Current Biology, Volume *22*

Supplemental Information

A Genome-wide SNP Genotyping Array Reveals Patterns of Global and Repeated Species-Pair Divergence in Sticklebacks

Felicity C. Jones, Yingguang Frank Chan, Jeremy Schmutz, Jane Grimwood, Shannon D. Brady, Audrey M. Southwick, Devin M. Absher, Richard M. Myers, Thomas E. Reimchen, Bruce E. Deagle, Dolph Schluter, and David M. Kingsley

Supplemental Inventory

Table S1 *(Related to Figure 1).* Details of Populations Used in the Present Study

Population Number (Figures 1, 4, and S3)	Code	Population Name	Ecotype	Basin	Geographic Region	GPS N	GPS E	Sample Size
1	JAMA	Japanese Pacific Marine	Marine	Pacific	Japanese Pacific	42.973	144.329	6
2	RABS	Rabbit Slough	Marine	Pacific	Alaska	61.556	149.249	6
3	LITC	Little Campbell River	Marine	Pacific	British Columbia	49.016	-122.779	6
4	BIGD	Big River Downstream	Marine	Pacific	California	39.289	-123.747	6
5	NAKA	Nakagawa Creek, Japan	Freshwater	Pacific	Japan	35.092	137.019	6
6	HUMP	Hump Lake	Freshwater	Pacific	Alaska	60.769	-151.170	6
7	MUDL	Mud Lake	Freshwater	Pacific	Alaska	61.593	-149.340	6
8	DUNS	Daniels Lake	Freshwater	Pacific	Alaska	60.735	-151.186	6
9	BOOT	Boot Lake	Freshwater	Pacific	Alaska	61.718	-150.13	6
10	ROUG	Rouge Lake	Freshwater	Pacific	British Columbia (Haida Gwaii)	54.033	-131.876	6
11	BOUL	Boulton Lake	Freshwater	Pacific	British Columbia (Haida Gwaii)	53.782	-132.097	6
12	DRIZ	Drizzle Lake	Freshwater	Pacific	British Columbia (Haida Gwaii)	53.935	-132.073	5
13	MAYR	Mayer Lake	Freshwater	Pacific	British Columbia (Haida Gwaii)	53.692	-132.044	6
14	HUTU	Humptulips	Freshwater	Pacific	Washington	47.231	-123.957	6
15	FTC	Fish Trap Creek	Freshwater	Pacific	Washington	48.931	-122.487	6
16	BIGU	Big River Upstream	Freshwater	Pacific	California	39.317	-123.686	6
17	MATA	Matadero Creek	Freshwater	Pacific	California	37.393	-122.162	6
18	WMSO	Santa Clara River	Freshwater	Pacific	California	34.435	-118.198	6
19	ANTL	Antigonish Landing	Marine	Atlantic	Nova Scotia	45.698	-61.878	5
20	BITJ	Bitrufjörður	Marine	Atlantic	Iceland	65.457	-21.440	5
21	RSHR	River Shiel Rockpools	Marine	Atlantic	Scotland	56.771	-5.831	4

22	JMRP	River Tyne Site 1	Marine	Atlantic	Scotland	55.999	-2.520	6
23	NEU	Neustadt	Marine	Atlantic	Germany	54.058	10.877	6
24	URRI	Urriðakotsvatn	Freshwater	Atlantic	Iceland	64.067	-21.914	6
25	SHEL	River Shiel	Freshwater	Atlantic	Scotland	56.748	-5.698	4
26	ABW	River Tyne Site 8	Freshwater	Atlantic	Scotland	55.943	-2.785	6
27	NOST	Norway Stream	Freshwater	Atlantic	Norway	59.765	5.712	6
28	SCX	Schwalle River	Freshwater	Atlantic	Germany	54.080	10.083	6
29	PAXB	Paxton Lake, Benthic	Freshwater	Pacific	British Columbia	49.703	-124.522	6
30	PRIB	Priest Lake, Benthic	Freshwater	Pacific	British Columbia	49.745	-124.566	6
31	QRYB	Little Quarry Lake, Benthic	Freshwater	Pacific	British Columbia	49.663	-124.110	6
32	PAXL	Paxton Lake, Limnetic	Freshwater	Pacific	British Columbia	49.703	-124.522	6
33	PRIL	Priest Lake, Limnetic	Freshwater	Pacific	British Columbia	49.745	-124.566	6
34	QRYL	Little Quarry Lake, Limnetic	Freshwater	Pacific	British Columbia	49.663	-124.110	5

Table S2 *(Related to Figure 1; see separate Excel file).* Details of the 1,159 SNPs used in this study. Further information including the flanking sequence of the SNP can be found by searching NCBI dbSNP with the appropriate submitted SNP (ss) or reference SNP (rs) identifier. SNPs were selected for the array to improve the assembly by tying in unlinked scaffolds, to obtain even coverage relative to local recombination rate, or to target genes regions of interest (*SNP Groups 0, 1 and 2 respectively).

Table S3 *(Related to Figure 2; see separate Excel file).* Summary of F_{ST} outlier SNPs identified in marine-freshwater BayEnv analysis of global populations (Bayes Factor Global BF_G >=1.5); Atlantic populations (Bayes Factor Atlantic BF_A >= 2.5); and Pacific populations (Bayes Factor Pacific BF_P >=2.5) and the Ensembl predicted genes within ±100 kb of SNPs.

Table S4 *(Related to Figures 3 and 4).*

SNP Name (Chromosome: Position)	Outlier in Number of Lakes	Paxton Lake $\log_{10}$(PO)	Priest Lake $\log_{10}$(PO)	Quarry Lake $\log_{10}$(PO)
chrX:14456479	3	1.7568	1.3187	1.9955
chrX:14549101	3	1.6813	1.2451	2.0227
chrUn:2776586	3	1.762	1.2643	0.93892
chrXIX:10552047	3	1.6018	1.2678	1.7888
chrII:14991358	2	0.55873	1.2643	~
chrXX:8905625	2	0.59762	1.252	~
chrXII:13045611	2	1.727	0.7233	~
chrIV:11367975	2	1.7175	1.266	~
chrXI:9039275	2	1.7779	1.2121	~
chrUn:2632376	2	1.7779	1.2916	~
chrVII:17995892	2	1.8114	1.2316	~
chrVII:18353106	2	~	1.2678	1.9082
chrI:7955458	2	~	1.2417	0.90172
chrIV:15721538	2	~	1.2842	1.7888
chrIV:15737291	2	~	1.339	2.1044
chrVII:13373948	1	0.58214	~	~
chrXXI:1893294	1	0.61455	~	~
chrVII:17369940	1	1.7128	~	~
chrUn:7381868	1	1.7417	~	~
chrXXI:7904439	1	1.7222	~	~
chrXIX:11573473	1	1.8292	~	~
chrII:418094 *	1	~	1.2714	~
chrIV:21232476	1	~	1.2251	~
chrXI:5715882 *	1	~	1.2153	~
chrUn:2474754	1	~	1.2607	~
chrXX:12436776	1	~	1.2916	~
chrI:15145305	1	~	1.2234	~
chrXI:7370453	1	~	1.2954	~
chrI:14261764	1	~	1.2503	~
chrXXI:7544041	1	~	1.275	~
chrI:7955618	1	~	1.3128	~
chrII:14611516	1	~	1.3011	~
chrI:11963492	1	~	1.2572	~
chrXVIII:5765162	1	~	1.3474	~
chrXX:12810044	1	~	1.3108	~
chrXVIII:4836241	1	~	1.2805	~

chrXI:1680578		1	~	1.266	~
chrXIX:3309372		1	~	1.2625	~
chrIII:13911180		1	~	1.2898	~
chrVIII:8450169		1	~	1.3088	~
chrVII:13205977		1	~	~	1.7082
chrXII:14353450		1	~	~	1.7725
chrVII:13525838		1	~	~	2.072
chrIV:23937349	*	1	~	~	2.0617
chrUn:498491		1	~	~	2.1518
chrVII:13452516		1	~	~	2.1775

Summary of F_{ST} outlier SNPs identified in BayeScan genome scans of lakes containing sympatric benthic and limnetic ecotypes. SNPs passing a false discovery rate FDR threshold of 0.05 for each lake analysis are shown with $\log_{10}$ Posterior Odds. ~ indicates SNPs that were not significant outliers for a given lake. * indicates SNPs that also showed Bayes Factor scores >1.5 in the global analysis of SNPs correlated with marine-freshwater environments.

Table S5 *(Related to Figure 4; see separate Excel file).* 15 SNPs identified as BayeScan F_{ST} outliers in two or more lakes containing benthic and limnetic ecotypes and Ensembl predicted genes within ±100 kb.

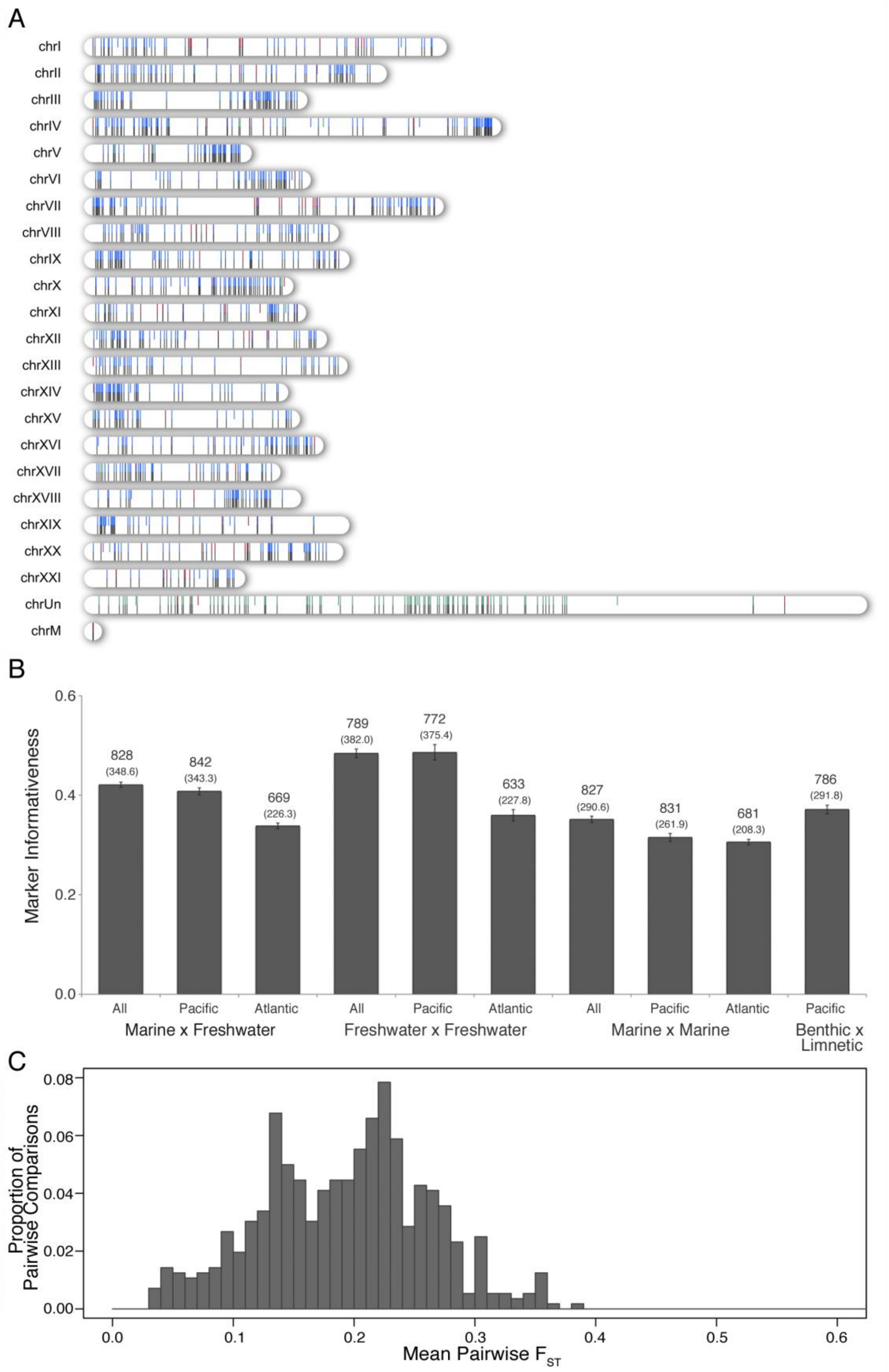
A
chrI
chrII
chrIII
chrIV
chrV
chrVI
chrVII
chrVIII
chrIX
chrX
chrXI
chrXII
chrXIII
chrXIV
chrXV
chrXVI
chrXVII
chrXVIII
chrXIX
chrXX
chrXXI
chrUn
chrM
B
Marker Informativeness
0.6
0.4
0.2
0.0
828
(348.6)
842
(343.3)
669
(226.3)
789
(382.0)
772
(375.4)
633
(227.8)
827
(290.6)
831
(261.9)
681
(208.3)
786
(291.8)
All
Pacific
Atlantic
All
Pacific
Atlantic
All
Pacific
Atlantic
Pacific
Marine x Freshwater
Freshwater x Freshwater
Marine x Marine
Benthic x Limnetic
C
Proportion of Pairwise Comparisons
0.08
0.06
0.04
0.02
0.00
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Mean Pairwise F_{ST}

Figure S1.

Figure S1 *(Related to Figure 1).*

(A) Genome graph showing the distribution of SNPs across the genome. 3,072 candidate SNPs are shown as tick marks on the top half of chromosome bars (blue – SNPs spaced evenly relative to recombination rate; red – SNPs chosen to tag genes of interest; green – SNPs chosen to improve assembly). 1,159 SNPs that passed filtering and were used in global population analyses are shown in black on the bottom half of the chromosome bars.

(B) Estimated marker informativeness for an F2 intercross between wild caught sticklebacks genotyped at 1,159 cleaned filtered SNPs [25]. Y-axis represents the proportion of alleles that would be fully informative allowing assignment of the grandparental line of origin. Above each bar, numbers in bold show the estimated total number of polymorphic markers, with numbers in parentheses showing the proportion that are estimated to be fully informative.

(C) The distribution of average genome-wide genetic distance (F_{ST}) for all pairwise comparisons between global populations.

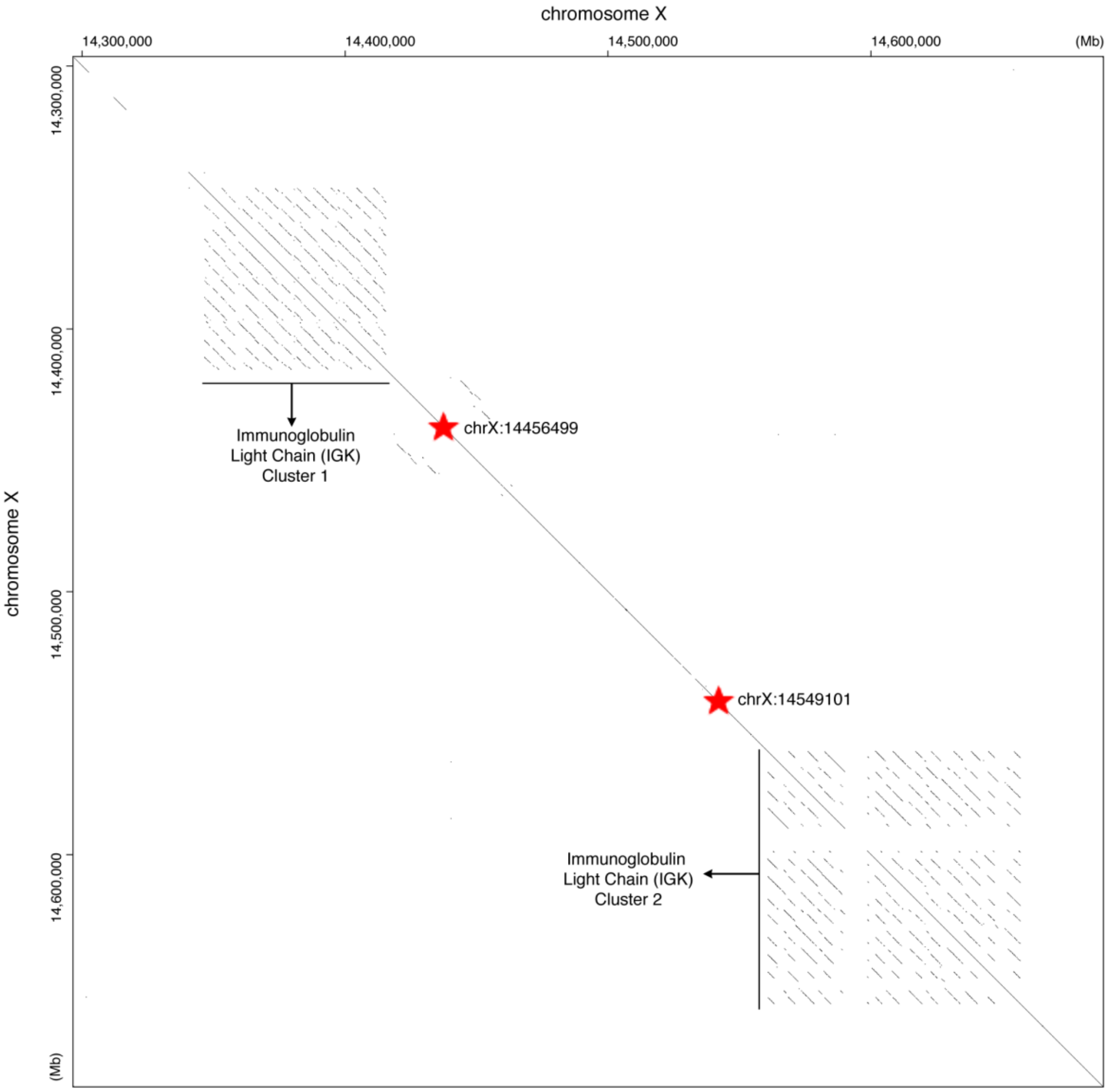

Figure S2 *(Related to Figure 3).*

A self dot-plot showing sequence similarity of a 390kb region on chromosome X that spans 2 outlier SNPs (red stars) found to have fixed differences in allele frequency between benthic and limnetic ecotypes in all three lakes. Flanking the SNPs, large stretches of repeats are evident. These correspond to two gene clusters containing a total of 24 immunoglobulin light chain (*IGK*) genes.

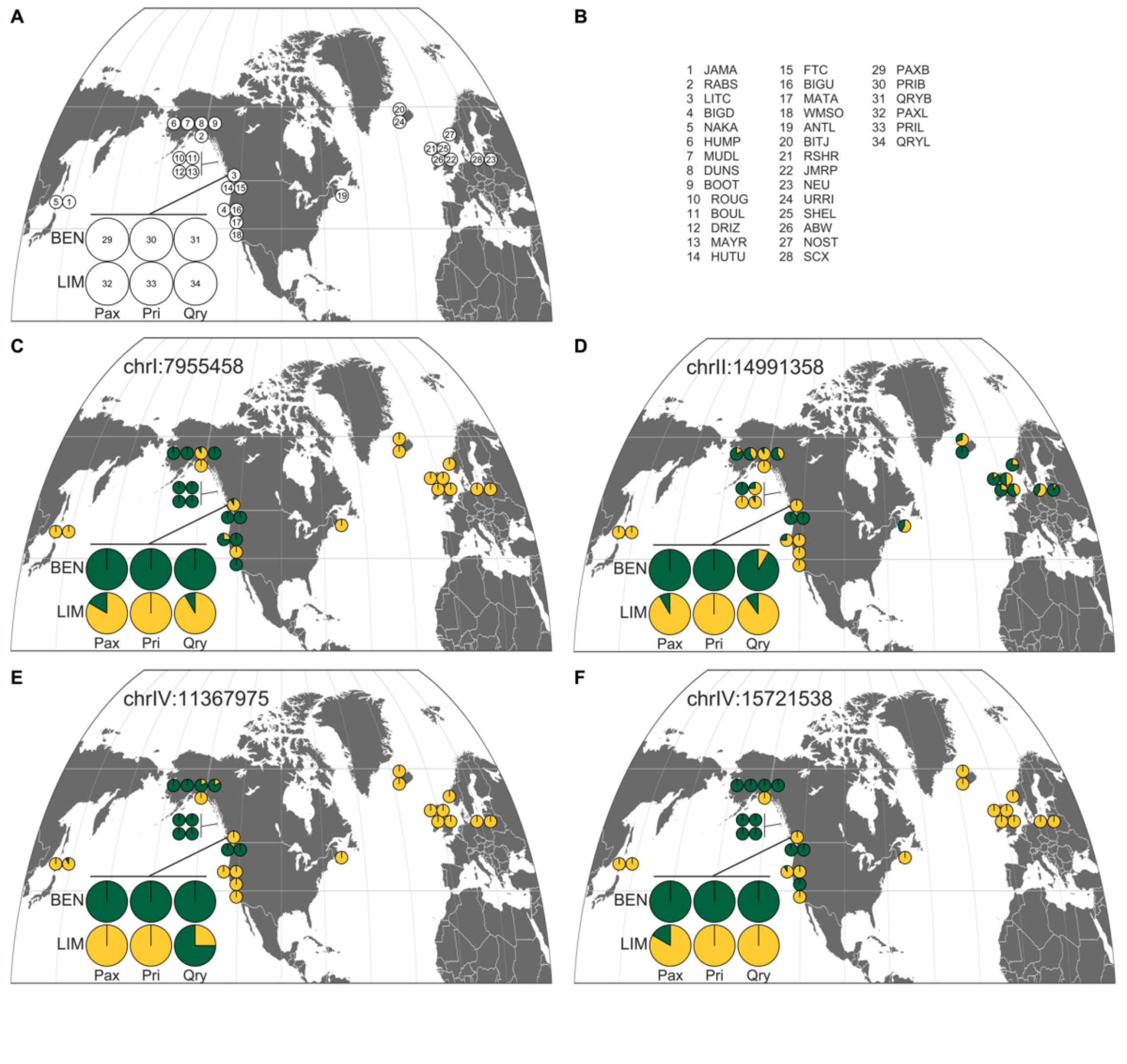
A
B
1 JAMA
2 RABS
3 LITC
4 BIGD
5 NAKA
6 HUMP
7 MUDL
8 DUNS
9 BOOT
10 ROUG
11 BOUL
12 DRIZ
13 MAYR
14 HUTU
15 FTC
16 BIGU
17 MATA
18 WMSO
19 ANTL
20 BITJ
21 RSHR
22 JMRP
23 NEU
24 URRI
25 SHEL
26 ABW
27 NOST
28 SCX
29 PAXB
30 PRIB
31 QRYB
32 PAXL
33 PRIL
34 QRYL
BEN
LIM
Pax
Pri
Qry
C
chrI:7955458
D
chrII:14991358
E
chrIV:11367975
F
chrIV:15721538

Figure S3.

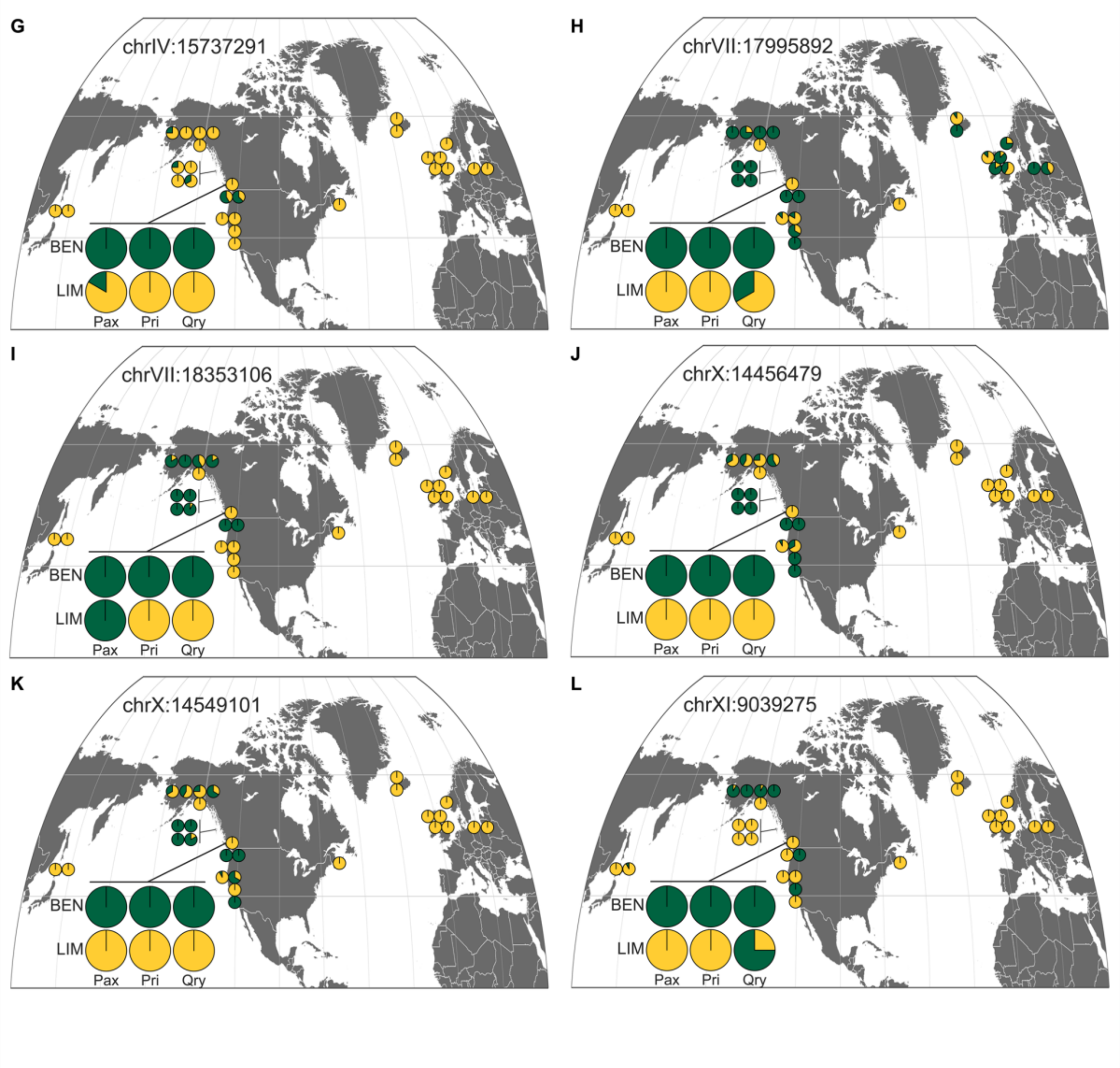

Figure S3 *(con't).*

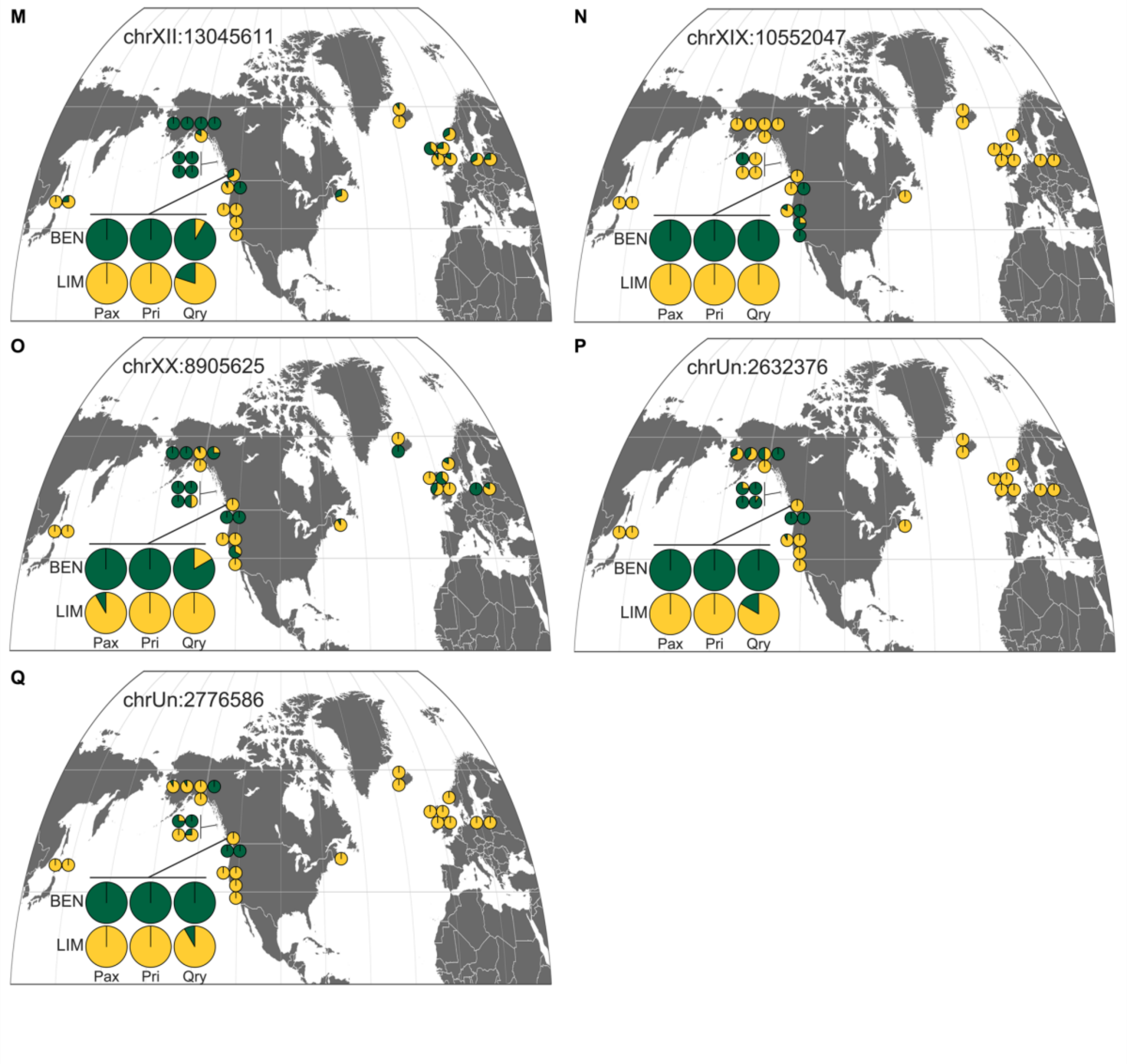

Figure S3 *(Related to Figure 4).*

Maps showing the global distribution of alleles at SNPs linked to putatively adaptive loci in benthic and limnetic fish from two or more lakes. Although adaptive alleles show a heterogeneous distribution throughout the world, allopatric freshwater populations containing the complete (or almost complete) set of benthic alleles are evident (Pops 14 and 15). Similarly, marine populations with the complete set of limnetic alleles can be seen (1, 2 and 3). This data argues for a more important role of allopatric adaptive divergence and reuse of standing genetic variation than perhaps previously recognized. Geographic location of populations and their corresponding number and letter codes are provided in (A) and (B). See also Table S1 for more population information.

Supplemental Experimental Procedures

SNP Discovery and Array Design

SNPs were ascertained by BLAT alignment [51] of 336,730 Sanger sequenced clones from EST libraries made from 4 different populations (2 marine: BDGB-California, BITJ-Iceland; 2 freshwater: COND-Washington, SALS-California; see [52] for more information) against the reference genome (Broad, gasAcu1.0 [5]). After removing 19,169 sequences that had multiple BLAT hits, 1,575,891 putative SNPs were identified by sequence comparisons among EST sequences and the reference genome. To enrich for high-quality putative SNPs, filtering was performed to remove: 1,173,073 SNPs with <2 read support for the alternate allele, 52,452 SNPs where the alternate allele was found solely in the reference genome, and 313,588 SNPs with other putative polymorphisms located within +/- 40bp of the focal SNP. This resulted in a set of 36,778 putative SNPs widely distributed throughout the genome, of which 2,664 SNPs were chosen for testing on Illumina custom genotyping arrays. A further 353 SNPs were ascertained from Illumina short-read whole genome sequences of one marine and one freshwater fish, and 55 from Sanger sequencing of BAC clones [11, 13, 44] or amplicons from a mixture of marine or freshwater fish, bringing the total to 3,072 SNPs on the genotyping platform (see Table S2 for more details). 2,479 of these SNPs were selected at approximately 0.6cM intervals, where the local genomic recombination rate was estimated by LOESS regression of genetic position (centimorgan) against the physical assembly (megabase). Three hundred and ninety-nine SNPs were added to tag previously un-oriented or unlinked scaffolds in the genome assembly, making a total of 2,878 total SNPs chosen based on overall genome position ("general class"). A further 194 SNPs were chosen in or near (within average 4.4kb, median 0kb) regions of biological interest ("candidate class") (see also Figure S1 and Table S2). The allele frequencies of general and candidate class SNPs within and among global stickleback populations were not known prior to the study.

SNP Genotyping and Data Cleaning

SNP genotyping was performed on 250ng of genomic DNA prepared from caudal fin clips of 196 individuals from 34 global stickleback populations (Table S1), using Illumina GoldenGate 1,536 custom SNP arrays (Illumina) according to manufacturer's protocols. Genotype calls were made using BeadStudio v3.0 software, with results from each SNP individually inspected to verify genotype clusters. Genotype data for the 196 individuals was filtered by excluding 1,913 SNPs that showed either fewer or more than the expected proportion in the three genotype clusters (AA, AB, BB); had more than 10% missing data; more than 50% heterozygosity; or a minor allele frequency less than 3 percent. Two individual fish samples with more than 10% missing data were excluded, leaving a data set comprising of 1,159 SNPs in 194 individuals. All 1,159 successful SNP positions have been deposited in Genbank's dbSNP database (accession numbers: Table S2).

Analysis of Polymorphism Information Content for Genetic Mapping

The polymorphism information content statistic reflecting marker informativeness for genetic mapping in outbred F2 cross designs was calculated for marine x marine, freshwater x freshwater and marine x freshwater crosses within and among Atlantic and Pacific fish following Rocha et al [53], and ranged from 0.31 to 0.49 (Figure S1B). This indicates that for any given F2 cross between wild sticklebacks a large proportion of markers will be informative, and will have alleles that can be unambiguously assigned to the grandparental line of origin. Based on estimates of marker informativeness, the density of polymorphic markers in a typical cross is estimated to be 1 SNP per 550-730kb (1.6-2.2 centimorgans), matching or improving the density in most previously published mapping studies based on microsatellites. At this marker density, the number of F2 individuals in a cross (i.e. the number of recombination events in a given clutch size) is now likely to be the limiting factor for many laboratory genetic mapping experiments in sticklebacks.

Data Analysis

Analyses of allele frequency, heterozygosity (2pq) and F_{ST} [50] were performed using Perl scripts. Principal components analysis was performed in R (v2.11.1) using the function '*prcomp*' from the base '*stats*' package. Any missing data was replaced with a genotype with probability determined by the frequencies of genotypes in all 194 individuals. Principal components analysis of global populations was performed using the set of 1,159 SNPs, while benthic-limnetic principal components analyses were performed on 925, 879 and 46 SNPs for each of Figure 4A, 4B, and 4C respectively. Biases in the populations from which SNPs are ascertained have the potential to skew/emphasize axes of genetic variation among populations/samples, an issue common to array-based SNP genotyping platforms in any organism. We explored the extent to which the major PC axes were likely to reflect SNP ascertainment by calculating an 'ascertainment vector' for each of the populations used in SNP discovery. Since SNPs were identified by the presence of an alternate allele in one (or more) population(s), the overall contribution of SNPs discovered from a given population to a specific PC axis of genetic variation can be calculated as the sum of the loadings of all SNPs at which the source population contributed the minor allele. We found three of the vectors of SNP ascertainment to be correlated with the first two major axes of variation (Figure 1C): SNPs ascertained from Atlantic marine population (BITJ) load heavily on PC1, while SNPs ascertained from the Pacific freshwater population (COND) and marine population (BDGB) load heavily on PC2. These ascertainment vectors are also likely to reflect true biological axes of genetic variation (e.g. the deep split between Atlantic and Pacific sticklebacks [13, 54]), and we note that when all SNPs where the alternate allele was detected only in Atlantic discovery fish are removed from the analysis, Pacific and Atlantic populations still separate along the first principal component of variation detected by PCA with the remaining markers .

Genome Scans for Marine-Freshwater Adaptive Loci

Although different freshwater populations are thought to have evolved independently from marine ancestors, the non-uniformly sampled populations in the current data set may show varying degrees of phylogenetic non-independence due to shared demographic history and gene flow. To account for the potential non-independence of populations in the current dataset, we used the Bayesian linear model approach of Coop *et al.* that corrects for shared demographic history and genome-wide patterns of population structure by including a matrix of covariance in allele frequencies among populations [15]. Genome scans for SNPs with allelic correlation to marine-freshwater environments were performed using BayEnv with all 34 study populations. We used 50,000 iterations for both the estimation of the covariance matrix, and subsequent Bayes factor estimation. Loci with $\log_{10}$(Bayes factors) >=1.5 are shown in column B of Table S3.

Genome Scans for Benthic-Limnetic Adaptive Loci

Since our benthic-limnetic species pairs are restricted to only three lakes in British Columbia, we performed genome scans for F_{ST} outliers separately for each lake using BayeScan 2.01 [42, 43]. The ongoing low levels of gene flow between ecotypes in sympatry should allow neutral loci to introgress readily, while loci subject to divergent selection will remain distinct. BayeScan was run with prior odds of 10, a false discovery rate of 0.05, and default chain parameters (number of iterations = 5,000; thinning interval = 10; number of pilot runs = 20; length of pilot run = 50,000; burn-in length = 50,000). Six hundred and seventy-six, 793, and 727 SNPs were found to be polymorphic and used in the analysis of Paxton, Priest and Quarry lakes, respectively. Forty-six outlier loci linked to genomic regions subject to directional selection in one or more of the lakes were identified (see Table S4). None of the loci were found to be candidate SNPs subject to balancing selection under the model analyzed. It should be noted that although only 4 SNPs were formally identified as F_{ST} outliers in all three lakes, as many as 20 show elevated frequency differences between species pairs in all three lakes, and an F_{ST} outlier analysis pooling ecotypes across lakes identified a total of 39 outliers (data not shown). As expected from the different axes of ecological differentiation, we see little overlap in SNPs linked to putatively adaptive loci in benthic-limnetic species pairs and marine-freshwater species pairs. Only three of the 46 benthic-limnetic F_{ST} outlier SNPs were also identified in our analysis of global populations as showing allele frequencies correlated with marine-freshwater environments (Table S4), and none of these were benthic-limnetic outlier SNPs shared across all three lakes. Several SNPs detected as outliers in one given lake show only moderate or no differences in allele frequency between benthic-limnetic ecotypes in other lakes, despite substantial variation available to detect differentiation in other lakes (Table S4). These loci could underlie phenotypes that have evolved in some lakes but not others, including known differences in degree of armor plate or pelvic reduction [36, 37]. Mutations completely private to a single lake would not be included on the SNP array itself, although selected regions might still be detected by linkage to other markers. However, the size of selective sweeps surrounding favorable mutations may differ across lakes, leading to variable detection with linked markers. Finally, genetic changes may occur at different loci that produce similar overall phenotypes, a phenomenon known to exist in other systems and also likely in sticklebacks, based on previous studies of modifier genes for pelvic and armor plate phenotypes [55, 56].

Genes located within 100 kb genomic windows of outlier SNPs were identified from the draft Stickleback Genome Assembly, gasAcu1.0 [5], using Ensembl gene predictions included in the annotated assembly. Self dot-plot of the genomic region with immunoglobulin light chain (*IGK*) clusters (Figure S2) was constructed in Geneious Pro v5.0.4.

Supplemental References

51. Kent, W. J. (2002). BLAT – the BLAST-like alignment tool. Genome Res. *12*, 656-664.

52. Kingsley, D. M., Zhu, B., Osoegawa, K., De Jong, P. J., Schein, J., Marra, M., Peichel, C., Amemiya, C., Schluter, D., Balabhadra, S., et al. (2004). New genomic tools for molecular studies of evolutionary change in threespine sticklebacks. Behaviour *141*, 1331-1344.

53. Rocha, J. L., Pomp, D., Vleck, L. D. V., and Nielsen, M. K. (2001). Predictors of marker-informativeness for an outbred F2 design. Anim. Genet. *32*, 365-370.

54. Ortí, G., Bell, M. A., Reimchen, T. E., and Meyer, A. (1994). Global survey of mitochondrial DNA sequences in the threespine stickleback: evidence for recent migrations. Evolution *48*, 608–622.

55. Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jónsson, B., Schluter, D., and Kingsley, D. M. (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature *428*, 717-723.

56. Colosimo, P. F., Peichel, C. L., Nereng, K., Blackman, B. K., Shapiro, M. D., Schluter, D., and Kingsley, D. M. (2004). The genetic architecture of parallel armor plate reduction in threespine sticklebacks. PLoS Biol. *2*, 635-641.