Supporting Information

Lamichhaney et al. 10.1073/pnas.1216128109

SI Materials and Methods

Transcriptome Analysis. A male spring-spawning herring was caught in June 2010, in the archipelago of Stockholm. A piece of white skeletal muscle was collected and stored in RNAlater (Invitrogen). mRNA was extracted from the tissue using polyA selection. The library was generated and sequenced at the Science for Life Laboratory (SciLifeLab) in Stockholm. The fragmented library with an insert size of 200 bp was sequenced on an Illumina HiSeq-2000 machine (Illumina) using 101 cycles per run, which yielded ~58 million paired-end reads. All lanes were spiked with 1% of phiX DNA. The last 25 nt at the 3′ end of most reads had PHRED33 quality values lower than 20, and these were, therefore, trimmed in all reads. We removed phiX reads and those contaminated with adapters, linkers, and primers by SeqClean (released on February 2, 2011) using the UniVec database (downloaded on February 2, 2011).

We fed ~98 million trimmed and cleaned reads to Trinity (1) (released on May 19, 2011) and assembled a transcriptome with 121× average depth of sequence, using default parameters. In a final filtering step, we retained only those transcripts with at least 200 bp. Trinity is a three-step assembler of unreferenced transcriptome that enables the assembly of alternative variants. Trinity yielded 76,107 contigs with a combined length of ~61.4 Mbp and with N50 equal to 1,420 bp (Table S1).

Whole-Genome Sequencing. Genomic DNA was isolated of muscle tissue from 400 fish collected from eight locations during 1978–1980 (Fig. 1*A* and Table 1); the samples had been stored at −20 °C until DNA was isolated. The DNA from 50 individuals per sampling location was pooled in equimolar concentrations. For each DNA pool, two libraries were constructed with 250- and 320-bp insert sizes at the SciLifeLab SNP&SEQ Technology Platform at the Uppsala Academy Hospital. These libraries were sequenced at 100 cycles using HiSeq-2000 sequencer (Illumina). The number of reads yielded a theoretical depth of coverage ranging from 42 to 53× for a genome size of 900 Mbp (2).

Previous to the alignment of the data, we selected only high-quality reads by using a trimming algorithm that kept pairs that met certain quality measures. We trimmed the reads from the 3′ end, removing all nucleotides that did not reach a PHRED33 value of 20. Only pairs where both reads were at least 75 bp were kept. We also discarded read pairs that did not have an overall quality value of 20 in PHRED33 scale for at least 80% of the bases or any base with quality less than 10.

Because of the variable length of the trimmed region for each read and the initial variable yield of the sequencing, the final depth of coverage for the eight populations was between 21 and 27×.

Exome Assembly. To cover a larger part of the herring genome and detect more genetic variants, we extended our transcriptome contigs using genomic reads (i.e., to generate what, henceforth, is referred to as an "exome assembly"). We aligned all of the transcriptome contigs against each other using BLASTn (e value, $\leq 10^{-10}$) to select the nonredundant (nr) part of the transcriptome. All contigs that were only aligned to itself were regarded as nr. In cases where contigs had multiple hits to other contigs of the same transcript, only the largest contig of each group was included in the nr set. In total, 56,699 contigs were retained as nr sequences for exome assembly (Table S1). We then aligned the trimmed genomic reads from the Kalix (BHK) sample (Table 1) against the nr transcriptome contigs using bwa (3) (version 0.5.9) with default parameters. We queried the resulting read alignments (5.245% of the total genomic reads) and extracted their unmapped pairs from the raw read files in cases where a genomic read resulted in a unique alignment to the transcriptome. With the genomic reads mapped and their unmapped pairs saved, we used Trinity (1) (released on November 26, 2011) with default parameters to de novo assemble short contigs, which will be extended around the exons with either intronic or intergenic sequence in a process we called exome assembly. By using genomic reads to assemble an exome, we may introduce reads from multiple gene copies and thus create assembly artifacts or chimeric sequences. To correct for this redundancy in the exome contigs, we clustered all sequences with at least 95% sequence identity together using UCLUSTAL (4) and kept the longest representative. In a last effort to remove any putative misassembled contigs, we realigned all of the exome contigs against the nr transcriptome using BLASTn (e value, $\leq 10^{-5}$), resulting in 166,873 exome contigs uniquely matched (Table S1).

SNP Detection. We called SNPs following three separate steps. First, we used bwa (version 0.5.9) (3) with default parameters to map the trimmed genomic reads (see above), separately from each of the eight sampled herring populations (Table 1), to our exome assembly. Then, we called variants with a Bayesian algorithm implemented in FreeBayes (version 0.9.4) with default settings. Alignments with mapping quality lower than 30 were excluded (-m flag). In addition, SNPs with base quality adjusted to PHRED33 less than 20 (-q flag) and posterior probability less than 0.0001 (-pvar flag) were discarded. We were only interested in single nucleotide substitutions ignoring indels, multinucleotide polymorphisms, or any other complex events (–no-indels,–no-mnps,–no-complex flags in FreeBayes). Furthermore, to exclude false-positive SNPs that may happen around indels, we used the indel realignment parameter (–left-align-indels) to perform left-realignment of reads and to merge gaps. FreeBayes also reported triallelic variants. Only 707 of 440,817 SNPs were reported as triallelic. The majority of these third alleles were only supported by a few reads, and many of these are, therefore, expected to be sequencing errors. In these cases, the two most common alleles were used as biallelic variants in the downstream analysis. We filtered the resulting SNPs in the VCF output file as follows:

i) We began by filtering positions with read coverage larger than 100, to avoid calling SNPs in parts of the exome having excess coverage. The reason for this step is to avoid false SNP calls caused by duplicated sequences.

ii) SNP positions with a read depth of at least 50 in the union of all reads from all populations were kept for further analyses. In this set, we also filtered positions that had less than 10 supporting reads for the variant allele.

Statistical Analysis. After the SNP calling, 8 × 2 contingency χ^2 analysis was performed to identify SNPs showing highly significant allele frequency differences among the eight populations (Table S2). The χ^2 test was performed at each SNP position by comparing expected and observed read counts for the reference and variant allele among the eight samples. After examining the outliers in a quantile–quantile (Q-Q) plot (not shown), $P \leq 10^{-10}$ was selected as a highly significance threshold for calling genetically differentiated SNPs given the number of tests performed.

Simulations aimed at describing the expected sampling distribution of F_{ST} values under a perfectly neutral model were conducted using a slightly modified version of the Powsim software (1). This program mimics sampling from populations at a predefined level of expected divergence through random number simulations under a classic Wright–Fisher model without migration or mutation. An infinitely large base population segregating for a specified number of independent, selectively neutral loci with defined allele frequencies is divided into s subpopulations of equal effective size (N_e) through random sampling of $2N_e$ genes. Each of the subpopulations is allowed to drift for t generations, and the expected degree of divergence in generation t is then $F_{ST} = 1 - (1 - 1/2N_e)^t$ (e.g., ref. 2, p. 359).

To reduce the sampling variance of F_{ST} and avoid unnecessary "noise," we restricted the analysis to the 36,794 SNPs that had a minimum of 40 reads from each of the eight populations sampled. Over all these SNPs, the average frequency of the most common allele was ~0.8, and the average F_{ST} (3) was 0.0223. We simulated an infinitely large base population where a single biallelic locus segregated at a frequency of 0.8, split this base population into eight subpopulations of effective size N_e = 10,000, let them drift apart for t = 451 generations to arrive at an expected F_{ST} = 0.0223 for neutral loci, sampled 40 alleles from each subpopulation, calculated F_{ST}, and repeated this process 36,794 times. Finally, the distribution of simulated F_{ST} values was compared with the observed one. The largest simulated F_{ST} value was F_{ST} = 0.1883, and we considered SNPs exhibiting $F_{ST} > 0.1883$ as under directional selection.

The effective sizes of the populations sampled is unknown (4), but as discussed in this paper, there is reason to believe that they may be "large," maybe of the order of millions or more. The divergence of such large populations is difficult to simulate because of the very long time needed for them to drift apart to an F_{ST} of about 0.02 and the large number of genes that must be sampled each generation during the simulation process. As it appears in the present case, however, the exact value of N_e (and the corresponding value of t for the populations to drift to and $F_{ST} \approx 0.02$) is not overly important for the distribution of F_{ST} for "large" N_e and small F_{ST} values. After trying various combinations of N_e and t for reaching $F_{ST} \approx 0.02$, we settled for N_e = 10,000 and t = 451 because these parameters resulted in a relatively modest simulation time (a few days) and produced an F_{ST} distribution that was almost identical to that obtained with N_e = 1,000 and t = 45 generations of drift (a combination that also yields an expected $F_{ST} \approx 0.02$). This striking similarity between the F_{ST} distributions obtained for the different N_e values is actually expected in situations with large populations and small F_{ST} values when alternate fixation for different alleles can be ignored. By approximating the change of allele frequency within subpopulations by a diffusion process (e.g., ref. 5, chap. 4), it can be shown that the variation of allele frequencies among subpopulations, and thereby the distribution of F_{ST}, is independent of N_e.

SNP Genotyping. We selected 5,000 SNPs (called from an earlier version of exome assembly) for genotyping of individual fish. These 5,000 SNPs were selected based on the following criteria:

i) All SNPs from the contigs containing the top 200 outlier SNPs (total SNPs: 673)

ii) All remaining outlier SNPs so that the total number from this set and the one above constitutes 3,000 SNPs (total SNPs: 2,327)

iii) A total of 1,000 SNPs with P values in the range 0.5–0.9 from the χ^2 analysis. (The average frequency across all populations of the rare allele at each locus was in the range 0.3–0.5 to select highly informative SNPs.)

iv) A total of 1,000 SNPs with P values in the range 0.1–0.5 from the χ^2 analysis. (The average frequency across all populations of the rare allele at each locus was in the range 0.3–0.5 to select highly informative SNPs.)

Fifty nucleotides at each side of the SNPs were used to design two assay probes per SNP following instructions provided by Roche NimbleGen. A total of 153 SNPs were discarded from the final design because fewer than 50 nt were available at one of their flanks in the exome assembly. The assay probes (Table S3) together with DNA samples representing 50 fish from each population (400 in total) were used to successfully genotype 3,024 of 4,847 SNPs in 380 (of 400) individuals by an AccuSNP custom array (Roche NimbleGen).

Phylogeny Analysis. The SNPs called from the resequencing data were used to generate two phylogenetic trees: one using all called SNPs (440,817 SNPs) and one using only the 3,847 significant SNPs. First, we generated a genetic distance matrix (10) from the allele frequencies of all SNPs in the eight populations, and these were used to construct neighbor-joining trees using POPTREE (11).

Of the 3,024 successfully genotyped SNPs in 380 individual fish using AccuSNP, we further removed 197 SNPs that were heterozygous in more than 90% of the individuals in each population indicating that they represented paralogous sequence variants rather than alleles at a single locus. In the downstream analyses, we also omitted genotypes of 14 individuals that had high proportions of missing genotypes (>15% of genotyped SNPs).

We calculated a pairwise identical by state (IBS) similarity matrices for each pair of individuals using PLINK (24), both for the 1,583 outlier and the 1,244 neutral SNPs. We converted the IBS matrices into genetic distance matrices using PLINK (12). Using these distance matrices, we generated two different UPGMA trees using Phylogeny Inference Package (PHYLIP) (13) (Fig. S3).

We also used the recently developed fineSTRUCTURE software (14) for cluster analysis. fineSTRUCTURE uses a Bayesian approach to scan for patterns of haplotype similarity and capture information about the underlying population structure. We used the linked model algorithm of fineSTRUCTURE because we expected some of our significant SNPs to be in linkage. We used the resample procedure of fineSTRUCTURE to evaluate the statistical support for the final tree (Fig. 3*C*). In this analysis, using Markov Chain Monte Carlo (MCMC), 100,000 burn steps were used, and 100,000 further iterations were sampled keeping every 100 samples.

Identifying Patterns of Strong Selection. With the resequencing and genotype data from each population, we looked for selection patterns according to the traditional knowledge of the herring population structure, the phylogeny constructed based on the significant markers and the salinity of the Baltic Sea (Fig. 1*A*). We conducted each pairwise interpopulation allele frequency comparison at each SNP position using 2 × 2 contingency χ^2 analysis and identified sets of SNPs with significant allele frequency differences for each comparison ($P \leq 10^{-10}$).

Furthermore, to investigate the genomic distribution of the SNPs showing strong genetic differentiation in our study, we aligned the 1,072 transcriptome contigs containing the significant SNPs against the stickleback genome (BROAD S1 assembly, downloaded from Ensembl database version 61) with tBLASTx. The genome of zebrafish is heavily rearranged compared with other sequenced teleost fishes; we, therefore, mapped our transcripts against the stickleback genome. We then filtered the output, looking for the unique hits that yielded ≥50% identity to the stickleback genome and an e value of $\leq 10^{-5}$. The resulting hits had a genome-wide distribution pattern with all of the chromosomes of the stickleback genome covered (Fig. 3*A*). We also clustered all of the hits in blocks using 500-kb nonoverlapping sliding windows along the stickleback genome. We used allele frequencies to construct heat maps for all blocks and manually inspected them to identify putative sweep regions. SNPs clustering within the same block in stickleback and showing similar frequency pattern indicate a sweep region.

Exome contigs that displayed an interesting pattern regarding frequency differences between populations were manually annotated. Both exome contigs and their corresponding transcriptome contigs were used as queries in nucleotide searches with BLASTn against the nr database at the National Center for Biotechnology Information (NCBI) website (December 2011 to May 2012). Simultaneously, we also performed translated nucleotide searches using tBLASTx against the genome of zebrafish (*Danio rerio*) available at the Ensembl database (15) (versions 65–67). With this strategy, we identified zebrafish proteins that were subsequently used as queries in translated nucleotide searches with tBLASTn against the herring transcriptome to reaffirm and strengthen our initial BLASTn annotation. In the particular cases where no annotated zebrafish orthologs to the herring contigs were found, the same procedure was repeated in stickleback (*G. aculeatus*).

In some cases, our herring transcriptome contigs did not match a full-length gene from zebrafish. Another issue affecting the annotation of transcripts is the difficulty to assign the correct name to a certain transcript matching one of the copies of certain families generated in the teleost specific whole-genome duplication. To account for these multicopies, the zebrafish gene names include generally an extra a and b at their end. For the gene names annotated in Fig. 3, we have, thus, preferred to use the corresponding gene name in human (if available) instead of assigning the partial herring contigs the specific a- or b-copy nomenclature from zebrafish.

1. Grabherr MG, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644–652.
2. Hardie DC, Hebert PDN (2004) Genome-size evolution in fishes. *Can J Fish Aquatic Sci* 61:1636–1646.
3. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
4. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461.
5. Ryman N, Palm S (2006) POWSIM - a computer program for assessing statistical power when testing for genetic differentiation. *Mol Ecol Notes* 6(3):600–602.
6. Nei M (1987) *Molecular Evolutionary Genetics* (Columbia Univ Press, New York).
7. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38(6):1358–1370.
8. Larsson LC, Laikre L, André C, Dahlgren TG, Ryman N (2010) Temporally stable genetic structure of heavily exploited Atlantic herring (*Clupea harengus*) in Swedish waters. *Heredity* 104:40–51.
9. Ewens WJ (2004) *Mathematical Population Genetics. I. Theoretical Introduction* (Springer, New York).
10. Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. *J Mol Evol* 19:153–170.
11. Takezaki N, Nei M, Tamura K (2010) POPTREE2: Software for constructing population trees from allele frequency data and computing other population statistics with Windows interface. *Mol Biol Evol* 27:747–752.
12. Purcell S, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
13. Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5: 164–166.
14. Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS Genet* 8:e1002453.
15. Flicek P, et al. (2011) Ensemble 2011. *Nucleic Acids Res* 39:D800–D806.

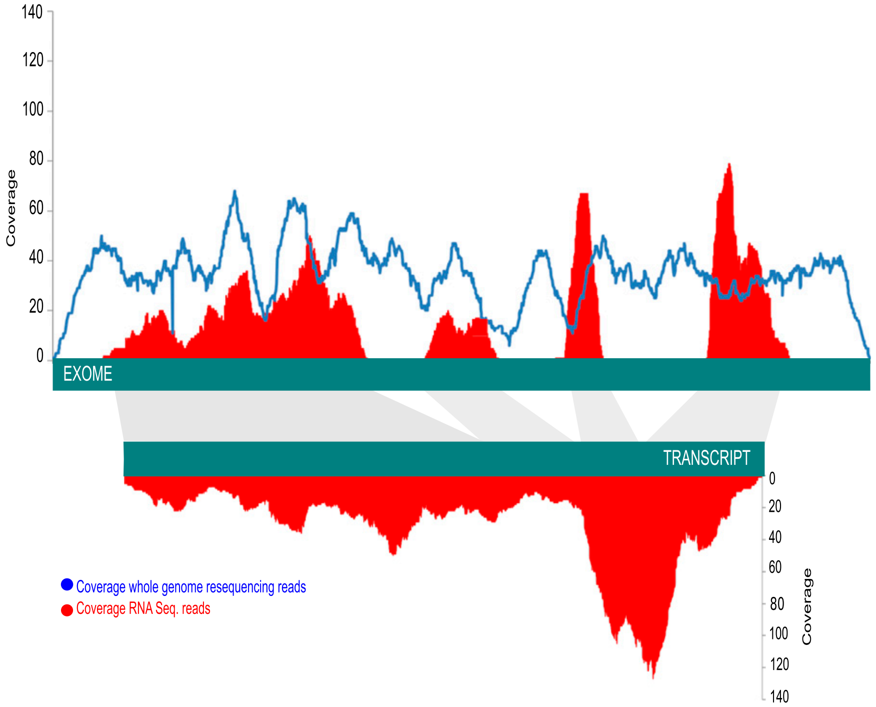

Fig. S1. Example of how exon/intron borders are deduced after aligning genomic reads to a transcript assembled using Trinity.

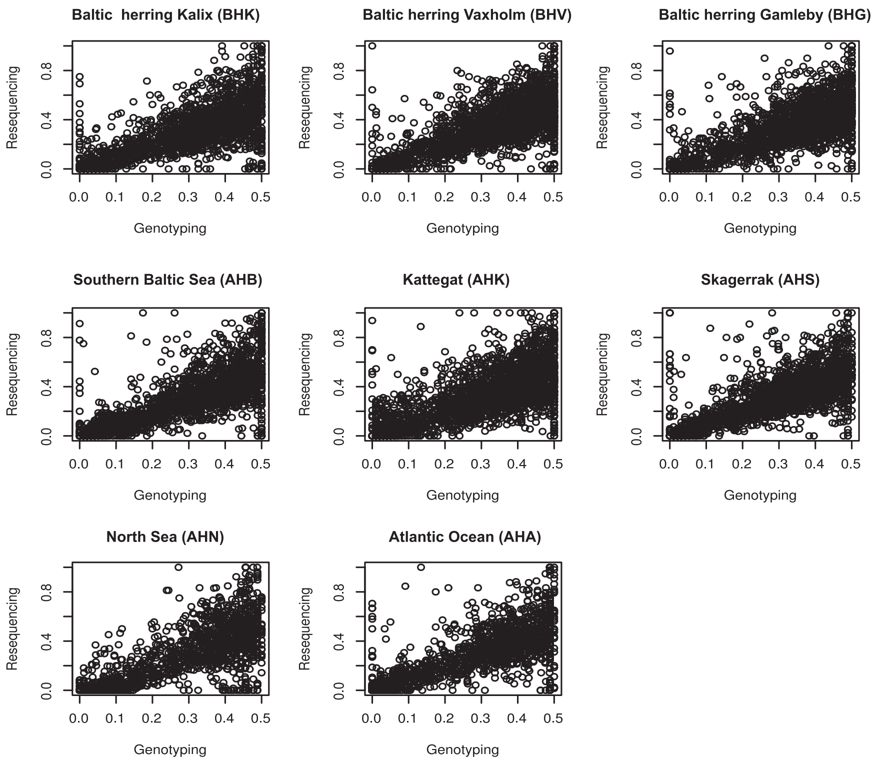

Fig. S2. Strong correlation between allele frequency estimates by genome sequencing of pooled samples and individual SNP genotyping.

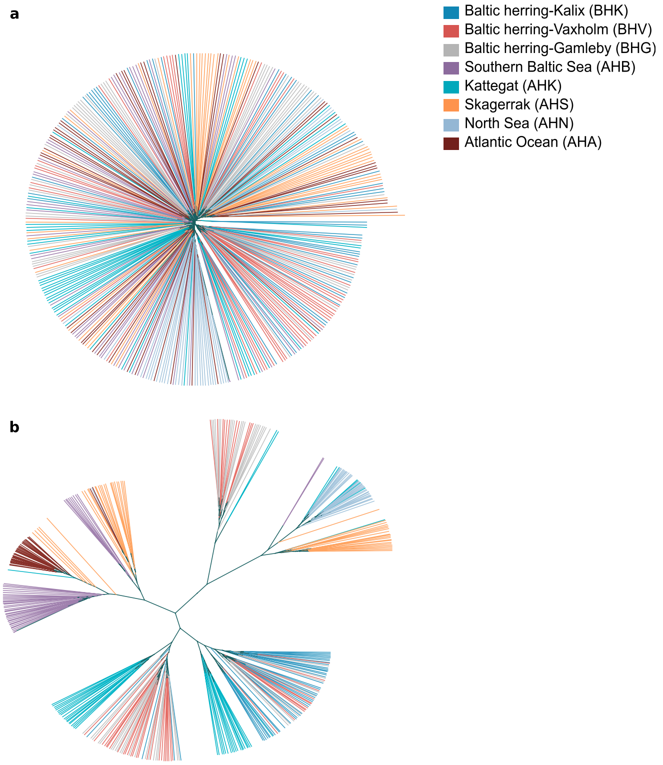
a
Baltic herring-Kalix (BHK)
Baltic herring-Vaxholm (BHV)
Baltic herring-Gamleby (BHG)
Southern Baltic Sea (AHB)
Kattegat (AHK)
Skagerrak (AHS)
North Sea (AHN)
Atlantic Ocean (AHA)
b

Fig. S3. Phylogenetic analysis of individual herring based on individual SNP analysis. (*A*) Phylogenetic analysis based on 1,244 loci showing no significant differentiation. (*B*) Phylogenetic analysis based on 1,583 loci showing significant genetic differentiation.

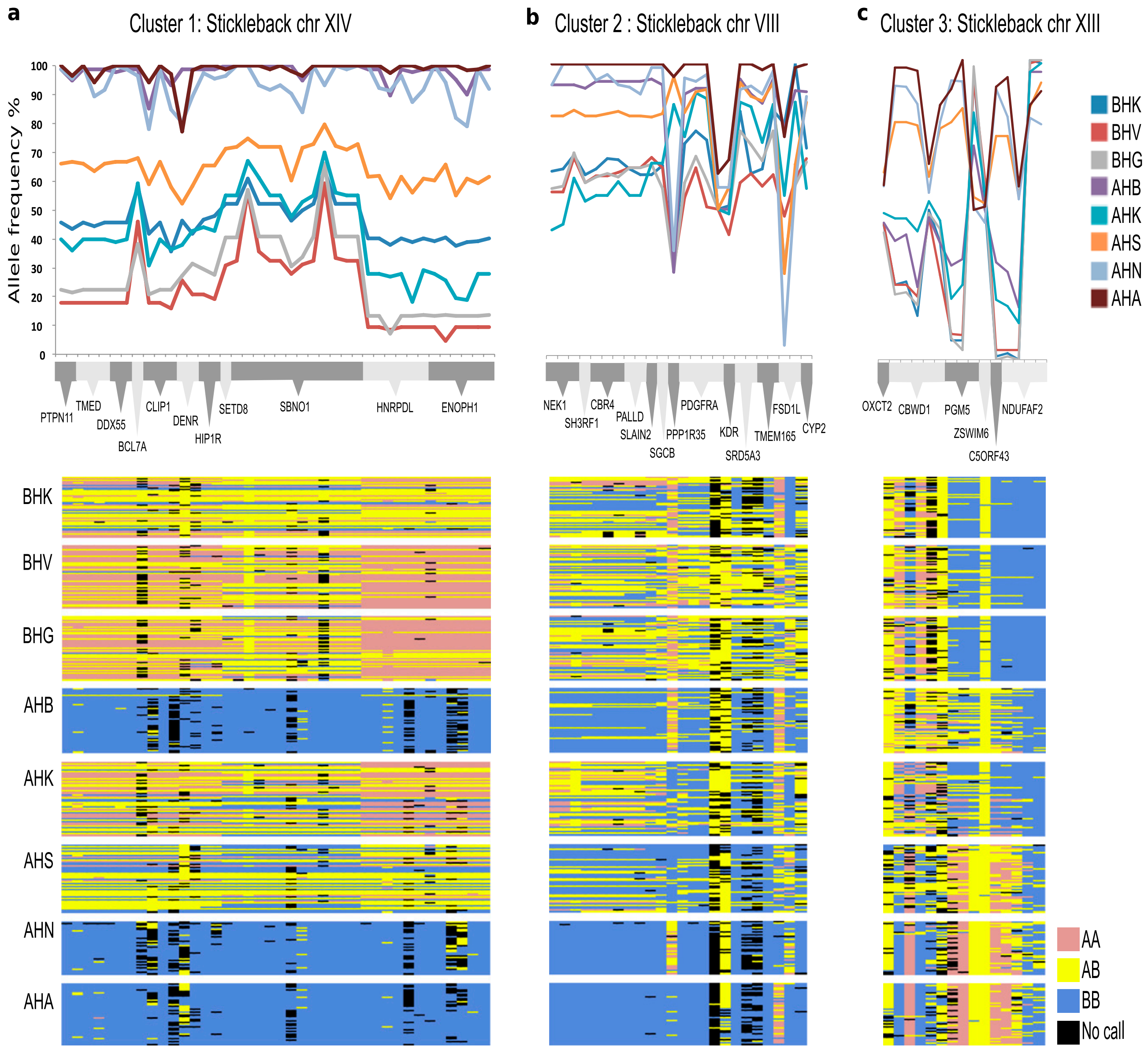

Fig. S4. Allele frequencies and individual SNP genotype data at three cluster of loci showing strong genetic differentiation in Atlantic and Baltic herring. (*A*) Cluster 1 corresponding to stickleback chromosome XIV. (*B*) Cluster 2 corresponding to stickleback chromosome VIII. (*C*) Cluster 3 corresponding to stickleback chromosome XIII. Transcript names are given below the allele frequency graphs. The most common allele in Atlantic Ocean was used as the reference allele at all loci. AHA, Atlantic herring Atlantic Ocean; AHB, Atlantic herring Southern Baltic Sea; AHK, Atlantic herring Kattegat; AHN, Atlantic herring North Sea; AHS, Atlantic herring Skagerrak; BHG, Baltic herring Gamleby; BHK, Baltic herring Kalix; BHV, Baltic herring Vaxholm.

Table S1. Summary statistics for the transcriptome and exome assembly based on Baltic herring muscle RNA sequencing data combined with genomic read alignments

Table S1

Table S2. Loci showing highly significant genetic differentiation ($P < 10^{-10}$) among populations of Atlantic and Baltic herring

Table S2

Table S3. SNP loci genotyped in individual Atlantic and Baltic herring

Table S3