) were weighted and averaged across all nucleotide sites; \( F_{ST} \) and private allele density \( \rho \) were weighted and averaged across all SNPs.

We also estimated the allele frequency spectrum within populations or groups of populations using Tajima’s D [102], applied to the nucleotide diversity \( \pi \) and number of SNPs within \( \sigma \) bp of the center of each window (i.e. \( 2\sigma = 300 \) bp windows). Sample size \( n \) was taken to be the mean of \( n_k \) across all sites within the window.
We assessed statistical significance at two levels. At individual SNPs, we estimated the significance of $F_{ST}$ values with a goodness-of-fit $G$ test statistic [157]. We corrected for false discovery rate in multiple tests using the Benjamini-Hochberg correction [158]. We assume that population differentiation at linked SNPs may be positively correlated, so this method of correction is still valid [159].

To assign significance values to moving average values of $\pi$, $H$, $F_{ST}$, and $p$, as well as window values of Tajima’s $D$, we used bootstrap resampling within each population comparison. For each nucleotide position (for $\pi$, or $H$, or $F_{ST}$ or $p$) within each truncated Gaussian window described above, we randomly sampled with replacement from across the entire genome a value for the statistic ($\pi$, $H$, $F_{ST}$, or $p$) and the corresponding sample size ($n$). We calculated the weighted average as above for each replicate. For Tajima’s $D$, for each nucleotide position within the 2σ window we randomly sampled with replacement from across the genome and calculated the overall $D$ for the re-sampled dataset. For computational efficiency, at each region we began with 100 (for $\pi$ or $H$), 1,000 (for $D$), or 10,000 (for $F_{ST}$ or $p$) replicates and stepped up to 1 million ($\pi$, $H$, or $D$) or 10 million ($F_{ST}$ or $p$) replicates as necessary to provide accuracy in the tails of the distribution. Essentially this bootstrapping technique gives a null distribution of expected genomic region averages, accounting for the observed genome-wide average of each statistic in a given population or population comparison, but assuming no correlation among neighboring positions. It thus indicates genomic regions that differ significantly from the genome-wide average as a result of the combination of linkage disequilibrium and evolutionary or demographic processes. Significance values ($p$) given in the text and tables represent proportions of these bootstrap distributions exceeding the particular statistic.

We used these significance values to delineate regions of interest for identification of candidate genes. For nucleotide diversity, two regions on LG III and XIII were delineated to include all regions with $p<10^{-3}$ for $\pi$ in the combined 5-population dataset, including positions within 2σ ($\pm 300$ kb) of the outer positions. For $F_{ST}$, we identified all genomic regions for which $p<10^{-3}$ in the overall freshwater-oceanic comparison as well as in all six of the pairwise freshwater-oceanic comparisons. We then delineated the region of interest using the overall freshwater-oceanic comparison, $+/-2\sigma$ as above. Note that this 2σ margin indicates locations that may contribute to a highly significant average value of a statistic, even if the value for the genomic region directly over the gene is not as significant (examples in Table 3). We took this approach in order to cast a wide net for selection on potential candidate genes, including their associated cis-regulatory regions.

For several reasons, we believe that our method may provide an underestimate of nucleotide diversity within populations. First, we expect polymorphism in RAD sites, such that the restriction enzyme recognition site is missing in some haplotypes and a RAD tag sequence will not be obtained for this allele. Individuals homozygous for absence of a RAD site will lack any sequence information for those two RAD tags; individuals heterozygous for the presence of a RAD site will be represented by one of only two possible sequences for each tag, so they will likely be scored as homozygous for all nucleotide positions in those tags. (It is intuitive to use the total number of reads to identify such RAD-site heterozygotes, although the sampling process and other sources of variation in read counts may make such inferences tenuous). We removed sequence data within the restriction enzyme recognition site prior to analysis. However, to the extent that presence/absence of a RAD site is in linkage disequilibrium with SNPs in the adjacent RAD tag sequence, this polymorphism will be underestimated. Second, RAD tags with low coverage are not assigned a genotype by the method above if the likelihood ratio test is not significant. Because of the multinomial sampling process, true heterozygotes may be more likely to go unscored than true homozygotes at the same, low level of sequencing depth. Third, we have some evidence that there is bias in number of reads and read quality between alternative alleles at heterozygous sites during library construction and/or Illumina sequencing (unpublished data). As described above, our method does not account for these unknown sources of bias, but they could also lead to the analysis assigning homozygous genotypes to heterozygous sites. We are currently exploring ways to account for all of these issues in the analysis (Hohenlohe and Cresko, in preparation). In any case, we believe that while our method may lead to an underestimate of nucleotide diversity measures within groups (i.e., $\pi$ and $H$), these issues are not likely to bias the distribution of these measures along the genome. Also, they should not bias measures of population differentiation ($F_{ST}$), assuming that these sources of error affect different population samples equally.

### Supporting Information

**Figure S1** Nucleotide diversity within single and groups of populations. Nucleotide diversity ($\pi$) across the genome, with colored bars indicating significantly elevated ($p<10^{-3}$, blue) and reduced ($p<10^{-3}$, green) values. Vertical gray shading indicates boundaries of the 21 linkage groups and unassembled scaffolds, and gold shading indicates two consistent peaks of elevated nucleotide diversity. (A) RS. (B) RB. (C) OC (RS + RB). (D) BP. (E) BL. (F) ML. (G) FW (BP + BL + ML). Found at: doi:10.1371/journal.pgen.1000862.s001 (2.35 MB TIF)

**Figure S2** Private allele density in the overall freshwater-oceanic comparison. Each plot shows density of private alleles ($p$), with colored bars indicating regions of significantly elevated ($p<10^{-3}$, blue; $p<10^{-3}$, red) and reduced ($p<10^{-3}$) values, assessed by bootstrap resampling. Vertical gray shading indicates the 21 linkage groups and unassembled scaffolds, and gold shading indicates the nine consistent peaks of population differentiation. (A) Private allele density in FW compared to OC. (B) Private allele density in OC compared to FW. Found at: doi:10.1371/journal.pgen.1000862.s002 (1.38 MB TIF)

**Figure S3** Density of annotated and predicted genes along the stickleback genome. Count of genes in each 1-Mb window, taking into account the nine consistent peaks of population differentiation (A) RS. (B) RB. (C) OC (RS + RB). (D) BP. (E) BL. (F) ML. (G) FW (BP + BL + ML). Found at: doi:10.1371/journal.pgen.1000862.s003 (0.66 MB TIF)

**Table S1** Illumina sequencing runs used in this analysis. Found at: doi:10.1371/journal.pgen.1000862.s004 (0.06 MB DOC)

**Table S2** A complete list of the protein coding genes that fall in genomic regions associated with differences between oceanic and freshwater populations. Gene names are listed, where available from Ensembl (release 55.1j). Where gene names were lacking, ortholog names are listed for candidate genes from Table 3. Orthology for unnamed genes was extracted from the Ensembl annotation for each gene or determined by a BLAST search of the NCBI protein database using the predicted protein/s for each gene. Broad orthology groups for candidates are denoted by red text (those listed under the heading “Morphology” in Table 3) or blue text (those listed under “Osmoregulation” in Table 3). Found at: doi:10.1371/journal.pgen.1000862.s005 (0.10 MB XLS)
Table S3  Candidate genes related to skeletal morphology and osmoregulation in additional regions of differentiation on Linkage Groups IV, VII, and XII.

References


