Widespread but spotty: Atlantic cod is a complex of five species through ecological speciation

Katrín Halldórsdóttir[1], Alejandro Llanos-Garrido[2], Axel Wilhelm Einarsson[1], Tim Sackton[3] and Einar Árnason[1,4*]

[1*]Institute of Life- and Environmental Sciences, University of Iceland, Sturlugata 7, Reykjavík, IS-102, , Iceland.
[2]Evolution and Conservation Biology Research Group, Complutense University, Av. Séneca, 2, Madrid, 28015, Madrid, Spain.
[3]Informatics Group, Faculty of Arts and Sciences, Harvard University, 26 Oxford Street, Cambridge, 02138, Massachusetts, USA.
[4]Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, 02138, Massachusetts, USA.

*Corresponding author(s). E-mail(s): einararn@hi.is;
Contributing authors: katrinh@hi.is; a.llanos@ucm.es; awe1@hi.is; tsackton@g.harvard.edu;

Abstract

Marine environments are home to a significant portion of fish and other organisms' biodiversity. However, our knowledge and empirical evidence of speciation in the marine realm is quite limited. Here we use genomics and study population differentiation and speciation in the Atlantic cod, a marine teleost of biological and economic importance. The Atlantic cod genome is liable to recurrent selective sweeps showing the importance of selective processes [1]. We show that the non-inversion part of the genome separates individuals into five clusters. At the four chromosomes carrying large inversions that define ecotypic variation, individuals also cluster into the same five clusters in conjunction with individuals' inversion genotypes. Within each cluster, individuals homozygous or heterozygous for

the inversions are divergent and there is an adaptive divergence of each type of inversion homozygote among clusters. These interchromosomal correlations imply the reproductive isolation of five cryptic species. We show that the inversion haplotypes are ancient [2] trans-species balanced polymorphisms that would evolve towards homozygous lethality under Muller's ratchet [3–5]. This and recombination suppression of inversions. However, gene-flux [6] from gene conversion [7] rescues homozygous fitness, and recombination and selection lead to the evolution of habitat specialists through ecological speciation [8, 9]. The specialized genotypes actively seek out their preferred habitat resulting in a widespread but patchy distribution that contradicts the traditional geographical speciation model and the stock concept in fisheries biology. Our results highlight the power of genomics in studying selection and speciation while informing conservation and management of essential fisheries resources.

Keywords: Speciation, ecology, selection, recombination, adaptation, Atlantic cod, trans-species balanced polymorphism

1 Introduction

Geography and geographical barriers to gene flow play an important role in models of population differentiation and geographic speciation where reproductive isolation arises as a by-product of divergence under allopatry [10]. In contrast, ecological speciation models [8, 9] propose that the primary drivers of phenotypic divergence are natural and sexual selection occurring in either allopatric or sympatric populations. In these models, the divergence of reproductive traits occurs as a consequence of the divergence and specialization of ecological traits facilitated by mechanisms such as pleiotropy or genetic linkage. Genomics often shows 'genomic islands of differentiation', clusters of loci that show large differentiation among populations in a sea of less differentiated genetic regions. Some are thought to be 'genomic islands of speciation', divergence under strong selection in the face of gene flow. The islands are also often associated with chromosomal structural variation, such as inversions that suppress recombination. Recombination in meiosis of inversion heterozygotes may lead to aneuploid gametes and lower fitness of heterozygotes (hybrids). Inversion may thus facilitate the evolution of post-mating isolation. Inversions can be ancestral polymorphisms associated with reproductive isolation via incomplete lineage sorting [11] or they can be trans-species balanced polymorphisms [12, 13].

The marine environment has few if any, geographic barriers, but isolation by distance is often thought to be important. Eggs and larvae can be dispersed over large areas by currents and some organisms, such as fish, are highly mobile and can actively search for suitable habitats, even over large distances. However, cryptic species are common in the sea [14, 15] and speciation may involve differences in chemical mate recognition, gametic incompatibility, and

specialization in adaptive habitat utilization [16] influenced by sexual selection [9].

Marine fish species of significant commercial value often contribute to important fisheries managed as distinct stocks. The stock structure [17] is a geographical concept derived in one way or another from the geographical (allopatric) speciation model. Fish stocks are considered subpopulations of a particular species with a particular ecology and life-history parameters (growth, reproduction, and mortality). The concept implies that a particular stocks is discrete and isolated from other stocks of the species and, therefore, it is independent and self-sustaining. Potential cryptic units are largely ignored, and stocks are managed using a geographical stock concept with regional or nation-state boundaries.

The genome of Atlantic cod (*Gadus morhua*) is subject to recurrent selective sweeps [1] showing that selective processes are of great importance. Atlantic cod and its related Pacific and Arctic gadid relatives [15] exhibit many features related to the above considerations of population differentiation. Four chromosomes carry large inversions related to ecotypic variation [18, 19] of deep-water-migratory and shallow-water-stationary behavioral ecotypes [20] as well as behavioral ecotypes of habitat choice related to salinity, oxygen, and temperature [21]. The chromosomal inversions are 'islands of divergence' [22] that are likely ancient [2] supergenes under balancing selection [15, 23, 24]. Additional complexity is also seen. For example, the presence of a distinct hybrid zone between Baltic cod and Danish Belt Sea cod [25] and with natural selection and adaptation to the specific conditions of the Baltic Sea contribute to reproductive isolation between the two populations [21]. Furthermore, certain localities have unusually high genetic diversity as demonstrated by variations in microsatellites in the North Sea [26] and biocomplexity of SNP variation at Western Greenland [27]. There are reports of parallel clines in the western and eastern Atlantic suggestive of selection [28] and of cryptic speciation [15, 29]. But no clear patterns have emerged for the variation in general.

Here we report findings from our 'two thousand-genomes cod project.' We have sequenced at medium to high coverage whole genomes of Atlantic cod from throughout its distributional range in the North Atlantic Ocean as well as other gadid taxa, the Pacific cod (*G. macrocephalus*) and walleye pollock (*G. chalcogrammus*) of the Pacific Ocean, Greenland cod (*G. ogac*) at western Greenland, and Polar cod (*Arctogadus glacialis*) and Arctic cod (*Boreogadus saida*) of the Arctic ocean. We provide conclusive answers to patterns of variation within Atlantic cod and lay the foundation for further studies of selection, recombination, speciation, and adaptation. We also reveal patterns and processes important for the conservation and management of important fisheries resources.

2 Results

2.1 Population resequencing across the codfish range

We sequenced 1,668 Atlantic cod from across the North Atlantic. We had reasonably large samples from most localities ranging clockwise from Cape Cod to the Faroe Islands (Fig. B1), including the largest sample of 828 individuals from Iceland (marginal sum in Table A1). Average sequencing depth was 10.8 (average of mean coverage) and ranged from 2.3 to 94.2; individuals with less than 2× depth or more than 0.1 missing variant calls were removed. We genotyped 17,886,971 SNPs in total across the genome after filtering with GATK best practices [30], for an estimated average pairwise nucleotide diversity (in 1 Mb windows) $\theta_\pi = 0.00203$ (0.00112 and 0.00257 for the 5% and 95% quantiles) of the 19 non-inversion chromosomes combined implying an effective population size of 1.4×10^5 (7.7×10^4 to 1.8×10^5 for 5% and 95% quantiles respectively; assuming a mutation rate 3.62×10^{-9} per nucleotide per generation [1]). This is similar to effective population size 1.3×10^5 to 4.3×10^5 for a locally circumscribed population of *R* fish on the South/south-east coast of Iceland [1]. The effective population size is many orders of magnitude lower than census size [1] but large enough that weak selection is effective.

2.2 Local structure of genome

Previous work has characterized large inversions on four chromosomes, Chr01, Chr02, Chr07, and Chr12 [18, 19]. To verify the presence of these inversions and detect new inversions, we computed local structure [31] across the genome. Briefly, the method divides each chromosome into segments, performs PCA on each segment, estimates the pairwise distance between all segment-pair PCA results, and performs multidimensional scaling on the resulting pairwise distance matrix [31]. Every chromosome showed a local structure (Fig. B2). The `lostruct` analysis [31] showed that pairwise distances between chromosomal segments clustered in two or more groups. Four chromosomes stood out, Chr01, Chr02, Chr07, and Chr12, showed evidence for large inversions that have already been described [18, 19] and are related to ecotypic variation. Other chromosomes showed many peaks covering smaller chromosomal segments (Fig. B2). These may represent smaller chromosomal structural variation, such as small inversions, or they may represent population structure due to selection or drift. The chromosomal segments outside of the large inversions also showed multiple peaks for multidimensional scaling coordinate 2 (MDS2, Fig. B2) possibly related to adaptive differentiation of the haplotypes of each inversion.

2.3 Non-inversion chromosomes and population structure

In this study, we classified the four chromosomes containing large inversions as 'inversion chromosomes,' while the remaining 19 chromosomes not carrying large inversions (Fig. B2) were referred to as 'non-inversion chromosomes.' To assess population differentiation, we utilized principal component analysis (PCA) and admixture analysis [32]. We specifically distinguished between the

two sets of chromosomes mentioned earlier. The Atlantic cod genome is subject to recurrent selective sweeps [1]. Variation at inversion chromosomes may be maintained at intermediate frequencies due to balancing selection [1, 15, 24], while variation at non-inversion chromosomes may better reflect genome-wide effects of demographic processes. Although the non-inversion chromosomes may be distorted due to linked selection from the recurrent sweeps in the Atlantic cod genome, the decay of linkage disequilibrium is also very high [1].

2.3.1 Principal components analysis of non-inversion chromosomes

Principal components analysis (Fig. 1) showed that individuals clustered into five groups that we refer to as PCA group *R*, *S*, *SS*, *T*, *U*. The *SS* group is associated with shell-sand seabed (hence *SS*) on the Faroe Bank [33], and it is also a predominant group in the Breidafjordur Bay in Iceland close to the fishing ground for the Iceland clam (*Chlamys islandica*). This apparent correlation with a type of seabed may indicate specific habitat requirements or choices. The other groups were named by single letters around the *SS*. The groups were defined by clustering individuals on the first four principal components. Thus *U* was separate from the rest on PC2 and *S*, *SS*, and *T* separated on PC3 and PC4. The largest *R* group comprising 76% of individuals was very tight-knit. On PC3 and PC4 (but not on PC1), this group was centered on coordinate 0,0, which has been thought to indicate missingness of data [34]. However, even 90% missingness can allow recovery of population structure [35]. Furthermore, the centering of the largest cluster to the center of the multidimensional space may be but the pull to the center by the large size of the *R* cluster, as our data were already screened for missingness (0.1). A PCA with a subsample of the *R* group was not centered on coordinate 0,0 (Fig. B3) suggesting that the size of the cluster affects the location. The admixture analysis of latent clusters (below) also supported the reality of these groups. Overall the first four PCs explained 76% of the variation (Fig. 1) and PC5 and PC6 added another 9%. The patterns of five groups were also seen for each of the individual non-inversion chromosomes (Fig. B4) so the observed pattern of the combined data (Fig. 1) was not due to averaging out individual differences but a general genomic pattern.

2.3.2 Admixture analysis of non-inversion chromosomes

Admixture analysis [32] also revealed the five groups detected by PCA with additional details (Fig. A1 and Fig. B5 with less data thinning). Importantly the tight knit *R* group seen in PCA splits into six well-defined clusters at $k = 10$ ancestral populations while the other PCA groups remained distinct. All ten replicate runs of admixture showed similar results and pong [36] detected only a single mode for all replicate runs at each k. Individual chromosomes showed much the same pattern (Fig. B6). The differentiation among admixture populations measured by F_{ST} (Table A2) was 0.07 to 0.12 among the five PCA

groups (with the T group most divergent) and F_{ST} of 0.01 to 0.02 among the population clusters within the R group (with pop6 frequent in the western Atlantic most divergent; Fig. B6).

2.3.3 Spotty but widespread geographical distribution of PCA groups

The geographical distribution of individuals belonging to the five PCA groups was widespread but spotty (Fig. 2). For example the S group fish were found in the Baltic (particularly the eastern Baltic), the North, Celtic, and Irish Seas, the White Sea, and the Barents Sea, a spot in south Iceland and at Cape Cod on the east coast of north America. Similarly, the SS group was found in the western Baltic Sea, on the Faroe Bank (where it is associated with a seabed of shell sand) but not on the Faroe Plateau, in one spot on the south coast of Iceland and the Breidafjord Bay on the western part of the island (where the seabed has an abundance of Iceland scallop *Chlamys islandica*) as well as in several other spots in the North Atlantic. A similar pattern holds for the other groups. R was the most abundant group and was found in all localities except the White Sea where only S and SS were found. The western Atlantic (Nova Scotia and Newfoundland) were depauperate, predominantlty R and very rare T group (Table A2). Fishing has decimated populations in these areas [37] and the relative abundance of the groups may be affected by fishing pressure. The widespread but spotty distribution suggests either habitat fragmentation or very specific habitat requirements or habitat choice, a specialization and adaptation to narrow niches. Estimating the relative abundance of different habitat specialists would require the same sampling effort in different patches. The sampling was done without prior knowledge of spottyness and, therefore, the observed relative frequency (Table A2) may not represent the relative abundance in nature.

2.3.4 Geographic distribution of clusters within PCA groups

Admixture showed that populations (latent clusters) within the R group were widely distributed and spotty but also unevenly distributed (Fig. A1, bottom). Thus, pop6, found in Iceland and several eastern localities (including Borgenfjord but not in Verrabotn (two adjacent infjords in Norway), was the most common type in the western localities. In contrast, pop5, found in the West, was more frequent in the eastern localities (Table A1). Asking if there is evidence within the R group for simple isolation by distance, under which frequency of a type decreases monotonically with geographic distance from the source, the answer is no. For example, the relative frequency of pop6 is U-shaped going from Nova Scotia/Newfoundland via Iceland to Borgenfjord Norway (Fig. A1; Table A1). Admixture analysis at a finer scale with $k = 2$ to $k = 20$ reveals more groups within PCA group R but is similarly ambiguous with respect to patterns related to geography (Figs. B7 and B8 and lack of a clear signal of isolation by distance Table B1).

Principal component analysis of the PCA groups also showed a complex connection to geography (Fig. A2). Thus within the *R* group, the western cod individuals (Cape Cod, Nova Scotia, and Newfoundland) were somewhat separate. Still, a fair number also overlapped with the Barents Sea and Iceland samples. Verrabotn, an in-fjord way inside the long Trondheim fjord in Norway, overlaps with the Faroe Islands across the Norwegian Sea and with Irish Sea samples but is distinct from Borgenfjord, which is a neighboring in-fjord in the Trondheim fjord. Borgenfjord, in turn, overlaps with some Icelandic samples (Fig. A2, **a**). Similarly, in the *S* group, Iceland samples are surrounded by Cape Cod samples, and Baltic and Barents Sea samples are close together (Fig. A2, **b**). The *SS*, *T*, and *U* groups also show the complexity of the connection between geography and genetic variation of separation and overlap. Overall, there was some effect of geography. However, there are no clear signs of expansion tracts from refugia [38] or isolation by distance in a regular manner. Only very convoluted migration paths could account for patterns with distance. Therefore, it is unclear whether geography or distance is an isolating factor or if the geographic locality is a proxy for some specific environmental conditions important for habitat choice.

2.4 Inversion chromosomes and genotypic variation

Principal component analysis of variation inside of the inversions of the inversion chromosomes (genetic variation that defines ecotypic variation) showed a complex pattern of clusters. Using a novel but simple method, we generated the composite genotypes of the inversion haplotypes or haplotypic genotypes (Appendix B.4, Figs. B9, B10, B11, B12, B13, and B14) in a robust way (Fig. B15). The five PCA groups defined by variation at the non-inversion chromosomes (Fig. 1) along with the respective haplotypic genotypes fully explained the clustering observed at the four inversion chromosomes (Figs. 3 and A3). Admixture analysis also showed the complexity seen in the PCA (Figs. B16, B17, B18, and B19). The homozygotes for the reference haplotype (blue) and the homozygotes for the alternative haplotype (red) clustered by PCA group (for Chr01, Fig. 3) showing divergence of genetic variation defining ecotypes (and similarly for Chr02, Chr07, and Chr12, Fig. A3). Thus the PCA groups each had its own unique variation and there was differentiation within and among haplotypic genotypes of the PCA groups and heterozygous genotypes were frequently not exactly intermediate between the respective homozygotes (Fig. A4).

2.4.1 Recombinational breakup of inversions

There was an apparent recombinational breakup of the inversions, a gene-flux [6] possibly from double-crossovers but more likely gene conversion [7]. The genotyping (Appendix B.4) showed that individuals that were called with a composite homozygous genotype frequently were heterozygous for individual sites. Similarly, individuals that were called with a composite heterozygous

genotype frequently were homozygous for individual sites (Fig. B9). Often adjacent or close sites in individuals that were called heterozygotes were homozygous for both the reference and alternative homozygote (red and blue, Fig. B9, top left) suggesting independent recombination events. There were 643 diagnostic sites covering 17 Mb of inversion on Chr01 (Fig. B9, top left), and thus, a site covers roughly a 25 kb segment on average. By the size of segments and independence of events, we suggest that these are signs of gene conversion tracts rather than double crossovers. In *Drosophila* the estimated size of gene conversion tracts is tens to a couple of thousand bp [Table S4 7]. If the size distribution of Atlantic cod gene conversion tracts is similar to *Drosophila* our observations, taking a peek at roughly 25 kb intervals, has revealed only the tip of the iceberg. There was a continuous distribution of numeric genotypic score with three modes (Fig. B9, bottom right). Principal component analysis showed that the numeric genotype score of heterozygotes of the four inversion chromosomes covered the genic space between the homozygotes and also extended beyond intermediateness (Fig. 4). There was a general heterozygote excess at the four inversions chromosomes (Table B2, although not significant for Chr12) overriding any potential Wahlund effect of population structure expected from large genotype frequency differences among localities (Appendix B.5, Figs. B20, B21, B22, B23, and B24). Overall, there were significant genotypic associations among the inversion chromosomes (Table B3). This effect was, however, not found considering a local Icelandic population that had been screened for population structure [1] (Table B4). Therefore, the the overall association was likely due to population structure rather than direct epistatic interactions. The frequency differences of the various genotypes among localities (Appendix B.5, Figs. B20, B21, B22, B23, and B24) supports this interpretation. This analysis also shows how heterozygotes can be close to the predominant homozygote but also extend beyond intermediateness for all inversion chromosomes and the five PCA groups.

2.4.2 Trans-species and trans-oceanic supergenes

Maximum likelihood phylogenetic trees of individuals homozygous for alternative inversion haplotypes of the five PCA groups and the Arctic and Pacific gadid taxa revealed a trans-species [2] and trans-oceanic relationships (Fig. 5 and Appendix B.6 and Fig. B25) that was not observed among representative non-inversion chromosomes (Fig. B26). Reference-allele homozygotes were in one clade and alternative-allele homozygotes in another clade for all four inversion chromosomes. Yet there was also divergence among the PCA groups and gadid taxa within each homozygous genotype representing an inversion haplotype (Fig. 5). This is similar to the divergence among homozygotes seen in PCA (Figs. 3 and A3). There were many other noteworthy patterns such as apparent paraphyly of Pacific cod and Greenland cod, the clustering of PCA group T with Pacific cod and Greenland cod, and the clustering of PCA group U and walleye pollock. The walleye pollock was the least polymorphic, and only a single homozygote for the reference allele was found for Chr07. The

walleye pollock thus has a reduced frequency of the reference allele for Chr01, Chr02, Chr07 but not for the Chr12 inversion (Fig. 5).

Treemix analysis showed similar trans-species and trans-ocean relationships of the inversion chromosomes (Figs. B27, B28, B29, and B30) and not for the sample of non-inversion chromosomes (Figs. B31, B32, and B33). Here the individuals homozygous for alternative inversion haplotypes (Fig. 5) are lumped together to form a population. Treemix also showed evidence for introgression, often between the two homozygotes within a PCA group or species, but also between taxa. For example, there was evidence for migration between U and walleye pollock, between T and Pacific cod/Greenland cod, and polar cod and PCA group S, some of which were also seen for the non-inversion chromosomes.

3 Discussion

We show that the non-inversion part of the Atlantic cod genome, which is not known to be associated with ecotypic differences, is differentiated into (at least) five groups. Population variation and differentiation of the four inversion chromosomes that code for ecotypic variation important for differential niche utilization is accounted for by these five groups account in conjunction with inversion genotypes. The evolution of these groups and genome-wide interchromosomal correlations can only have happened in reproductively isolated units. Thus, the Atlantic cod is a complex of at least five hitherto cryptic species [14, 39, 40]. We further show that the inversion supergenes [15, 23, 24], 'islands of divergence' related to ecotypic variation, are ancient trans-species and trans-ocean balanced polymorphisms that are older than speciation of the Atlantic, Arctic, and Pacific gadid clade [2].

Previously we showed that recurrent selective sweeps permeate the Atlantic cod genome, a selection that may be chasing ephemeral optima but also driving longer-duration adaptive traits such as those of balanced polymorphisms [1]. The following ecological speciation [8, 9] scenario is fully compatible with our results. Inversion haplotypes (e.g. A and B) will rapidly diverge due to recurrent selective sweeps within each [1]. Deleterious mutations will hitch-hike, drift due to lowered effective population size of a homokaryotypic population of chromosomes, fix, and accumulate in both homozygotes (e.g. AA and BB) sinking them to lethality with Muller's ratchet [3–5]. This process leads to increasing overdominance [5] as suggested by a general heterozygote excess. In the absence of recombination this system would evolve to become a balanced lethal system [41]. However, rampant gene-flux [6] likely due to gene conversion [7] rescues homozygous fitness and allows heterozygotes and homozygotes, in turn to explore fitness space providing material for divergent selection leading to divergent haplotypes of ecologically relevant variation. The ecotypic haplotypes branch out, for example, with haplotype structuring of bimodal fitness distribution of homokaryotypes [5] (with free recombination within homokaryotypes), and diverge either allopatrically or sympatrically [42] by

natural selection and possibly also by sexual selection [43–45] into specialized niches and habitat specialization by depth, temperature, oxygen, salinity, type of bottom substrate, and so on. They then breed within their habitat of choice defined by fine-scaled environmental differences [46] so that their ecology defines reproductive isolation, a case of ecological speciation [8, 42, 47]. The specialized types actively seek out the preferred habitat that they are adapted to, resulting in a widespread but spotty geographic distribution.

Speciation-with-gene-flow [48, 49] was proposed to explain population differentiation in Atlantic cod [29, 50]. The model assumes low levels of differentiation in large parts of the genome because of ongoing gene flow homogenizing those genomic regions. Under this model, 'islands of genomic divergence,' genomic regions of high differentiation, are loci connected to isolating traits between the incipient species, speciation genes in effect. Introgression of islands is prevented by selection, but recombination homogenizes the rest of the genome. This model assumes the establishment of interchromosomal linkage disequilibrium, although chromosomes assort independently. Maintaining such a strong association at unlinked loci at different chromosomes implies that most recombinant genotypes are less viable than non-recombinants. Only genotypes with the correct allelic combination survive. Even with a model of only few islands participating the number of gametes that can form viable zygotes will be very small [51]. The high-fecundity Atlantic cod has high reproductive excess and and is to withstand substantial selection pressures [1, 52], however, such strong selective forces should lead to the evolution of effective reproductive barriers between the forms. A model of divergence-after-speciation [51] is an alternative to the model of speciation-with-gene-flow. Under this model, there is no differential gene flow among loci. Instead, reproductive isolation is rapid, nearly instantaneous, and complete, as in the scenario above. The divergence-after-speciation model explains the low level of divergence throughout most parts of the genome as shared ancestral polymorphisms. The more recent the split, the more homogenized the genomes. The elevated differentiation in the 'islands of divergence.' is explained as being due to heterogeneous natural selection over the genome but not due to heterogeneous migration rates. The genomic islands are loci of ecological adaptations. This model also explains linkage disequilibrium between islands on different chromosomes because ecological adaptation within species involving genes of different chromosomes necessarily results in inter-chromosomal correlations.

The above speciation scenario settles the major features of the results. But there are several other patterns in the data that will be the basis for further analysis. The F_{ST} differentiation of non-inversion chromosomes within the R group is similar to the differentiation of neutral loci in a sensory drive speciation in cichlids [53]. A rhodopsin polymorphism important for Atlantic cod vision at different depths and under turbid conditions [54] is inside the Chr01 inversion and several other opsin genes are located within the inversion as well. Is the differentiation of the R group a case of incipient speciation along similar lines [53]? The widespread, partly overlapping, but spotty distribution of the

species we have found, has a bearing on several previous findings. For example, a study of a hybrid zone of Baltic cod [25] possibly or even likely sampled primarily *S* in the eastern Baltic, *SS* in the transition zone, along with *R* in both localities. Unusually high genetic diversity of microsatellites in the North Sea [26] and pronounced biocomplexity of SNP variation at Western Greenland [27] are localities where four of our species co-occur. Parallel clines in the western and eastern Atlantic suggestive of selection [28] may have sampled cryptic units to different extents and suggestions of cryptic speciation [15, 29] accord well with our findings. The claim of 'Two races of cod at Faroe' [55] are understandable as the *SS* and *R* species found at the Faroe Bank. Our results may shed light on the enigma of the walleye pollock niche shift [15] which may have all but sorted out certain chromosomes. Paraphyly and trans-ocean relationship of deep lineages within the Pacific cod/Greenland cod duo suggests a more complex situation than the notion that Greenland cod is a recent re-invasion of Pacific cod into the Atlantic at Greenland [56]. The results of widespread and spotty distributions are at variance with an allopatric geographic mode of differentiation and also are a challenge for the geographical stock concept of fisheries management for cod. The relative abundance of the species detected here may be affected by fisheries [37] or sampling strategies, and fisheries-induced selection [57–59] may affect the groups to different extent and influence their relative abundance.

4 Conclusion

We conclude that the Atlantic cod is a complex of five cryptic species that have adapted to specific habitats. Trans-species balanced inversion polymorphisms subject to recurrent selective sweeps [1] and recombination are important in this process. One species split into five highlights the importance of genomics for elucidating cryptic biodiversity in the marine environment.

The results are important for conservation and management of important fisheries resources. Fisheries management based on a geographical stock concept will need revision taking habitat specificity and choice and multi-species into account. However, much remains to be done. What is the origination and evolutionary age of the trans-species balanced polymorphisms? Does hybridization and introgression play a role? Can we quantify selection and recombination involved in the processes described here [60, 61]? What specific genes are involved in the adaptation and speciation, and what is the role of potential structural variation at the non-inversion chromosomes? How old are these speciation events, and what is the role of migration and introgression in these processes? How can we simulate the complex selection and demography involved [62] to test the hypothesis quantitatively? Answering these and other such questions is beyond the scope of this paper. We will address them subsequently.

5 Figures

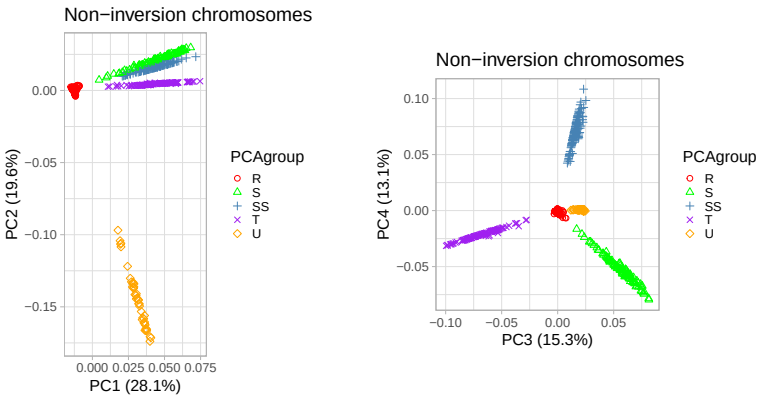

Fig. 1 Principal components analysis (PCA) of non-inversion chromosomes in Atlantic cod. PCA of the 19 non-inversion chromosomes combined (Chr03 to Chr06, Chr08-Chr11, and Chr13-Chr23) shows five groups *R*, *S*, *SS*, *T*, and *U*. PC1 to PC4 explain 76% of the variation. Bi-allelic SNPs with singletons excluded, LD pruned, and thinned to 450,953 SNPs.

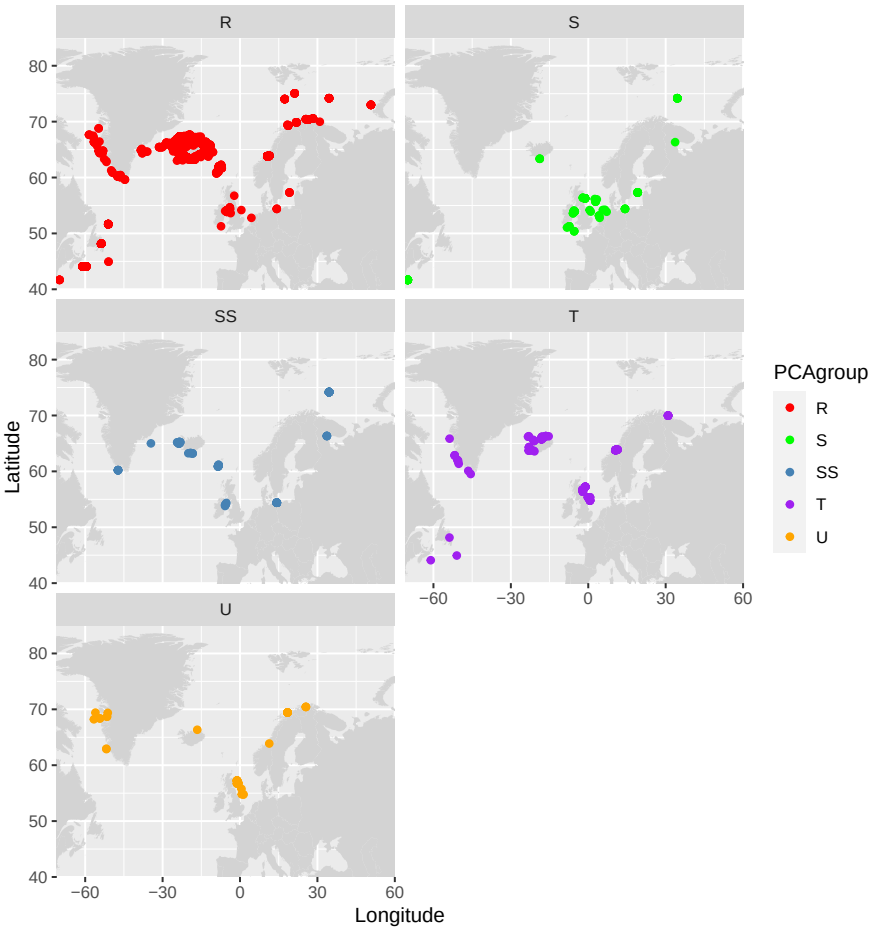

Fig. 2 Widespread but spotty. The geographic distribution of individuals from the five PCA groups on a map of the North-Atlantic.

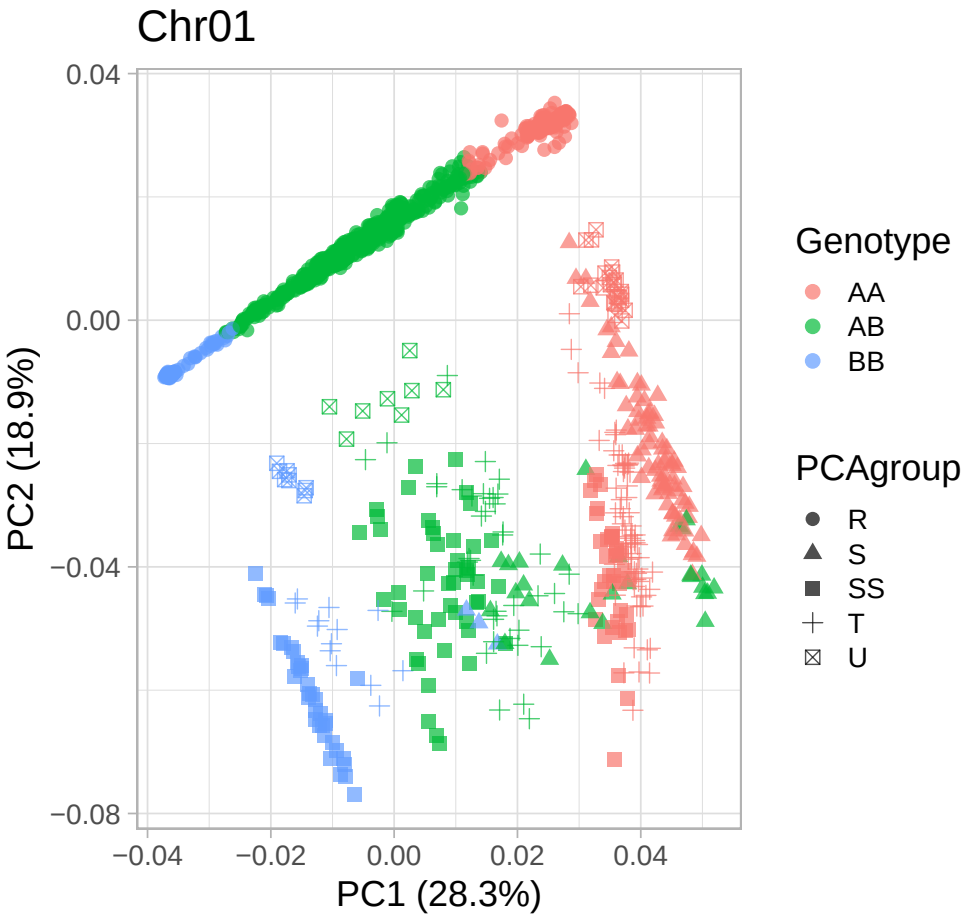

Fig. 3 Principal components analysis of ecotypic defining inversions on Chr01. PCA of inversion region of Chr01 between co-ordinates 11,763,336 and 27,936,178 labeled by PCA group (Fig. 1) and inversion genotype (Fig. B9). Bi-allelic SNPs with singletons excluded and LD pruned 1,260,547 SNPs.

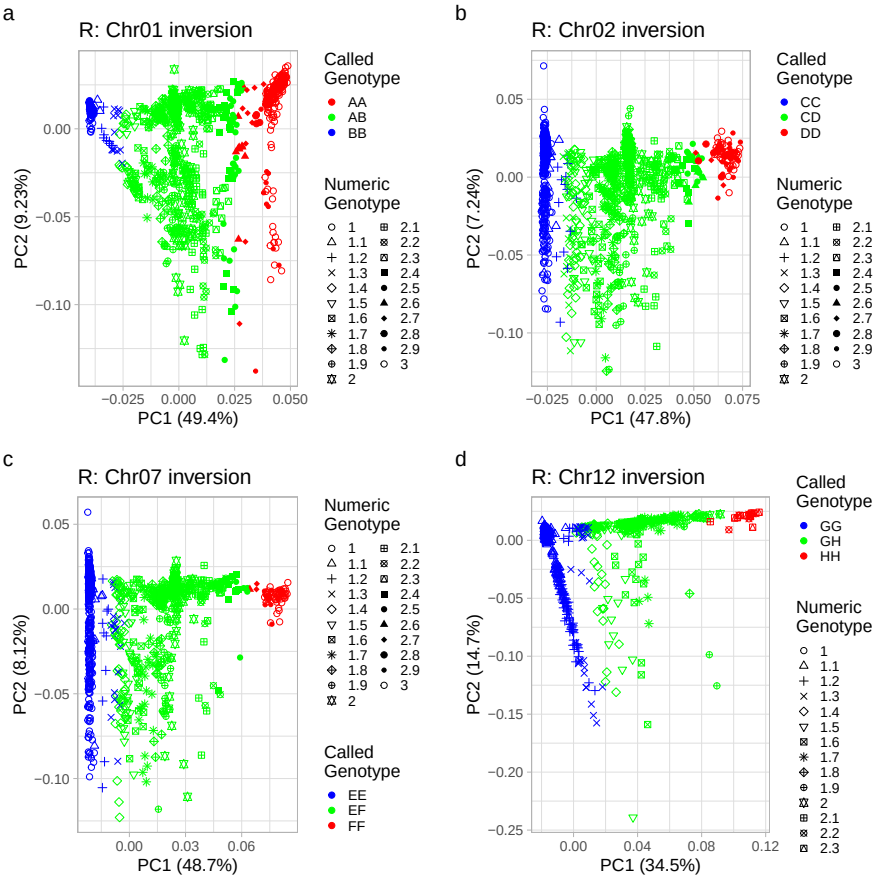

Fig. 4 Principal component analysis of the four inversion chromosomes of PCA group *R* labelled by called and numeric genotype. Called and numeric genotypes based on inversion genotyping (Fig. B9).

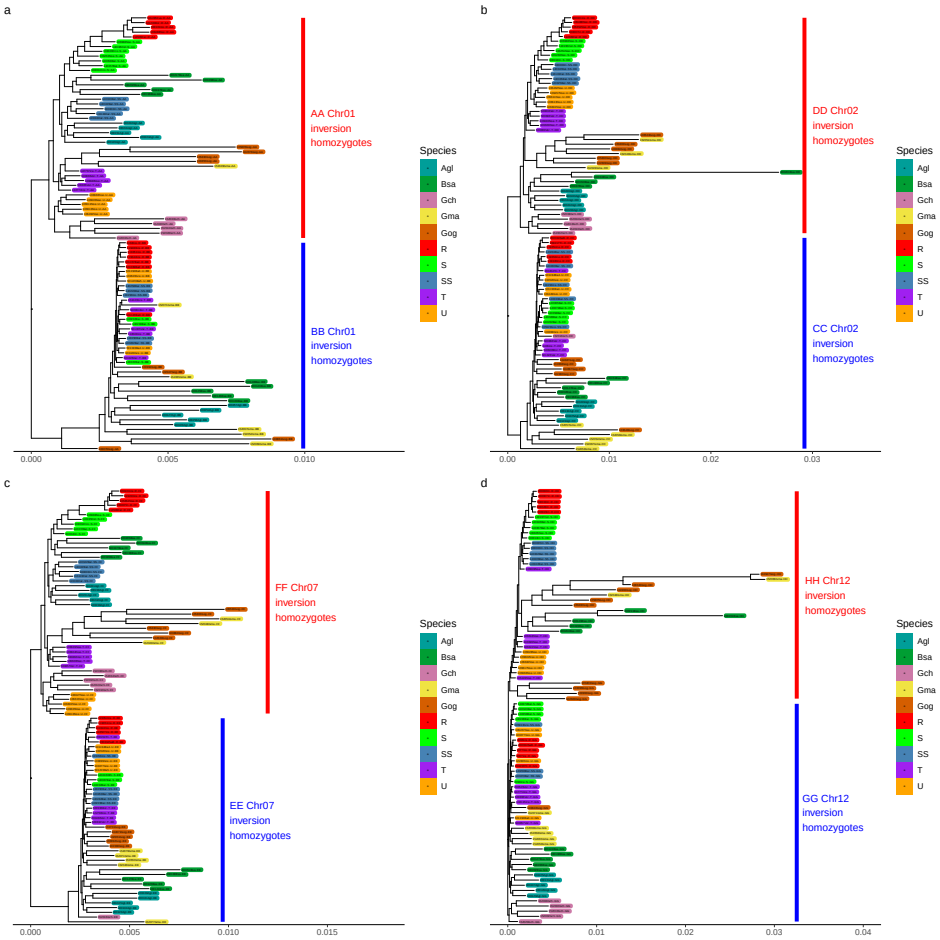

Fig. 5 Trans-species polymorphism of inversion haplotypes. Maximum likelihood phylogenetic trees of inversion homozygotes of PCA groups and related gadid taxa for Chr01 (**a**), Chr02 (**b**), Chr07 (**c**), and Chr12 (**d**). Best substitution model estimated with `ModelFinder` [63] and trees estimated with `iqtree` [64]. The PCA groups *R*, *S*, *SS*, *T*, and *U* of Atlantic cod (*Gadus morhua*) and the gadid taxa Polar cod (Agl, *Arctogadus glacialis*), Arctic cod (Bsa *Boreogadus saida*), walleye pollock (Gch *G. chalcogrammus*), Pacific cod (Gma *G. macrocephalus*), and Greenland cod (Gog *G. ogac*). The blue and red bars indicate homozygotes for the reference and the alternative inversion haplotype, respectively. Chromosomal coordinates inside inversions: 11,763,336 to 27,936,178 for Chr01, 579,097 to 4,096,513 for Chr02, 17,004,745 to 25,979,379 for Chr07, and 1,001,809 to 13,965,277 for Chr12. All trees are subtrees excluding outgroups from trees in Fig. B25.

6 Methods

6.1 Sampling

We obtained tissue samples from our large DNAfish tissue sample base that has been already been described [1, 15]. We sampled $n = 1688$ Atlantic cod from throughout its distributional range in the North Atlantic Ocean (Fig. 2). of Atlantic cod from throughout its distributional range in the North For comarison and phylogenetic analysis of inversion polymorphisms we sampled Pacific cod and walleye pollock of the Pacific Ocean, Greenland cod at western Greenland, and Polar cod and Arctic cod of the Arctic ocean. We used single specimens of haddock (*Melanogrammus aeglefinus*) and whiting (*Merlangius merlangus*) as outgroups for phylogenetic analysis.

6.2 Ethics statement

The DNAfish tissue sample collection was built with ethical board review as already described [1, 15].

6.3 Molecular analysis

Genomic DNA was isolated and shipped to the Bauer Core Facility at Harvard University. Libraries were made with Kapa HyperPrep Plus kit (Kapa Biosystems) for enzymatic DNA fragmentation and adapter ligation indexed using IDT (Integrated DNA Technologies) unique dual 8 bp indexes for Illumina, and sequenced on NovaSeq (Illumina) S4 flow cells with a 2×150 bp read format as already described [1, 15].

6.4 Bioinformatic analysis

The sequencing returned de-multiplexed `fastq` files for different runs of each individual. Data processing followed the Genome Analysis Toolkit (GATK) best practices [30] as implemented in the fastq_to_vcf pipeline of Alison Shultz (github.com/ajshultz/comp-pop-gen). Using the pipeline the raw reads were adapter trimmed using `NGmerge` [65], the trimmed `fastq` files aligned to the gadMor3.0 chromosome-level reference genome assembly (NCBIaccessionID:GCF_902167405.1) using `bwa mem` [66], and the resulting `bam` files deduplicated, sorted, and indexed with `gatk` [30] as already described [1, 15]. To go from `bam` to `VCF` files we used snpArcher https://github.com/harvardinformatics/snpArcher, which implements GATK best practices for variant calling [30].

6.4.1 Population analysis

For subsequent population-level analysis, we relied heavily on `samtools` and `BCFtools` [67], `VCFtools` [68], `plink` [69, 70], and `linux` command line tools (`bash`, `awk`, and `sed`) to manipulate data, select or combine individual chromosomes or inversion regions. We selected bi-allelic SNPs (`bcftools view`

`-m2 -M2 -v snps`). For analysis of the population structure and admixture, we first removed singletons (with minor allele count greater than 2 `plink --mac 2`) followed by LD pruning (`plink --indep-pairwise 50 10 0.1`) 50 kb segments with 10 kb window and $r^2 < 0.1$) followed by random thinning of remaining SNPs for computationally heavy analysis (`plink --thin 0.10`, keeping 10% of remaining SNPs, or `--thin 0.01` for even more thinning). We studied local heterogeneity in the genome and looked for inversion using the `lostruct` `R` package [31]. We studied population structure performing principal component analysis of various chromosomes or chromosomal regions, such as inversions. (PCA, `plink --pca`) and analyzed and plotted results with the `tidyverse` environment [71] in `R`. We inferred population structure and individual ancestries with `admixture` [32]. We ran 10 replicates for each $k = 2 \ldots 10$ ancestral populations. We analyzed and visualized the `admixture` runs with `pong` [36].

6.4.2 Inversion and non-inversion chromosomes

Four chromosomes (Chr01, Chr02, Chr07, and Chr12) carry large inversions that are clearly visible in the genome and have been described and tied to ecotypic variation and selection [1, 18, 19, 72, 73]. We refer to them as the inversion chromosomes. The remaining 19 chromosomes may have smaller structural variations and are subject to selective sweeps and rapid LD decay [1]. We refer to them as the non-inversion chromosomes.

6.4.3 Genotyping inversions

To genotype inversions, we used a well-defined population from the the south coast of Iceland, which we already had genotyped using PCA and analysis of site frequency spectra [1]. We took x individuals of one homozygote and y individuals of the other, used `plink --count-geno` to count the number of individuals of each genotype at all sites within the inversion or the whole chromosome, and defined as diagnostic sites, all sites that returned exactly x homozygotes for the reference allele and y individuals homozygous for the alternative allele (or y and x if unsure which group is which homozygote). We then used the diagnostic sites thus defined for the South/south-east population in Iceland [1], which belongs to the R PCA group, to genotype the inversions among the PCA groups and among the other gadid taxa. Some of the sites that were diagnostic for R were not diagnostic for, say, the SS group, but the genotyping works. We gave homozygous reference allele (0/0), a numeric score of 1, heterozygotes (0/1), a score of 2, and homozygous alternative allele (1/1), a score of 3. We found the numeric genotype score as the marginal mean over sites for each individual and plotted a heatmap with hierarchical clustering of individuals using the `pheatmap` `R` package [74]. Because of the population structure within taxa, differences among taxa, and the recombinational breakup of inversion, the diagnostic sites based on the local Icelandic population did not yield 'pure' composite genotypes in some cases. To call the composite genotype of an inversion in an individual we inspected the mean

numerical genotype, the heatmap, and the hierarchical clustering and drew a line to separate homozygotes from heterozygote. To determine the robustness of the method we took individuals from the *SS* PCA group called as homozygous based on the *R* PCA group and used the method to find sites that were diagnostic for the inversion within *SS*. We then genotyped *SS* using the sites diagnostic for *SS*. The differences were minor and always involved individuals that were on the boundary of being called homozygous or heterozygous.

6.4.4 Analysis of hybridization and phylogeny

We studied the relationships of inversion haplotypes among the PCA groups and the various gadid taxa. We randomly selected five homozygotes for each haplotype with the highest sequence coverage (fewer than five if the haplotype frequency was low). We used the `vcf_to_treemix.py` script from the Pop-Gen Pipeline Platform `PPP` [75] to generate input for studying population splits and a mixture of the inversion regions of the inversion chromosomes (Chr01, Chr02, Chr07, Chr12) for the PCA groups and gadid taxa with `treemix` [76]. In some instances, five homozygous individuals were not found in the sample, so we used what we had. We ran maximum likelihood `treemix` analysis with migration edges $m = 0 \ldots 5$ treating each inversion genotype as a population of five individuals. We used the individuals selected for Chr01 inversion to do the analysis for non-inversion chromosomes Chr04, Chr17, and Chr23. We analyzed and visualized the results with the `plot_treemix` `R` function from the `popcorn` package [77].

We generated nucleotide sequences for the individuals of PCA groups and gadid taxa selected for the `treemix` analysis with `vcf_to_seq.py` script from `PPP`. We also included sequences for the haddock and whiting outgroups. We estimated the most appropriate evolutionary model with `ModelFinder` [63] and maximum likelihood phylogenetic trees with `iqtree2` [64] and 1000 replicates with the ultrafast bootstrap [78]. We analyzed and visualized the trees with `ggtree` [79, 80].

6.5 Computations

The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. Computations were run on the Mimir bioinformatics server and the Assa bioinformatics computer at the University of Iceland.

Supplementary information. This paper has Supplementary Material

Acknowledgments. We thank Kristján Kristinsson and the Icelandic Marine Research Institute, Grant Pogson, Mike Canino, and Darin Jones and Mike Levine at NOAA/NMFS/AFSC, Snæbjörn Pálsson, JS Christiansen, SE Fevolden, and the TUNU-I Expedition (the University of Tromsø, Norwegian College of Fishery Science, Institute of Aquatic BioSciences), Greenland Institute of Natural Resources and crew of R/V Paamiut, Lilja Jansen, Svava

Ingimarsdóttir, Gudni Magnús Eiríksson, the Faroese Fisheries Laboratory and crew of R/V Magnus Heinason, Remment ter Hofstede and the Netherlands Institute for Fisheries Research, and Óskar Gudmundsson on Sigrún RE-303 for help with sampling and Jarle Mork for samples from the J.Mork collection. We thank Matthew Haynsen for help with molecular work. We acknowledge technical help with the molecular analysis by The Bauer Core Facility at Harvard University.

Declarations

- Funding: The work was supported by an Icelandic Research Fund Grant of Excellence no. 185151-051 to EÁ, KH, Alison Etheridge, Wolfgang Stephan, and Bjarki Eldon.
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use): The authors declare no conflict of interest or competing interests.
- Ethics approval: The samples used for this study were from our large DNAfish tissue collection, the sampling of which was approved by appropriate ethics review boards as appropriate and already described [1, 15].
- Consent to participate: Not applicable
- Consent for publication: Not applicable
- Availability of data and materials: The `bam` alignment files will be available in the NCBI short read sequence archive as bioproject XXX at the time of publication.
- Code availability: Code for snpArcher variant calling is available on `github`: https://github.com/harvardinformatics/snpArcher.
- Authors' contributions: EÁ, KH: designed experiment, secured funding, performed molecular analysis, analyzed data, wrote and revised the paper. AWE performed data analysis and read and commented on the paper. ALG performed data analysis and read and commented on the paper. TS secured computing, analyzed data, read and revised the paper.

Appendix A Extended Data

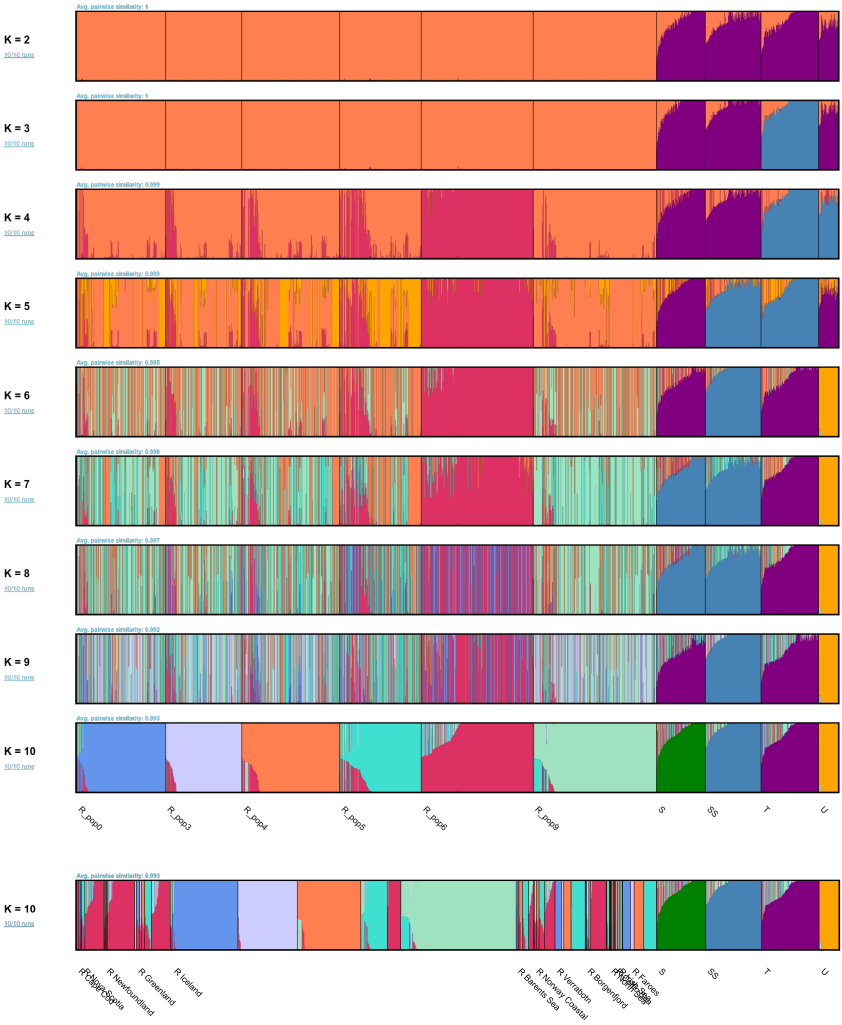

Fig. A1 Admixture analysis of non-inversion chromosomes. Pong [36] analysis and display of ten `admixture`[32] runs for non-inversion chromosomes. Admixture runs for $k = 2 \ldots 10$ with individuals assigned to populations detected by admixture (pop0 ... pop9) by maximum membership at $k = 10$ and labeled according to the admixture population (top panel). The PCA groups S, SS, T, and U (pop1, pop7, pop8, and pop2 respectively) are separate and different from R. The R group is heterogeneous and admixture detects six clusters of R at $k = 10$. All ten admixture runs agree for every $k = 2 \ldots 10$. Color indicates population revealed by admixture for $k = 10$ and R populations are each labeled by populations revealed by admixture (top) and ordered by geographic locality (ranging clockwise from Cape Cod in the west to the Faroe Islands in the east) and admixture population within the locality indicated by color (bottom). Bi-allelic SNPs with singletons excluded, LD pruned, and thinned to 450,953 SNPs.

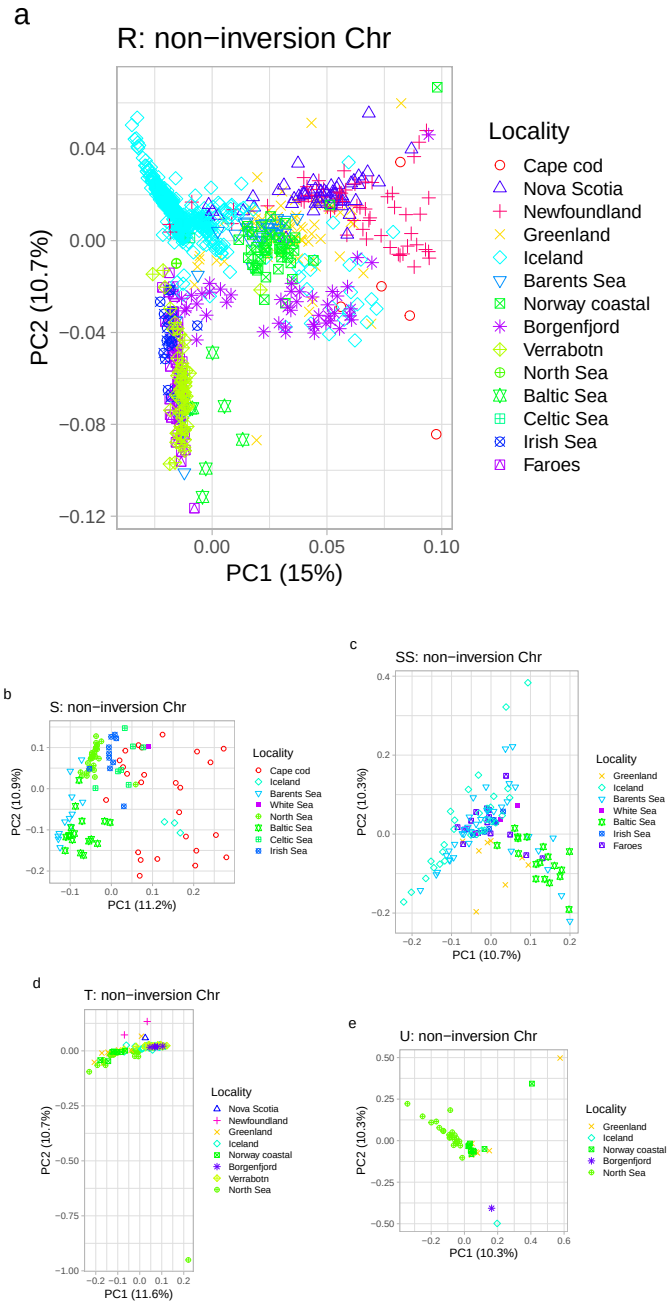

Fig. A2 Principal component analysis of non-inversion chromosomes by PCA group and geographic locality. PCA of PCA group *R* (**a**), *S* (**b**), *SS* (**c**), *T* (**d**), and *U* (**e**) labeled by geographic locality ranging clockwise from Cape Cod to Faroe Islands. The *R* group PCA was based on 1218 individuals after removing 52 outliers that compress the PCA results.

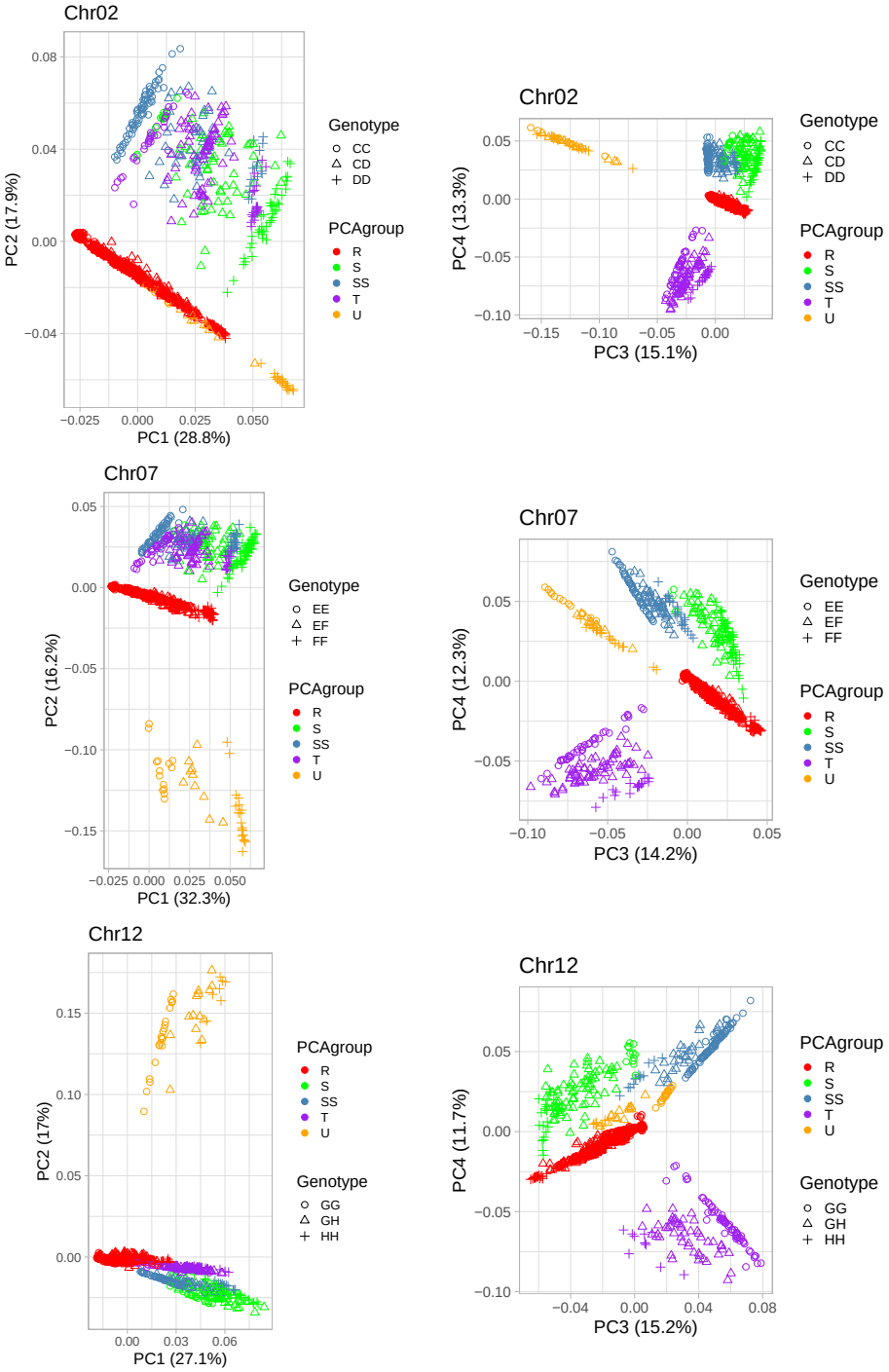

Fig. A3 Principal components analysis of ecotypic defining inversions on chromosomes 02, 07, and 12 classified by PCA group and inversion genotype. PC2 on PC1 and PC4 on PC3 explain 75, 75, and 71% of variation respectively. Chromosomal coordinates inside inversions: 579,097 to 4,096,513 for Chr02, 17,004,745 to 25,979,379 for Chr07, and 1,001,809 to 13,965,277 for Chr12.

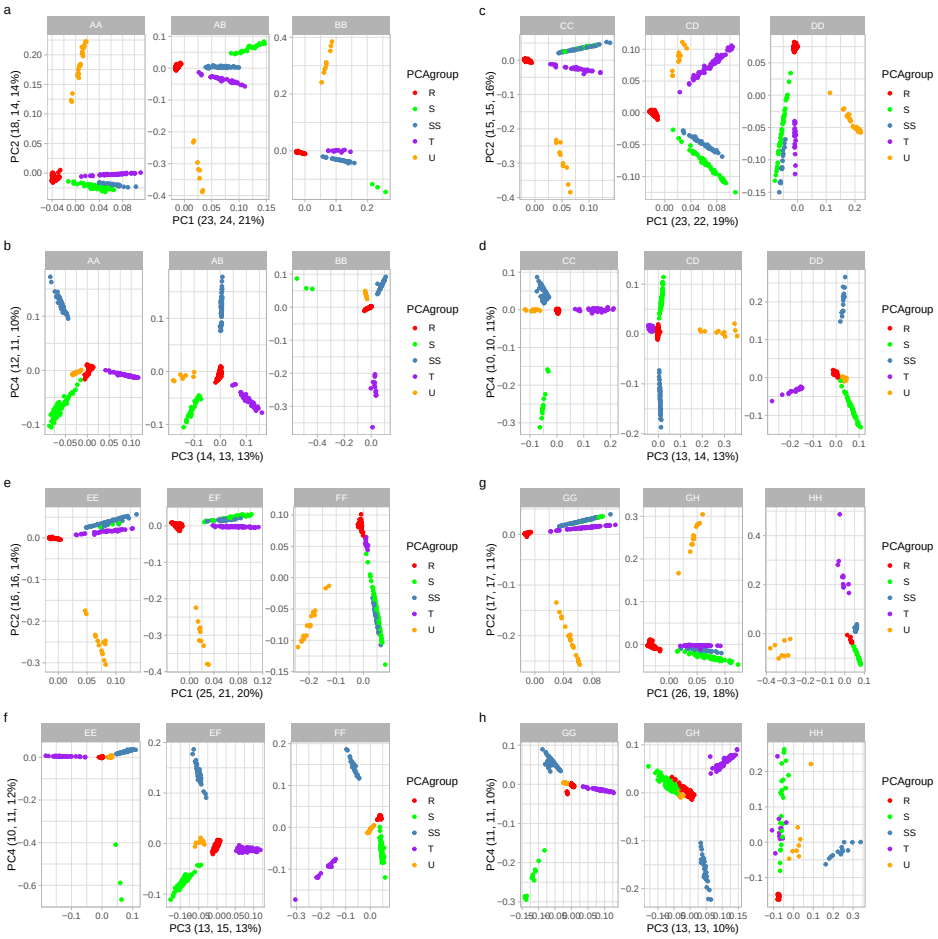

Fig. A4 Differentiation of haplotypic genotypes of inversion chromosomes. Principal components analysis of inversion genotypes labeled by PCA group. PC2 on PC1 and PC4 on PC3 for Chr01 (**a** and **b**), Chr02 (**c** and **d**), Chr07 (**e** and **f**), and Chr12 (**g** and **h**) respectively. The percentage of variation explained for the three genotypes, and each PC is given in parentheses. Chromosomal coordinates inside inversions: 11,763,336 to 27,936,178 for Chr01, 579,097 to 4,096,513 for Chr02, 17,004,745 to 25,979,379 for Chr07, and 1,001,809 to 13,965,277 for Chr12. Bi-allelic SNPs with singletons excluded and LD pruned.

Table A1 Frequency of types revealed by principal component analysis and admixture in relation to geographic locality.

	Principal Component Analysis (PCA) group										
	R						*S*	*SS*	*T*	*U*	
	Admixture population (pop)										
	0	3	4	5	6	9	2	7	8	1	Sum
Cape cod	0	0	0	0	6	0	24	0	0	0	30
Nova Scotia	4	2	0	7	42	2	0	0	1	0	58
Newfoundland	0	2	2	1	61	4	0	0	1	0	71
Greenland	8	4	8	15	41	4	0	8	8	9	105
Iceland	144	130	139	58	29	253	3	30	41	1	828
Barents Sea	4	2	8	13	11	2	12	40	0	0	92
White Sea	0	0	0	0	0	0	1	4	0	0	5
Norway coastal	0	4	6	12	23	0	0	0	8	10	63
Verrabotn	14	5	17	31	0	1	0	0	27	0	95
Borgenfjord	2	5	4	6	34	1	0	0	5	1	58
North Sea	0	0	1	1	0	1	31	0	31	22	87
Baltic Sea	0	2	3	2	0	0	18	18	0	0	43
Celtic Sea	0	0	1	0	0	0	7	0	0	0	8
Irish Sea	4	3	7	5	0	1	8	6	0	0	34
Faroes	17	8	21	28	0	1	0	16	0	0	91
Sum	197	167	217	179	247	270	104	122	122	43	1668

Table A2 Differentiation of populations revealed by principal component analysis and admixture. Pairwise F_{ST} for non-inversion chromosomes among populations revealed by PCA and `admixture`[32] analysis for $k = 10$ ancestral populations (Fig. A1) and labeled by PCA group (Fig. 1).

	Principal Component Analysis (PCA group)								
	R						*S*	*SS*	*T*
	Admixture population (pop)								
	pop0	pop3	pop4	pop5	pop6	pop9	pop2	pop7	pop8
*R*_pop3	0.007								
*R*_pop4	0.006	0.007							
*R*_pop5	0.008	0.008	0.007						
*R*_pop6	0.019	0.019	0.018	0.017					
*R*_pop9	0.006	0.007	0.006	0.009	0.019				
*S*_pop2	0.096	0.096	0.095	0.092	0.092	0.098			
*SS*_pop7	0.083	0.084	0.081	0.080	0.082	0.085	0.052		
*T*_pop8	0.114	0.115	0.113	0.111	0.110	0.117	0.076	0.067	
*U*_pop1	0.073	0.074	0.072	0.071	0.073	0.075	0.083	0.073	0.085

Appendix B Supplementary Information

B.1 Sampling localities

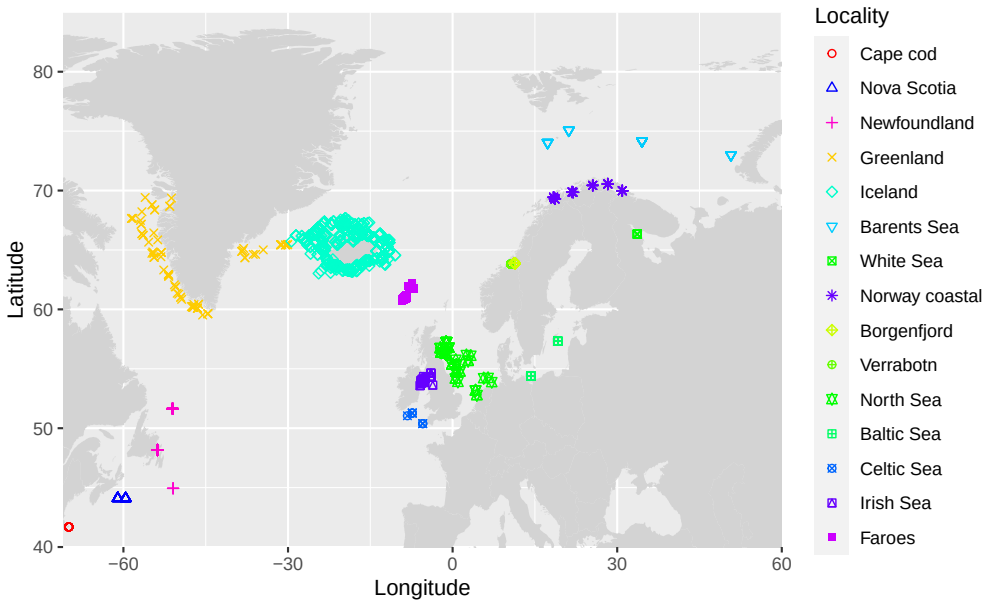

Fig. B1 Sample localities of Atlantic cod *Gadus morhua* on a map of the North Atlantic Ocean. Localities range from Cape Cod in the west to the Faroe Islands in the east.

B.2 Local structure of the genome

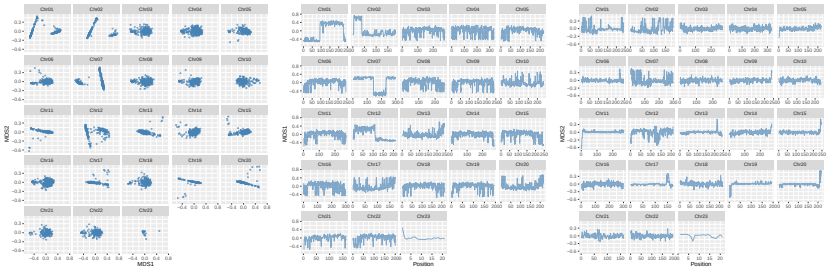

Fig. B2 Local structure of the genome. Local PCA using `lostruct` [31] of all chromosomes of Atlantic cod. Matrix of distances of principal component analysis of local chromosomal segments analyzed with multidimensional scaling (MDS). MDS coordinate 2 on MDS coordinate 1 (left), and MDS1 and MDS2 on chromosomal location (middle and right).

B.3 Population structure

B.3.1 Principal components analysis and subsampling

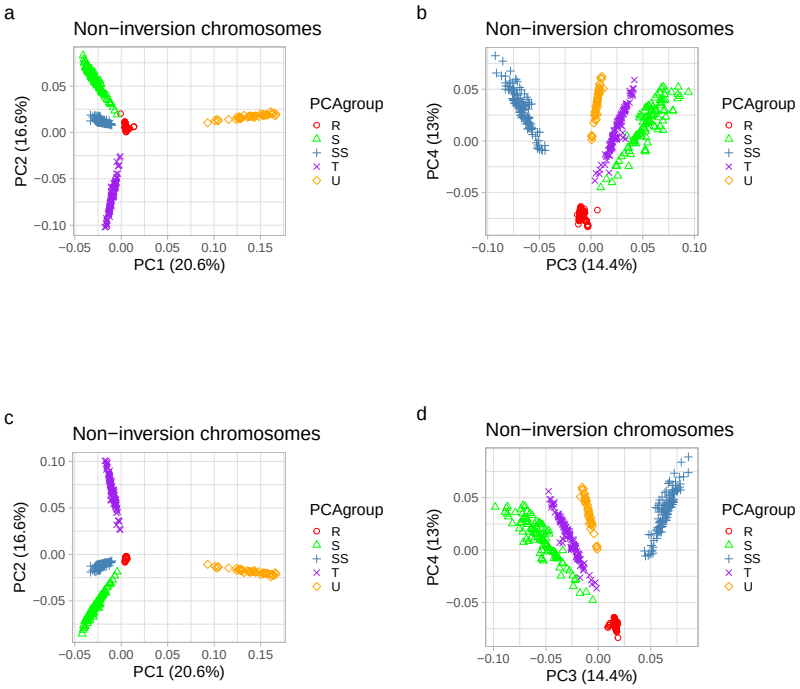

Fig. B3 Principal component analysis of non-inversion chromosomes by subsampling *R* group. Random subsample of 122 individuals from the *R* group (**a** and **b**), and same size subsample of *R* group individuals from Iceland only (**c** and **d**).

B.3.2 Principal components analysis of individual non-inversion chromosomes

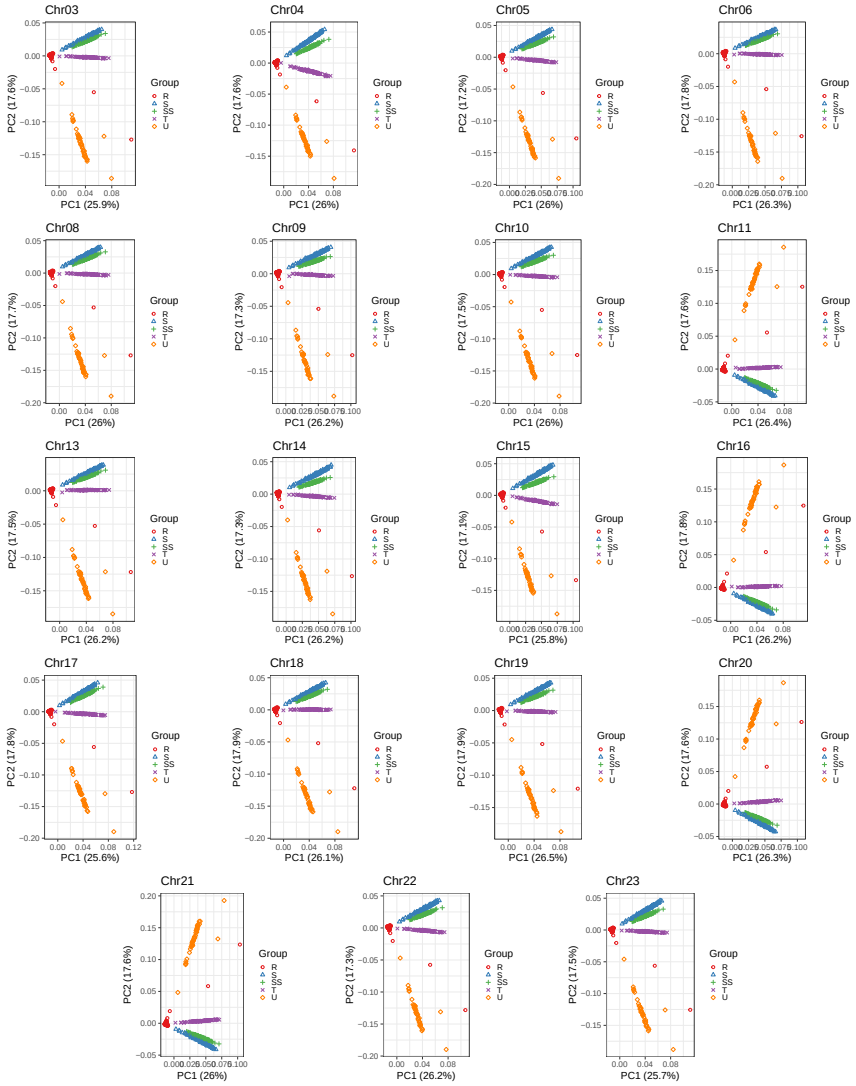

Fig. B4 Principal component analysis of individual non-inversion chromosomes.

B.3.3 Admixture analysis of non-inversion chromosomes

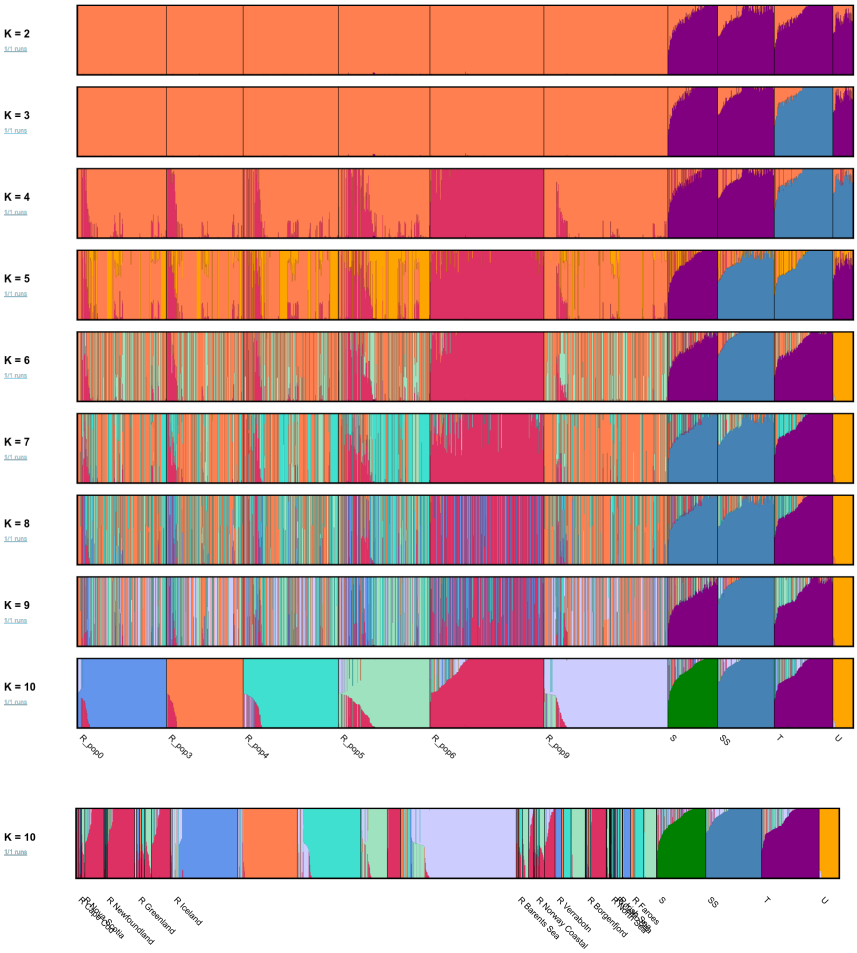

Fig. B5 Admixture analysis of non-inversion chromosomes with fuller data. Pong [36] analysis and display of a single `admixture`[32] run for non-inversion chromosomes with less thinning than in Fig. A1. Admixture runs for $k = 2 \ldots 10$ with individuals assigned to populations detected by admixture (pop0 ... pop9) by maximum membership at $k = 10$ and labeled according to the admixture population (top panel). The PCA groups *S*, *SS*, *T*, and *U* (pop1, pop7, pop8, and pop2 respectively) are separate and different from *R*. The *R* group is heterogeneous and admixture detects six clusters of *R* at $k = 10$. Color indicates population revealed by admixture for $k = 10$ and *R* populations are each labeled by populations revealed by admixture (top) and ordered by geographic locality (ranging clockwise from Cape Cod in the west to the Faroe Islands in the east) and admixture population within the locality indicated by color (bottom). Bi-allelic SNPs with singletons excluded, LD pruned, and thinned to 4,516,652 SNPs.

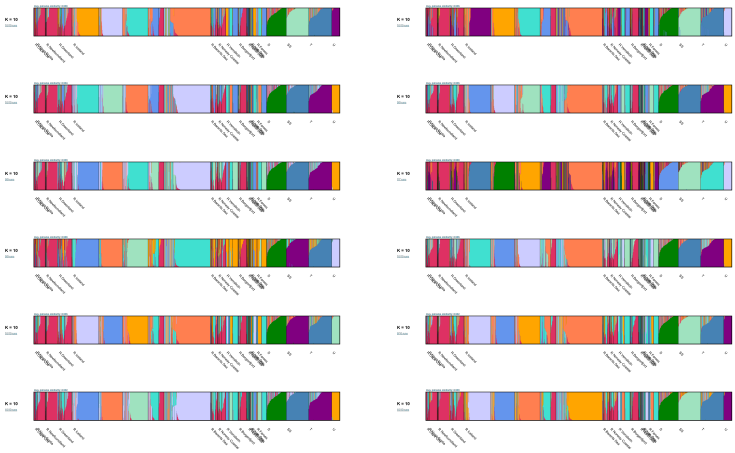

Fig. B6 Admixture analysis of a sample of individual non-inversion chromosomes. Pong [36] display of (usually) ten admixture runs for representative non-inversion chromosomes 03, 05 06, 08, 09, 17, 18, 19, 20, 21, 22, and 23 (from top left to bottom right). The PCA groups *S*, *SS*, *T*, and *U* are clearly separate and different from *R*. The *R* group is heterogeneous. *R* labelled with admixture population and geographic location ranging clockwise from Cape Cod to the Faroe Islands. All ten replicate admixture runs agree for every $k = 2 \ldots 10$ in most cases. Individuals assigned to a population and ordering of admixture populations within geographic locality is that based on all non-inversion chromosomes depicted in Fig. A1. The color scheme is not the same in different panels. Individual non-inversion chromosomes show the same pattern of multicolored geographic localities as seen for the non-inversion chromosomes combined (Fig. A1).

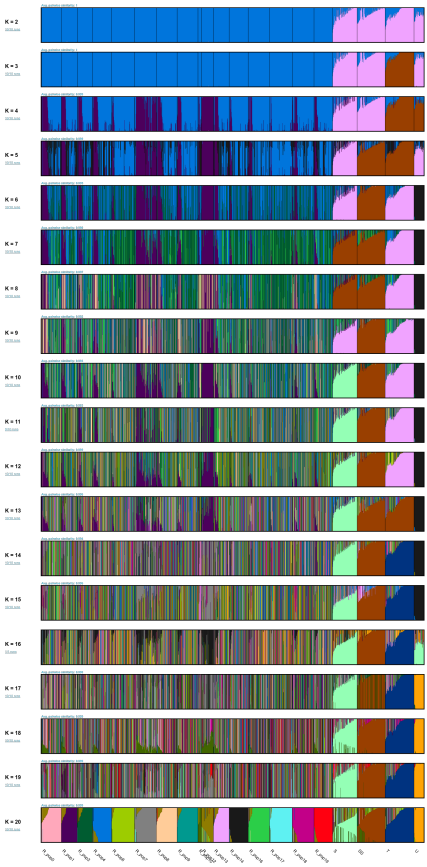

Fig. B7 Admixture analysis of non-inversion chromosomes for $k = 2$ to $k = 20$. Pong [36] analysis and display of ten `admixture`[32] runs for non-inversion chromosomes with data thinned as in Fig. A1. Admixture runs for $k = 2 \ldots 20$ with individuals assigned to populations detected by admixture (pop0 ... pop19) by maximum membership at $k = 20$ and labeled according to the admixture population (top panel). The PCA groups S, SS, T, and U are separate and different from R. A few individuals, particularly from S, had maximum membership in an admixture population other than the major admixture population of that group; they were kept with the PCA group for this analysis. The R group is heterogeneous and admixture detects 16 clusters of R at $k = 20$. All ten admixture runs mostly agree for every $k = 2 \ldots 20$. At $k = 11$ there were two modes with 9 runs agreeing on one mode and one run on another mode. Color indicates population revealed by admixture for $k = 20$ and R populations are each labeled by populations revealed by admixture. Bi-allelic SNPs with singletons excluded, LD pruned, and thinned to 450,953 SNPs.

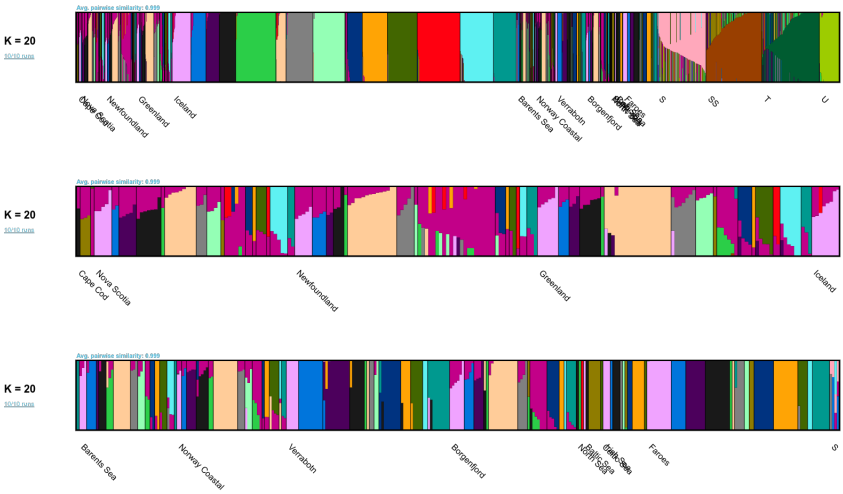

Fig. B8 Admixture analysis of non-inversion chromosomes for $k = 20$ by geographic locality. Pong [36] analysis and display of ten `admixture`[32] runs for non-inversion chromosomes with data thinned as in Fig. A1. Admixture runs for $k = 20$ with individuals assigned to populations detected by admixture (pop0 ... pop19) by maximum membership at $k = 20$, ordered by locality (ranging clockwise from Cape Cod in the west to the Faroe Islands in the east), and labeled according to the admixture population (color) within locality (top panel). A zoom-in of the western localities from Cape Cod to Greenland (middle panel), and a zoom-in of the eastern localities (Barents Sea to Faroes) (bottom panel). Biallelic SNPs with singletons excluded, LD pruned, and thinned to 450,953 SNPs.

Table B1 Frequencies of admixture populations within the *R* group by geographic locality. Absolute and relative frequencies of populations detected by `admixture` [32] ordered by geographic locality arranged clockwise from Cape Cod in the west to the Faroe Islands in the east to facilitate comparison of frequency of genetic types (admixture populations: color) with geographic distance along potential migration routes. Each locality is multi-colored.

	Admixture population of R group at $k = 20$																
	0	1	3	4	6	7	8	9	10	12	13	14	16	17	18	19	Total
Cape cod	0	0	0	1	0	0	0	0	3	1	0	0	0	0	0	0	5
Nova Scotia	5	2	5	7	1	9	3	4	1	6	2	1	3	1	5	2	57
Newfoundland	5	4	2	3	1	14	5	1	0	22	3	1	2	1	2	3	69
Greenland	6	3	3	6	1	19	7	5	1	6	4	1	5	2	6	3	78
Iceland	41	33	29	36	88	22	59	70	1	4	33	55	65	94	73	50	753
Barents Sea	3	4	1	3	3	7	3	3	0	2	2	2	3	0	3	1	40
Norway coastal	2	2	4	5	2	10	3	3	0	4	1	2	4	0	1	2	45
Verrabotn	5	10	11	6	1	1	2	2	0	0	9	4	5	0	2	9	67
Borgenfjord	6	4	4	1	1	12	4	0	1	7	5	2	0	0	1	4	52
North Sea	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	3
Baltic Sea	1	0	0	1	0	0	0	0	5	0	0	0	0	0	0	0	7
Celtic Sea	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
Irish Sea	3	1	0	3	0	0	1	1	1	0	2	5	1	0	0	0	18
Faroes	10	6	8	10	1	1	4	2	2	0	8	10	4	0	2	7	75
	Relative frequency at locality																
Cape cod	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.60	0.20	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Nova Scotia	0.09	0.04	0.09	0.12	0.02	0.16	0.05	0.07	0.02	0.11	0.04	0.02	0.05	0.02	0.09	0.04	1.00
Newfoundland	0.07	0.06	0.03	0.04	0.01	0.20	0.07	0.01	0.00	0.32	0.04	0.01	0.03	0.01	0.03	0.04	1.00
Greenland	0.08	0.04	0.04	0.08	0.01	0.24	0.09	0.06	0.01	0.08	0.05	0.01	0.06	0.03	0.08	0.04	1.00
Iceland	0.05	0.04	0.04	0.05	0.12	0.03	0.08	0.09	0.00	0.01	0.04	0.07	0.09	0.12	0.10	0.07	1.00
Barents Sea	0.07	0.10	0.03	0.07	0.07	0.17	0.07	0.07	0.00	0.05	0.05	0.05	0.07	0.00	0.07	0.03	1.00
Norway coastal	0.04	0.04	0.09	0.11	0.04	0.22	0.07	0.07	0.00	0.09	0.02	0.04	0.09	0.00	0.02	0.04	1.00
Verrabotn	0.07	0.15	0.16	0.09	0.01	0.01	0.03	0.03	0.00	0.00	0.13	0.06	0.07	0.00	0.03	0.13	1.00
Borgenfjord	0.12	0.08	0.08	0.02	0.02	0.23	0.08	0.00	0.02	0.13	0.10	0.04	0.00	0.00	0.02	0.08	1.00
North Sea	0.00	0.00	0.00	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	1.00
Baltic Sea	0.14	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Celtic Sea	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Irish Sea	0.17	0.06	0.00	0.17	0.00	0.00	0.06	0.06	0.06	0.00	0.11	0.28	0.06	0.00	0.00	0.00	1.00
Faroes	0.13	0.08	0.11	0.13	0.01	0.01	0.05	0.03	0.03	0.00	0.11	0.13	0.05	0.00	0.03	0.09	1.00

B.4 Inversion genotypes

B.4.1 Genotyping inversions based on a local Icelandic population

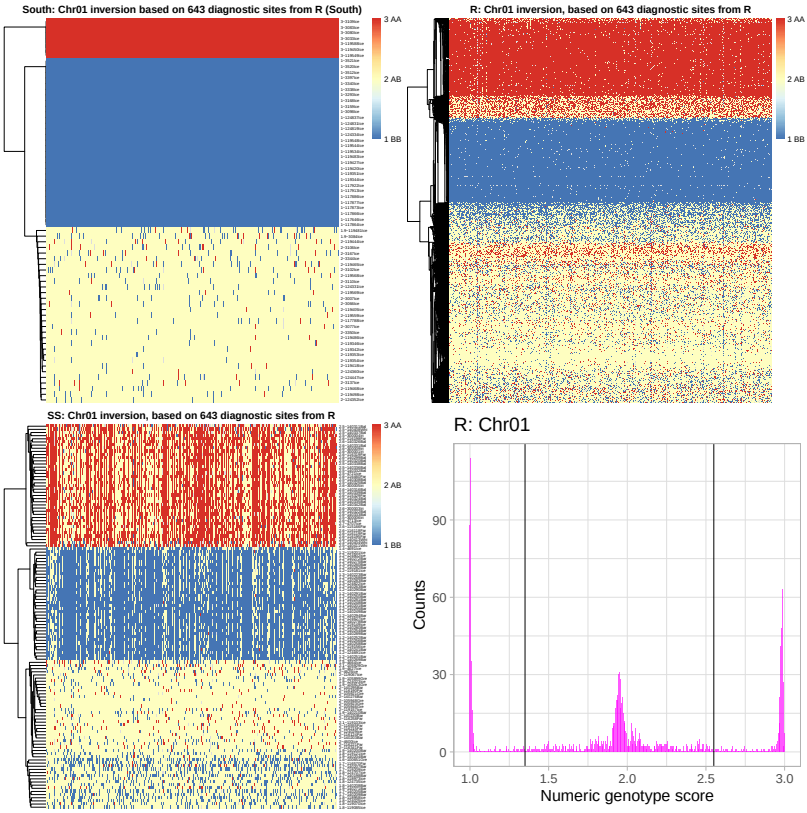

Fig. B9 Diagnostic sites and method of inversion genotyping. Diagnostic sites obtained based on a local population from south/south-east coast of Iceland that is *R* group and already genotyped by PCA and site-frequency spectra [1] (top left). For each site, homozygotes for reference allele are given a numeric score 1 (blue), homozygotes for alternative allele a score of 3 (red), and heterozygotes a 2 (yellow) score. A numeric genotype score is the average over all sites (columns). Results plotted with `pheatmap` [74] and individuals clustered with hierarchical clustering. Diagnostic sites from the south population [1] used to genotype all *R* individuals (top right), and all *SS* individuals (bottom left). Distribution of numeric genotype score for all *R* individuals (bottom right). Vertical lines delimiting called genotypes are determined by inspection of hierarchical clustering and color.

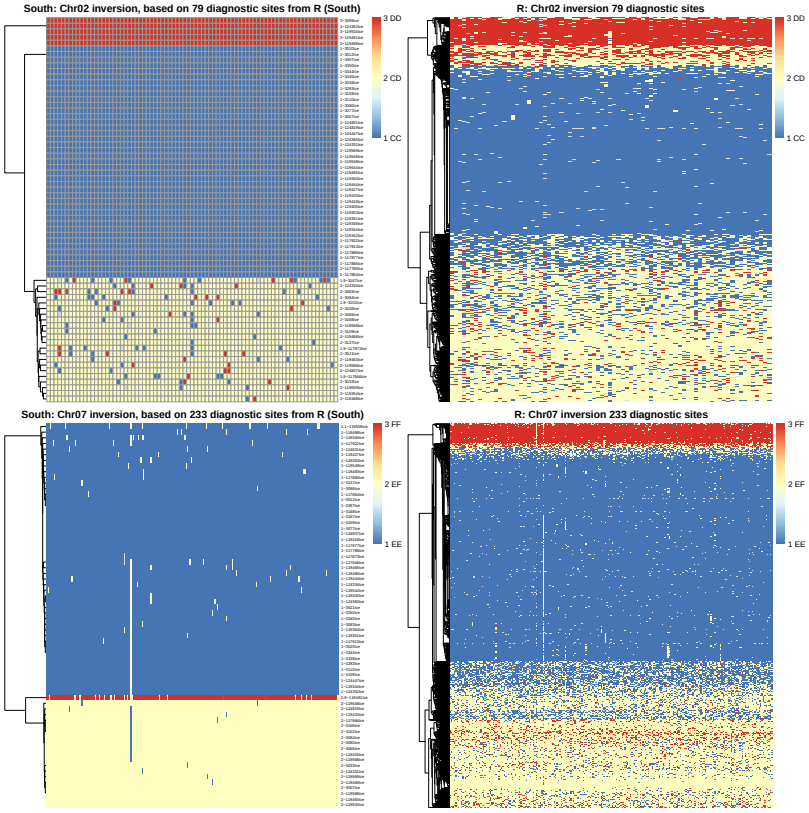

Fig. B10 Diagnostic sites and genotyping of inversions on Chr02 and Chr07. Diagnostic sites obtained for Chr02 (top left) and Chr07 (bottom left) based on a local population from south/south-east coast of Iceland that is R group and already genotyped by PCA and site-frequency spectra [1]. Genotyping of al R individuals for Chr02 (top right) and Chr07 (bottom right).

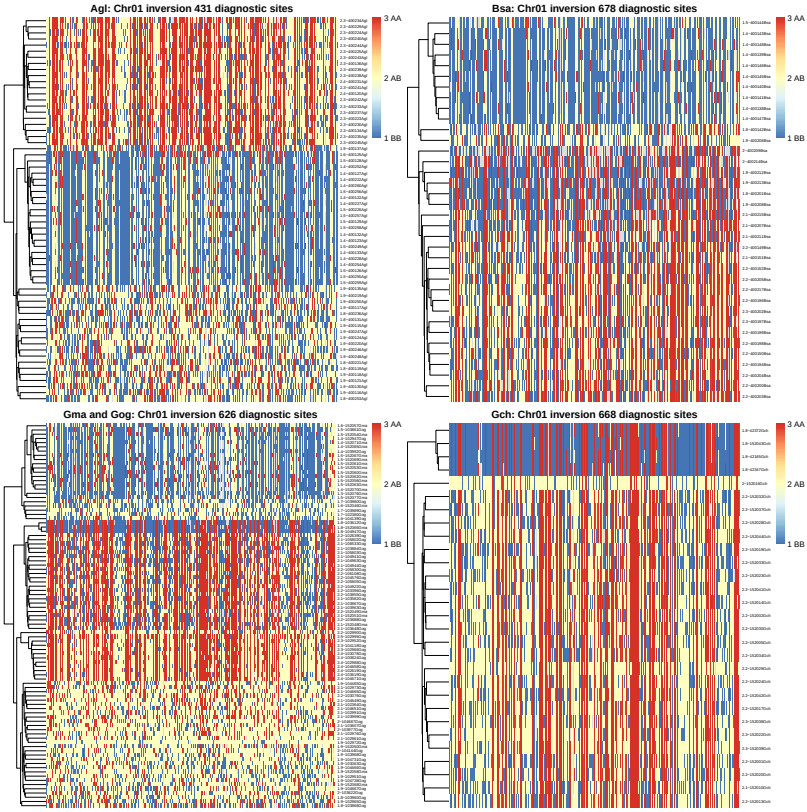

Fig. B11 Genotyping samples of other gadids for Chr01. Samples of Polar cod (Agl), Arctic cod (Bsa), Pacific cod (Gma), Greenland cod (Gog), and walleye pollock (Gch) genotypes for Chr01 inversion using diagnostic sites from *R*.

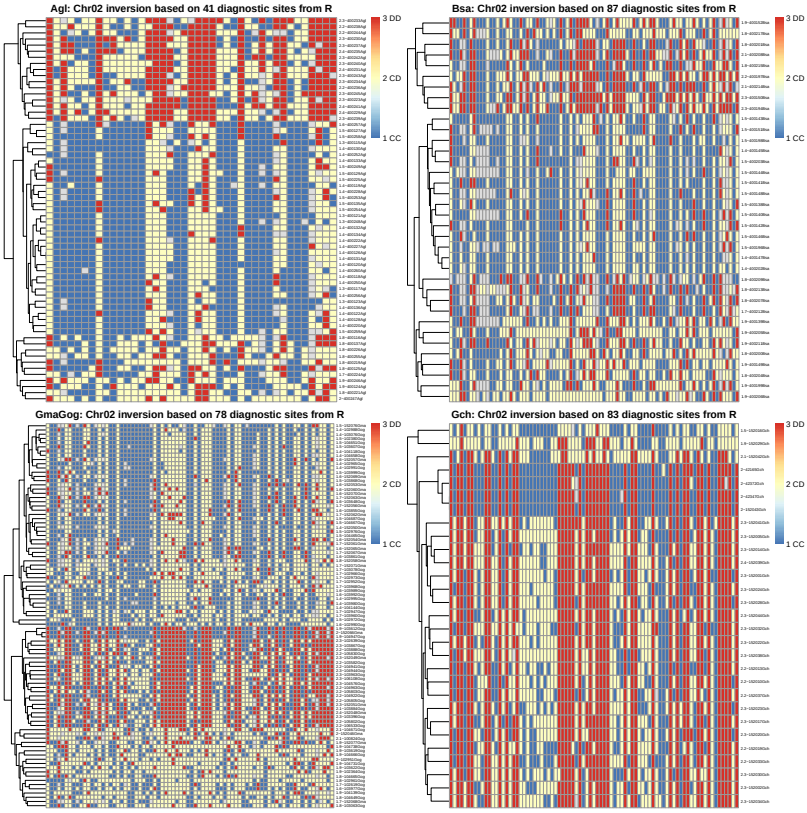

Fig. B12 Genotyping samples of other gadids for Chr02. Samples of Polar cod (Agl), Arctic cod (Bsa), Pacific cod (Gma), Greenland cod (Gog), and walleye pollock (Gch) genotypes for Chr02 inversion using diagnostic sites from *R*.

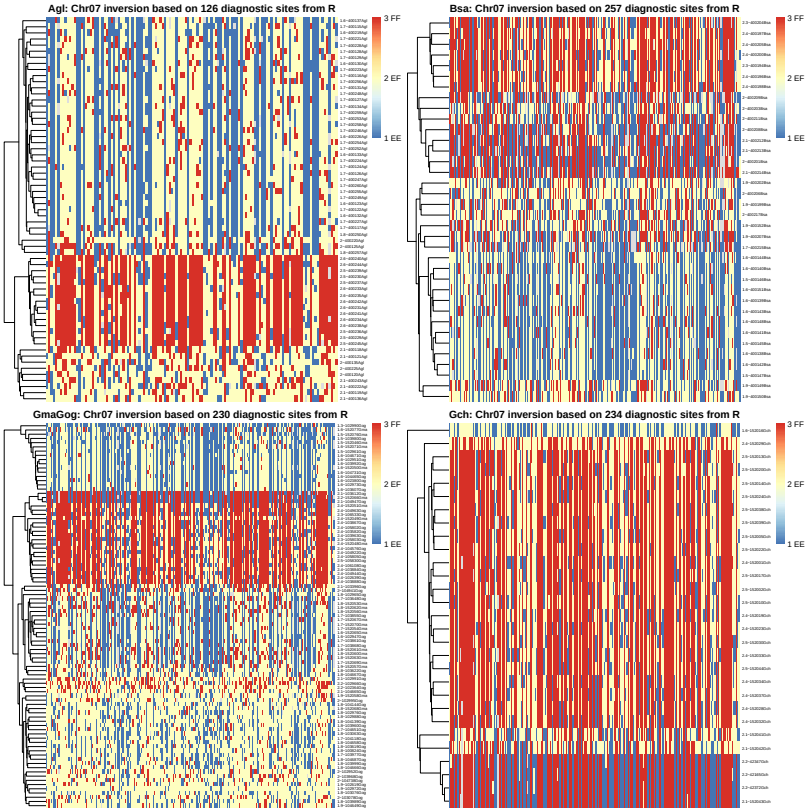

Fig. B13 Genotyping samples of other gadids for Chr07. Samples of Polar cod (Agl), Arctic cod (Bsa), Pacific cod (Gma), Greenland cod (Gog), and walleye pollock (Gch) genotypes for Chr07 inversion using diagnostic sites from *R*.

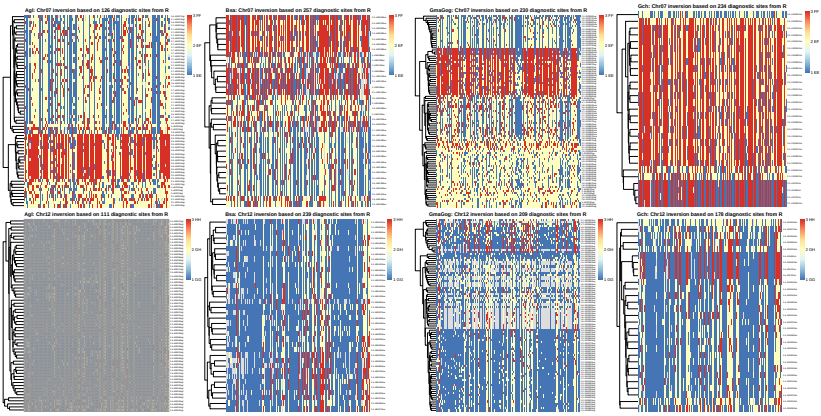

Fig. B14 Test Genotyping samples of other gadids for Chr12. Samples of Polar cod (Agl), Arctic cod (Bsa), Pacific cod (Gma), Greenland cod (Gog), and walleye pollock (Gch) genotypes for Chr12 inversion using diagnostic sites from *R*.

B.4.2 Robustness of genotyping

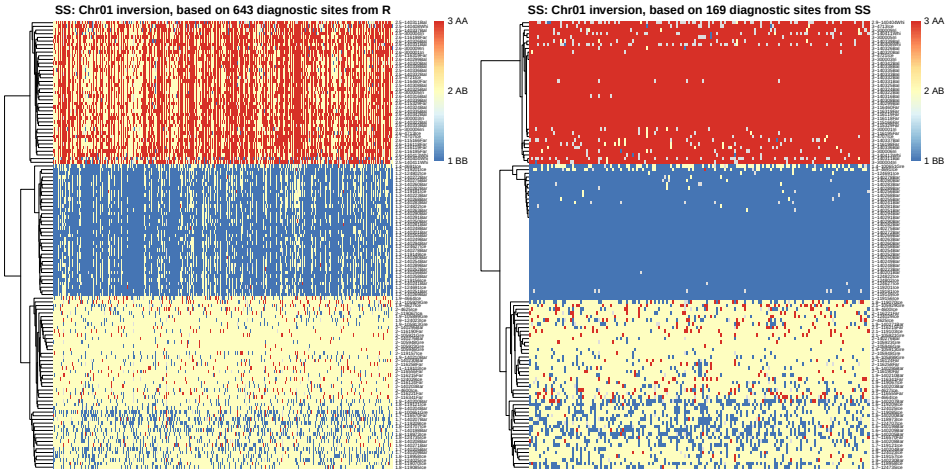

Fig. B15 Robustness of genotyping of Chr01 in *SS*. The *SS* group genotyped with diagnostic sites based on *R* (left) and with diagnostic sites based on *SS* itself (right).

B.4.3 Admixture analysis of inversion chromosomes

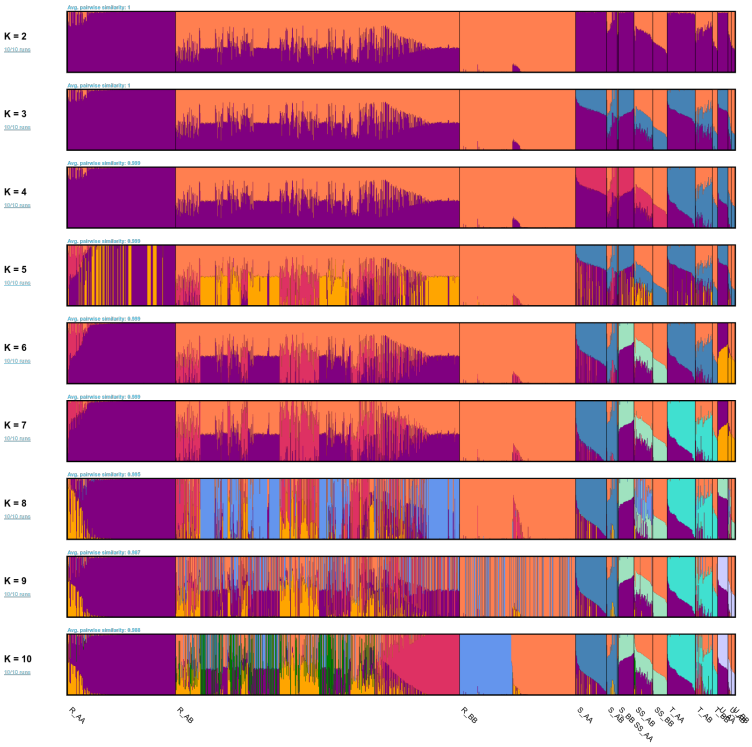

Fig. B16 Admixture analysis of inversion chromosome 01 by genotype. Pong [36] display of ten replicate admixture runs for $k = 2 \ldots 10$ ordered by PCA group and by genotype of the inversion. The PCAgroups *S*, *SS*, *T*, and *U* are clearly separate and different from *R*. The *R* group is heterogeneous. All PCA groups are segregating for the inversion. All ten admixture runs agree for every $k = 2 \ldots 10$. Based on 241,645 SNPs for the entire Chr01 (inside and outside of the inversion; Fig. B2). The genotype of the inversion is based on co-ordinates 11,763,336 to 27,936,178 inside the inversion.

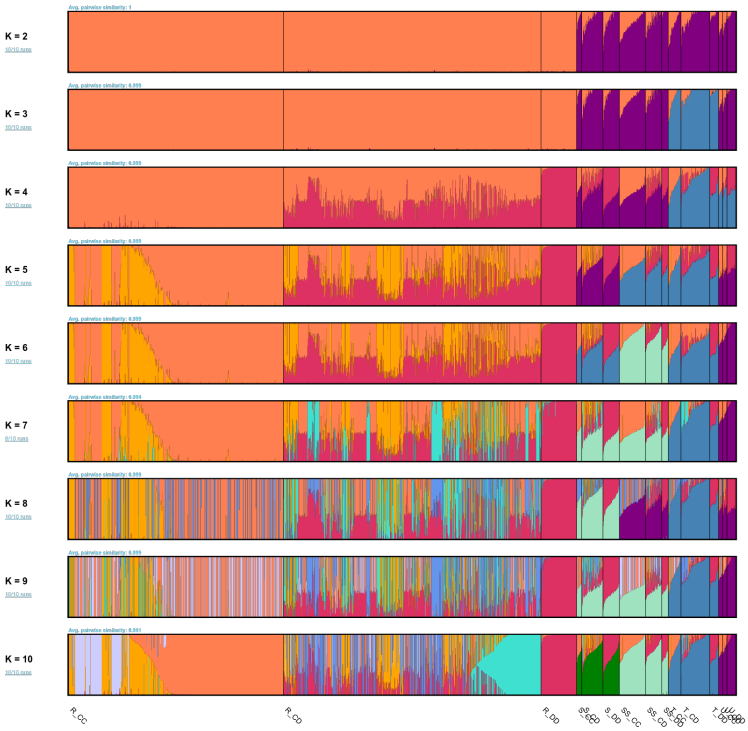

Fig. B17 Admixture analysis of inversion chromosome 02 by genotype. Pong [36] display of ten replicate admixture runs for the entire Chr02 (inside and outside of the inversion Fig. B2) for $k = 2 \ldots 10$ (top). The PCAgroups S, SS, T, and U are clearly separate and different from R. The R group is heterogeneous. All groups are segregating for the inversion. All ten admixture runs agree for every $k = 2 \ldots 10$. Based on 182,774 SNPs for the entire Chr02 (inside and outside of the inversion; Fig. B2). The genotype of the inversion is based on co-ordinates 579,097 to 4,096,513 inside the inversion.

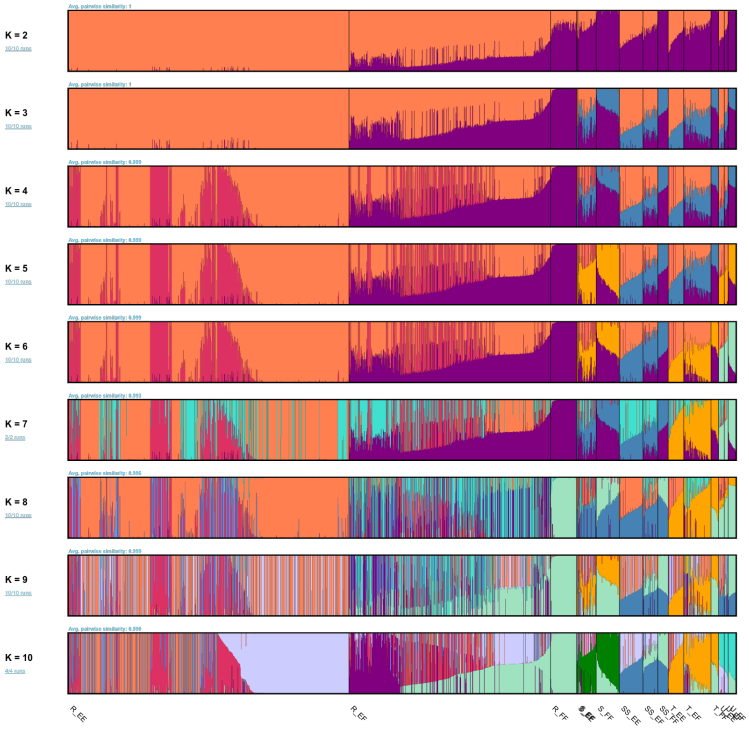

Fig. B18 Admixture analysis of inversion chromosome 07 by genotype. Pong [36] display of (usually) ten replicate admixture runs for the entire Chr07 (inside and outside of the inversion Fig. B2) for $k = 2 \ldots 10$ (top). The PCAgroups S, SS, T, and U are clearly separate and different from R. The R group is heterogeneous. All groups are segregating for the inversion. All ten admixture runs agree for every $k = 2 \ldots 10$. Based on 297,416 SNPs for the entire Chr07 (inside and outside of the inversion; Fig. B2). The genotype of the inversion is based on co-ordinates 17,004,745 to 25,979,379 inside the inversion.

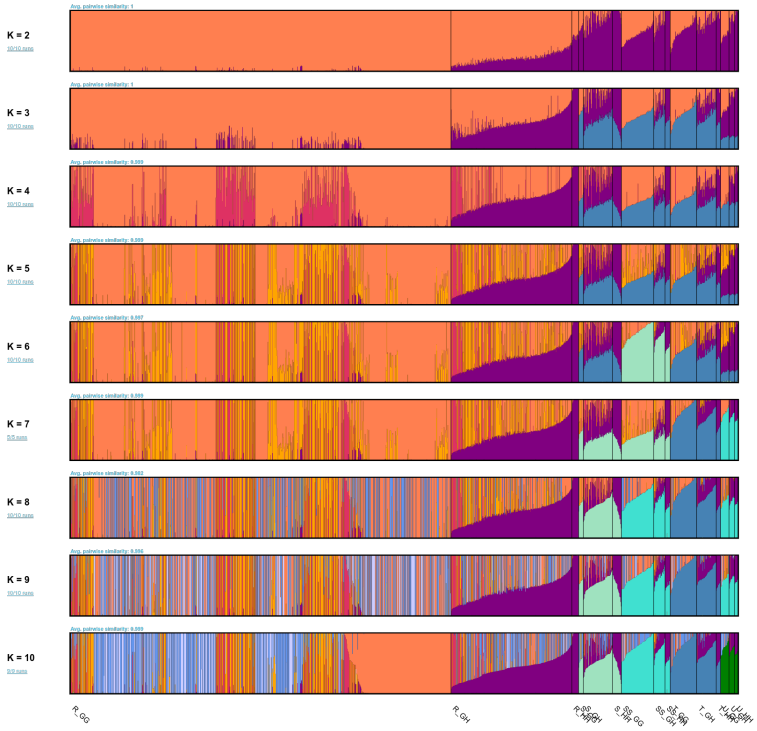

Fig. B19 Admixture analysis of inversion chromosome 12 by genotype. Pong [36] display of (usually) ten replicate admixture runs for the entire Chr12 (inside and outside of the inversion Fig. B2) for $k = 2 \ldots 10$ (top). The PCAgroups S, SS, T, and U are clearly separate and different from R. The R group is heterogeneous. All groups are segregating for the inversion. All ten admixture runs agree for every $k = 2 \ldots 10$. Based on 209,863 SNPs for the entire Chr07 (inside and outside of the inversion; Fig. B2). The genotype of the inversion is based on co-ordinates 1,001,809 to 13,965,277 inside the inversion.

B.4.4 Interchromosomal genotypic associations of inversions

Table B2 Test of Hardy-Weinberg equilibrium of haplotypic genotype frequencies at inversion chromosomes in PCA group *R*. Observed (Obs.) and expected (Exp.) genotype frequencies, X^2 test statistic distributed as χ^2, significance P, and allele frequency of alternative allele, p.

	Chr01			Chr02			Chr07			Chr12		
	AA	*AB*	*BB*	*DD*	*CD*	*CC*	*FF*	*EF*	*EE*	*HH*	*GH*	*GG*
Obs.	270	710	290	89	644	537	65	504	701	17	302	951
Exp.	307.6	634.8	327.6	133.0	556.0	581.0	79.1	475.8	715.1	22.2	291.6	956.2
X^2	17.8, $P < 0.001$			31.8, $P < 0.001$			4.48, $P < 0.001$			1.63, $P > 0.2$		
p	0.49			0.32			0.25			0.13		

Table B3 Association of haplotypic genotypes across chromosomes in PCAgroup *R*. All comparisons of chromosomal association were highly significant by X^2 test ($P << 0.001$).

		Chr01			Chr02			Chr07		
		AA	*AB*	*BB*	*DD*	*CD*	*CC*	*FF*	*EF*	*EE*
Chr02	*DD*	71	11	7						
	CD	150	408	86						
	CC	49	291	197						
Chr07	*FF*	59	4	2	44	15	6			
	EF	147	311	46	31	372	101			
	EE	64	395	242	14	257	430			
Chr12	*HH*	16	1	0	11	5	1	11	4	2
	GH	161	130	11	58	190	54	45	180	77
	GG	93	579	279	20	449	482	9	320	622

Table B4 Association of haplotypic genotypes across chromosomes in the South/south-east population in Iceland [1]. The association comparing Chr01 and Chr12 was statistically significant by Fisher's exact test ($P = 0.001$); none of the other comparisons were significant.

		Chr01			Chr02			Chr07		
		AA	*AB*	*BB*	*DD*	*CD*	*CC*	*FF*	*EF*	*EE*
Chr02	*DD*	0	3	2						
	CD	4	11	7						
	CC	3	17	21						
Chr07	*FF*	0	1	0	1	0	0			
	EF	4	10	5	1	8	10			
	EE	3	20	25	3	14	31			
Chr12	*GH*	4	1	1	0	4	2	0	1	5
	GG	3	30	29	5	18	39	1	18	43

B.5 Recombinational spectrum of inversion genotypes and geographical distribution

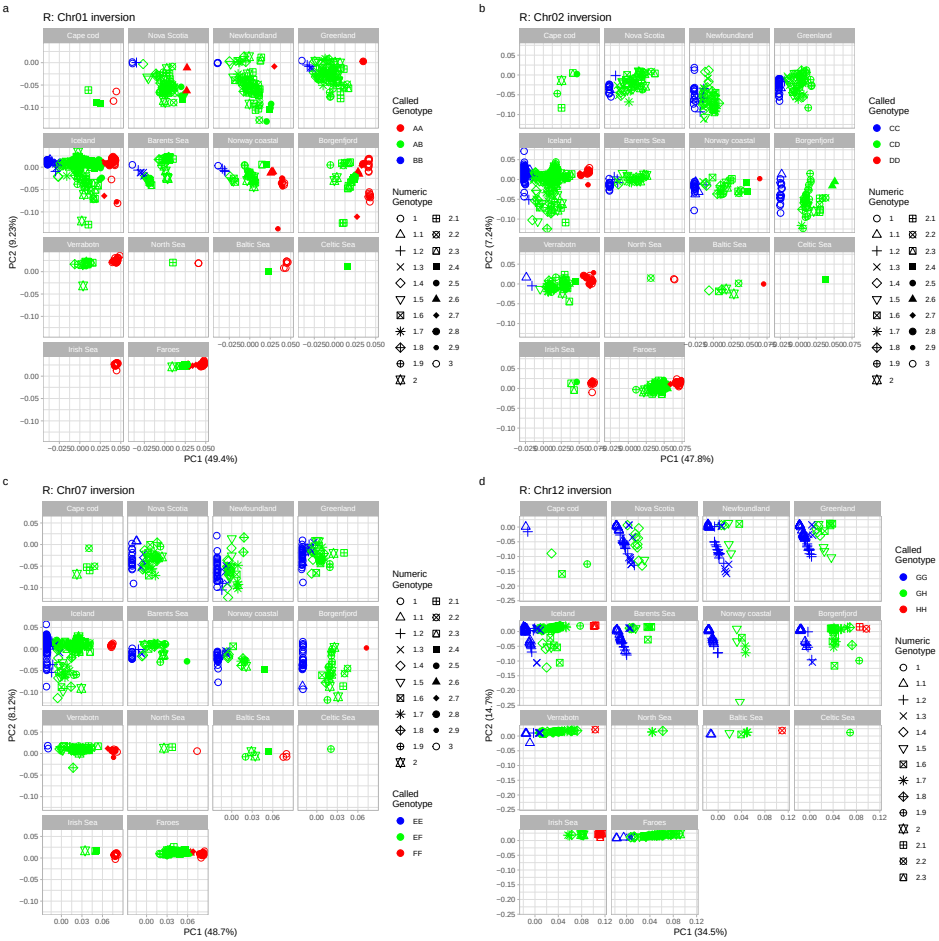

Fig. B20 Principal component analysis of inversion chromosomes of *R* group labeled by called and numeric genotype faceted by locality. Genotypes of inversions on Chr01 (**a**), Chr02 (**b**), Chr07 (**c**), and Chr12 (**d**), labelled [1] and called as described B.4.1. Localities arranged from Cape Cod in the west to the Faroe Islands in the east.

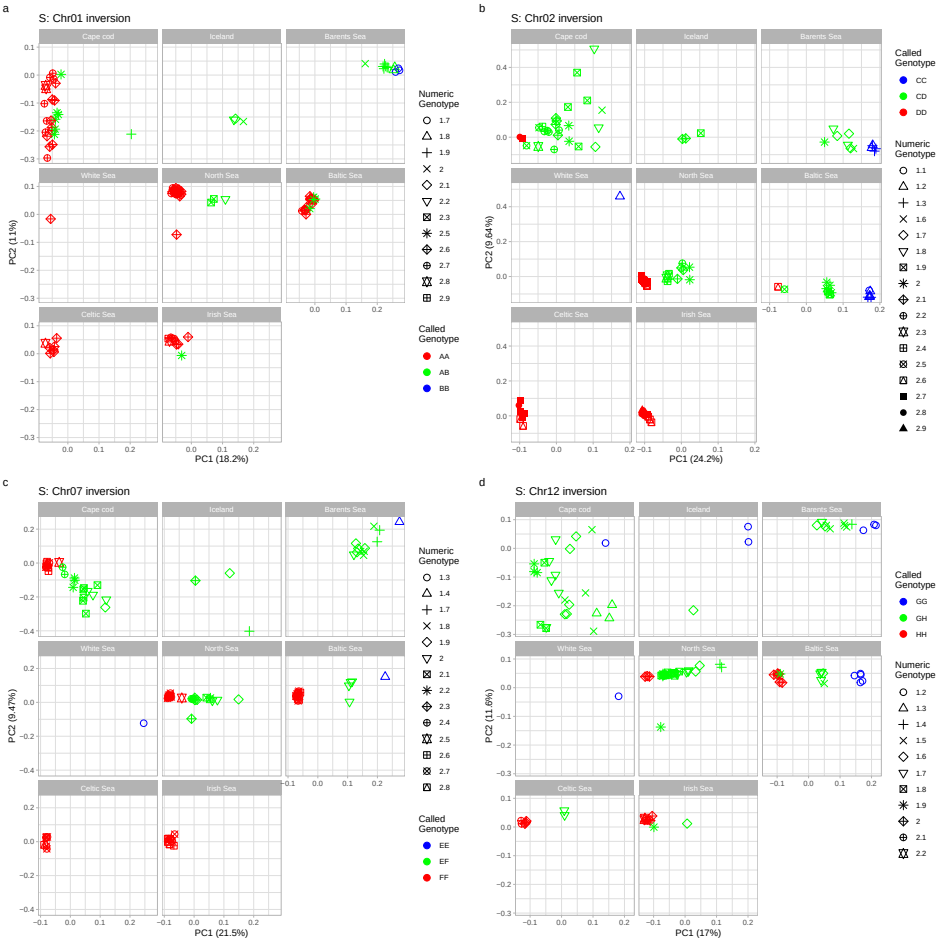

Fig. B21 Principal component analysis of inversion chromosomes of *S* group labeled by called and numeric genotype faceted by locality. Genotypes of inversions on Chr01 (**a**), Chr02 (**b**), Chr07 (**c**), and Chr12 (**d**), labelled [1] and called as described B.4.1. Localities arranged from Cape Cod in the west to the Faroe Islands in the east.

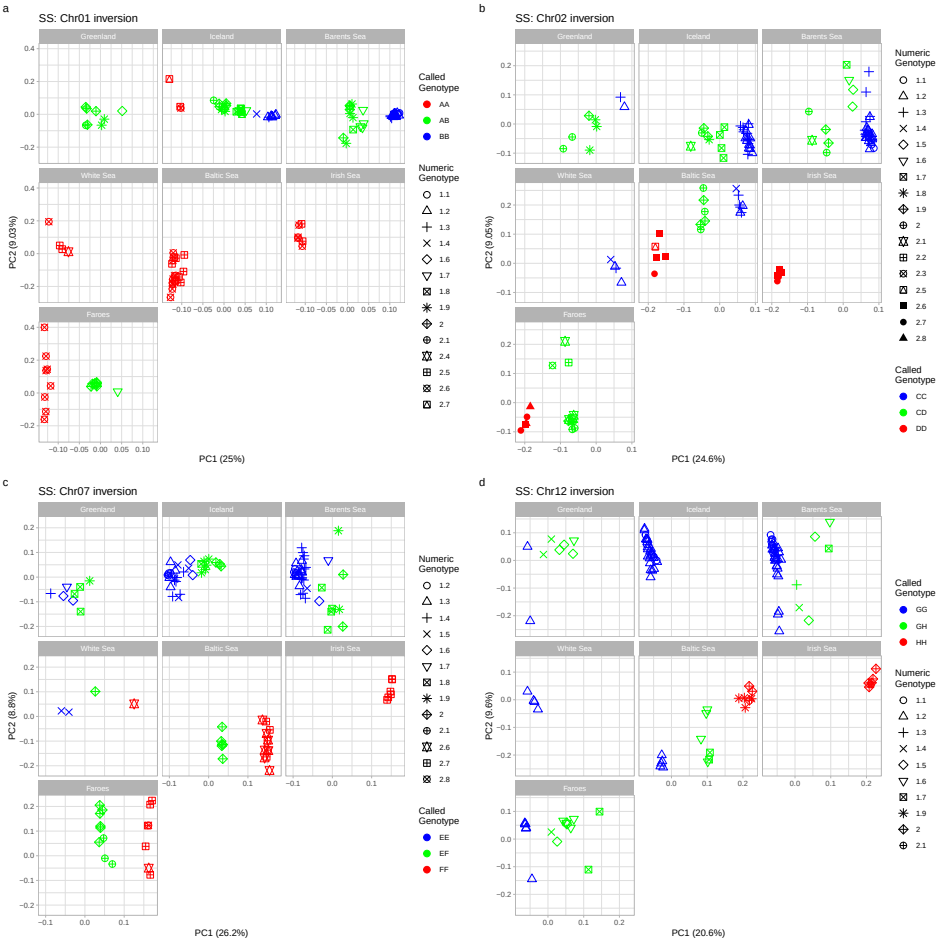

Fig. B22 Principal component analysis of inversion chromosomes of *SS* group labeled by called and numeric genotype faceted by locality. Genotypes of inversions on Chr01 (**a**), Chr02 (**b**), Chr07 (**c**), and Chr12 (**d**), labelled [1] and called as described B.4.1. Localities arranged from Cape Cod in the west to the Faroe Islands in the east.

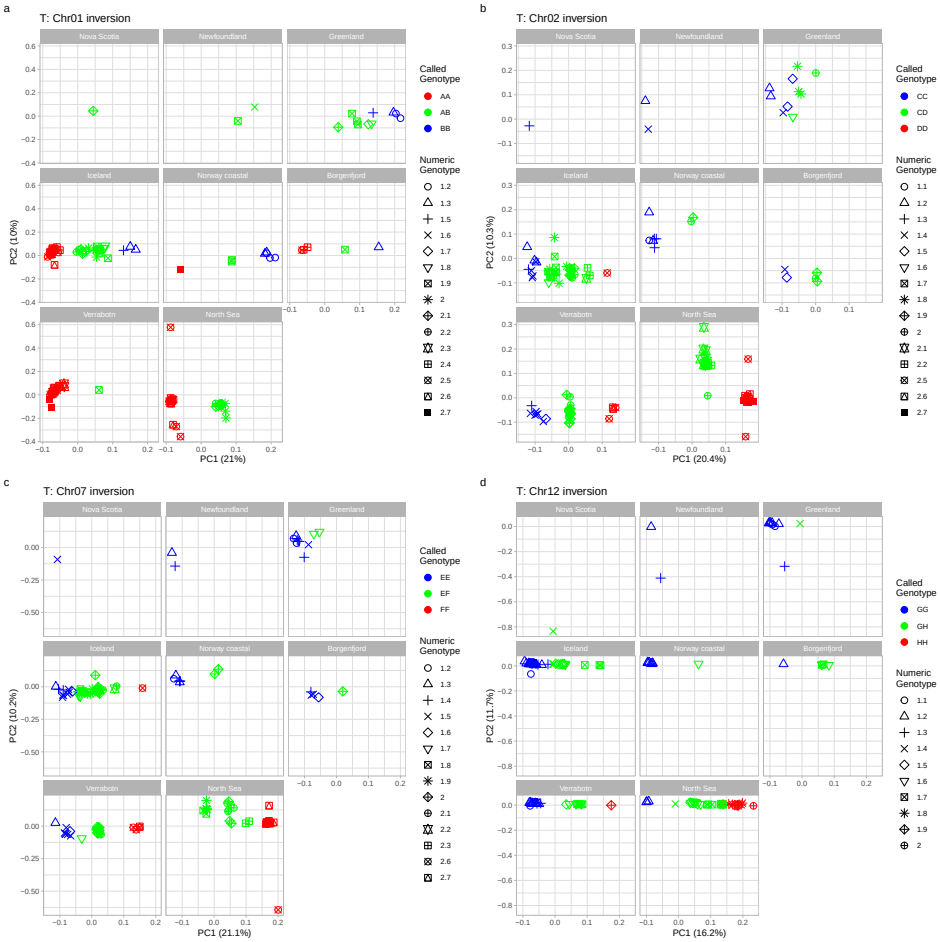

Fig. B23 Principal component analysis of inversion chromosomes of *T* group labeled by called and numeric genotype faceted by locality. Genotypes of inversions on Chr01 (**a**), Chr02 (**b**), Chr07 (**c**), and Chr12 (**d**), labelled [1] and called as described B.4.1. Localities arranged from Cape Cod in the west to the Faroe Islands in the east.

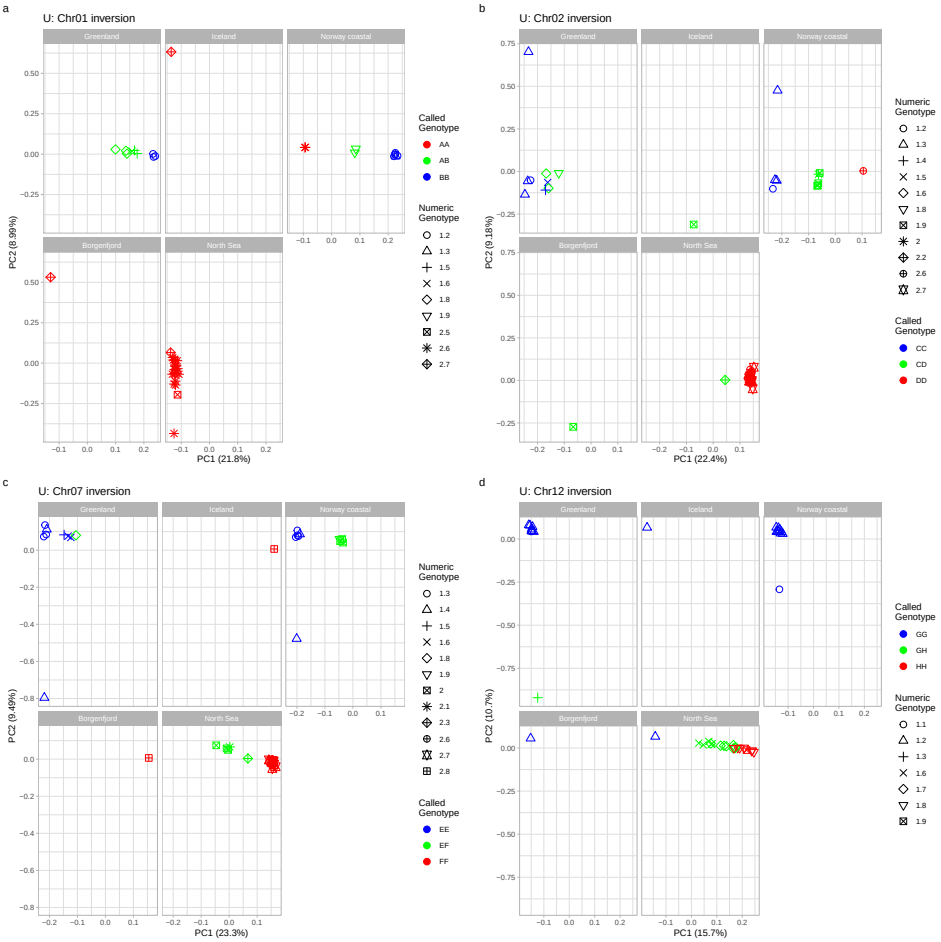

Fig. B24 Principal component analysis of inversion chromosomes of *U* group labeled by called and numeric genotype faceted by locality. Genotypes of inversions on Chr01 (**a**), Chr02 (**b**), Chr07 (**c**), and Chr12 (**d**), labelled [1] and called as described B.4.1. Localities arranged from Cape Cod in the west to the Faroe Islands in the east.

B.6 Divergence of inversion haplotypes related to ecotypic variation

B.6.1 Trans-species and trans-oceanic supergenes

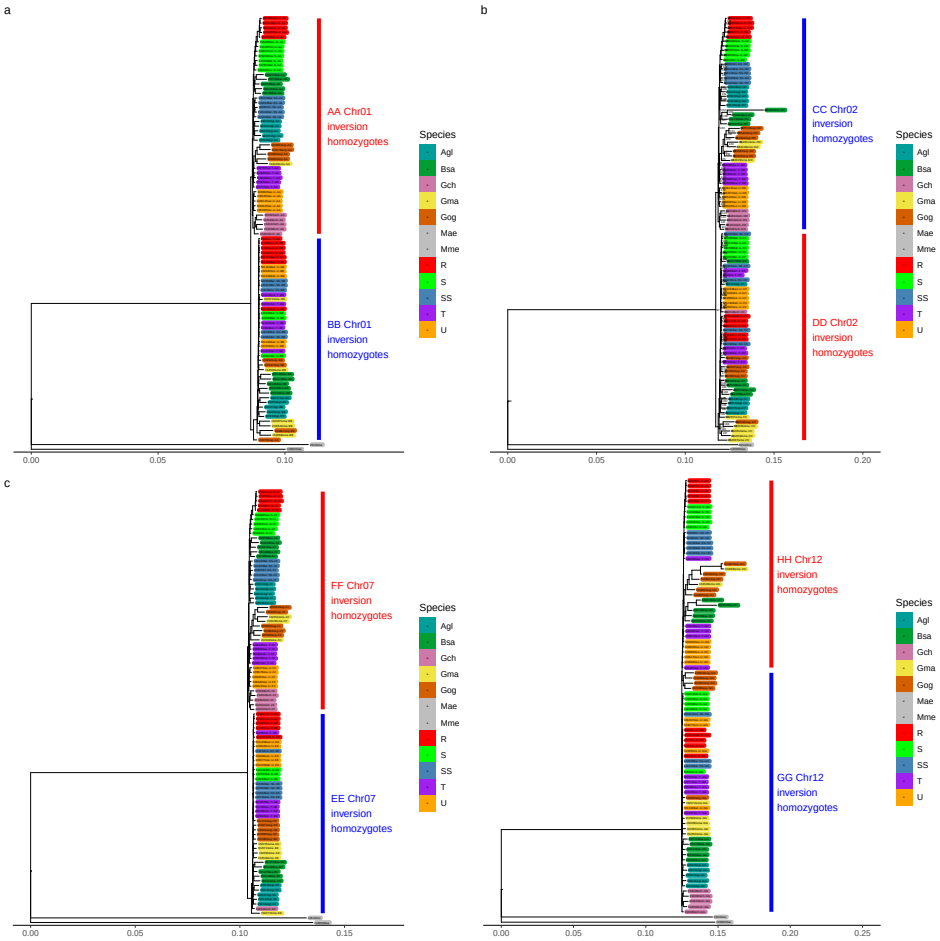

Fig. B25 Trans-species polymorphism of inversion genotypes with outgroup. Maximum likelihood phylogenetic trees of inversion homozygotes of PCAgroups and related gadid taxa for Chr01 (**a**), Chr02 (**b**), Chr07 (**c**), and Chr12 (**d**). Best substitution model estimated with `ModelFinder` [63] and trees estimated with `iqtree` [64]. The PCA groups *R*, *S*, *SS*, *T*, and *U* of Atlantic cod (*Gadus morhua*) and the gadid taxa Polar cod (Agl, *Arctogadus glacialis*), Arctic cod (Bsa *Boreogadus saida*), walleye pollock (Gch *G. chalcogrammus*), Pacific cod (Gma *G. macrocephalus*), and Greenland cod (Gog *G. ogac*). The outgroup taxa are haddock (Mae *Melanogrammus aeglefinus*) and whiting (Mme *Merlangius merlangus*). The blue and red bars indicate homozygotes for the reference and the alternative allele, respectively. Chromosomal coordinates inside inversions: 11,763,336 to 27,936,178 for Chr01, 579,097 to 4,096,513 for Chr02, 17,004,745 to 25,979,379 for Chr07, and 1,001,809 to 13,965,277 for Chr12.

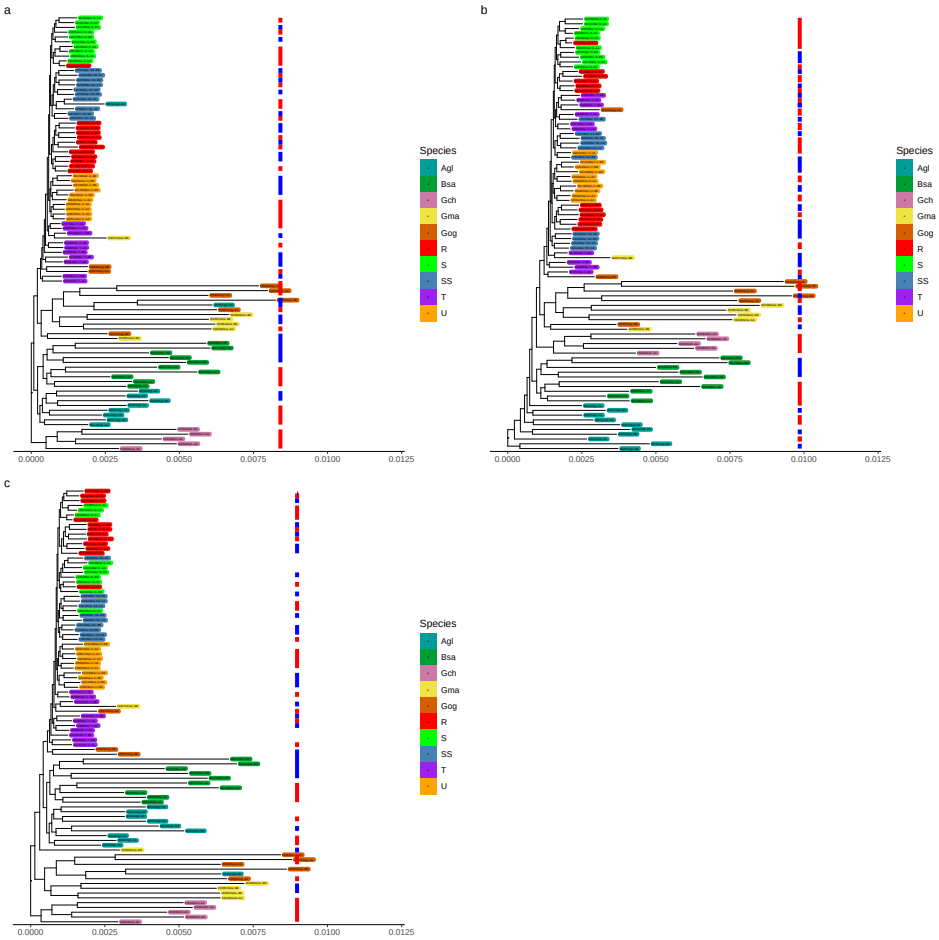

Fig. B26 Phylogeny of non-inversion chromosomes. Maximum likelihood phylogenetic trees of Chr04, Chr17, and Chr23 (**a**, **b**, and **c** respectively) for individuals homozygous for *AA* and *BB* genotypes at Chr01 (Fig. 5 **a**). Best substitution model estimated with `ModelFinder` [63] and trees estimated with `iqtree` [64]. The blue and red bars indicate homozygotes for the reference and the alternative allele respectively at Chr01.

B.6.2 Treemix network analysis of trans-species inversion genotypes

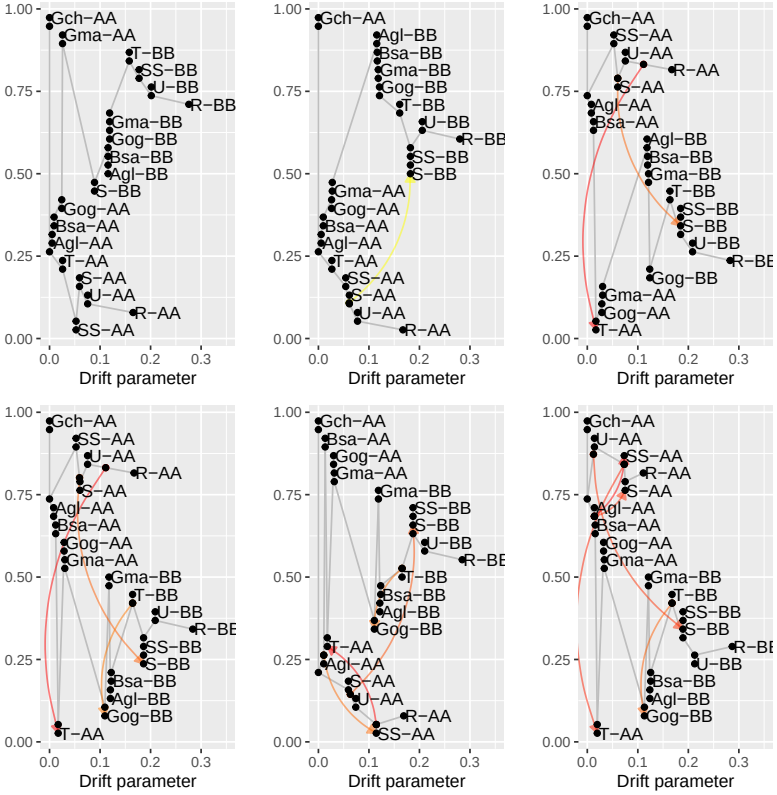

Fig. B27 Treemix network analysis of species and PCA groups of inversions on Chr01. Maximum likelihood treemix network analysis [76] of homozygous genotypes of PCA groups and gadid taxa for Chr01 (individuals studied in Fig. B25 **a**), usually five to six individuals per group. The PCA groups *R*, *S*, *SS*, *T*, and *U* of Atlantic cod (*Gadus morhua*) and the gadid taxa Polar cod (Agl, *Arctogadus glacialis*), Arctic cod (Bsa *Boreogadus saida*), walleye pollock (Gch *G. chalcogrammus*), Pacific cod (Gma *G. macrocephalus*), and Greenland cod (Gog *G. ogac*). The number of migration edges $m = 0 \ldots 5$ (top left to bottom right).

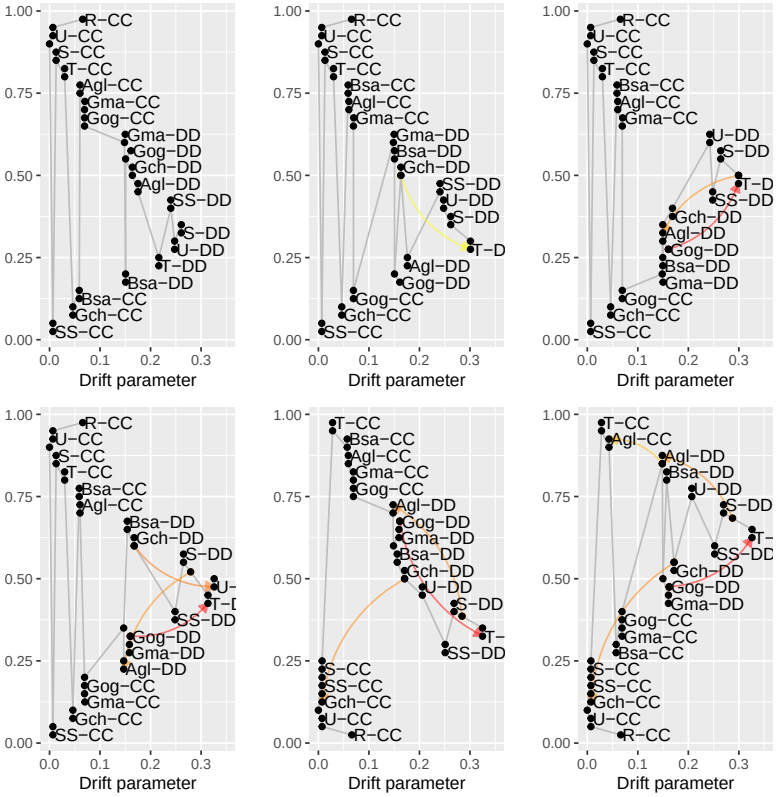

Fig. B28 Treemix network analysis of species and PCA groups of inversions on Chr02. Maximum likelihood treemix network analysis [76] of homozygous genotypes of PCA groups and gadid taxa for Chr02 (individuals studied in Fig. B25 **b**), usually five to six individuals per group. The PCA groups *R*, *S*, *SS*, *T*, and *U* of Atlantic cod (*Gadus morhua*) and the gadid taxa Polar cod (Agl, *Arctogadus glacialis*), Arctic cod (Bsa *Boreogadus saida*), walleye pollock (Gch *G. chalcogrammus*), Pacific cod (Gma *G. macrocephalus*), and Greenland cod (Gog *G. ogac*). The number of migration edges $m = 0 \ldots 5$ (top left to bottom right).

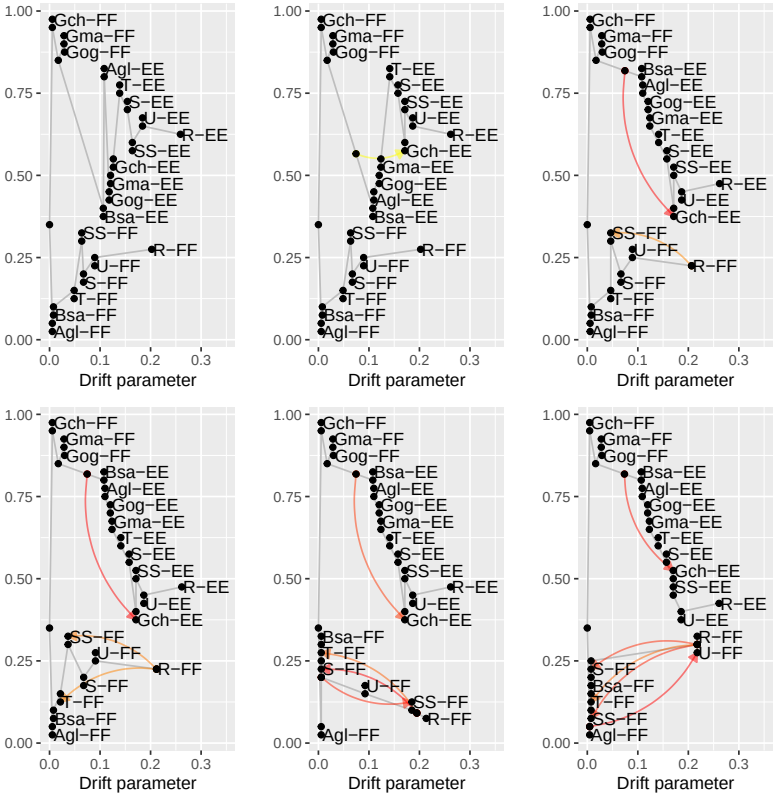

Fig. B29 Treemix network analysis of species and PCA groups of inversions on Chr07. Maximum likelihood treemix network analysis [76] of homozygous genotypes of PCA groups and gadid taxa for Chr07 (individuals studied in Fig. B25 **c**), usually five to six individuals per group. The PCA groups *R*, *S*, *SS*, *T*, and *U* of Atlantic cod (*Gadus morhua*) and the gadid taxa Polar cod (Agl, *Arctogadus glacialis*), Arctic cod (Bsa *Boreogadus saida*), walleye pollock (Gch *G. chalcogrammus*), Pacific cod (Gma *G. macrocephalus*), and Greenland cod (Gog *G. ogac*). The number of migration edges $m = 0 \ldots 5$ (top left to bottom right).

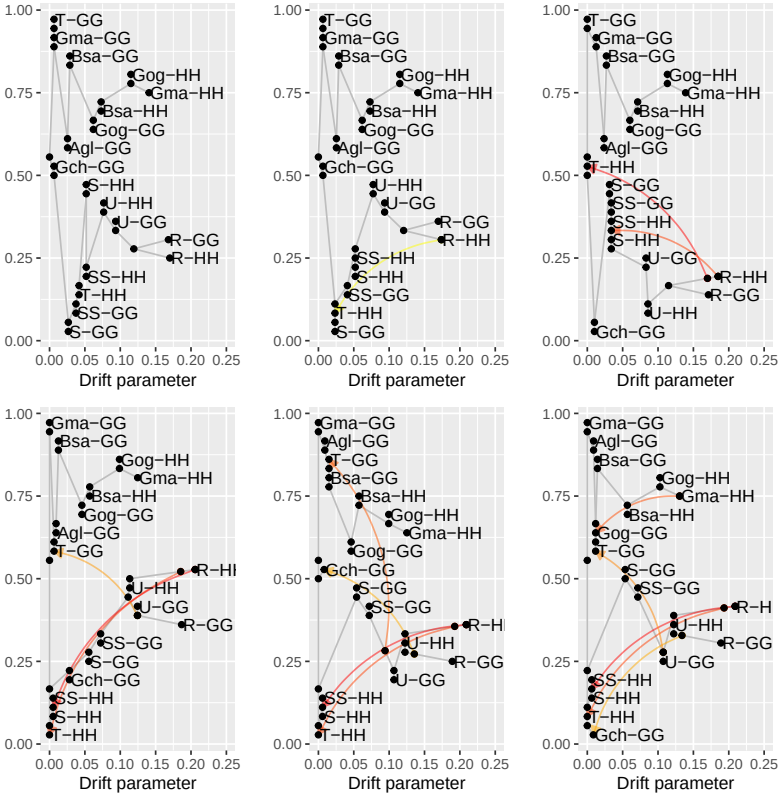

Fig. B30 Treemix network analysis of species and PCA groups of inversions on Chr12. Maximum likelihood treemix network analysis [76] of homozygous genotypes of PCA groups and gadid taxa for Chr12 (individuals studied in Fig. B25 **d**), usually five to six individuals per group. The PCA groups *R*, *S*, *SS*, *T*, and *U* of Atlantic cod (*Gadus morhua*) and the gadid taxa Polar cod (Agl, *Arctogadus glacialis*), Arctic cod (Bsa *Boreogadus saida*), walleye pollock (Gch *G. chalcogrammus*), Pacific cod (Gma *G. macrocephalus*), and Greenland cod (Gog *G. ogac*). The number of migration edges $m = 0 \ldots 5$ (top left to bottom right).

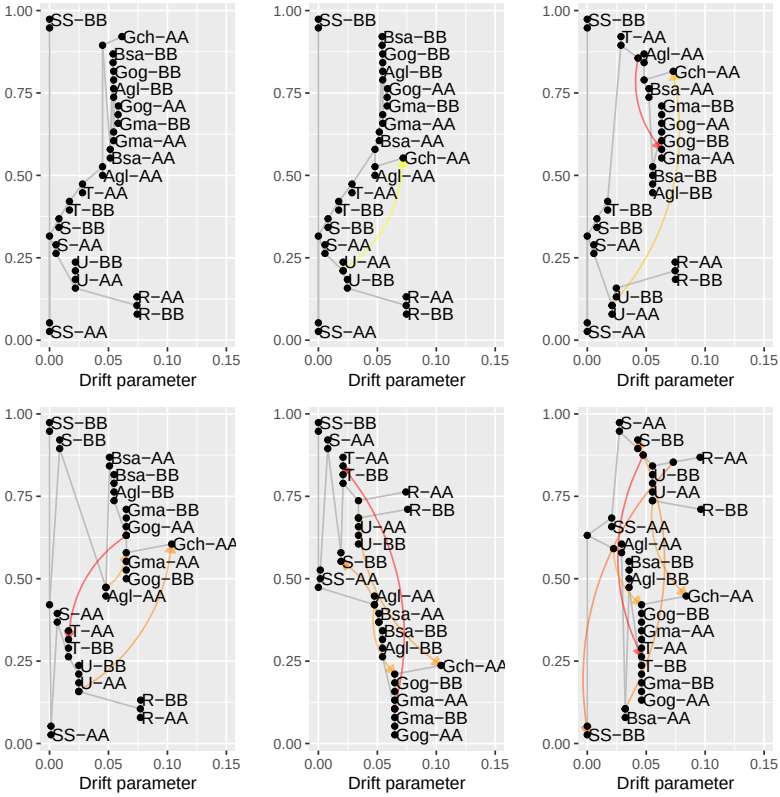

Fig. B31 Treemix network analysis of species and PCA groups of non-inversions Chr04. Maximum likelihood treemix network analysis [76] of non-inversion Chr04 using individuals of homozygous genotypes for Chr01 (individuals studied in Fig. B25 **a**), usually five to six individuals per group. The PCA groups *R*, *S*, *SS*, *T*, and *U* of Atlantic cod (*Gadus morhua*) and the gadid taxa Polar cod (Agl, *Arctogadus glacialis*), Arctic cod (Bsa *Boreogadus saida*), walleye pollock (Gch *G. chalcogrammus*), Pacific cod (Gma *G. macrocephalus*), and Greenland cod (Gog *G. ogac*). The number of migration edges $m = 0 \ldots 5$ (top left to bottom right).

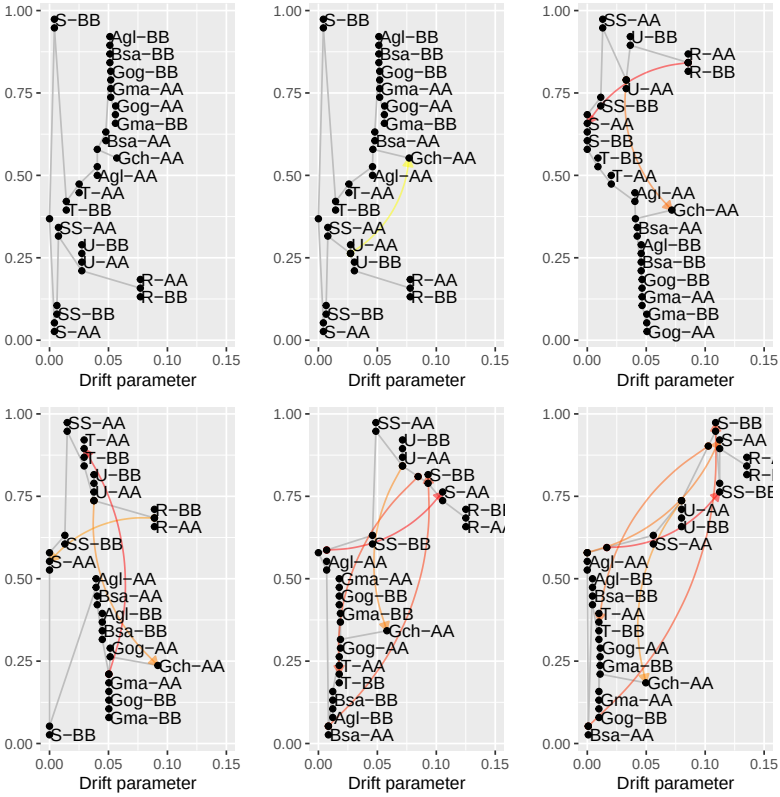

Fig. B32 Treemix network analysis of species and PCA groups of non-inversions Chr17. Maximum likelihood treemix network analysis [76] of non-inversion Chr04 using individuals of homozygous genotypes for Chr01 (individuals studied in Fig. B25 **a**), usually five to six individuals per group. The PCA groups *R*, *S*, *SS*, *T*, and *U* of Atlantic cod (*Gadus morhua*) and the gadid taxa Polar cod (Agl, *Arctogadus glacialis*), Arctic cod (Bsa *Boreogadus saida*), walleye pollock (Gch *G. chalcogrammus*), Pacific cod (Gma *G. macrocephalus*), and Greenland cod (Gog *G. ogac*). The number of migration edges $m = 0 \ldots 5$ (top left to bottom right).

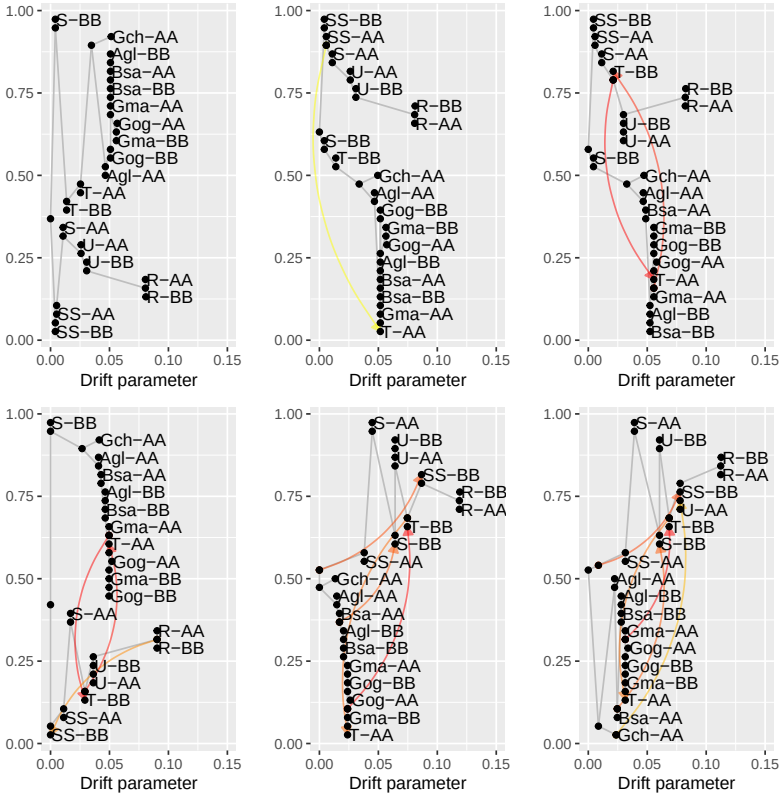

Fig. B33 Treemix network analysis of species and PCA groups of non-inversions Chr23. Maximum likelihood treemix network analysis [76] of non-inversion Chr04 using individuals of homozygous genotypes for Chr01 (individuals studied in Fig. B25 **a**), usually five to six individuals per group. The PCA groups *R*, *S*, *SS*, *T*, and *U* of Atlantic cod (*Gadus morhua*) and the gadid taxa Polar cod (Agl, *Arctogadus glacialis*), Arctic cod (Bsa *Boreogadus saida*), walleye pollock (Gch *G. chalcogrammus*), Pacific cod (Gma *G. macrocephalus*), and Greenland cod (Gog *G. ogac*). The number of migration edges $m = 0 \ldots 5$ (top left to bottom right).

References

[1] Árnason, E., Koskela, J., Halldórsdóttir, K. & Eldon, B. Sweepstakes reproductive success via pervasive and recurrent selective sweeps. *eLife* **12** (2023). URL https://doi.org/10.7554/elife.80781. https://doi.org/10.7554/elife.80781 .

[2] Einarsson, A. W., Árnason, E. & Halldórsdóttir, K. Genomic evidence for large inversions in the polar cod *Boreogadus saida* at similar chromosomal locations as found in Atlantic cod *Gadus morhua*. *Research Square Platform LLC* (2023). URL https://doi.org/10.21203/rs.3.rs-3097434/v1. https://doi.org/10.21203/rs.3.rs-3097434/v1 .

[3] Muller, H. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **1** (1), 2–9 (1964). URL https://doi.org/10.1016/0027-5107(64)90047-8. https://doi.org/10.1016/0027-5107(64)90047-8 .

[4] Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78** (2), 737–756 (1974). URL https://doi.org/10.1093/genetics/78.2.737. https://doi.org/10.1093/genetics/78.2.737 .

[5] Berdan, E. L., Blanckaert, A., Butlin, R. K. & Bank, C. Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLOS Genetics* **17** (3), e1009411 (2021). URL https://doi.org/10.1371/journal.pgen.1009411. https://doi.org/10.1371/journal.pgen.1009411 .

[6] Pegueroles, C., Aquadro, C. F., Mestres, F. & Pascual, M. Gene flow and gene flux shape evolutionary patterns of variation in *Drosophila subobscura*. *Heredity* **110** (6), 520–529 (2013). URL https://doi.org/10.1038/hdy.2012.118. https://doi.org/10.1038/hdy.2012.118 .

[7] Korunes, K. L. & Noor, M. A. F. Pervasive gene conversion in chromosomal inversion heterozygotes. *Molecular Ecology* **28** (6), 1302–1315 (2018). URL https://doi.org/10.1111/mec.14921. https://doi.org/10.1111/mec.14921 .

[8] Schluter, D. Ecology and the origin of species. *Trends in Ecology & Evolution* **16** (7), 372–380 (2001). URL https://doi.org/10.1016/s0169-5347(01)02198-x. https://doi.org/10.1016/s0169-5347(01)02198-x .

[9] Rundle, H. D. & Nosil, P. Ecological speciation. *Ecology Letters* **8** (3), 336–352 (2005). URL https://doi.org/10.1111/j.1461-0248.2004.00715.x. https://doi.org/10.1111/j.1461-0248.2004.00715.x .

[10] Nosil, P. Ernst Mayr and the integration of geographic and ecological factors in speciation. *Biological Journal of the Linnean Society* **95** (1), 26–46 (2008). URL https://doi.org/10.1111/j.1095-8312.2008.01091.x. https://doi.org/10.1111/j.1095-8312.2008.01091.x .

[11] Fuller, Z. L., Leonard, C. J., Young, R. E., Schaeffer, S. W. & Phadnis, N. Ancestral polymorphisms explain the role of chromosomal inversions in speciation. *PLOS Genetics* **14** (7), e1007526 (2018). URL https://doi.org/10.1371/journal.pgen.1007526. https://doi.org/10.1371/journal.pgen.1007526 .

[12] Charlesworth, D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics* **2** (4), e64 (2006). URL https://dotheymayi.org/10.1371/journal.pgen.0020064. https://doi.org/10.1371/

journal.pgen.0020064 .

[13] Halldórsdóttir, K. & Árnason, E. Trans-species polymorphism at antimicrobial innate immunity cathelicidin genes of Atlantic cod and related species. *PeerJ* **3**, e976 (2015). URL https://doi.org/10.7717/peerj.976. https://doi.org/10.7717/peerj.976 .

[14] Knowlton, N. Sibling species in the sea. *Annual Review Ecology Systematics* **24** (1), 189–216 (1993). URL http://dx.doi.org/10.1146/annurev.es.24.110193.001201. https://doi.org/10.1146/annurev.es.24.110193.001201 .

[15] Árnason, E. & Halldórsdóttir, K. Codweb: Whole-genome sequencing uncovers extensive reticulations fueling adaptation among Atlantic, Arctic, and Pacific gadids. *Science Advances* **5** (3), eaat8788 (2019). URL https://doi.org/10.1126/sciadv.aat8788. https://doi.org/10.1126/sciadv.aat8788 .

[16] Palumbi, S. R. Genetic divergence, reproductive isolation, and marine speciation. *Annual Review of Ecology and Systematics* **25** (1), 547–572 (1994). URL http://dx.doi.org/10.1146/annurev.es.25.110194.002555. https://doi.org/10.1146/annurev.es.25.110194.002555 .

[17] MacLean, J. A. & Evans, D. O. The stock concept, discreteness of fish stocks, and fisheries management. *Canadian Journal of Fisheries and Aquatic Sciences* **38** (12), 1889–1898 (1981). URL https://doi.org/10.1139/f81-235. https://doi.org/10.1139/f81-235 .

[18] Kirubakaran, T. G. *et al.* Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology* **2016**, 2130–2143 (2016). URL http://dx.doi.org/10.1111/mec.13592. https://doi.org/10.1111/mec.13592 .

[19] Berg, P. R. *et al.* Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Scientific Reports* **6**, 23246 (2016). URL http://dx.doi.org/10.1038/srep23246. https://doi.org/10.1038/srep23246 .

[20] Thorsteinsson, V., Pálsson, Ó., Tómasson, G., Jónsdóttir, I. & Pampoulie, C. Consistency in the behaviour types of the Atlantic cod: Repeatability, timing of migration and geo-location. *Marine Ecology Progress Series* **462**, 251–260 (2012). URL http://dx.doi.org/10.3354/meps09852. https://doi.org/10.3354/meps09852 .

[21] Berg, P. R. *et al.* Adaptation to low salinity promotes genomic divergence in Atlantic cod (*Gadus morhua* L.). *Genome Biology and Evolution* **7** (6), 1644–1663 (2015). URL http://dx.doi.org/10.1093/gbe/evv093. https:

//doi.org/10.1093/gbe/evv093 .

[22] Sodeland, M. *et al.* "Islands of Divergence" in the Atlantic cod genome represent polymorphic chromosomal rearrangements. *Genome Biology and Evolution* **8** (4), 1012–1022 (2016). URL http://dx.doi.org/10.1093/gbe/evw057. https://doi.org/10.1093/gbe/evw057 .

[23] Hernandez, U. B. & Árnason, E. Living in each other's pockets: Nucleotide variation inside a genomic island harboring *Pan* I and its neighbors in Atlantic cod. PeerJ PrePrints (2015). URL http://dx.doi.org/10.7287/peerj.preprints.956v1.

[24] Matschiner, M. *et al.* Supergene origin and maintenance in Atlantic cod. *Nature Ecology & Evolution* (2022). URL https://doi.org/10.1038/s41559-022-01661-x. https://doi.org/10.1038/s41559-022-01661-x .

[25] Nielsen, E. E., Hansen, M. M., Ruzzante, D. E., Meldrup, D. & Grønkjær, P. Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea revealed by individual admixture analysis. *Molecular Ecology* **12** (6), 1497–1508 (2003). URL https://doi.org/10.1046/j.1365-294x.2003.01819.x. https://doi.org/10.1046/j.1365-294x.2003.01819.x .

[26] Hutchinson, W., Carvalho, G. & Rogers, S. Marked genetic structuring in localised spawning populations of cod *Gadus morhua* in the North Sea and adjoining waters, as revealed by microsatellites. *Marine Ecology Progress Series* **223**, 251–260 (2001). URL https://doi.org/10.3354/meps223251. https://doi.org/10.3354/meps223251 .

[27] Therkildsen, N. O. *et al.* Spatiotemporal SNP analysis reveals pronounced biocomplexity at the northern range margin of Atlantic cod *Gadus morhua*. *Evolutionary Applications* **6** (4), 690–705 (2013). URL http://dx.doi.org/10.1111/eva.12055. https://doi.org/10.1111/eva.12055 .

[28] Bradbury, I. R. *et al.* Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society, B.* **277**, 3725–3734 (2010). https://doi.org/doi:10.1098/rspb.2010.0985 .

[29] Bradbury, I. R. *et al.* Long distance linkage disequilibrium and limited hybridization suggest cryptic speciation in Atlantic cod. *PLoS ONE* **9** (9), e106380 (2014). URL http://dx.doi.org/10.1371/journal.pone.0106380. https://doi.org/10.1371/journal.pone.0106380 .

[30] Van der Auwera, G. A. *et al.* From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current*

Protocols in Bioinformatics **43** (1) (2013). URL https://doi.org/10.1002/0471250953.bi1110s43. https://doi.org/10.1002/0471250953.bi1110s43 .

[31] Li, H. & Ralph, P. Local PCA shows how the effect of population structure differs along the genome. *Genetics* **211** (1), 289–304 (2018). URL https://doi.org/10.1534/genetics.118.301747. https://doi.org/10.1534/genetics.118.301747 .

[32] Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19** (9), 1655–1664 (2009). URL https://doi.org/10.1101/gr.094052.109. https://doi.org/10.1101/gr.094052.109 .

[33] Magnussen, E. Food and feeding habits of cod (*Gadus morhua*) on the Faroe Bank. *ICES Journal of Marine Science* **68** (9), 1909–1917 (2011). URL https://doi.org/10.1093/icesjms/fsr104. https://doi.org/10.1093/icesjms/fsr104 .

[34] Yi, X. & Latch, E. K. Nonrandom missing data can bias Principal Component Analysis inference of population genetic structure. *Molecular Ecology Resources* **22** (2), 602–611 (2021). URL https://doi.org/10.1111/1755-0998.13498. https://doi.org/10.1111/1755-0998.13498 .

[35] Elhaik, E. Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports* **12** (1) (2022). URL https://doi.org/10.1038/s41598-022-14395-4. https://doi.org/10.1038/s41598-022-14395-4 .

[36] Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P. & Ramachandran, S. pong: Fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* **32** (18), 2817–2823 (2016). URL https://doi.org/10.1093/bioinformatics/btw327. https://doi.org/10.1093/bioinformatics/btw327 .

[37] Hutchings, J. A. Collapse and recovery of marine fishes. *Nature* **406** (6798), 882–885 (2000). URL https://doi.org/10.1038/35022565. https://doi.org/10.1038/35022565 .

[38] Hewitt, G. M. Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society* **68** (1-2), 87–112 (1999). URL https://doi.org/10.1111/j.1095-8312.1999.tb01160.x. https://doi.org/10.1111/j.1095-8312.1999.tb01160.x .

[39] Laakkonen, H. M., Hardman, M., Strelkov, P. & Väinölä, R. Cycles of trans-Arctic dispersal and vicariance, and diversification of the amphiboreal marine fauna. *Journal of Evolutionary Biology* **34** (1), 73–96 (2020). URL https://doi.org/10.1111/jeb.13674. https://doi.org/10.

1111/jeb.13674 .

[40] Faria, R., Johannesson, K. & Stankowski, S. Speciation in marine environments: Diving under the surface. *Journal of Evolutionary Biology* **34** (1), 4–15 (2021). URL https://doi.org/10.1111/jeb.13756. https://doi.org/10.1111/jeb.13756 .

[41] Wielstra, B. Balanced lethal systems. *Current Biology* **30** (13), R742–R743 (2020). URL https://doi.org/10.1016/j.cub.2020.05.011. https://doi.org/10.1016/j.cub.2020.05.011 .

[42] Nosil, P. *Ecological Speciation* (Oxford University Press, Oxford, 2012).

[43] Hutchings, J. A., Bishop, T. D. & McGregor-Shaw, C. R. Spawning behaviour of Atlantic cod, *Gadus morhua*: Evidence of mate competition and mate choice in a broadcast spawner. *Canadian Journal of Fisheries and Aquatic Sciences* **56** (1), 97–104 (1999). URL https://doi.org/10.1139/f98-216. https://doi.org/10.1139/f98-216 .

[44] Nordeide & Folstad. Is cod lekking or a promiscuous group spawner? *Fish and Fisheries* **1** (1), 90–93 (2000). URL https://doi.org/10.1046/j.1467-2979.2000.00005.x. https://doi.org/10.1046/j.1467-2979.2000.00005.x .

[45] Rudolfsen, G., Figenschou, L., Folstad, I., Nordeide, J. T. & Soreng, E. Potential fitness benefits from mate selection in the Atlantic cod (*Gadus morhua*). *Journal of Evolutionary Biology* **18** (1), 172–179 (2005). URL https://doi.org/10.1111/j.1420-9101.2004.00778.x. https://doi.org/10.1111/j.1420-9101.2004.00778.x .

[46] Grabowski, T. B., Thorsteinsson, V., McAdam, B. J. & Marteinsdóttir, G. Evidence of segregated spawning in a single marine fish stock: Sympatric divergence of ecotypes in Icelandic cod? *PLoS ONE* **6** (3), e17528 (2011). URL http://dx.doi.org/10.1371/journal.pone.0017528. https://doi.org/10.1371/journal.pone.0017528 .

[47] Schluter, D. & Conte, G. L. Genetics and ecological speciation. *Proceedings of the National Academy of Sciences* **106** (Supplement_1), 9955–9962 (2009). URL https://doi.org/10.1073/pnas.0901264106. https://doi.org/10.1073/pnas.0901264106 .

[48] Turner, T. L., Hahn, M. W. & Nuzhdin, S. V. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology* **3** (9), e285 (2005). URL http://dx.doi.org/10.1371/journal.pbio.0030285. https://doi.org/10.1371/journal.pbio.0030285 .

[49] Nosil, P. Speciation with gene flow could be common. *Molecular Ecology* **17** (9), 2103–2106 (2008). URL http://dx.doi.org/10.1111/j.1365-294X.2008.03715.x. https://doi.org/10.1111/j.1365-294X.2008.03715.x .

[50] Hemmer-Hansen, J. *et al.* A genomic island linked to ecotype divergence in Atlantic cod. *Molecular Ecology* **22**, 2653–2667 (2013). URL http://dx.doi.org/10.1111/mec.12284. https://doi.org/10.1111/mec.12284 .

[51] Cruickshank, T. E. & Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology* **23** (13), 3133–3157 (2014). URL http://dx.doi.org/10.1111/mec.12796. https://doi.org/10.1111/mec.12796 .

[52] Árnason, E. & Halldórsdóttir, K. Nucleotide variation and balancing selection at the *Ckma* gene in Atlantic cod: analysis with multiple merger coalescent models. *PeerJ* **3**, e786 (2015). URL http://dx.doi.org/10.7717/peerj.786. https://doi.org/10.7717/peerj.786 .

[53] Seehausen, O. *et al.* Speciation through sensory drive in cichlid fish. *Nature* **455** (7213), 620–626 (2008). URL https://doi.org/10.1038/nature07285. https://doi.org/10.1038/nature07285 .

[54] Pampoulie, C. *et al.* Rhodopsin gene polymorphism associated with divergent light environments in Atlantic cod. *Behavior Genetics* **45** (2), 236–244 (2015). URL https://doi.org/10.1007/s10519-014-9701-7. https://doi.org/10.1007/s10519-014-9701-7 .

[55] Jamieson, A. & Jones, B. W. Two races of cod at Faroe. *Heredity* **22** (4), 610–612 (1967). URL https://doi.org/10.1038/hdy.1967.76. https://doi.org/10.1038/hdy.1967.76 .

[56] Coulson, M. W., Marshall, H., Pepin, P. & Carr, S. Mitochondrial genomics of gadine fishes: implications for taxonomy and biogeographic origins from whole-genome data sets. *Genome* **49** (9), 1115–1130 (2006). URL http://dx.doi.org/10.1139/g06-083. https://doi.org/10.1139/g06-083 .

[57] Olsen, E. M. *et al.* Maturation trends indicative of rapid evolution preceded the collapse of northern cod. *Nature* **428**, 932–935 (2004). https://doi.org/10.1038/nature02430 .

[58] Árnason, E., Hernandez, U. B. & Kristinsson, K. Intense habitat-specific fisheries-induced selection at the molecular *Pan* I locus predicts imminent collapse of a major cod fishery. *PLoS ONE* **4** (5), e5529 (2009). URL http://dx.doi.org/10.1371/journal.pone.0005529. https://doi.org/10.1371/journal.pone.0005529 .

[59] Reid, B. N., Star, B. & Pinsky, M. L. Detecting parallel polygenic adaptation to novel evolutionary pressure in wild populations: a case study in Atlantic cod (*Gadus morhua*). *Philosophical Transactions of the Royal Society B* **378** (1881) (2023). URL https://doi.org/10.1098/rstb.2022.0190. https://doi.org/10.1098/rstb.2022.0190 .

[60] Johri, P. *et al.* Recommendations for improving statistical inference in population genomics. *PLOS Biology* **20** (5), e3001669 (2022). URL https://doi.org/10.1371/journal.pbio.3001669. https://doi.org/10.1371/journal.pbio.3001669 .

[61] Charlesworth, B. & Jensen, J. D. Effects of selection at linked sites on patterns of genetic variability. *Annual Review of Ecology, Evolution, and Systematics* **52** (1), 177–197 (2021). URL https://doi.org/10.1146/annurev-ecolsys-010621-044528. https://doi.org/10.1146/annurev-ecolsys-010621-044528 .

[62] Lauterbur, M. E. *et al.* Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *eLife* (2023). URL https://doi.org/10.7554/elife.84874.2. https://doi.org/10.7554/elife.84874.2 .

[63] Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods* **14** (6), 587–589 (2017). URL https://doi.org/10.1038/nmeth.4285. https://doi.org/10.1038/nmeth.4285 .

[64] Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* **37** (5), 1530–1534 (2020). URL https://doi.org/10.1093/molbev/msaa015. https://doi.org/10.1093/molbev/msaa015 .

[65] Gaspar, J. M. NGmerge: Merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics* **19** (1) (2018). URL https://doi.org/10.1186/s12859-018-2579-2. https://doi.org/10.1186/s12859-018-2579-2 .

[66] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25** (14), 1754–1760 (2009). URL http://dx.doi.org/10.1093/bioinformatics/btp324. https://doi.org/10.1093/bioinformatics/btp324 .

[67] Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10** (2) (2021). URL https://doi.org/10.1093/gigascience/giab008. https://doi.org/10.1093/gigascience/giab008 .

[68] Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27** (15), 2156–2158 (2011). URL https://doi.org/10.1093/bioinformatics/

btr330. https://doi.org/10.1093/bioinformatics/btr330 .

[69] Purcell, S. *et al.* PLINK a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81** (3), 559–575 (2007). URL https://doi.org/10.1086/519795. https://doi.org/10.1086/519795 .

[70] Purcell, S. & Chang, C. PLINK v1.90b6.16. URL www.cog-genomics.org/plink/1.9/.

[71] Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* **4** (43), 1686 (2019). URL https://doi.org/10.21105/joss.01686. https://doi.org/10.21105/joss.01686 .

[72] Refoyo-Martínez, A. *et al.* Identifying loci under positive selection in complex population histories. *bioRχiv* (2018). https://doi.org/10.1101/453092 .

[73] Kirubakaran, T. G. *et al.* A nanopore based chromosome-level assembly representing Atlantic cod from the Celtic Sea. *bioRxiv* (2019). URL https://www.biorxiv.org/content/early/2019/11/22/852145. https://doi.org/10.1101/852145, https://www.biorxiv.org/content/early/2019/11/22/852145.full.pdf .

[74] Kolde, R. *pheatmap: Pretty Heatmaps* (2019). URL https://CRAN.R-project.org/package=pheatmap. R package version 1.0.12.

[75] Webb, A. *et al.* The pop-gen pipeline platform: A software platform for population genomic analyses. *Molecular Biology and Evolution* **38** (8), 3478–3485 (2021). URL https://doi.org/10.1093/molbev/msab113. https://doi.org/10.1093/molbev/msab113 .

[76] Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* **8** (11), e1002967 (2012). URL https://doi.org/10.1371/journal.pgen.1002967. https://doi.org/10.1371/journal.pgen.1002967 .

[77] Morgan, A. *popcorn: File-Format Wrangler for Population-Based Genetic Data* (2023). R package version 0.0.2.

[78] Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35** (2), 518–522 (2017). URL https://doi.org/10.1093/molbev/msx281. https://doi.org/10.1093/molbev/msx281 .

[79] Yu, G. Using ggtree to visualize data on tree-like structures. *Current Protocols in Bioinformatics* **69** (1) (2020). URL https://doi.org/10.1002/

cpbi.96. https://doi.org/10.1002/cpbi.96 .

[80] Yu, G., Lam, T. T.-Y., Zhu, H. & Guan, Y. Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular Biology and Evolution* **35** (12), 3041–3043 (2018). URL https://doi.org/10.1093/molbev/msy194. https://doi.org/10.1093/molbev/msy194 .