OPEN ACCESS Freely available online

PLoS GENETICS

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

Paul A. Hohenlohe[1☯], Susan Bassham[1☯], Paul D. Etter[2], Nicholas Stiffler[3], Eric A. Johnson[2], William A. Cresko[1]*

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.

Citation: Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, et al. (2010) Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. PLoS Genet 6(2): e1000862. doi:10.1371/journal.pgen.1000862

Editor: David J. Begun, University of California Davis, United States of America

Received October 20, 2009; **Accepted** January 28, 2010; **Published** February 26, 2010

Funding: This work was funded by grants from the National Science Foundation (IOS-0642264) and from the National Institutes of Health (1R24GM079486-01A1 and Ruth L. Kirschstein National Research Service Award F32 GM078949). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: wcresko@uoregon.edu

☯ These authors contributed equally to this work.

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 [65–75]. 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 [98,99]. 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 [53,55,79]. 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 F_{ST} [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 SbfI 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

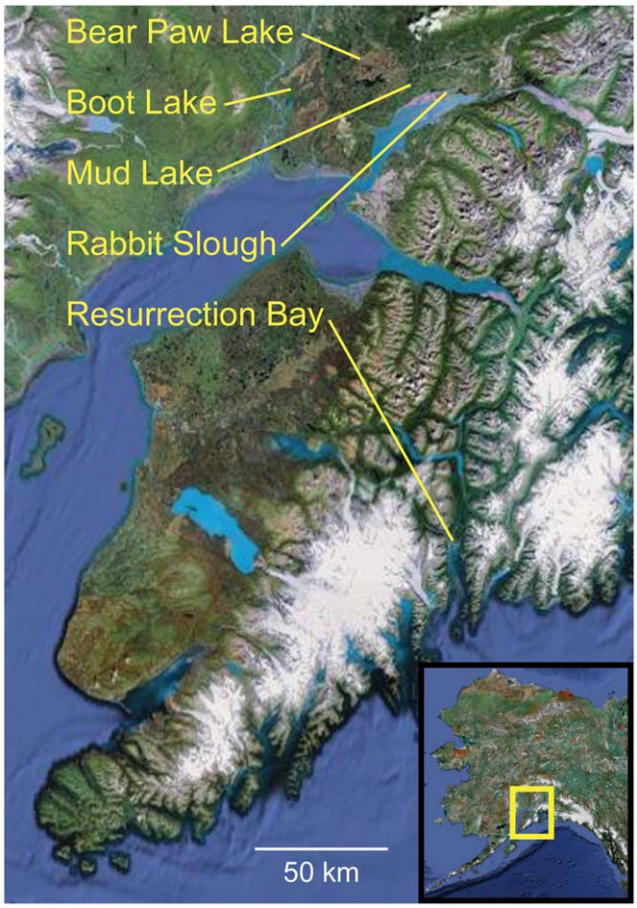

Figure 1. Location of oceanic and freshwater populations examined. Threespine stickleback were sampled from three freshwater (Bear Paw Lake [BP], Boot Lake [BL], Mud Lake [ML]) and two oceanic (Rabbit Slough [RS], Resurrection Bay [RB]) populations in south central Alaska, USA (see inset). The three freshwater populations occur in different drainages and are separated by barriers to dispersal, and previous evidence supports the hypothesis that they represent independent colonization events from ancestral oceanic populations [49].
doi:10.1371/journal.pgen.1000862.g001

conservative (and unbiased) nature with which SNPs are called using our methodology (see Methods), and additional sequencing of these samples may increase the number of SNPs identified. Furthermore, in agreement with the hypothesis that freshwater populations in this region have been derived post-glacially from oceanic populations [49,55,65,79], global genetic diversity measures are increased only slightly when combining pairs of populations whether they are both oceanic, both freshwater, or one of each (Table 2).

Our data support the hypothesis that oceanic stickleback populations have few barriers to dispersal, relatively large amounts of gene flow, and little population genetic subdivision [55,57,59,60,103,104]. Rabbit Slough and Resurrection Bay, the two oceanic populations in our study, are the most geographically distant from one another (>1000 km as the fish swims). Despite this distance, the oceanic populations show the least amount of differentiation between them ($F_{ST} = 0.0076$; Table 2). In contrast, higher values of F_{ST} were observed in pairwise comparisons among freshwater populations and between freshwater and oceanic populations (0.05–0.15), which is generally interpreted as low to moderate amounts of population structuring (Table 2).

The freshwater populations, despite their younger age, are more divergent both from the oceanic ancestral populations and from each other, consistent with our supposition that they represent independent colonizations from the ancestral oceanic population. These results are remarkably similar to results obtained previously from some of these same populations using a small number of microsatellite and mtDNA markers [55]. This combination of large amounts of genetic variation and overall low-to-moderate differentiation between populations, coupled with recent and rapid phenotypic evolution in the freshwater populations, presents an ideal situation for identifying genomic regions that have responded to various kinds of natural selection.

Patterns of genetic diversity distributed across the genome

To assess genome-wide patterns we examined mean nucleotide diversity (π) 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 0.0003 to over 0.01, whereas heterozygosity values range from 0.0001 to 0.0083. This variation in diversity across the genome provides important clues to the evolutionary processes that are maintaining genetic diversity. For example, while expected (π) 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 580 kb outside the interval

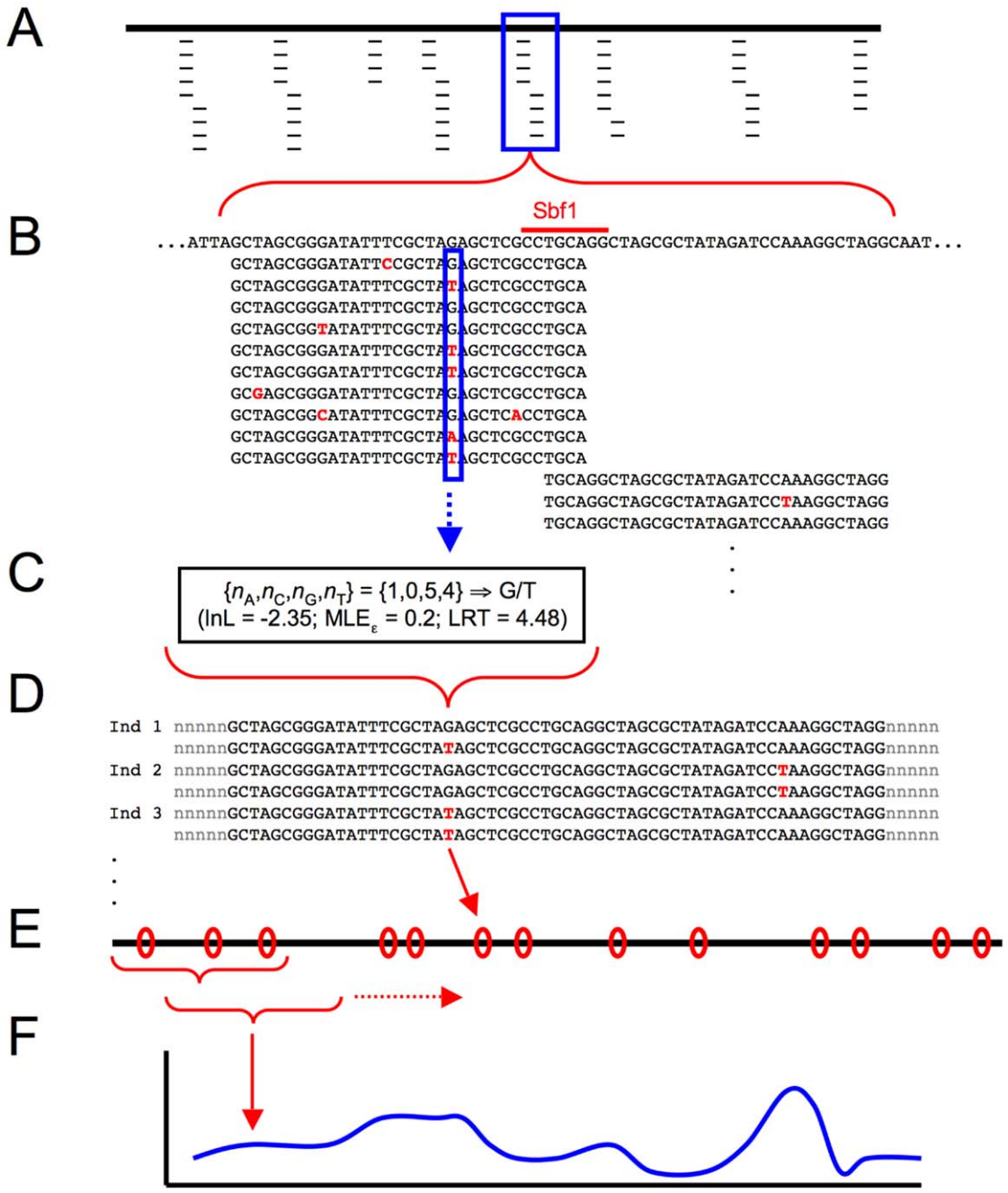

Figure 2. Schematic diagram of population genomic data analysis using RAD sequencing. (A) Following Illumina sequencing of barcoded fragments, sequence reads (thin lines) are aligned to a reference genome sequence (thick line). Depth of coverage varies across tags. Reads that do not align to the genome, 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 ε, 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

of maximum nucleotide diversity on LG III, although both π and *H* are moderately elevated at this region as well (π = 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, ENSGACG0000000-

Table 1. Nucleotide sites and SNPs identified on each linkage group.

Linkage group[1]	Length[2]	Sites[3]	RS[4]	RB	BP	BL	ML	OC	FW	ALL
I	28,185,914	125,496	994	1,316	688	812	1,025	1,694	1,549	2,417
II	23,295,652	100,502	764	1,074	566	620	893	1,336	1,329	1,979
III	16,798,506	84,770	840	1,191	697	763	1,035	1,499	1,574	2,257
IV	32,632,948	138,898	999	1,408	749	865	1,278	1,774	1,842	2,871
V	12,251,397	59,631	497	656	347	394	561	851	813	1,243
VI	17,083,675	77,914	688	907	440	512	799	1,140	1,082	1,615
VII	27,937,443	115,092	838	1,092	677	739	984	1,429	1,489	2,312
VIII	19,368,704	87,664	700	933	456	589	774	1,188	1,141	1,736
IX	20,249,479	91,100	731	971	511	560	787	1,250	1,171	1,798
X	15,657,440	69,574	602	827	427	477	661	1,040	979	1,490
XI	16,706,052	82,787	699	948	495	586	763	1,215	1,172	1,801
XII	18,401,067	74,887	634	806	473	535	703	1,055	1,063	1,630
XIII	20,083,130	91,333	794	998	538	634	847	1,307	1,255	1,897
XIV	15,246,461	73,639	611	874	462	505	773	1,072	1,084	1,560
XV	16,198,764	75,415	618	837	414	476	645	1,041	938	1,438
XVI	18,115,788	74,669	653	795	392	464	642	1,039	981	1,519
XVII	14,603,141	65,431	606	772	401	427	598	1,004	882	1,370
XVIII	16,282,716	80,526	678	923	484	544	799	1,170	1,156	1,709
XIX	20,240,660	89,505	582	919	594	664	814	1,118	1,180	1,689
XX	19,732,071	78,669	558	777	463	472	659	988	996	1,538
XXI	11,717,487	51,484	428	552	339	359	526	730	751	1,169
Other	60,744,953	303,308	2,536	3,891	2,692	2,940	4,507	4,767	6,618	8,751
TOTAL	461,533,448	2,092,294	16,870	23,467	13,305	14,937	21,073	29,707	31,045	45,789

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

0348, ENSGACG00000000350). Nucleotide diversity ($\pi = 0.0046$, $p<0.02$), heterozygosity ($H = 0.0030$, pH 0.01), and freshwater-oceanic differentiation ($F_{ST} = 0.0218$, pH 0.05) averaged over this scaffold are somewhat consistent with a hypothesis of balancing selection.

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-oceanic comparison (Figure 6E) and the comparison among freshwater populations (Figure 6F).

Examining more closely the height and location of peaks in F_{ST} across these comparisons, we can discern a set of general patterns to generate hypotheses about the modes of genetic variation and selective forces operating in the adaptation to freshwater, and to identify putative candidate genes. Single linkage groups illustrating examples of these distinctive patterns are shown in Figure 7 and Figure 8. First, the large majority of genomic regions of elevated F_{ST} are shared across the three freshwater populations. This pattern suggests independent, parallel evolution in the form of similar 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 XXI (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, $\pi<0.001$, $p = 0.0003$; $H = 0.0006$, $p = 0.0003$).

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 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 $p<6.85\times10^{-23}$) following false discovery rate correction of individual *G*-tests (see Methods). These highly significant SNPs largely correspond with the genomic

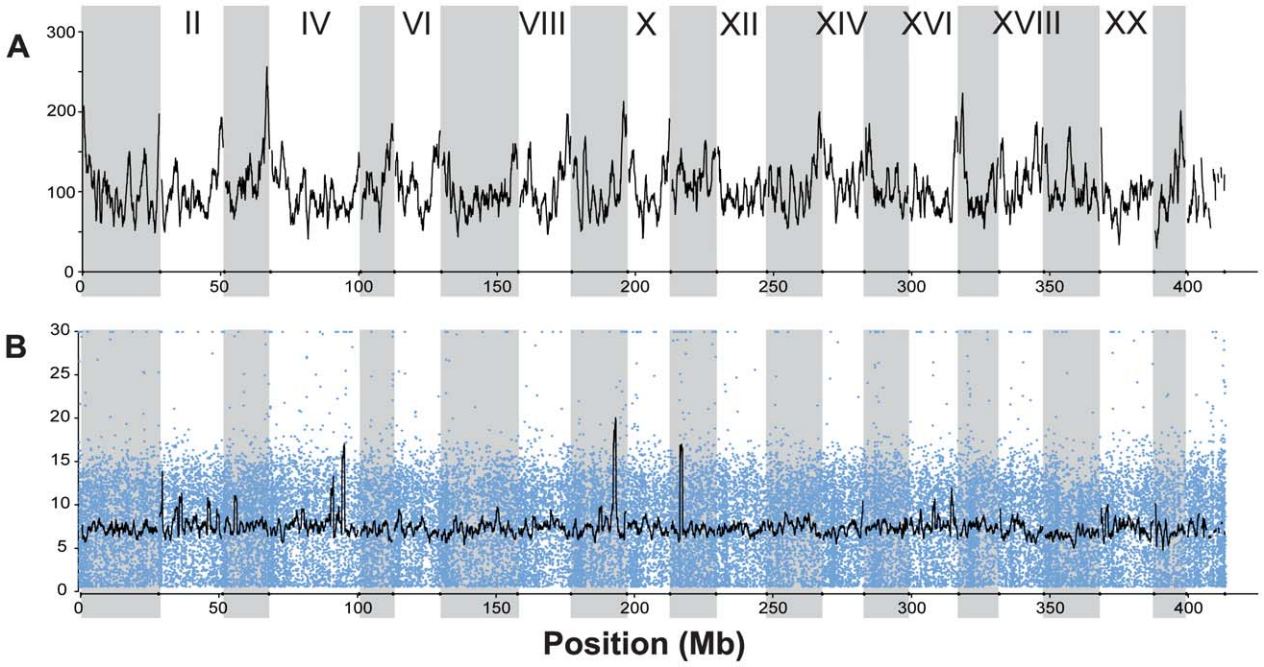

Figure 3. Depth of RAD sequencing coverage. (A) Number of RAD tags sequenced per 1-Mb sliding window across the genome. Each RAD tag represents either 30 or 47 bp of sequence data (see Table S1). Vertical gray shading indicates Linkage Groups I through XXI, followed by all unassembled scaffolds greater than 1 Mb in length. Not all RAD tags were sequenced in all individuals, because of both random sampling in the sequencing process and polymorphism in the restriction enzyme recognition site. (B) Sequencing depth per RAD tag per individual from one sample run (22 May 2009, lane 7; see Table S1). Blue dots represent the average number of reads per individual across 16 individuals sampled for each RAD tag. The black line shows the mean depth per individual in a 1-Mb sliding window. A total of 5,597,895 barcoded and aligned sequence reads from 16 individuals were generated from this run.
doi:10.1371/journal.pgen.1000862.g003

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 ~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 ($F_{ST} = 0.3092$, $p<0.02$ vs. RS; $F_{ST} = 0.2737$, $p<0.01$ vs. RB) that is not observed to the same extent in the other two populations. Correspondingly, there is a peak in differentiation among the freshwater populations at this location. This contrasts with the peak at ~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.[1]

	RS	RB	BP	BL	ML
RS	**0.00216**	0.00267	0.00277	0.00290	0.00308
RB	0.0076	**0.00250**	0.00291	0.00296	0.00308
BP	0.1391	0.0650	**0.00203**	0.00269	0.00295
BL	0.1040	0.0462	0.1310	**0.00227**	0.00299
ML	0.1252	0.0849	0.0798	0.0868	**0.00268**

1 Above the diagonal is average nucleotide diversity (π) in each combined pair of populations; along the diagonal is π within each single population; below the diagonal is average F_{ST} between the two populations. Population abbreviations are as in Table 1.
doi:10.1371/journal.pgen.1000862.t002

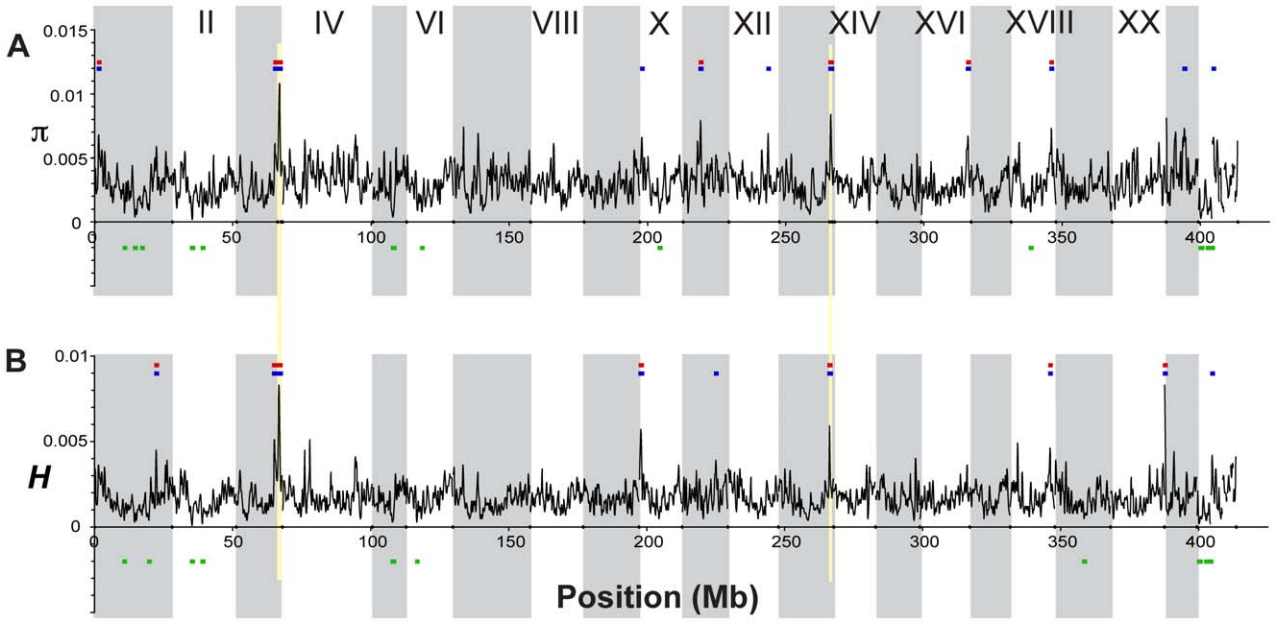

Figure 4. Genome-wide patterns of nucleotide diversity. Each plot shows a smoothed distribution of the statistical measure across the genome (black lines). Colored bars above and below the distributions indicate regions of significantly elevated ($p \leq 10^{-5}$, blue; $p \leq 10^{-7}$, red) and reduced ($p \leq 10^{-5}$, 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 (π) across all five stickleback populations sampled. (B) Heterozygosity (*H*) across all five populations.
doi:10.1371/journal.pgen.1000862.g004

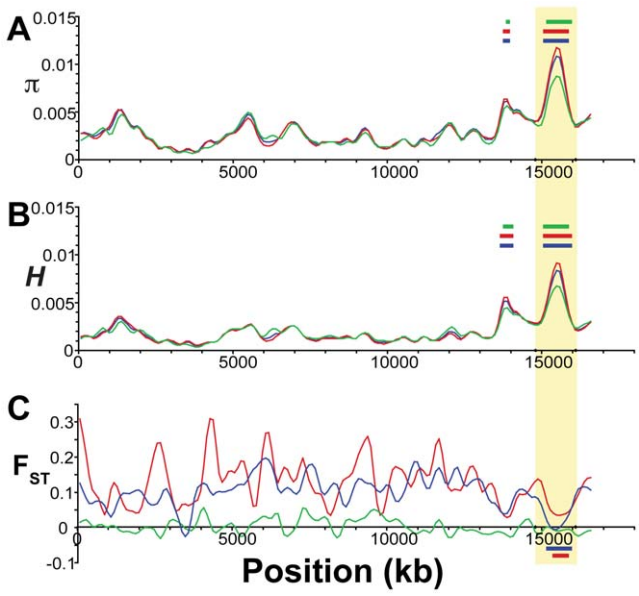

Figure 5. Evidence for balancing selection on Linkage Group III. Population genetic measures plotted along Linkage Group III. (A) Nucleotide diversity (π) and (B) heterozygosity (*H*) across all five (blue), the three freshwater (red), and the two oceanic (green) populations. (C) Population differentiation (F_{ST}) between oceanic and freshwater (blue), among freshwater (red), and between oceanic (green) populations. Colored bars indicate significant ($p \leq 10^{-5}$) 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

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 (ρ) 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.

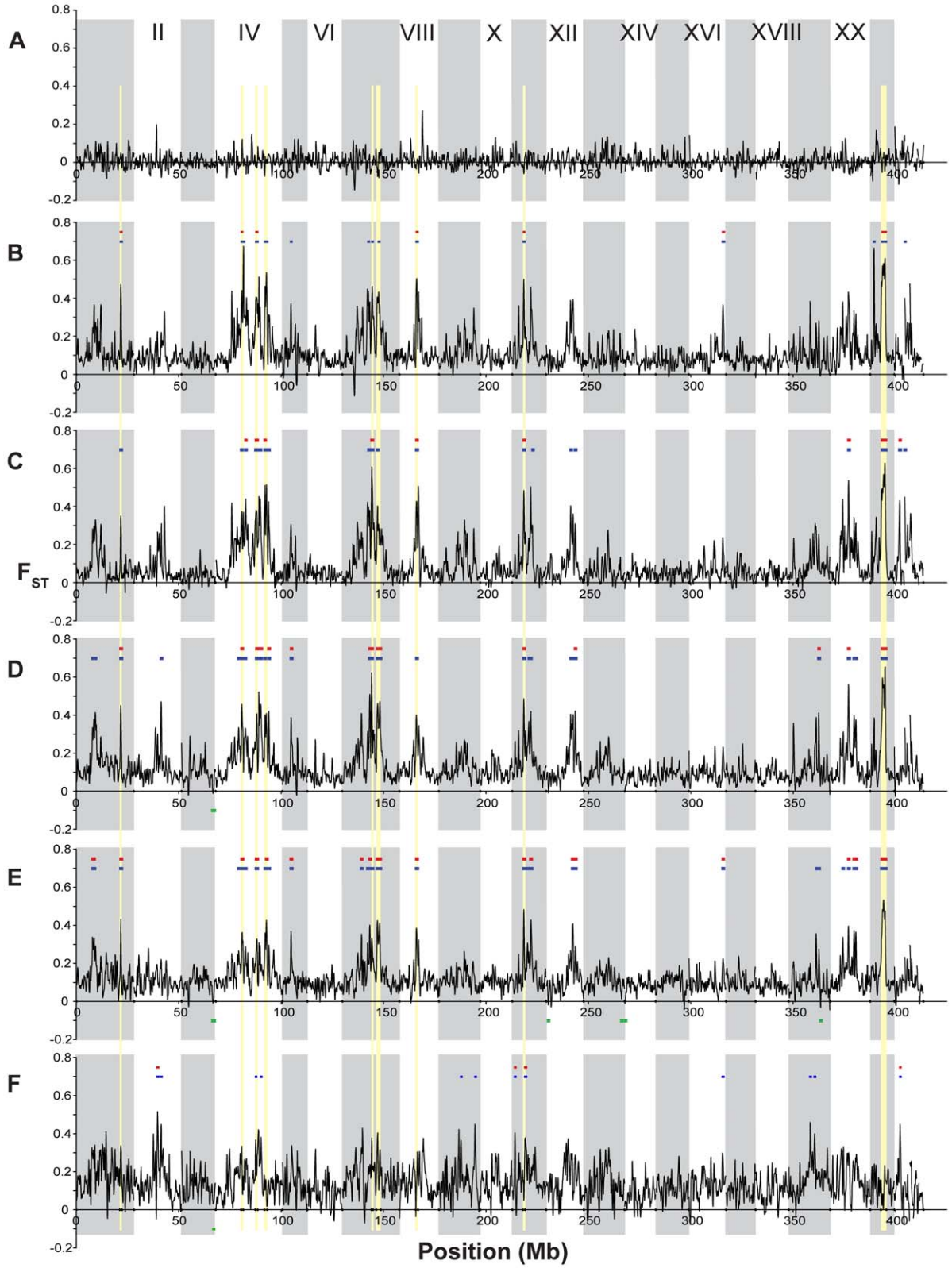

Figure 6. Genome-wide differentiation among populations. F_{ST} across the genome, with colored bars indicating significantly elevated ($p \leq 10^{-5}$, blue; $p \leq 10^{-7}$, red) and reduced ($p \leq 10^{-5}$, green) values. Vertical gray shading indicates boundaries of the linkage groups and unassembled 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

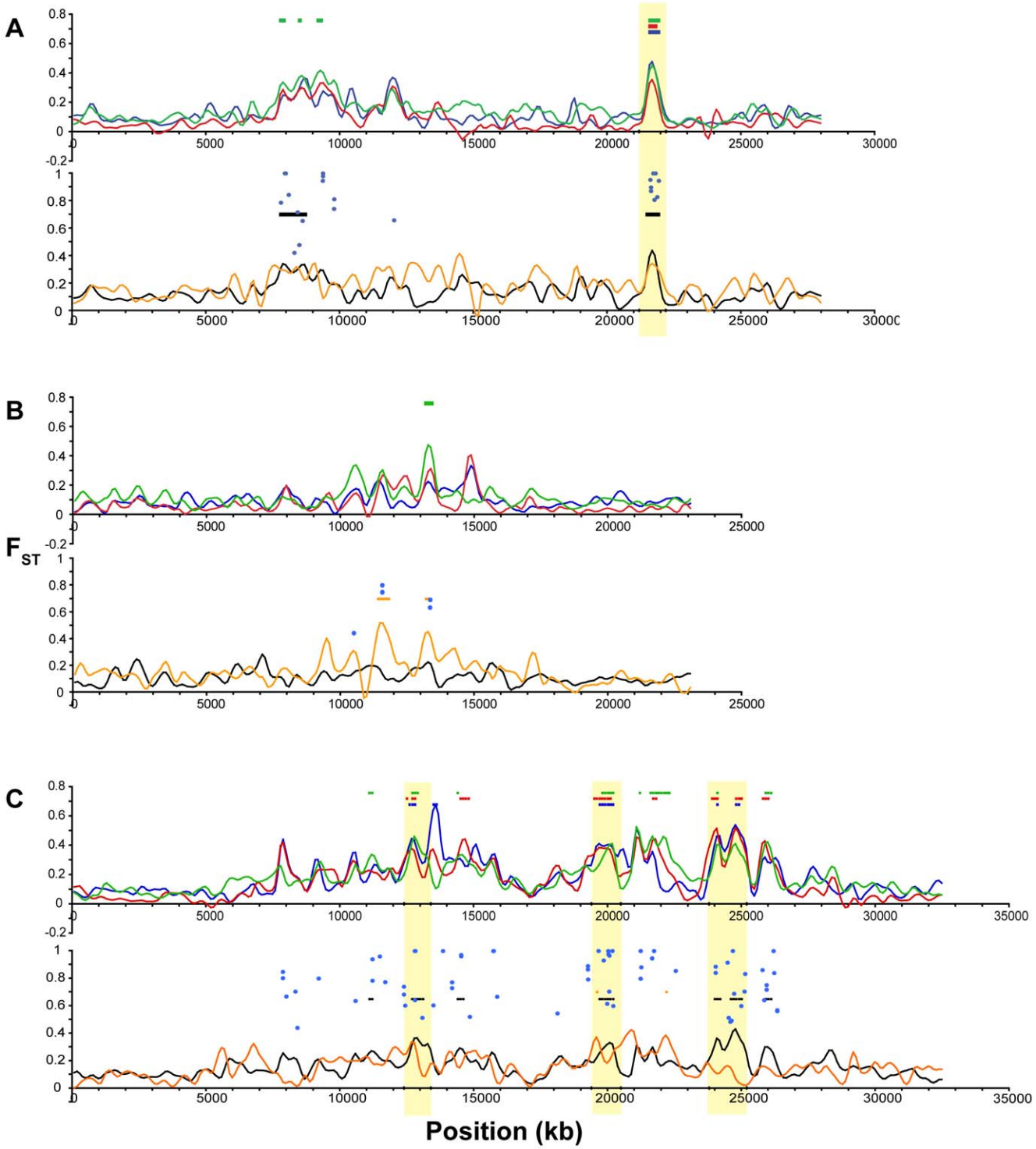

Figure 7. Differentiation among oceanic and freshwater populations on Linkage Groups I, II, and IV. For each linkage group, the upper panel shows population differentiation (F_{ST}) of each freshwater population from the two oceanic populations, plotted as the mean of the two freshwater versus oceanic 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

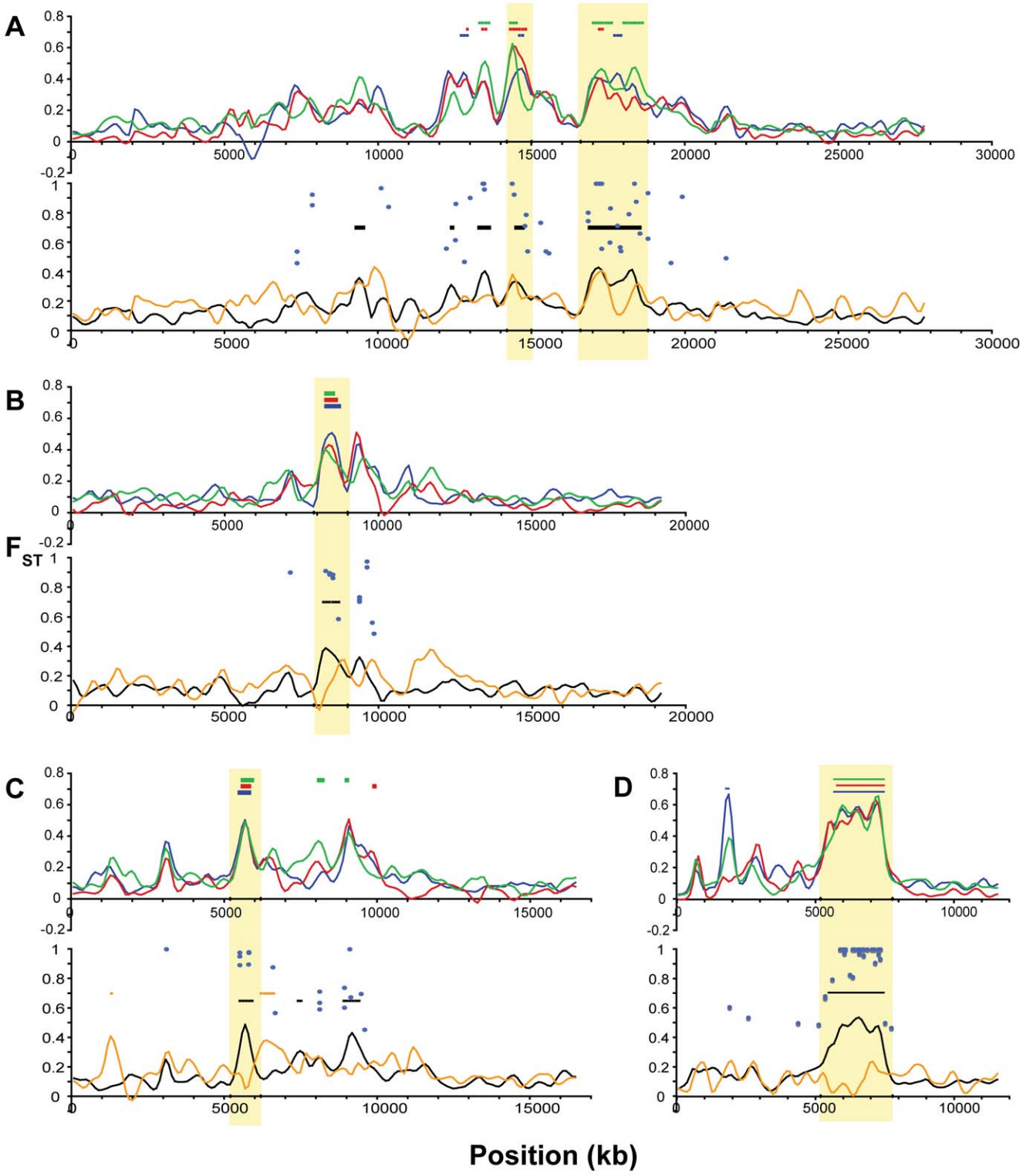

Figure 8. Differentiation among oceanic and freshwater populations on Linkage Groups VII, VIII, XI, and XXI. All panels show population differentiation as in Figure 7. (A) LG VII. (B) LG VIII. (C) LG XI. (D) LG XXI.
doi:10.1371/journal.pgen.1000862.g008

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, π 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

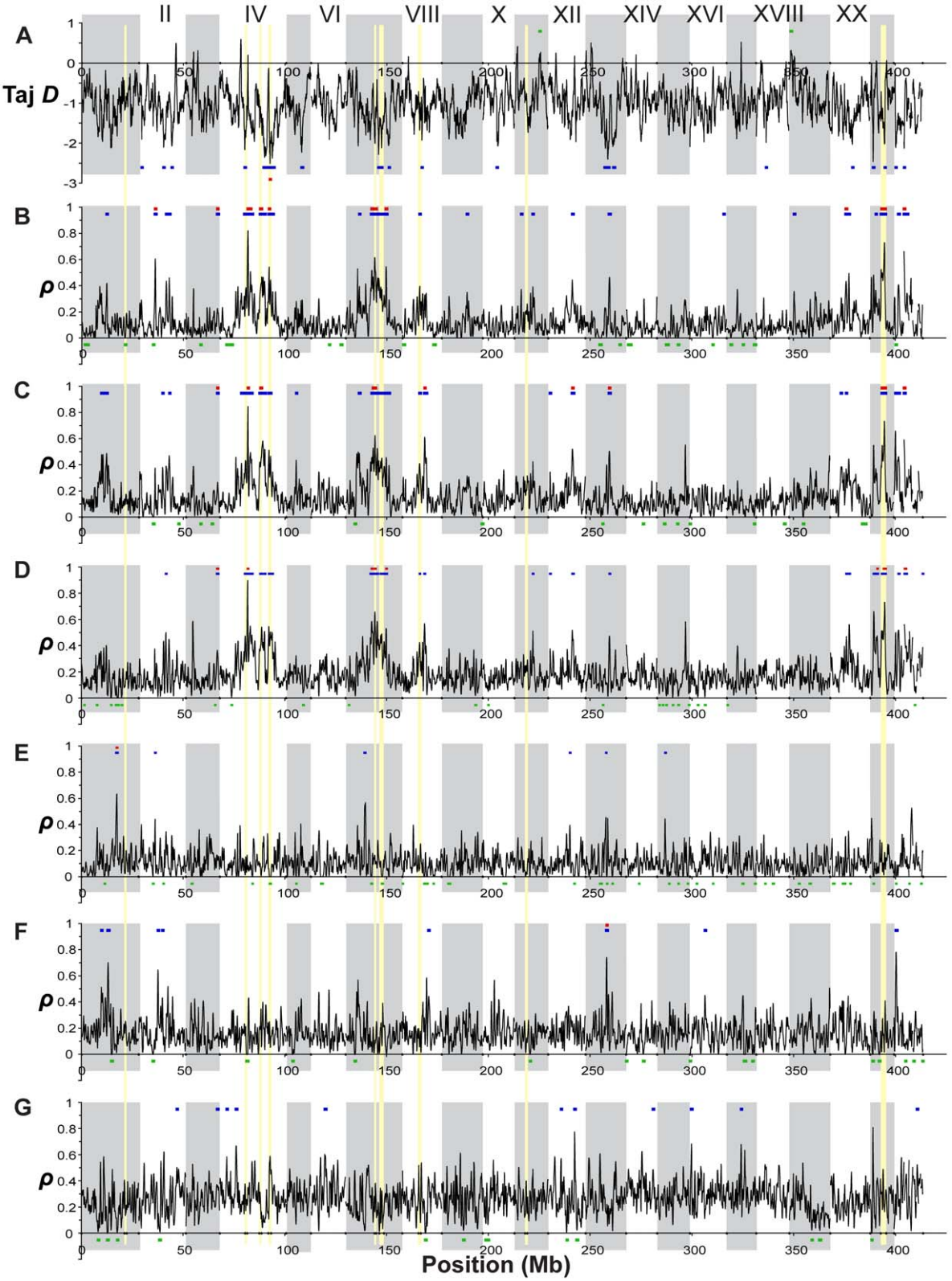

Figure 9. Genome-wide distributions of allele frequency spectrum and private allele density. (A) Tajima's *D*, a measure of allele frequency spectrum, within the combined oceanic population (RS and RB). Colored bars above and below the distribution indicate regions of significantly elevated ($p \leq 10^{-2}$, green) or reduced ($p \leq 10^{-2}$, blue; $p \leq 10^{-4}$, red) values, assessed by bootstrap resampling. (B–G) Private allele density (ρ) in single freshwater populations. Colored bars indicate regions of significantly elevated ($p \leq 10^{-3}$, blue; $p \leq 10^{-5}$, red) or reduced ($p \leq 10^{-3}$) 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 Ectodysplasin A (*Eda*), implicated in the loss of the lateral plate phenotype [93], 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 skeletogensis 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 zygosities 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_2 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]

Location	Gene	p-value	OD	BD	TO	CF	OS	KF	IG	References
LG 1: 1 Mb, 52 genes										
21,543,442	TNS1	$<10^{-7}$				Yes				[160]
21,583,240	IGFBP5	$<10^{-7}$	Yes	Yes		Yes				[160,161]
21,589,378	IGFBP2	$<10^{-7}$	Yes	Yes		Yes				[160,162]
LG IV Peak 1: 1 Mb, 43 genes										
12,800,220	EDA	$<10^{-7}$		Yes/T	Yes/T					[78,93,139,163]
12,904,952	FLT4	$<10^{-7}$	Yes	Yes						[164]
13,220,801	PDLIM7	2.6×10^{-5}	Yes	Yes						[165]
13,375,789	ANXA6	0.0043		Yes						[166]
LG IV Peak 2: 1.1 Mb, 31 genes										
19,899,773	WNT7B	$<10^{-7}$	Yes		Yes/T					[163,167]
19,916,813	FBLN1	$<10^{-7}$				Yes				[142]
LG IV Peak 3: 1.4 Mb, 55 genes										
23,792,283	LEMD3	$<10^{-7}$		Yes						[144]
23,839,219	PRL	0.0073					Yes/T			[148]
24,111,028	SCUBE1	1.2×10^{-6}			Yes	Yes				[138]
24,342,759	NFYB	0.0005				Yes/T				[137]
24,367,757	PODXL	0.0006						Yes		[168]
24,652,574	SLC26A3	$<10^{-7}$							Yes/T	[169,170]
24,662,013	SLC26A3	$<10^{-7}$							Yes/T	[169,170]
24,994,302	OSBPL8	10^{-5}	Yes							[171]
LG VII Peak 1: 0.8 Mb, 42 genes										
14,464,316	CAMKK1	2.1×10^{-6}	Yes							[172]
14,824,723	CA4	9×10^{-5}		Yes			Yes/T		Yes/T	[146,152]
LG VII Peak 2: 2.2 Mb, 143 genes										
16,871,846	HRH2	1.6×10^{-6}							Yes	[173]
17,113,900	AR	$<10^{-7}$		Yes						[174]
18,769,519	ADRB2	0.044		Yes						[115]
18,798,063	IL12B	0.044		Yes						[114]
LG VIII: 1.1 Mb, 50 genes										
8,049,501	LEPR	0.0012		Yes						[175]
8,625,098	ADAMTS10	$<10^{-7}$				Yes				[176]
LG XI: 1 Mb, 55 genes										
5,644,968	FZD2	$<10^{-7}$				Yes/T				[177]
5,736,635	STAT3	$<10^{-7}$		Yes						[178]
LG XXI: 2.6 Mb, 119 genes										
5,618,122	BMI1	1.1×10^{-6}	Yes							[179]
6,648,367	RDH10	$<10^{-7}$				Yes				[141]
6,826,891	EYA1	$<10^{-7}$			Yes	Yes/T		Yes		[140,180,181]
7,262,834	SGK3	$<10^{-7}$						Yes		[182]
7,305,661	CRH	$<10^{-7}$					Yes			[183]
7,519,692	FLT1	2×10^{-7}	Yes	Yes				Yes		[149–151,184,185]
7,575,402	LNX2	0.0017	Yes							[186]
7,736,424	ATP6V1A	0.076					Yes/T		Yes/T	[147]

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 F_{ST} 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 differentiation among populations. This pattern indicates balancing selection. This situation is best known for the vertebrate Major HistoCompatability (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 one 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 F_{ST} 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.

Genome scans are inherently comparative, and as with all correlative studies conclusions about adaptive evolution drawn from observed population genetic patterns should be accepted provisionally. These patterns provide support for signatures of selection, but are also the source of testable hypotheses for future studies. For example, although the clear expectation in genomic comparisons between ancestral and derived populations is that extreme values of the population genetic parameters we examined will be due to selection, combinations of non-selective processes may in some instances generate similar patterns. Variation across populations in mutation and recombination rates of homologous genomic regions may lead to a pattern similar to those that occur under selection. Although we do not expect this sort of variation in mutation or recombination to occur among these closely related stickleback populations, this hypothesis deserves exploration through subsequent comparative and manipulative studies. For example, the nature of the data we present here - SNP genotypes spread throughout the genome - does not allow the use of the full battery of molecular evolution tools developed recently for the analysis of sequence data [132]. However, regions that have been identified in our frequency-based genome scan can be the focus of subsequent re-sequencing research, or studies to test the association between the identified genomic region and fitness (e.g. [74]). Nonetheless, the particular stickleback system examined here–replicate, independently and recently derived freshwater populations that exhibit little neutral divergence from their extant ancestral stock–allows for uniquely strong inferences from comparative genomic data about the adaptive basis of parallel phenotypic evolution.

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 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 the 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 *Pitx1* 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 morphological 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]. Complementary 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 populations 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, NFYB, RDH10, 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 osteopoikilosis. These genes include LEMD3, LEPR, ARHGEF3 and RHOA (Table 3 and Table S3) [143–145].

Anadromous 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) [146–148]. 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. 88 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 *SbfI* (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/GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCAGAACAA3′; P2-2 bottom oligo 5′ CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT 3′) 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, 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).

Inferring 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

trimmed 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:

$$L(1/1) = P\left(n_1,n_2,n_3,n_4 \middle| 1/1\right) = \frac{n!}{n_1!n_2!n_3!n_4!}\left(1-\frac{3\varepsilon}{4}\right)^{n_1}\left(\frac{\varepsilon}{4}\right)^{n_2+n_3+n_4} \text{ and} \quad (1a)$$

$$L(1/2) = P\left(n_1,n_2,n_3,n_4 \middle| 1/2\right) = \frac{n!}{n_1!n_2!n_3!n_4!}\left(0.5-\frac{\varepsilon}{4}\right)^{n_1+n_2}\left(\frac{\varepsilon}{4}\right)^{n_3+n_4} \quad (1b)$$

where ε 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 ε 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 π (equivalent to expected heterozygosity) as

$$\pi = 1 - \sum_i \binom{n_i}{2} \Big/ \binom{n}{2} \quad (2)$$

where n_i is the count of allele i in the sample, and $n = \sum n_i$. Observed heterozygosity H was calculated as the proportion of diploid genotypes in the sample that are heterozygotes. To estimate differentiation among populations, we adapted a formula for F_{ST} from [155] that accounts for unequal sample sizes among populations by weighting:

$$F_{st} = 1 - \frac{\sum_j \binom{n_j}{2}\pi_j}{\pi_\bullet \sum_j \binom{n_j}{2}} \quad (3)$$

where n_j is the number of alleles sampled in population j, π_j is the nucleotide diversity within population j from equation (2), and $\pi_\bullet$ 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). Finally, 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 p to the region average was weighted by the Gaussian function $\exp\left(-(p-c)^2 \Big/ 2\sigma^2\right)$, where $\sigma = 150$ kb. For computational efficiency, we truncated this weighted average at 3σ in each direction (beyond which nucleotide sites have a relative weight less than ~0.01). We evaluated multiple choices for the width σ 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 6σ 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 π and heterozygosity H were weighted and averaged across all nucleotide sites; F_{ST} and private allele density ρ 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 π and number of SNPs within σ 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 π, *H*, F_{ST}, and ρ, as well as window values of Tajima's *D*, we used bootstrap resampling within each population comparison. For each nucleotide position (for π, or *H*) or SNP position (for F_{ST} or ρ) within each truncated Gaussian window described above, we randomly sampled with replacement from across the entire genome a value for the statistic (π, *H*, F_{ST}, or ρ) and the corresponding sample size (n_k). 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 π or *H*), 1,000 (for *D*), or 10,000 (for F_{ST} or ρ) replicates and stepped up to 1 million (π, *H*, or *D*) or 10 million (F_{ST} or ρ) 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 $p<10^{-5}$ for π in the combined 5-population dataset, including positions within 2σ (= 300 kb) of the outer positions. For F_{ST}, we identified all genomic regions for which $p<10^{-5}$ 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 includes 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., π 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 (π) across the genome, with colored bars indicating significantly elevated ($p \leq 10^{-5}$, blue) and reduced ($p \leq 10^{-5}$, 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 (2.85 MB TIF)

Figure S2 Private allele density in the overall freshwater-oceanic comparison. Each plot shows density of private alleles (ρ), with colored bars indicating regions of significantly elevated ($p \leq 10^{-3}$, blue; $p \leq 10^{-5}$, red) or reduced ($p \leq 10^{-3}$) 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 each gene's position to be its lower bound as given in the *Gasterosteus aculeatus* genome database (Ensembl, database version 56.1j, assembly Broad S1). Vertical gray shading indicates the 21 linkage groups and unassembled scaffolds.
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 ontology 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.
Found at: doi:10.1371/journal.pgen.1000862.s006 (0.11 MB DOC)

Acknowledgments

We thank A. Amores, R. Brown, J. Catchen, M. Currey, P. Phillips, and members of the Cresko, Phillips, and Thornton labs at the University of Oregon for assistance and helpful comments on this research. We also thank P. Petraitis for directing us to the kernel smoothing approach to sliding windows, and three anonymous reviewers for valuable comments on the manuscript.

Author Contributions

Conceived and designed the experiments: PAH SB EAJ WAC. Performed the experiments: PAH SB PDE. Analyzed the data: PAH SB NS WAC. Contributed reagents/materials/analysis tools: PAH SB PDE EAJ WAC. Wrote the paper: PAH SB WAC.

References

1. Fisher RA (1958) The Genetical Theory of Natural Selection. New York: Dover.
2. Wright S (1978) Evolution and the genetics of populations. Chicago: University of Chicago Press.
3. Kimura M, Ota T (1971) Theoretical aspects of population genetics. Monogr Popul Biol 4: 1–219.
4. Gillespie JH (1984) The status of the Neutral Theory: The Neutral Theory of Molecular Evolution. Science 224: 732–733.
5. Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10: 195–205.
6. Hurst LD (2009) Genetics and the understanding of selection. Nat Rev Genet 10: 83–93.
7. Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol 13: 969–980.
8. Bowcock AM (2007) Genomics: guilt by association. Nature 447: 645–646.
9. Bonin A (2008) Population genomics: a new generation of genome scans to bridge the gap with functional genomics. Mol Ecol 17: 3583–3584.
10. Brinkman FS, Parkhill J (2008) Population genomics: modeling the new and a renaissance of the old. Curr Opin Microbiol 11: 439–441.
11. Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9: 387–402.
12. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133–141.
13. Imelfort M, Duran C, Batley J, Edwards D (2009) Discovering genetic polymorphisms in next-generation sequencing data. Plant Biotechnol J 7: 312–317.
14. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4: 981–994.
15. Rockman MV, Hahn MW, Soranzo N, Goldstein DB, Wray GA (2003) Positive selection on a human-specific transcription factor binding site regulating IL4 expression. Curr Biol 13: 2118–2123.
16. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15: 1566–1575.
17. Flowers JM, Purugganan MD (2008) The evolution of plant genomes: scaling up from a population perspective. Curr Opin Genet Dev 18: 565–570.
18. Li YF, Costello JC, Holloway AK, Hahn MW (2008) "Reverse ecology" and the power of population genomics. Evolution 62: 2984–2994.
19. Flowers JM, Hanzawa Y, Hall MC, Moore RC, Purugganan MD (2009) Population genomics of the *Arabidopsis thaliana* flowering time gene network. Mol Biol Evol 26: 2475–2486.
20. Rokas A, Abbot P (2009) Harnessing genomics for evolutionary insights. Trends Ecol Evol 24: 192–200.
21. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38: 1358–1370.
22. Holsinger KE, Weir BS (2009) Genetics in geographically structure populations: defining, estimating and interpreting Fst. Nat Rev Genet 10: 639–650.
23. Baer CF (1999) Among-locus variation in Fst: fish, allozymes and the Lewontin-Krakauer test revisited. Genetics 152: 653–659.
24. Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol Ecol 14: 671–688.
25. Foll M, Beaumont MA, Gaggiotti O (2008) An approximate Bayesian computation approach to overcome biases that arise when using amplified fragment length polymorphism markers to study population structure. Genetics 179: 927–939.
26. Holloway AK, Lawniczak MK, Mezey JG, Begun DJ, Jones CD (2007) Adaptive gene expression divergence inferred from population genomics. PLoS Genet 3: e187. doi:10.1371/journal.pgen.0030187.
27. Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180: 977–993.
28. Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity 103: 285–298.
29. Pariset L, Joost S, Marsan PA, Valentini A (2009) Landscape genomics and biased Fst approaches reveal single nucleotide polymorphisms under selection in goat breeds of North-East Mediterranean. BMC Genet 10: 7.
30. Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175–195.
31. Schlotterer C (2003) Hitchhiking mapping–functional genomics from the population genetics perspective. Trends Genet 19: 32–38.
32. Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169: 2335–2352.
33. Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, et al. (2005) Ancient and recent positive selection transformed opioid *cis*-regulation in humans. PLoS Biol 3: e387. doi:10.1371/journal.pbio.0030387.
34. Pennings PS, Hermisson J (2006) Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet 2: e186. doi:10.1371/journal.pgen.0020186.
35. Pennings PS, Hermisson J (2006) Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol 23: 1076–1084.
36. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8: 857–868.
37. Thornton KR, Jensen JD, Becquet C, Andolfatto P (2007) Progress and prospects in mapping recent selection in the genome. Heredity 98: 340–348.
38. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, et al. (2008) Natural selection on genes that underlie human disease susceptibility. Curr Biol 18: 883–889.
39. Pavlidis P, Hutter S, Stephan W (2008) A population genomic approach to map recent positive selection in model species. Mol Ecol 17: 3585–3598.
40. Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100: 158–170.
41. Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19: 711–722.
42. Campbell D, Bernatchez L (2004) Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. Mol Biol Evol 21: 945–956.
43. Rogers SM, Bernatchez L (2005) Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). Mol Ecol 14: 351–361.
44. Rockman MV, Kruglyak L (2006) Genetics of global gene expression. Nat Rev Genet 7: 862–872.
45. Roberge C, Guderley H, Bernatchez L (2007) Genomewide identification of genes under directional selection: gene transcription Qst scan in diverging Atlantic salmon subpopulations. Genetics 177: 1011–1022.
46. Rockman MV (2008) Reverse engineering the genotype-phenotype map with natural genetic variation. Nature 456: 738–744.
47. Rockman MV, Kruglyak L (2009) Recombinational landscape and population genomics of *Caenorhabditis elegans*. PLoS Genet 5: e1000419. doi:10.1371/journal.pgen.1000419.
48. Wootton RJ (1976) The Biology of the Sticklebacks. New York: Academic Press.
49. Bell MA, Foster SA (1994) The Evolutionary Biology of the Threespine Stickleback. Oxford: Oxford University Press.
50. Cresko W, McGuigan K, Phillips P, Postlethwait J (2007) Studies of threespine stickleback developmental evolution: progress and promise. Genetica 129: 105–126.
51. Bell MA (1995) Intraspecific systematics of *Gasterosteus aculeatus* populations: implications for behavioral ecology. Behaviour 132: 15–16.
52. O'-Reilly P, Reimchen TE, Beech R, Strobeck C (1993) Mitochondrial DNA in *Gasterosteus* and Pleistocene glacial refugium on the Queen Charlotte Islands, British Columbia. Evolution 47: 678–684.
53. Orti G, Bell MA, Reimchen TE, Meyer A (1994) Global survey of mitochondrial DNA sequences in the threespine stickleback - Evidence for recent migrations. Evolution 48: 608–622.
54. Thompson CE, Taylor EB, McPhail JD (1997) Parallel evolution of lake-stream pairs of threespine sticklebacks (*Gasterosteus*) inferred from mitochondrial DNA variation. Evolution 51: 1955–1965.
55. Cresko WA (2000) The Ecology and Geography of Speciation: A Case Study Using an Adaptive Radiation of Threespine Stickleback in Alaska. Ph.D thesis.

56. Taylor EB, McPhail JD (2000) Historical contingency and ecological determinism interact to prime speciation in sticklebacks, *Gasterosteus*. Proc Roy Soc Lond B 267: 2375–2384.
57. Reusch TB, Wegner KM, Kalbe M (2001) Rapid genetic divergence in postglacial populations of threespine stickleback (*Gasterosteus aculeatus*): the role of habitat type, drainage and geographical proximity. Mol Ecol 10: 2435–2445.
58. Hendry AP, Taylor EB, McPhail JD (2002) Adaptive divergence and the balance between selection and gene flow: lake and stream stickleback in the misty system. Evolution 56: 1199–1216.
59. McKinnon JS, Mori S, Blackman BK, David L, Kingsley DM, et al. (2004) Evidence for ecology's role in speciation. Nature 429: 294–298.
60. Raeymaekers JA, Maes GE, Audenaert E, Volckaert FA (2005) Detecting Holocene divergence in the anadromous-freshwater three-spined stickleback (*Gasterosteus aculeatus*) system. Mol Ecol 14: 1001–1014.
61. Jones FC, Brown C, Pemberton JM, Braithwaite VA (2006) Reproductive isolation in a threespine stickleback hybrid zone. J Evol Biol 19: 1531–1544.
62. Makinen HS, Cano JM, Merila J (2006) Genetic relationships among marine and freshwater populations of the European three-spined stickleback (*Gasterosteus aculeatus*) revealed by microsatellites. Mol Ecol 15: 1519–1534.
63. Malhi RS, Rhett G, Bell AM (2006) Mitochondrial DNA evidence of an early Holocene population expansion of threespine sticklebacks from Scotland. Mol Phylogenet Evol 40: 148–154.
64. Raeymaekers JA, Van Houdt JK, Larmuseau MH, Geldof S, Volckaert FA (2007) Divergent selection as revealed by Pst and QTL-based Fst in three-spined stickleback (*Gasterosteus aculeatus*) populations along a coastal-inland gradient. Mol Ecol 16: 891–905.
65. Bell MA (1976) Evolution of phenotypic diversity in *Gasterosteus aculeatus* superspecies on the Pacific coast of North America. Syst Zool 25: 211–227.
66. Bell MA (1981) Lateral plate polymorphism and ontogeny of the complete plate morph of threespine sticklebacks (*Gasterosteus aculeatus*). Evolution 35: 67–74.
67. Schluter D (1993) Adaptive radiation in sticklebacks - size, shape, and habitat use efficiency. Ecology 74: 699–709.
68. Bell MA, Orti G (1994) Pelvic reduction in threespine stickleback from Cook Inlet lakes: geographical distribution and intrapopulation variation. Copeia. pp 314–325.
69. Schluter D (1995) Adaptive radiation in sticklebacks - trade-offs in feeding performance and growth. Ecology 76: 82–90.
70. Baker JA, Foster SA, Heins DC, Bell MA, King RW (1998) Variation in female life-history traits among Alaskan populations of the threespine stickleback, *Gasterosteus aculeatus* L. (Pisces: Gasterosteidae). Biol J Linn Soc 63: 141–159.
71. Foster SA (1995) Understanding the evolution of behavior in threespine stickleback: the value of geographic variation. Behaviour 132: 15–16.
72. Foster SA, Baker JA (1995) Evolutionary interplay between ecology, morphology and reproductive behavior in threespine stickleback, *Gasterosteus aculeatus*. Environ Biol Fish 44: 1–3.
73. Bell MA (2001) Lateral plate evolution in the threespine stickleback: getting nowhere fast. Genetica 112: 445–461.
74. Barrett RD, Rogers SM, Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. Science 322: 255–257.
75. McKinnon JS, Rundle HD (2002) Speciation in nature: the threespine stickleback model systems. Trends Ecol Evol 17: 480–488.
76. Walker JA (1997) Ecological morphology of lacustrine threespine stickleback *Gasterosteus aculeatus* L. (Gasterosteidae) body shape. Biol J Linn Soc 61: 3–50.
77. Walker JA, Bell MA (2000) Net evolutionary trajectories of body shape evolution within a microgeographic radiation of threespine sticklebacks (*Gasterosteus aculeatus*). J Zool 252: 293–302.
78. Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, et al. (2004) The genetic architecture of parallel armor plate reduction in threespine sticklebacks. PLoS Biol 2: e109. doi:10.1371/journal.pbio.0020109.
79. Cresko WA, Amores A, Wilson C, Murphy J, Currey M, et al. (2004) Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. Proc Natl Acad Sci U S A 101: 6050–6055.
80. Schluter D (2004) Parallel evolution and inheritance of quantitative traits. Am Nat 163: 809–822.
81. Cresko WA, Baker JA (1996) Two morphotypes of lacustrine threespine stickleback, *Gasterosteus aculeatus*, in Benka Lake, Alaska. Environ Biol Fish 45: 343–350.
82. von Hippel FA, Weigner H (2004) Sympatric anadromous-resident pairs of threespine stickleback species in young lakes and streams at Bering Glacier. Behaviour 141: 1441–1464.
83. Baker JA, Cresko WA, Foster SA, Heins DC (2005) Life-history differentiation of benthic and limnetic ecotypes in a polytypic population of threespine stickleback (*Gasterosteus aculeatus*). Evol Ecol Res 7: 121–131.
84. McPhail JD (1984) Ecology and evolution of sympatric sticklebacks (*Gasterosteus*): morphological and genetic evidence for a species pair in Enos Lake, British Columbia. Can J Zool 62: 1402–1408.
85. McPhail JD (1992) Ecology and evolution of sympatric sticklebacks (*Gasterosteus*): evidence for a species-pair in Paxton Lake, Texada Island, British Columbia. Can J Zool 70: 361–369.
86. Schluter D (1998) Ecological causes of speciation. In: Howard DJ, Berlocher SH, eds. Endless Forms: Species and Speciation. Oxford: Oxford University Press.
87. Schluter D (2000) The Ecology of Adaptive Radiations. Oxford: Oxford University Press.
88. Kitano J, Ross JA, Mori S, Kume M, Jones FC, et al. (2009) A role for a neo-sex chromosome in stickleback speciation. Nature 461: 1079–1083.
89. Bell MA, Aguirre WE, Buck NJ (2004) Twelve years of contemporary armor evolution in a threespine stickleback population. Evolution 58: 814–824.
90. Kingsley DM, Zhu BL, Osoegawa K, De Jong PJ, Schein J, et al. (2004) New genomic tools for molecular studies of evolutionary change in threespine sticklebacks. Behaviour 141: 1331–1344.
91. Kimmel CB, Ullmann B, Walker C, Wilson C, Currey M, et al. (2005) Evolution and development of facial bone morphology in threespine sticklebacks. Proc Natl Acad Sci U S A 102: 5791–5796.
92. Peichel CL, Nereng KS, Ohgi KA, Cole BL, Colosimo PF, et al. (2001) The genetic architecture of divergence between threespine stickleback species. Nature 414: 901–905.
93. Colosimo PF, Hosemann KE, Batabhadra S, Villarreal G Jr, Dickson M, et al. (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science 307: 1928–1933.
94. Peichel CL, Ross JA, Matson CK, Dickson M, Grimwood J, et al. (2004) The master sex-determination locus in threespine stickleback is on a nascent Y chromosome. Current Biology 14: 1416–1424.
95. Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, et al. (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428: 717–723.
96. Miller CT, Beleza S, Pollen AA, Schluter D, Kittles RA, et al. (2007) *cis*-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. Cell 131: 1179–1189.
97. Albert AY, Sawaya S, Vines TH, Knecht AK, Miller CT, et al. (2008) The genetics of adaptive shape shift in stickleback: pleiotropy and effect size. Evolution 62: 76–85.
98. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17: 240–248.
99. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3: e3376. doi:10.1371/journal.pone.0003376.
100. Beaumont MA (2005) Adaptation and speciation: what can Fst tell us? Trends Ecol Evol 20: 435–440.
101. Barrett RD, Schluter D (2008) Adaptation from standing genetic variation. Trends Ecol Evol 23: 38–44.
102. Tajima, F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
103. Makinen HS, Cano JM, Merila J (2008) Identifying footprints of directional and balancing selection in marine and freshwater three-spined stickleback (*Gasterosteus aculeatus*) populations. Mol Ecol 17: 3565–3582.
104. Makinen HS, Shikano T, Cano JM, Merila J (2008) Hitchhiking mapping reveals a candidate genomic region for natural selection in three-spined stickleback chromosome VIII. Genetics 178: 453–465.
105. Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, et al. (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol 5: e310. doi:10.1371/journal.pbio.0050310.
106. Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural populations. Nature 452: 169–175.
107. Andres AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, et al. (2009) Targets of balancing selection in the human genome. Mol Biol Evol 26: 2755–2764.
108. Woolhouse MEJ, Webster JP, Domingo E, Charlesworth B, Levin BR (2002) Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat Genet 32: 569–577.
109. Yokomizo T, Izumi T, Chang K, Takuwa Y, Shimizu T (1997) A G-protein-coupled receptor for leukotriene B-4 that mediates chemotaxis. Nature 387: 620–624.
110. Seymour RE, Hasham MG, Cox GA, Shultz LD, HogenEsch H, et al. (2007) Spontaneous mutations in the mouse Sharpin gene result in multiorgan inflammation, immune system dysregulation and dermatitis. Genes Immun 8: 416–421.
111. Litvak V, Ramsey SA, Rust AG, Zak DE, Kennedy KA, et al. (2009) Function of C/EBP delta in a regulatory circuit that discriminates between transient and persistent TLR4-induced signals. Nat Immun 10: 437–443.
112. Ozato K, Shin DM, Chang TH, Morse HCr (2008) TRIM family proteins and their emerging roles in innate immunity. Nat Rev Immunol 8: 849–860.
113. Newman RM, Hall L, Connole M, Chen GL, Sato S, et al. (2006) Balancing selection and the evolution of functional polymorphism in Old World monkey TRIM5alpha. Proc Natl Acad Sci U S A 103: 19134–19139.
114. Horwood NJ, Elliott J, Martin TJ, Gillespie MT (2001) IL-12 alone and in synergy with IL-18 inhibits osteoclast formation in vitro. J Immunol 166: 4915–4921.
115. Elefteriou F (2005) Neuronal signaling and the regulation of bone remodeling. Cell Mol Life Sci 62: 2339–2349.
116. Hesslein DGT, Fretz JA, Xi YG, Nelson T, Zhou SM, et al. (2009) Ebf1-dependent control of the osteoblast and adipocyte lineages. Bone 44: 537–546.
117. Laurie CC, Nickerson DA, Anderson AD, Weir BS, Livingston RJ, et al. (2007) Linkage disequilibrium in wild mice. PLoS Genet 3: e144. doi:10.1371/journal.pgen.0030144.
118. McVean G (2007) The structure of linkage disequilibrium around a selective sweep. Genetics 175: 1395–1406.

119. Drogemuller C, Karlsson EK, Hytonen MK, Perloski M, Dolf G, et al. (2008) A mutation in hairless dogs implicates FOXI3 in ectodermal development. Science 321: 1462–1462.
120. Meyer D, Thomson G (2001) How selection shapes variation of the human major histocompatibility complex: a review. Ann Hum Genet 65: 1–26.
121. Ferguson W, Dvora S, Gallo J, Orth A, Boissinot S (2008) Long-term balancing selection at the West Nile virus resistance gene, Oas1b, maintains transspecific polymorphisms in the house mouse. Mol Biol Evol 25: 1609–1618.
122. Yu X, Wang Z, Mertz JE (2007) ZEB1 Regulates the latent-lytic switch in infection by Epstein-Barr virus. PLoS Pathog 3: e194. doi:10.1371/journal.ppat.0030194.
123. Vanhollebeke B, Truc P, Poelvoorde P, Pays A, Joshi PP, et al. (2006) Brief report: Human *Trypanosoma evansi* infection linked to a lack of apolipoprotein L-I. New Eng J Med 355: 2752–2756.
124. Smith EE, Malik HS (2009) The apolipoprotein L family of programmed cell death and immunity genes rapidly evolved in primates at discrete sites of host-pathogen interactions. Genome Res 19: 850–858.
125. Wegner KM, Reusch TB, Kalbe M (2003) Multiple parasites are driving major histocompatibility complex polymorphism in the wild. J Evol Biol 16: 224–232.
126. van der Aa LM, Levraud JP, Yahmi M, Lauret E, Briolat V, et al. (2009) A large new subset of TRIM genes highly diversified by duplication and positive selection in teleost fish. BMC Biol 7: 7.
127. Carthagena L, Bergamaschi A, Luna JM, David A, Uchil PD, et al. (2009) Human TRIM gene expression in response to interferons. PLoS One 4: e4894. doi:10.1371/journal.pone.0004894.
128. Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? Genome Res 16: 702–712.
129. Boitard S, Schlotterer C, Futschik A (2009) Detecting selective sweeps: a new approach based on hidden Markov models. Genetics 181: 1567–1578.
130. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, et al. (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet 3: e90. doi:10.1371/journal.pgen.0030090.
131. Gaggiotti OE, Bekkevold D, Jorgensen HB, Foll M, Carvalho GR, et al. (2009) Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. Evolution 63: 2939–2951.
132. Jensen JD, Wong A, Aquadro CF (2007) Approaches for identifying targets of positive selection. Trends Genet 23: 568–577.
133. Cano JM, Matsuba C, Makinen H, Merila J (2006) The utility of QTL-Linked markers to detect selective sweeps in natural populations–a case study of the EDA gene and a linked marker in threespine stickleback. Mol Ecol 15: 4613–4621.
134. Shapiro MD, Bell MA, Kingsley DM (2006) Parallel genetic origins of pelvic reduction in vertebrates. Proc Natl Acad Sci U S A 103: 13753–13758.
135. Kimmel CB, Ullmann B, MC, Aguirre WE, Cresko WA (2008) Heterotopy explains opercular shape evolution in Alaskan threespine sticklebacks. Behaviour 145: 669–691.
136. Yamaguchi TP, Bradley A, McMahon AP, Jones S (1999) A Wnt5a pathway underlies outgrowth of multiple structures in the vertebrate embryo. Development 126: 1211–1223.
137. Chen YH, Lin YT, Lee GH (2009) Novel and unexpected functions of zebrafish CCAAT box binding transcription factor (NF-Y) B subunit during cartilages development. Bone 44: 777–784.
138. Xavier GM, Sharpe PT, Cobourne MT (2009) Scube1 is Expressed During Facial Development in the mouse. J Exp Zool B-Mol Dev Evol 312B: 518–524.
139. Srivastava AK, Pispa J, Hartung AJ, Du YZ, Ezer S, et al. (1997) The Tabby phenotype is caused by mutation in a mouse homologue of the EDA gene that reveals novel mouse and human exons and encodes a protein (ectodysplasin-A) with collagenous domains. Proc Natl Acad Sci U S A 94: 13069–13074.
140. Xu PX, Adams J, Peters H, Brown MC, Heaney S, et al. (1999) Eya1-deficient mice lack ears and kidneys and show abnormal apoptosis of organ primordia. Nat Genet 23: 113–117.
141. Sandell LL, Sanderson BW, Moiseyev G, Johnson T, Mushegian A, et al. (2007) RDH10 is essential for synthesis of embryonic retinoic acid and is required for limb, craniofacial, and organ development. Genes Dev 21: 1113–1124.
142. Cooley MA, Kern CB, Fresco VM, Wessels A, Thompson RP, et al. (2008) Fibulin-1 is required for morphogenesis of neural crest-derived structures. Dev Biol 319: 336–345.
143. Koh JM, Kim DJ, Hong JS, Park JY, Lee KU, et al. (2002) Estrogen receptor alpha gene polymorphisms Pvu II and Xba I influence association between leptin receptor gene polymorphism (Gln223Arg) and bone mineral density in young men. Eur J Endocrinol 147: 777–783.
144. Hellemans J, Preobrazhenska O, Willaert A, Debeer P, Verdonk PCM, et al. (2004) Loss-of-function mutations in LEMD3 result in osteopoikilosis, Buschke-Ollendorff syndrome and melorheostosis. Nat Genet 36: 1213–1218.
145. Mullin BH, Prince RL, Mamotte C, Spector TD, Hart DJ, et al. (2009) Further genetic evidence suggesting a role for the RhoGTPase-RhoGEF pathway in osteoporosis. Bone 45: 387–391.
146. Grosell M, Gilmour KM, Perry SF (2007) Intestinal carbonic anhydrase, bicarbonate, and proton carriers play a role in the acclimation of rainbow trout to seawater. Am J Physiol Regul Integr Comp 293: R2099–R2111.
147. Horng JL, Lin LY, Huang CJ, Katoh F, Kaneko T, et al. (2007) Knockdown of V-ATPase subunit A (atp6v1a) impairs acid secretion and ion balance in zebrafish (*Danio rerio*). Am J Physiol Regul Integr Comp 292: R2068–76.
148. Tomy S, Chang YM, Chen YH, Cao JC, Wang TP, et al. (2009) Salinity effects on the expression of osmoregulatory genes in the euryhaline black porgy Acanthopagrus schlegeli. General Comp Endocrinol 161: 123–132.
149. Mayr-Wohlfart U, Waltenberger J, Hausser H, Kessler S, Günther KP, et al. (2002) Vascular endothelial growth factor stimulates chemotactic migration of primary human osteoblasts. Bone 30: 472–477.
150. Chen JR, Chatterjee B, Meyer R, Yu JC, Borke JL, et al. (2004) Tbx2 represses expression of Connexin43 in osteoblastic-like cells. Calc Tiss Internat 74: 561–573.
151. Chen S, Kasama Y, Lee JS, Jim B, Marin M, et al. (2004) Podocyte-derived vascular endothelial growth factor mediates the stimulation of alpha 3(IV) collagen production by transforming growth factor-beta 1 in mouse podocytes. Diabetes 53: 2939–2949.
152. Riihonen R, Supuran CT, Parkkila S, Pastorekova S, Vaananen HK, et al. (2007) Membrane-bound carbonic anhydrases in osteoclasts. Bone 40: 1021–1031.
153. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.
154. Lynch M (2009) Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics 182: 295–301.
155. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, et al. (2009) Darwinian and demographic forces affecting human protein coding genes. Genome Res 19: 838–849.
156. Hunter JE, Schmidt FL (1990) Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. Newbury Park, CA: Sage Publications.
157. Goudet J, Raymond M, de Meeüs T, Rousset F (1996) Testing differentiation in diploid populations. Genetics 144: 1933–1940.
158. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B 57: 289–300.
159. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Annals Stat 29: 1165–1188.
160. Beaty TH, Hetmanski JB, Fallin MD, Park JW, Sull JW, et al. (2006) Analysis of candidate genes on chromosome 2 in oral cleft case-parent trios from three populations. Hum Genet 120: 501–518.
161. Richman C, Baylink DJ, Lang K, Dony C, Mohan S (1999) Recombinant human insulin-like growth factor-binding protein-5 stimulates bone formation parameters *in vitro* and *in vivo*. Endocrinology 140: 4699–4705.
162. Amin S, Riggs BL, Melton L Jr, Achenbach SJ, Atkinson EJ, et al. (2007) High serum IGFBP-2 is predictive of increased bone turnover in aging men and women. J Bone Miner Res 22: 799–807.
163. Fraser GJ, Bloomquist RF, Streelman JT (2008) A periodic pattern generator for dental diversity. BMC Biol 6.
164. Orlandini M, Spreafico A, Bardelli M, Rocchigiani M, Salameh A, et al. (2006) Vascular endothelial growth factor-D activates VEGFR-3 expressed in osteoblasts inducing their differentiation. J Biol Chem 281: 17961–17967.
165. Boden SD, Liu YS, Hair GA, Helms JA, Hu D, et al. (1998) LMP-1, a LIM-domain protein, mediates BMP-6 effects on bone formation. Endocrinology 139: 5125–5134.
166. Thouverey C, Strzelecka-Kiliszek A, Balcerzak M, Buchet R, Pikula S (2009) Matrix Vesicles Originate From Apical Membrane Microvilli of Mineralizing Osteoblast-Like Saos-2 Cells. J Cell Biochem 106: 127–138.
167. Hu HL, Hilton MJ, Tu XL, Yu K, Ornitz DM, et al. (2005) Sequential roles of Hedgehog and Wnt signaling in osteoblast development. Development 132: 49–60.
168. Doyonnas R, Kershaw DB, Duhme C, Merkens H, Chelliah S, et al. (2001) Anuria, omphalocele, and perinatal lethality in mice lacking the CD34-related protein podocalyxin. J Exp Med 194: 13–27.
169. Schweinfest CW, Spyropoulos DD, Henderson KW, Kim JH, Chapman JM, et al. (2006) slc26a3 (dra)-deficient mice display chloride-losing diarrhea, enhanced colonic proliferation, and distinct up-regulation of ion transporters in the colon. J Biol Chem 281: 37962–37971.
170. Perry SF, Vulesevic B, Grosell M, Bayaa M (2009) Evidence that SLC26 anion transporters mediate branchial chloride uptake in adult zebrafish (*Danio rerio*). Am J Physiol Regul Integr Comp 297: R988–97.
171. Barski A, Pregizer S, Frenkel B (2008) Identification of transcription factor target genes by ChIP display. Meth Mol Biol 455: 177–190.
172. Li X, Wang H, Touma E, Qi Y, Rousseau E, et al. (2007) TP508 accelerates fracture repair by promoting cell growth over cell death. Biochem Biophys Res Comm 364: 187–193.
173. Schultheiss G, Hennig B, Schunack W, Prinz G, Diener M (2006) Histamine-induced ion secretion across rat distal colon: Involvement of histamine H-1 and H-2 receptors. Eur J Pharm 546: 161–170.
174. Wiren KM, Sernirale AA, Zhang XW, Woo A, Tornmasini SM, et al. (2008) Targeting of androgen receptor in bone reveals a lack of androgen anabolic action and inhibition of osteogenesis - a model for compartment-specific androgen action in the skeleton. Bone 43: 440–451.
175. Shi Y, Yadav VK, Suda N, Liu XS, Guo XE, et al. (2008) Dissociation of the neuronal regulation of bone mass and energy metabolism by leptin in vivo. Proc Natl Acad Sci U S A 105: 20529–20533.

176. Dagoneau N, Benoist-Lasselin C, Huber C, Faivre L, Megarbane A, et al. (2004) ADAMTS10 mutations in autosomal recessive Weill-Marchesani syndrome. Am J Hum Genet 75: 801–806.
177. Sisson BE, Topczewski J (2009) Expression of five frizzleds during zebrafish craniofacial development. Gene Expr Patterns 9: 520–527.
178. Itoh S, Udagawa N, Takahashi N, Yoshitake F, Narita H, et al. (2006) A critical role for interleukin-6 family-mediated Stat3 activation in osteoblast differentiation and bone formation. Bone 39: 505–512.
179. Oguro H, Iwama A, Morita Y, Kamijo T, van Lohuizen M, et al. (2006) Differential impact of Ink4a and Arf on hematopoietic stem cells and their bone marrow microenvironment in Bmi1-deficient mice. J Exp Med 203: 2247–2253.
180. Xu PX, Woo I, Her H, Beier DR, Maas RL (1997) Mouse Eya homologues of the *Drosophila* eyes absent gene require Pax6 for expression in lens and nasal placode. Development 124: 219–231.
181. Kozlowski DJ, Whitfield TT, Hukriede NA, Lam WK, Weinberg ES (2005) The zebrafish dog-eared mutation disrupts eya1, a gene required for cell survival and differentiation in the inner ear and lateral line. Dev Biol 277: 27–41.
182. Friedrich B, Feng Y, Cohen P, Risler T, Vandewalle A, et al. (2003) The serine/threonine kinases SGK2 and SGK3 are potent stimulators of the epithelial Na+ channel alpha, beta, gamma-ENaC. Pfl Archiv-Europ J Phys 445: 693–696.
183. Aguilera G (1998) Corticotropin releasing hormone, receptor regulation and the stress response. Trends Endocrinol Metab 9: 329–336.
184. Tufro A, Norwood VF, Carey RM, Gomez RA (1999) Vascular endothelial growth factor induces nephrogenesis and vasculogenesis. J Am Soc Neph 10: 2125–2134.
185. Otomo H, Sakai A, Uchida S, Tanaka S, Watanuki M, et al. (2007) Flt-1 tyrosine kinase-deficient homozygous mice result in decreased trabecular bone volume with reduced osteogenic potential. Bone 40: 1494–1501.
186. Pregizer S, Barski A, Frenkel B (2006) Identification of novel Runx2 targets in osteoblasts. J Bone Min Res 21: S383–S383.