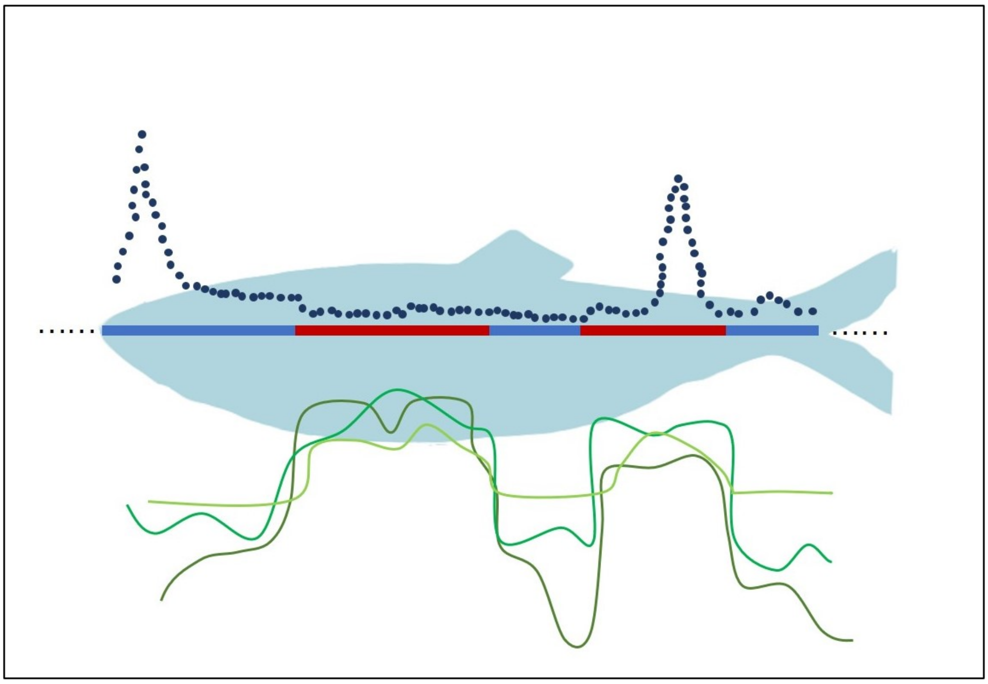
UPPSALA
UNIVERSITET

Evolutionary conservation analysis of sequence variants underlying ecological adaptation in Atlantic herring



Leyi Su

Degree project in biology, Master of science (2 years), 2023
Examensarbete i biologi 30 hp till masterexamen, 2023
Biology Education Centre and Department of Medical Biochemistry and Microbiology, Uppsala University
Supervisors: Leif Andersson, Sabine Felkel and Mats Pettersson
External opponent: Andreas Wallberg

Abstract

The Atlantic herring (*Clupea harengus*) is a model organism for exploring the genetic basis of ecological adaptation on account of its large effective population size which greatly restrains the disturbance of genetic drift. In this project, I conducted an evolutionary conservation analysis of the Atlantic herring genome using conservation scores generated by phastCon, phyloP and GERP++ in order to study the sequence variants related to salinity or spawning-time adaptation in Atlantic herring. Results of conservation score comparisons between SNPs and randomly selected bases showed that SNPs strongly related to ecological adaptation in Atlantic herring did not tend to have high conservation scores, suggesting that sequences showing high sequence conservation among species may contribute relatively little to ecological adaptation in herring. SNPs associated with ecological adaptation in Atlantic herring might be located in fast-evolving genomic regions involved in recurring adaptive evolution in Clupeiformes. Results of the enrichment analysis indicated that nonsynonymous coding variants was the most overrepresented variant group among those associated with ecological adaptation in Atlantic herring, followed by synonymous coding variants, 5kb-upstream and 5kb-downstream variants. Taken all the results together, the conclusion is that coding SNPs under positive selection were the most strongly enriched variant group underlying ecological adaptation in Atlantic herring, while non-coding SNPs, mostly neutral or under negative selection, did not show a similar enrichment.

Introduction

The Atlantic herring (*Clupea harengus*), one of the most abundant fish species on Earth, not only plays an important role in the marine food chain but also has been a major component of commercial fishery in the North Atlantic and in the Baltic Sea for centuries (Lamichhaney *et al.* 2012, Boyce *et al.* 2021). Its wide distribution including North Atlantic Ocean and neighboring water bodies suggests a great number of subpopulations colonizing different geographic regions with diverse environmental conditions, among which populations in the Gulf of Bothnia and Central Baltic sea are taxonomically classified as Baltic herring (*Clupea harengus membras*), a subspecies of Atlantic herring (Lamichhaney *et al.* 2012). In the past few decades, several genetic studies were carried out on Atlantic herring to reveal its population structure and local adaptation mechanism. Before the application of next-generation sequencing techniques, results from these studies consistently showed a lack of genetic differentiation among different populations of Atlantic herring (Ryman *et al.* 1984, Larsson *et al.* 2007, Larsson *et al.* 2010). This can be explained by the combined effect of the huge effective population size and gene flow among herring populations, leading to very low genetic drift. (Masel 2011)

However, with effort devoted on whole genome sequencing of Atlantic herring, hundreds of loci showing significant genetic differentiation were finally revealed to stand out from the clean background, and were suggested to be under selection (Martinez Barrio *et al.* 2016, Lamichhaney *et al.* 2017, Pettersson *et al.* 2019). The data cannot be explained by genetic drift because genetic drift should affect the entire genome more or less equally. Many of these differentiated regions show strong association with salinity or spawning time. In the most recent study with a great amount of sequencing data from 53 populations across most of the distribution of this species, researchers found up to 115 loci (physical positions of genes or specific DNA segments within a genome) to be associated with the adaptation to brackish water and as many as 31 loci to be related to the photoperiodic (daily or seasonal changes in duration of light and darkness) regulation of spawning time (Han *et al.* 2020). Some of the strongest associated missense mutations (a DNA change that results in different amino acids being encoded at a particular position in the resulting protein) have revealed several candidate genes underlying these ecologically adaptive traits, such as the *rhodopsin* (*rho*) gene contributing to adaptation from marine to brackish water (Hill *et al.* 2019) and the *thyroid stimulating hormone receptor* (*tshr*) gene contributing to the regulation of seasonal reproduction in herring (Chen *et al.* 2021). However, a large number of significant SNPs (single nucleotide polymorphisms, a variation at a single position in a DNA sequence among individuals) fall outside of exons of any known gene but in the non-coding sequences, within or between genes. Although the statistical support for the importance of these hundreds of loci is overwhelming, it is not known which sequence variants are functionally important because each locus often includes hundreds of sequence variants showing a similar strong association to ecological adaptation.

Better insight into which sequence variants may contribute to function can be gained by a comprehensive evolutionary conservation analysis of the Atlantic herring genome using so-called conservation scores. Sequences with important functions may not tolerate deleterious change and natural selection will act to remove mutations from such sequences from the population. Consequently, these sequences are less variable and share more similarity among species. These patterns are regarded as evolutionary conservation, which can be revealed by multiple sequence alignment and evaluated by tools like phastCon (Siepel *et al.* 2005), phyloP (Pollard *et al.* 2010) and GERP++ (Davydov *et al.* 2010). Conservation scores are very useful to understand genome biology and they have been thoroughly implemented in the Zoonomia project to discover shared and species-specific patterns of genome evolution in mammals (Christmas *et al.* 2023, Xue *et al.* 2023). Using the evolutionary-conserved inferences generated by a whole-genome alignment of 240 mammalian species, researchers found millions of significantly conserved bases without known function, illustrating the great potential for discovery of functional elements using evolutionary conservation analysis (Christmas *et al.* 2023). Although using conservation scores to discover novel functional elements is a well-established method, overlaying delta allele frequencies (dAFs, the difference of the frequencies of the allele in the two populations) with conservation scores to study which SNPs might be functionally important in ecological adaptation is still not a common approach in population genomic research. Therefore, this project might provide new insights into the application of conservation scores in ecological adaptation studies.

The three different conservation scores used in this project differed in algorithms applied and resolution of measuring evolutionary conservation. The phastCon score measures the probability of a specific base being in a conserved element with the given neutral model (Siepel *et al.* 2005), so bases belonging to the same conserved element share the same value of phastCon score. The possible value of phastCon score for each base is ranging from 0 to 1. The greater the value, the more conserved the element to which the base belongs. The phyloP score measures not only conservation but also acceleration of evolution at individual alignment sites and thus have higher resolution of evolutionary measure. Conservation is indicated by a positive score ranging from 0 to 1 while acceleration is indicated by a negative score. The absolute value of phyloP score is the (–log p) value under a null hypothesis of neutral evolution so a greater absolute value shows stronger evidence that the evolutionary rate of this base violates the neutral evolution hypothesis (Pollard *et al.* 2010). The GERP++ score also have high resolution at each individual alignment position like phyloP, but is calculated via a different algorithm, which measures conservation by comparing the number of substitutions expected under neutrality and the number of substitutions observed at the position. In other words, a positive GERP++ score represents a deficit in substitutions, suggesting an evolutionarily conserved base, while a negative GERP++ score suggests an accelerated evolution or neutral evolution (Davydov *et al.* 2010).

Aims

Using genome-wide estimates of phastCon, phyloP and GERP++ scores per base, I aimed to:

1. Investigate whether bases with high conservation scores (which may have important functions) are overrepresented among those loci strongly associated with ecological adaptation in herring
2. Explore the selective forces acting on SNPs related to the ecological adaptation in herring
3. Explore which category of SNPs (exons, introns, etc.) are most strongly overrepresented among those associated with ecological adaptation in herring.

Materials and methods

Alignment and SNP data

The main dataset of this project was made up of the chromosome-level genome assembly of Atlantic herring (*Clupea harengus*) (accession no. GCA_900700415.2) (Pettersson *et al.* 2019) and 8 other fish genomes also from Clupeiformes, which are *Alosa alosa* (accession no. GCF_017589495.1), *Alosa sapidissima* (accession no. GCF_018492685.1), *Coilia nasus* (accession no. GCA_007927625.1), *Denticeps clupeoides* (accession no. GCA_900700375.2), *Limnothrissa miodon* (accession no. GCA_017657215.1), *Sardina pilchardus* (accession no. GCA_003604335.1), *Sprattus sprattus* (not published), *Tenualosa ilisha* (accession no. GCA_015244755.2) . First of all, a multiple sequence alignment with the Atlantic herring assembly as the co-ordinate backbone was built by Sabine Felkel (Leif Andersson's research group) using progressiveCactus (Paten *et al.* 2011) and was converted into a PHAST-readable multiple alignment format (MAF) via hal2maf (Hickey *et al.* 2013). Secondly, the MAF file was filtered according to the following criteria: (i) any alignment block with less than three genomes aligned to should be removed; (ii) any alignment column with more than three gaps (missing data) should be removed from the specific block. To visualize the composition of this final Clupeiformes alignment which was used for all the downstream analyses, I measured the overlaps between the alignment and different annotation types using bedtools and generated bar plots using the matplotlib module in python.

The SNP data used in this project was from a previous study on the ecological adaptation in Atlantic herring (Han *et al.* 2020), including: (i) two lists of SNPs strongly related to salinity adaptation and spawning-time adaptation, respectively; (ii) the absolute allele frequency difference (delta allele frequency, dAF) for each SNP calculated from the two replicates of salinity contrast (Atlantic spring spawners vs Baltic spring spawners, Atlantic autumn spawners vs Baltic autumn spawners) and the two replicates of spawning-time contrast (Atlantic spring spawners vs Atlantic autumn spawners, Baltic spring spawners vs Baltic autumn spawners).

Estimate conservation scores

Conservation scores for each base were estimated by Sabine Felkel using phastCon (Siepel *et al.* 2005), phyloP (Pollard *et al.* 2010) and GERP++ (Davydov *et al.* 2010). The first two tools are implemented in PHAST package (Hubisz *et al.* 2011, Ramani *et al.* 2019). In order to eliminate an observed GC-bias in high phyloP scores, I used the MinMaxScaler function from the scikit-learns module in python to rescale all positive phyloP scores into the new range 0-1 for each nucleotide separately. Therefore, the phyloP scores used in the enrichment analyses consists of the original non-positive phyloP scores and the rescaled positive phyloP scores.

Conservation score comparisons and statistical tests

To investigate whether bases with high conservation scores are overrepresented among those SNPs strongly associated with ecological adaptation in herring, I calculated the mean values of conservation scores of (i) 1150 SNPs related to salinity, (ii) 522 SNPs related to spawning-time and (iii) 10,000 randomly selected bases using the numpy module in python and illustrated them by generating bar plots using the matplotlib module in python. I also generated histograms to compare the distribution of conservation scores using the seaborn module and matplotlib module in python. To test the significance of observed differences, I conducted a series of statistical tests using the stats package from the scipy module in python. First of all, I used the normality test and Levene test to find out if the datasets satisfy the required assumptions of parametric tests. Then, I performed the two-sample Kolmogorov-Smirnov test (KS test) and the two-sample Wilcoxon Rank Sum Test (also called Mann-Whitney U Test) in order to compare the underlying distributions of conservation scores of SNPs strongly associated with ecological adaptation and that of bases not associated with ecological adaptation.

Enrichment analysis of different SNP categories

SNPs were already functionally annotated using SnpEff (v.3.4) (Cingolani *et al.* 2012). I classified them into nine categories namely non-synonymous coding, synonymous coding, UTR, 5kb-upstream, 5kb-downstream, intronic, intergenic, splice-related and ORF-related (the gain or loss of start or stop codon). The last two categories were not shown in the final plots because sometimes these had zero counts, which would skew the plot. The first seven categories could be further grouped into three categories (non-synonymous coding, synonymous coding and non-coding) or two categories (coding and non-coding). The dAFs were sorted into seven bins (0-0.05, 0.05-0.10, 0.10-0.15, 0.15-0.20, 0.20-0.30, 0.30-0.50, 0.50-1.0). To measure the SNP enrichment, I calculated the M-value as the log2 fold change of the observed SNP count for each SNP category in each bin compared to the expected SNP count (calculated as the proportion of a specific SNP category in the entire genome times the total number of SNPs in a given bin) using python. For example, if we have twice the expected number of SNPs the M-value we get would be 1, and if we have only half the expected number of SNPs the M-value we get would be -1. Therefore, using logarithms allows a more intuitive interpretation of M-value that a positive M-value means overrepresentation while a negative M-value means underrepresentation. All the plots were generated by the matplotlib module in python.

Results

Overview of the alignment

The Clupeiformes alignment used for conservation scores estimation spans 406,121 kb, corresponding to about 50% of the herring genome. The composition of this alignment by various annotation groups is highly similar to the composition of the Atlantic herring genome itself (Figure 1A). Nearly half (45.6%) of the Clupeiformes alignment belongs to intronic region, which is also the largest group in the genome of Atlantic herring. The second largest group is intergenic region (excluding every 5kb region upstream or downstream of gene), which takes up a slightly smaller fraction (26.4%) compared with its percentage (29.4%) in the Atlantic herring genome. The most functionally important group, exons make up 13.2% of the alignment, which is 1.5 times higher than its percentage in the whole genome. The 5' and 3' untranslated regions of mRNAs (UTR), well-known for their regulatory roles in gene expression, whose ratio (3.2%) ranks the lowest among all six annotation groups, but is the other of the only two groups showing notable increase in percentage (1.3-fold) compared with that of herring genome. Furthermore, given the fact that the 406 Mb Clupeiformes alignment covers 56.0% of the whole genome assembly of Atlantic herring, it actually covers remarkably large portions of exons and UTRs, which are 82.8% and 75.2% respectively (Figure 1B). These observed coverages are significantly higher than the expected value if aligned regions are randomly distributed across the herring genome, suggesting that the exons and UTRs were relatively well-aligned and therefore more conserved among species compared with noncoding regions. These results are in line with the fact that exons code for proteins and UTRs regulate translation process both of which play important functional roles in maintaining life. Meanwhile the underrepresented intergenic regions were more variable and harder to align.

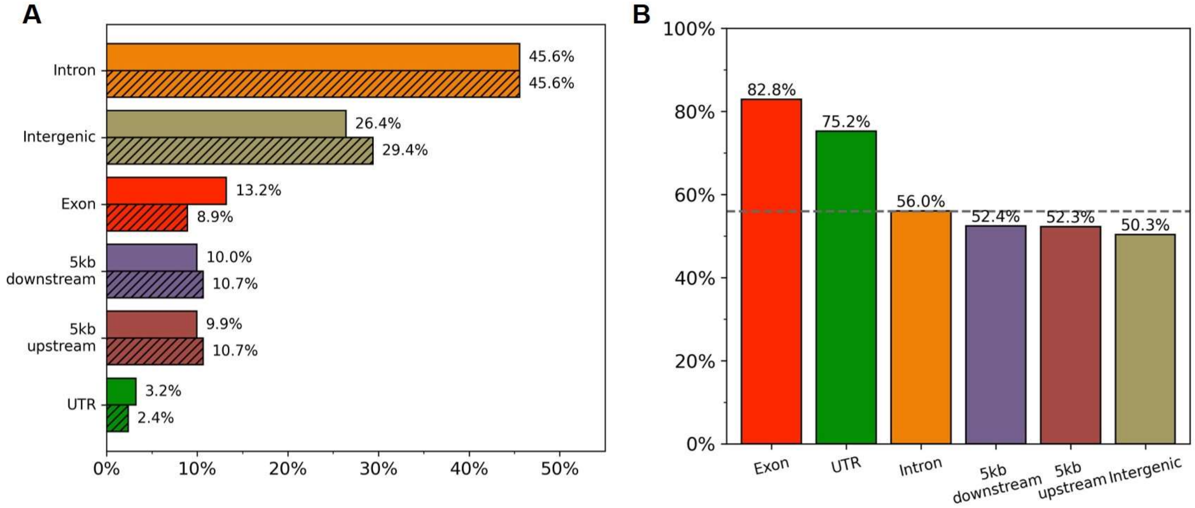


Figure 1. The composition of the Clupeiformes alignment measured in percentages
(A) Proportions of Clupeiformes alignment belonging to different annotation groups (open bars) in comparison to

proportions of whole genome of Atlantic herring belonging to the different annotation groups (hatched bars). The sum of all the percentage is greater than 100% due to overlaps between the 5kb upstream and 5kb downstream regions. **(B)** Proportions of various annotation groups covered by the Clupeiformes alignment. The dashed gray line indicates the expected coverage which is the percentage of Atlantic herring genome covered by the alignment. Bars are sorted from highest to lowest. The exact percentage for each annotation group is shown on the top of the corresponding bar.

Conservation scores comparisons and statistical tests

To study if SNPs related to ecological adaptation in Atlantic herring have a tendency to fall into conserved regions, I compared the mean value and the distribution of conservation scores of SNPs related to salinity, SNPs related to spawning-time and randomly selected bases. The result of mean value comparison indicated that randomly selected bases had higher mean values of phastCon scores, phyloP scores as well as GERP++ scores compared with those of SNPs related to salinity or spawning-time adaptation (Supplementary Figure 1).

In the distribution comparison of phastCon scores (Figure 2A), all three histograms showed highly similar patterns that mostly phastCon scores fell in bin 0-0.05 and only a minority of scores fell in bins with high phastCon scores. However, there were still some small but perceptible differences among these three histograms, which were that SNPs related to ecological adaptation had higher probability to fall in bins with low phastCon scores and had lower probability to fall in bins with high phastCon scores compared with randomly selected bases. Different from the homogenous distributions of phastCon scores, the distributions of phyloP scores showed two different patterns (Figure 2B). The distributions of phyloP scores of SNPs related to salinity or spawning-time adaptation were similar with each other and both had a highest peak located at bins with negative phyloP scores ranging from -0.2 to -0.6. On the contrary, the phyloP scores of randomly selected bases had a different distribution pattern with a highest peak of probability located in bin with positive scores ranging from 0.4 to 0.6. Besides, SNPs related to ecological adaptation had lower probability to reach high phyloP scores (see the magnified subplots of Figure 2B) than that of randomly selected bases. These two findings both supported that SNPs related to ecological adaptation had lower phyloP scores in comparison to randomly selected bases. In the distribution comparison of GERP++ scores (Figure 2C), the highest peak was located near zero (bin 0-0.1) for SNPs related to salinity or spawning-time, while the highest peak in randomly selected bases was located in bin with higher positive GERP++ scores (bin 0.4-0.5). Like the patterns observed in bins with high phastCon and phyloP scores, the SNPs related to ecological adaptation had lower probability to get high GERP++ scores than randomly selected bases (see the magnified subplots of Figure 2C). Taken together the results of comparisons of mean values as well as distributions of conservation scores, I could conclude that SNPs related to salinity or spawning-time adaptation had lower phastCon scores, phyloP scores and GERP++ scores compared with randomly selected bases.

To examine if the differences between conservation scores of SNPs related to ecological adaptation and randomly selected bases were significant or not, I conducted several statistical tests using six generated sets of bases: (A) 1150 SNPs strongly related to salinity adaptation with conservation scores, (B) 406,120 kb bases which did not belong to set A from the Clupeiformes alignment, (C) 522 SNPs strongly related to spawning time adaptation with conservation scores, (D) 406,120 kb bases which did not belong to set C from the Clupeiformes alignment, (E) 1507 SNPs strongly related to salinity or spawning time adaptation with conservation scores, (F) 406,120 kb bases which did not belong to set E from the Clupeiformes alignment. For each extremely large complement sets (B/D/F), I used random sampling to generate equal size of samples of phastCon, phyloP and GERP++ scores according to the size of set A/C/E so as to avoid unbalanced sample sizes in the following statistical tests. In addition, the sampling procedure was repeated ten times for each pair of datasets for comparison in order to perform ten independent replicate tests for reproducibility of the test results.

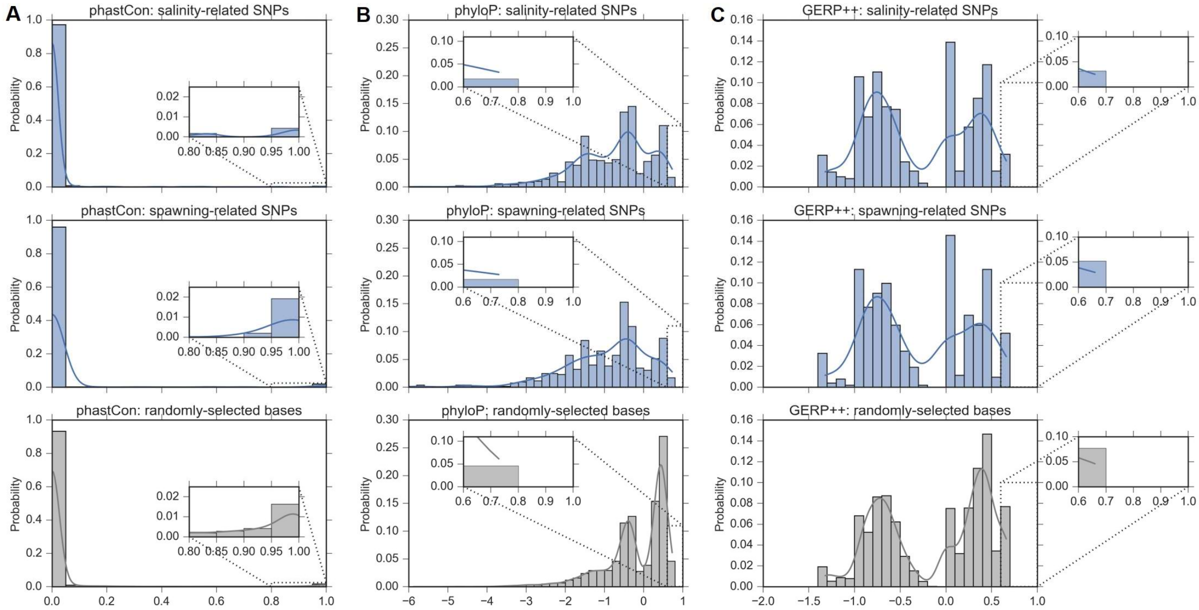


Figure 2. Distribution of conservation scores of SNPs related to ecological adaptation and randomly selected bases

Histograms illustrated the distribution of phastCon scores **(A)**, phyloP scores **(B)** and GERP++ scores **(C)** of 1150 SNPs related to salinity (1st row), 522 SNPs related to spawning-time (2nd row) and 10,000 randomly selected bases (3rd row) measured in probability. Bins with high conservation scores (close to 1) were magnified using inserted subplots with dashed linking lines. The three histograms of the same type of conservation score shared the same setting of x-axis as well as y-axis. Conservation scores of SNPs related to ecological adaptation were colored in blue while conservation scores of randomly selected bases were colored in grey.

To decide which statistical test could be taken, I conducted normality test on all of the samples in the first place. The results showed that all samples did not conform to normal distribution with very significant P-values ($P < 0.001$) (Supplementary Table 1), which indicated that parametric tests like t-test could not be performed on these samples due to unsatisfied basic assumption of normal distribution. Common statistical transformations as well as the powerful Box-Cox transformation have been taken to try to convert current non-normal samples into normal distributions but failed. Moreover, according to the results of Levene Test, the phastCon (and phyloP) scores of most pairs of samples do not have equal variance, which violated another important assumption of variance equality in many statistical tests (Supplementary Table 1).

Therefore, when comparing different samples of conservation score, I could only choose nonparametric tests which did not rely on any underlying statistical distribution in the data. To begin with, I used two-sample KS Test to compare the Cumulative Distribution Functions

(CDFs) of each pair of samples. Each comparison has three parallel KS tests to be done since there are three different null hypotheses (H_0) that could be tested. The results from parallel tests were very consistent with each other, indicating that for all three types of conservation scores, the CDF of sample from adaptation-related SNPs was greater than the CDF of sample from its complement (Table 1). In other words, SNPs related to ecological adaptation tend to have lower conservation scores than other bases since the sample with greater CDF has lower values. Moreover, I also conducted Wilcoxon Rank Sum Test on each pair of samples to further compare their underlying distributions, and the results were in great agreement with the results from KS Test, supporting that the distribution of conservation scores of SNPs related to ecological adaptation was less than that of other bases (Table 1). Thus, the results of statistical tests suggested that the differences between conservation scores of SNPs related to ecological adaptation and randomly selected bases were significant and these SNPs showed in general lower conservation scores than randomly selected bases.

Table 1. Nonparametric tests on conservation scores from different sample of SNPs

Dataset	Sample size	Score type	KS Test		Wilcoxon Rank Sum Test	
			H_0	Result	H_0	Result
Salinity-related (A) vs Others (B)	$n_a = n_b = 1150$	phastCon	$CDF_a = CDF_b$	reject, ***	$a = b$	reject, ***
			$CDF_a < CDF_b$	reject, ***	$a > b$	reject, ***
			$CDF_a > CDF_b$	cannot reject, NS	$a < b$	cannot reject, NS
		phyloP	$CDF_a = CDF_b$	reject, ***	$a = b$	reject, ***
			$CDF_a < CDF_b$	reject, ***	$a > b$	reject, ***
			$CDF_a > CDF_b$	cannot reject, NS	$a < b$	cannot reject, NS
		GERP++	$CDF_a = CDF_b$	reject, ***	$a = b$	reject, ***
			$CDF_a < CDF_b$	reject, ***	$a > b$	reject, ***
			$CDF_a > CDF_b$	cannot reject, NS	$a < b$	cannot reject, NS
Spawning-related (C) vs Others (D)	$n_c = n_d = 522$	phastCon	$CDF_c = CDF_d$	reject, **	$c = d$	reject, **
			$CDF_c < CDF_d$	reject, **	$c > d$	reject, **
			$CDF_c > CDF_d$	cannot reject, NS	$c < d$	cannot reject, NS
		phyloP	$CDF_c = CDF_d$	reject, ***	$c = d$	reject, ***
			$CDF_c < CDF_d$	reject, ***	$c > d$	reject, ***
			$CDF_c > CDF_d$	cannot reject, NS	$c < d$	cannot reject, NS
		GERP++	$CDF_c = CDF_d$	reject, ***	$c = d$	reject, ***
			$CDF_c < CDF_d$	reject, ***	$c > d$	reject, ***
			$CDF_c > CDF_d$	cannot reject, NS	$c < d$	cannot reject, NS
Salinity /Spawning-related (E) vs Others (F)	$n_e = n_f = 1507$	phastCon	$CDF_e = CDF_f$	reject, ***	$e = f$	reject, ***
			$CDF_e < CDF_f$	reject, ***	$e > f$	reject, ***
			$CDF_e > CDF_f$	cannot reject, NS	$e < f$	cannot reject, NS
		phyloP	$CDF_e = CDF_f$	reject, ***	$e = f$	reject, ***
			$CDF_e < CDF_f$	reject, ***	$e > f$	reject, ***
			$CDF_e > CDF_f$	cannot reject, NS	$e < f$	cannot reject, NS
		GERP++	$CDF_e = CDF_f$	reject, ***	$e = f$	reject, ***
			$CDF_e < CDF_f$	reject, ***	$e > f$	reject, ***
			$CDF_e > CDF_f$	cannot reject, NS	$e < f$	cannot reject, NS

NS, $P>0.1$; *, $0.05>P\geqslant0.01$; **, $0.01>P\geqslant0.001$; ***, $P< 0.001$.

A, B, C, D, E and F represent the populations of six sets of bases while a, b, c, d, e and f represent the samples

from the corresponding population, except for A=a, C=c and E=e.
The results with the highest number of occurrences out of ten replicate tests were displayed in the table.
CDF means the Cumulative Distribution Function, which describes the probabilities of a random variable having values less than or equal to x, so greater CDF means a negative shift in distribution compared with the other.

To sum up, the results of comparisons and statistical tests demonstrated that SNPs strongly related to ecological adaptation in Atlantic herring did not have a tendency to fall into conserved regions and in fact, they seemed to be less conserved among the Clupeiformes compared with other bases.

Enrichment analysis of different SNP categories

I carried out several enrichment analyses to explore which categories of SNPs were most overrepresented among those associated with ecological adaptation in herring. In the initial analysis no information on conservation scores were taken into account. In this analysis, the contrast between spring- and autumn-spawning herring in the Atlantic (Figure 3C) clearly showed a very different outcome compared to the other contrasts, so I disregarded it when summarizing common patterns. Non-synonymous coding variants was the most overrepresented group at dAF > 0.5 bin in the two replicates of salinity contrast and the Baltic replicate of spawning-time contrast (Figure 3ABD). Synonymous coding variants were also of great enrichment among SNPs with high dAF, especially in the spring-spawning replicate of salinity contrast and the Baltic replicate of spawning-time contrast (Figure 3AD). 5kb-upstream and 5kb-downstream variants with dAF > 0.5 mostly had small positive M-values, which indicated more observed counts than expected values, therefore suggesting that regulatory changes played a considerable role in the ecological adaptation in herring as well.

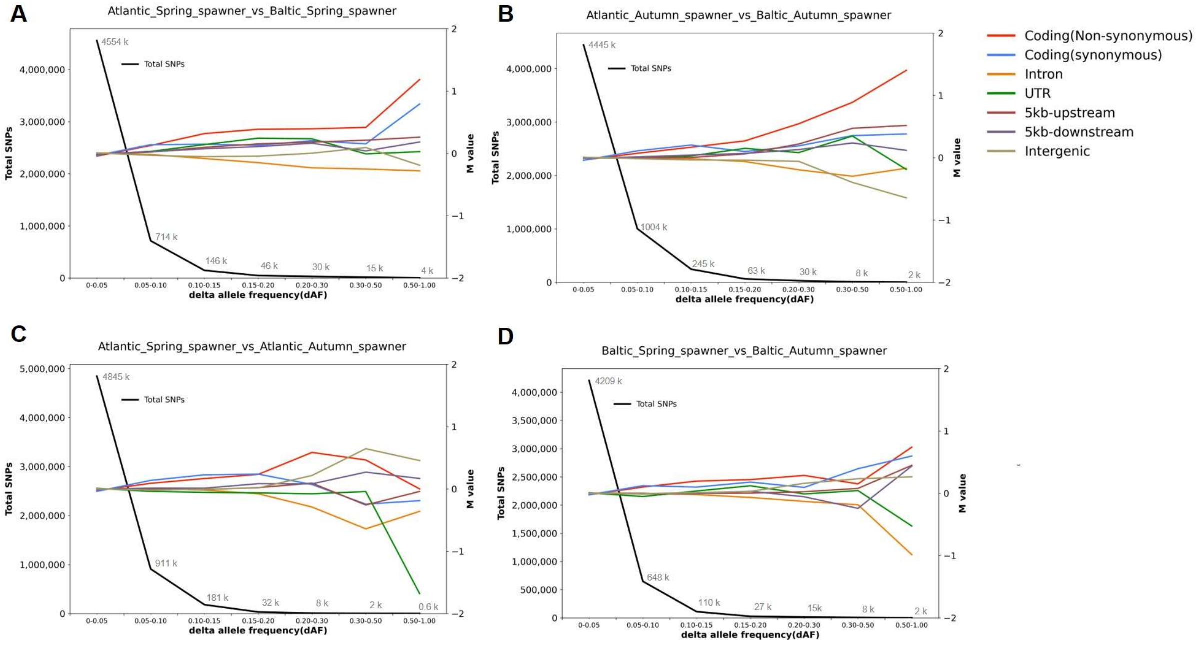


Figure 3. Genomic distribution of different categories of SNPs at different delta allele frequency (dAF)
The dAFs were calculated from the salinity contrast (marine vs brackish water) in spring spawners **(A)** and autumn spawners **(B)**, and from the spawning-time contrast (spring-spawning vs autumn-spawning) in Atlantic populations **(C)** and Baltic populations **(D)** respectively. Each colored line represents the M values (see Methods) of a specific category of SNPs using a consistent color scheme throughout the text. The total number of SNPs falling in each bin of dAF is shown by the black line and grey text. All the SNPs with dAF > 0 are included in the analysis.

Additionally, with annotations of conservation scores, I carried out a similar enrichment analysis using only the conserved SNPs. The results confirmed the above pattern of non-synonymous coding variants with even steeper increase in the enrichment among conserved SNPs with high dAFs, reaching higher M values (Figure 4). Moreover, at the bin of dAF > 0.5, synonymous coding variants were still overrepresented among conserved SNPs, leaving conserved non-coding SNPs as the only underrepresented category of SNPs.

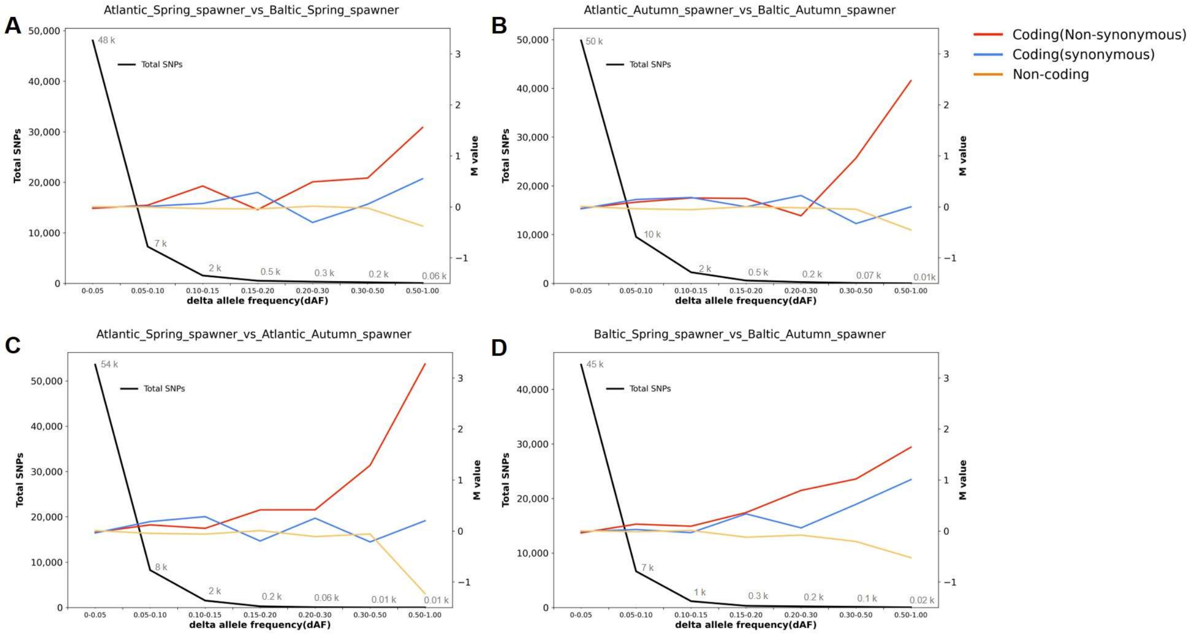


Figure 4. Distribution of different categories of conserved SNPs at different delta allele frequency (dAF) The dAFs were calculated from the salinity contrast (marine vs brackish water) in spring spawners **(A)** and autumn spawners **(B)**, and from the spawning-time contrast (spring-spawning vs autumn-spawning) in Atlantic populations **(C)** and Baltic populations **(D)** respectively. Each colored line represents the M values (see Methods) of a specific category SNPs. The total number of SNPs falling in each bin of dAF is shown by the black line and grey text. Only the SNPs with all three types of conservation scores in the top 10% are included in the analysis.

In order to further study the relative importance of different categories of SNPs in terms of conservation and whether they occurred in coding or non-coding regions, all the SNPs were classified into four new categories based on conserved or non-conserved and coding or non-coding. Although the ranking orders were not consistent with each other in these four analyses, some common patterns were worth pointing out. One is that either of the two groups of coding SNPs was always the most overrepresented group at dAF > 0.5 and for 3 out of 4, it was the Conserved Coding SNPs category (Figure 5ACD). The other common pattern was that Conserved Non-coding SNPs was always underrepresented at dAF > 0.5 except for the spring spawner replicate of salinity contrast, which showed a significantly different trajectory compared with the flat lines of Non-conserved Non-coding SNPs across all the bins (Figure 5BCD). All of these patterns implied that a considerable fraction of coding SNPs were under positive selection while non-coding SNPs were mostly neutral or under negative selection.

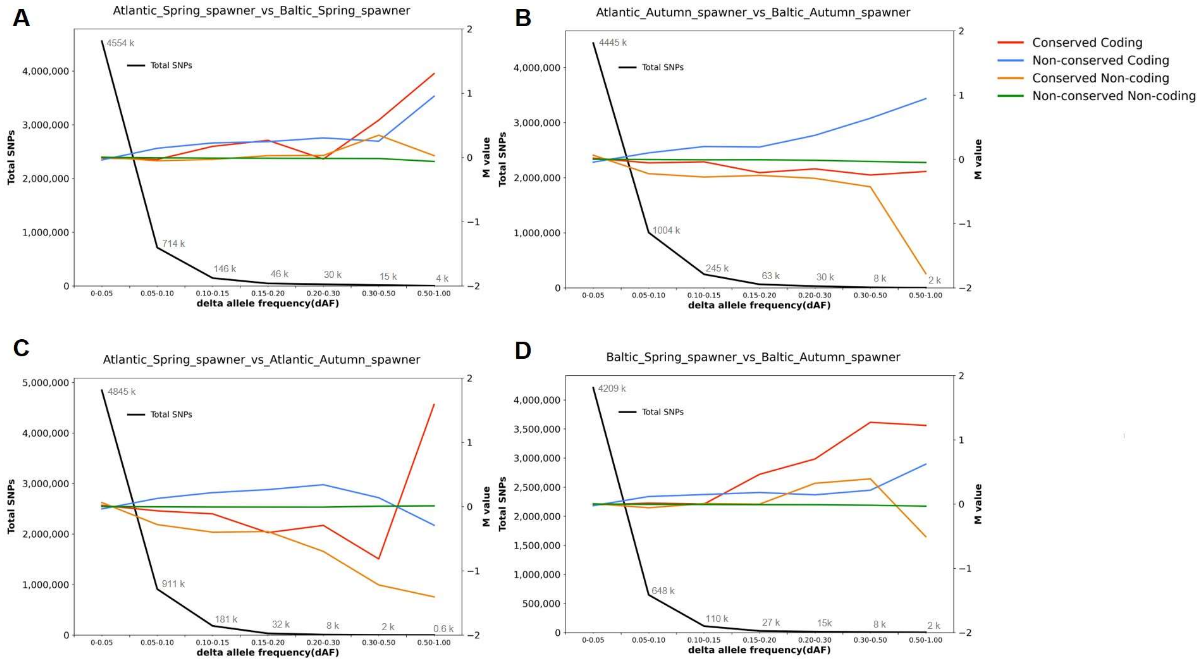


Figure 5. Distribution of different categories of SNPs at different delta allele frequencies (dAFs) in terms of conservation and whether they occurred in coding or non-coding regions

The dAFs were calculated from the salinity contrast (marine vs brackish water) in spring spawners **(A)** and autumn spawners **(B)**, and from the spawning-time contrast (spring-spawning vs autumn-spawning) in Atlantic populations **(C)** and Baltic populations **(D)** respectively. Each colored line represents the M values (see Methods) of a specific category SNPs. The total number of SNPs falling in each bin of dAF is shown by the black line and grey text. Conserved SNPs are those with all three types of conservation scores in the top 10% and the rest of SNPs are all defined as non-conserved SNPs. All the SNPs with dAF > 0 are included in the analysis.

Discussion

Identifying the biologically causal sequence variant(s) for any trait has always been a great challenge in the genetic research (Altshuler *et al.* 2008). Even though with the help of whole-genome sequencing, it is still hard to distinguish the casual variant(s) from other sequence variants showing similarly strong association with the phenotype of interest due to linkage disequilibrium and other reasons (Hormozdiari *et al.* 2014, Mountjoy *et al.* 2021). In the past decade, with the genome assembly of higher integrity and more resequencing data from different herring populations, researchers found more and more loci showing significant genetic differentiation in the contrast of different ecotypes or geographically distinct ecoregions (Martinez Barrio *et al.* 2016, Lamichhaney *et al.* 2017, Pettersson *et al.* 2019, Han *et al.* 2020). As in human and many other organisms, finding genetic evidence of functional importance of these hundreds of sequence variants has become the most challenging question in the research on local adaptation of Atlantic herring. Therefore, the main purpose of this project was to use the genomic annotations of conservation scores to better characterize the sequence variants contributing to ecological adaptation in Atlantic herring so as to find out which sequence variants are functionally important and controlling adaptation.

One important task is to investigate whether bases with high conservation scores are overrepresented among those loci strongly associated with ecological adaptation in herring. Based on the results of conservation scores comparisons (Supplementary Figure 1, Figure 2) and statistical tests (Table 1), I can conclude that SNPs strongly related to ecological adaptation in Atlantic herring did not in general had high conservation scores but showed indication of accelerated evolution among species, suggesting that sequences showing high sequence conservation among species contribute relatively little to ecological adaptation in herring. A possible interpretation could be that the SNPs associated with ecological adaptation were located in genomic regions that were undergoing continuous changes not only in herring but also in many other species to keep up with the changing environment. In other words, there is recurring adaptive evolution in these parts of the genome, like the reported recurrent missense mutation in *rhodopsin* (Phe261Tyr) contributing to adaptation from marine to brackish environment with red-shifted light (Hill *et al.* 2019). With different species using the same set of genes for ecological adaptation, these genomic regions became highly variable among species, resulting in the low conservation scores observed in this project. Therefore, this result demonstrated that not only highly conserved bases could be evidence of functionally important elements but fast-evolving bases may also shape adaptative traits in some species to support their survival facing environmental changes and be regarded as functionally important regions. Additionally, to further explore the relationship between delta allele frequency (dAF) and conservation score, I also tried plotting the three types of conservation scores against dAFs separately and none of them showed strong correlation.

A second main goal was to explore which category of SNPs are most overrepresented among those associated with ecological adaptation in Atlantic herring. Nonsynonymous coding variants ranked first at dAF > 0.5 in most cases, usually followed by synonymous coding variants, 5kb-upstream and 5kb-downstream variants (Figure 2). This pattern is in line with a similar analysis of dAF for different categories of SNPs in a previous study of Atlantic herring based on a smaller but partially overlapping dataset (Martinez Barrio *et al.* 2016). Therefore, although coding and regulatory regions both contributed, the coding changes played more important roles in the ecological adaptation in herring, which is not the cases in stickleback, where researchers concluded that regulatory changes accounting for a much larger proportion of the overall set of loci repeatedly selected during marine–freshwater divergence (Jones *et al.* 2012). The great enrichment of non-synonymous coding changes also suggested that the genetic architecture underlying ecological adaptation in herring deviates from the classical infinitesimal model for complex traits (Fisher 1919) and might be more consistent with a few mutations of large effects. Moreover, the overrepresentation of UTR in Martinez Barrio *et al.*'s study did not reappear in my results. One possible reason is that this project included a lot more SNPs and populations. Some of them were underrepresented in the dataset of the previous study, namely the spring-spawning populations and Atlantic populations.

Exploring the selective forces acting on SNPs related to the ecological adaptation in herring was also one of the major objectives of this study. Reading directly from the conservation scores, especially the sign of the phyloP or GERP++ score was the planned method. However, the three types of conservation scores were in disagreement with each other for some of the SNPs, making it risky to interpret the direction of evolution base by base. In spite of that, the enrichment analysis of different SNP categories gave some answers to this question: (i) the overrepresentation of coding SNPs at dAF > 0.5 suggested that positive selection was acting on the protein sequence of genes associated with ecological adaptation in Atlantic herring, (ii) the underrepresentation of conserved non-coding SNPs at dAF > 0.5 suggested negative selection was acting on these regions, which implied the conserved non-coding regions were probably not the major contributor to the ecological adaptation of Atlantic herring. Therefore, using conservation scores to reveal the role of those unannotated non-coding sequences variants associated with ecological adaptation was not fruitful.

There is a lot of room for improvement in this project. Firstly, it seems imprecise to regard all the SNPs which could not be safely defined as conserved SNPs (did not have top conservation scores) as non-conserved SNPs. Moreover, as described in the methods, the process of building and filtering this Clupeiformes alignment excluded many non-coding bases (about 44% of the herring genome) because there was no significant alignment across Clupeiformes species. This effect is documented in the bar plot (Figure 1B): the exons constitute the annotation group most disproportionately overrepresented in the Clupeiformes alignment, followed by UTRs, which is the same pattern as found in a research on evolutionarily conserved elements in vertebrate and many other genomes (Siepel *et al.* 2005). This appeared as a sound basis for this project but also brought out a question: would it be better to mark all bases without conservation scores as non-conserved instead of removing them completely from the dataset for downstream analyses as what I did in this project? According to calculation, I have lost almost 50% of SNPs due to the exclusion of bases outside the Clupeiformes alignment. These missing SNPs fall in highly variable genomic regions and are unlikely to have high conservation scores, so losing them might make the comparisons and statistical test between conservation scores of SNPs related to ecological adaptation and randomly selected bases conservative. To sum up, the classification of conserved and non-conserved SNPs might require more consideration and would probably have affected the results.

Acknowledgements

I am very thankful for my three supportive supervisors, Professor Leif Andersson, Dr. Sabine Felkel, and Researcher Mats Pettersson. Leif set clear goals for me and continuously guided me through them. Sabine built the multiple sequence alignment of Clupeiformes and generated phastCon, phyloP and GERP++ scores based on it, so I could focus on the downstream analyses using these great annotations of conservation scores. Not only did she prepare me well for this project, she also answered a lot of questions from me, solving my doubts patiently. Every time when I presented my results on the group meeting, Leif and Mats always helped me improve my interpretations and gave me great suggestions on further digging into them. Moreover, I am very grateful to Andreas Wallberg for giving me valuable comments and helping me revise my thesis. Finally, many thanks to my families, my boyfriend, my friends and my cat for all the love and strength they gave me.

References

Altshuler D, Daly MJ, Lander ES. 2008. Genetic Mapping in Human Disease. Science 322: 881–888.

Boyce DG, Petrie B, Frank KT. 2021. Fishing, predation, and temperature drive herring decline in a large marine ecosystem. Ecology and Evolution 11: 18136–18150.

Chen J, Bi H, Pettersson ME, Sato DX, Fuentes-Pardo AP, Mo C, Younis S, Wallerman O, Jern P, Molés G, Gómez A, Kleinau G, Scheerer P, Andersson L. 2021. Functional differences between TSHR alleles associate with variation in spawning season in Atlantic herring. Communications Biology 4: 795.

Christmas MJ, Kaplow IM, Genereux DP, Dong MX, Hughes GM, Li X, Sullivan PF, Hindle AG, Andrews G, Armstrong JC, Bianchi M, Breit AM, Diekhans M, Fanter C, Foley NM, Goodman DB, Goodman L, Keough KC, Kirilenko B, Kowalczyk A, Lawless C, Lind AL, Meadows JRS, Moreira LR, Redlich RW, Ryan L, Swofford R, Valenzuela A, Wagner F, Wallerman O, Brown AR, Damas J, Fan K, Gatesy J, Grimshaw J, Johnson J, Kozyrev SV, Lawler AJ, Marinescu VD, Morrill KM, Osmanski A, Paulat NS, Phan BN, Reilly SK, Schäffer DE, Steiner C, Supple MA, Wilder AP, Wirthlin ME, Xue JR, Zoonomia Consortium, Birren BW, Gazal S, Hubley RM, Koepfli K-P, Marques-Bonet T, Meyer WK, Nweeia M, Sabeti PC, Shapiro B, Smit AFA, Springer MS, Teeling EC, Weng Z, Hiller M, Levesque DL, Lewin HA, Murphy WJ, Navarro A, Paten B, Pollard KS, Ray DA, Ruf I, Ryder OA, Pfenning AR, Lindblad-Toh K, Karlsson EK. 2023. Evolutionary constraint and innovation across hundreds of placental mammals. Science 380: eabn3943.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6: 80–92.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLOS Computational Biology 6: e1001025.

Fisher RA. 1919. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth and Environmental Science Transactions of The Royal Society of Edinburgh 52: 399–433.

Han F, Jamsandekar M, Pettersson ME, Su L, Fuentes-Pardo AP, Davis BW, Bekkevold D, Berg F, Casini M, Dahle G, Farrell ED, Folkvord A, Andersson L. 2020. Ecological adaptation in Atlantic herring is associated with large shifts in allele frequencies at hundreds of loci. eLife 9: e61076.

Hickey G, Paten B, Earl D, Zerbino D, Haussler D. 2013. HAL: a hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics 29: 1341–1342.

Hill J, Enbody ED, Pettersson ME, Sprehn CG, Bekkevold D, Folkvord A, Laikre L, Kleinau G, Scheerer P, Andersson L. 2019. Recurrent convergent evolution at amino acid residue 261 in fish rhodopsin. Proceedings of the National Academy of Sciences 116: 18473–18478.

Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. 2014. Identifying causal variants at loci with multiple signals of association. Genetics 198: 497–508.

Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. Briefings in Bioinformatics 12: 41–51.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, Birney E, Searle S, Schmutz J, Grimwood J, Dickson MC, Myers RM, Miller CT, Summers BR, Knecht AK, Brady SD, Zhang H, Pollen AA, Howes T, Amemiya C, Lander ES, Di Palma F, Lindblad-Toh K, Kingsley DM. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484: 55–61.

Lamichhaney S, Barrio AM, Rafati N, Sundstrom G, Rubin C-J, Gilbert ER, Berglund J, Wetterbom A, Laikre L, Webster MT, Grabherr M, Ryman N, Andersson L. 2012. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proceedings of the National Academy of Sciences 109: 19345–19350.

Lamichhaney S, Fuentes-Pardo AP, Rafati N, Ryman N, McCracken GR, Bourne C, Singh R, Ruzzante DE, Andersson L. 2017. Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. Proceedings of the National Academy of Sciences 114: E3452–E3461.

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.

Larsson LC, Laikre L, Palm S, André C, Carvalho GR, Ryman N. 2007. Concordance of allozyme and microsatellite differentiation in a marine fish, but evidence of selection at a microsatellite locus. Molecular Ecology 16: 1135–1147.

Martinez Barrio A, Lamichhaney S, Fan G, Rafati N, Pettersson M, Zhang H, Dainat J, Ekman D, Höppner M, Jern P, Martin M, Nystedt B, Liu X, Chen W, Liang X, Shi C, Fu Y, Ma K, Zhan X, Feng C, Gustafson U, Rubin C-J, Sällman Almén M, Blass M, Casini M, Folkvord A, Laikre L, Ryman N, Ming-Yuen Lee S, Xu X, Andersson L. 2016. The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. eLife 5: e12081.

Masel J. 2011. Genetic drift. Current Biology 21: R837–R838.

Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, Fumis L, Hayhurst J, Buniello A, Karim MA, Wright D, Hercules A, Papa E, Fauman EB, Barrett JC, Todd JA, Ochoa D, Dunham I, Ghoussaini M. 2021. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nature Genetics 53: 1527–1533.

Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. 2011. Cactus: Algorithms for genome multiple sequence alignment. Genome research 21: 1512–28.

Pettersson ME, Rochus CM, Han F, Chen J, Hill J, Wallerman O, Fan G, Hong X, Xu Q, Zhang H, Liu S, Liu X, Haggerty L, Hunt T, Martin FJ, Flicek P, Bunikis I, Folkvord A, Andersson L. 2019. A chromosome-level assembly of the Atlantic herring genome—detection of a supergene and other signals of selection. Genome Research 29: 1919–1928.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research 20: 110–121.

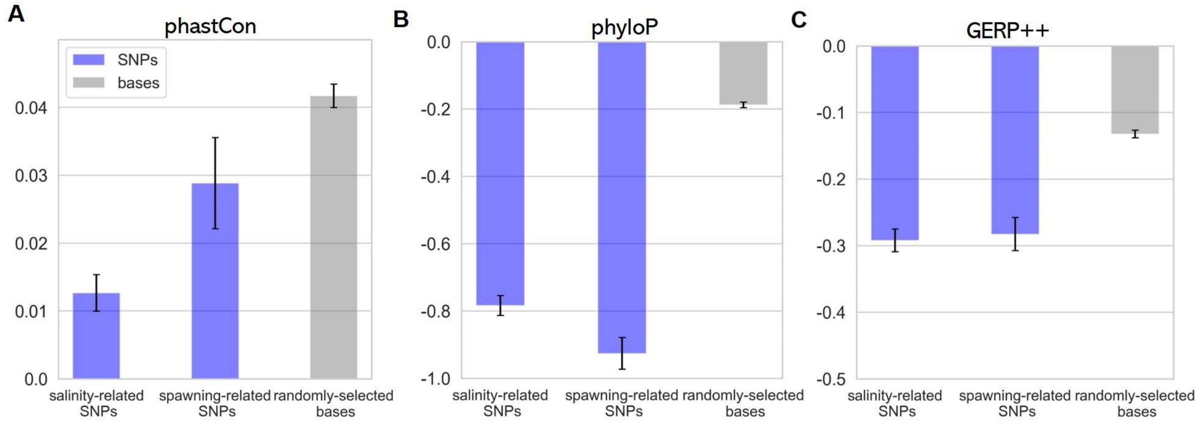
Ramani R, Krumholz K, Huang Y-F, Siepel A. 2019. PhastWeb: a web interface for evolutionary conservation scoring of multiple sequence alignments using phastCons and phyloP. Bioinformatics (Oxford, England) 35: 2320–2322.

Ryman N, Lagercrantz U, Andersson L, Chakraborty R, Rosenberg R. 1984. Lack of correspondence between genetic and morphologic variability patterns in Atlantic herring (*Clupea harengus*). Heredity 53: 687–704.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15: 1034–1050.

Xue JR, Mackay-Smith A, Mouri K, Garcia MF, Dong MX, Akers JF, Noble M, Li X, ZOONOMIA CONSORTIUM, Lindblad-Toh K, Karlsson EK, Noonan JP, Capellini TD, Brennand KJ, Tewhey R, Sabeti PC, Reilly SK. 2023. The functional and evolutionary impacts of human-specific deletions in conserved elements. Science 380: eabn2253.

Appendix



Supplementary Figure 1. Mean values of conservation scores of SNPs related to ecological adaptation and randomly selected bases

Bar plots illustrated the mean values of phastCon scores **(A)**, phyloP scores **(B)** and GERP++ scores **(C)** of 1150 SNPs related to salinity (left), 522 SNPs related to spawning-time (middle) and 10,000 randomly selected bases (right). Error bars showed standard error of the mean. Conservation scores of SNPs related to ecological adaptation were colored in blue while conservation scores of randomly selected bases were colored in grey.

Supplementary Table 1. Normality test and Levene test on conservation scores from different sample of SNPs

Dataset	Sample size	Score type	Normality Test		Levene Test	
			H_0	Result	H_0	Result
Salinity-related (A) vs Others (B)	$n_a = n_b =$ 1150	phastCon	a,b has a normal distribution	reject, ***	$Var_a = Var_b$	reject, ***
		phyloP		reject, ***		reject, ***
		GERP++		reject, ***		cannot reject, NS
Spawning-related (C) vs Others (D)	$n_c = n_d =$ 522	phastCon	c,d has a normal distribution	reject, ***	$Var_c = Var_d$	cannot reject, NS
		phyloP		reject, ***		reject, ***
		GERP++		reject, ***		cannot reject, NS
Salinity/ Spawning-related (E) vs Others (F)	$n_e = n_f =$ 1507	phastCon	e,f has a normal distribution	reject, ***	$Var_e = Var_f$	reject, ***
		phyloP		reject, ***		reject, ***
		GERP++		reject, ***		cannot reject, NS

NS, P>0.1; *, 0.05>P⩾0.01; **, 0.01>P⩾0.001; ***, P< 0.001.

A, B, C, D, E and F represent the populations of six sets of bases while a, b, c, d, e and f represent the samples from the corresponding population, except for A=a, C=c and E=e.

The results with the highest number of occurrences out of ten replicate tests were displayed in the table.

Var means variance.