Article

Origin and fate of supergenes in Atlantic cod

Michael Matschiner[1,2,3]*, Julia Maria Isis Barth[4], Ole Kristian Tørresen[1], Bastiaan Star[1], Helle Tessand Baalsrud[1], Marine Servane Ono Brieuc[1], Christophe Pampoulie[5], Ian Bradbury[6], Kjetill Sigurd Jakobsen[1], Sissel Jentoft[1]*

Addresses:
[1]Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway.
[2]Department of Palaeontology and Museum, University of Zurich, Zurich, Switzerland.
[3]Current address: Natural History Museum, University of Oslo, Oslo, Norway.
[4]Zoological Institute, Department of Environmental Sciences, University of Basel, Basel, Switzerland.
[5]Marine and Freshwater Research Institute, Hafnarfjördur, Iceland.
[6]Fisheries and Oceans Canada, St. John's, Canada.
*Corresponding authors: E-mail: michael.matschiner@nhm.uio.no, sissel.jentoft@ibv.uio.no

Abstract

Supergenes are sets of genes that are inherited as a single marker and encode complex phenotypes through their joint action. They are identified in an increasing number of organisms, yet their origins and evolution remain enigmatic. In Atlantic cod, four large supergenes have been identified and linked to migratory lifestyle and environmental adaptations. Here, we investigate the origin and fate of these four supergenes through analysis of whole-genome-sequencing data, including a new long-read-based genome assembly for a non-migratory Atlantic cod individual. We corroborate that chromosomal inversions underlie all four supergenes, and show that they originated separately, between 0.40 and 1.66 million years ago. While introgression was not involved in the origin of the four supergenes, we reveal gene flow between inverted and noninverted supergene haplotypes, occurring both through gene conversion and double crossover. Moreover, the presence of genes linked to salinity adaptations in a sequence transferred through double crossover indicates that these sequences exchanged between the haplotypes are subject to selection. Our results suggest that the fate of supergenes is comparable to that of hybridizing species, by depending on the degree to which separation is maintained through purging of introduced genetic variation.

Keywords: Supergene; Inversion; Introgression; Gene conversion; Double crossover; Selection; Atlantic cod

Introduction

Some of the most spectacular phenotypical variation within species, such as mimicry patterns in butterflies[1], social organization in ants[2], plumage morphs in birds[3, 4], and floral types in plants[5], is encoded by supergenes — tightly linked sets of coadapted genes that are inherited as a single Mendelian locus[6–9]. Even though supergenes have now been known for nearly a century[10], their origin remains a challenging question[9], because it requires that beneficially interacting mutations occur in at least two genes together with a reduction of recombination between these genes[7]. As a scenario in which these requirements can be met, recent research has pointed to chromosomal inversions arising in incompletely separated groups: either locally adapted populations that exchange migrants, or species that receive genetic material — introgression — through hybridization[11–13]. In these systems, the beneficial interaction between mutations in different genes can come from their joint adaptation to the same environment, and these mutations can become linked if they are captured by the same inversion[6–9, 14]. This linkage between mutations within inversions is the result of a loop formation that occurs when chromosomes with the inverted haplotype pair with chromosomes without it during meiosis. If a single crossover occurs between the two chromosomes within the loop region, the recombinant chromosomes are affected by both duplications and deletions and therefore unbalanced. The gametes carrying these unbalanced chromosomes are usually lethal, and thus do not contribute to the next generation[15, 16]. When viewed backwards in time, the recombination rate between inverted and noninverted haplotypes therefore appears reduced. On the other hand, crossovers among two inverted haplotypes do not affect the viability of gametes, so that their recombination rate remains unchanged. Since most inversions originate just once, in a single individual, the number of individuals in which inverted haplotypes can successfully recombine is initially very low, increasing only as the inversion becomes more frequent in the species. The origin of a supergene is therefore expected to be equivalent to a severe bottleneck (down to a single sequence) that affects part of the genome (the inversion region) in a part of the species (the carriers of the inversion)[17].

Once established, the fate of supergenes depends on the interaction of selection, mutation, and recombination. The spread of an inversion within a species can be halted — and the supergene can thus remain polymorphic — by frequency-dependent selection, by heterogeneous selection regimes in different populations, or by recessive deleterious mutations that accumulate in the inversion region[7, 11, 13, 18]. As mutations are added over time, the inverted and noninverted haplotypes diverge from each other, due to the suppression of recombination between them[18]. Owing to the reduced opportunity for recombination, mildly deleterious mutations are more likely to be fixed inside the inversion region compared to outside and can result in the accumulation of mutational load[18]. When the mutational load becomes high within a supergene, it can lead to its decay[19], similar to the degeneration observed in sex chromosomes[20, 21] (which are often considered a special case of supergenes[6, 7]). This decay, however, can be counteracted by two processes that can allow genetic exchange between inverted and noninverted haplotypes despite the suppression of recombination between them[2, 9, 19, 22]: Short fragments with lengths on the order of 50–1,000 bp[23, 24] can be copied through gene conversion, a process in which a homologous sequence is used as template during the repair of a double-strand break, without requiring crossover with that homologous sequence[25, 26]. Gene conversion was found to copy sequences at rates around 6×10^{-6} and $1.0 - 2.5 \times 10^{-5}$ per site and generation in humans and *Drosophila*, respectively[24, 27], and it is known to increase the GC-content of the involved sequences due to biased repair of A-C and G-T mismatches[26]. Longer fragments can be exchanged when double crossovers occur within the loop formed when chromosomes with and without inverted haplotypes pair during meiosis, in which case recombinant chromosomes are not unbalanced and gametes are viable[15]. Double crossovers have been observed at rates around 10^{-4} to 10^{-3} per generation in *Drosophila*[28], and are more likely to affect central regions of inversions than those near the inversion breakpoints[22, 29, 30]. Either alone or in tandem, the two processes could have the potential to erode differences between inverted and noninverted haplotypes if their per-site rates are high relative to the mutation rate[27]. However, outside of model systems like *Drosophila* that allow the genetic analysis of crosses produced in the laboratory, the rates of gene conversion and double crossovers are largely unknown, so that the fate of supergenes could differ from species to species.

In Atlantic cod (*Gadus morhua*), large genomic regions with tight linkage over 4–17 Megabase-pairs (Mbp) and strong differentiation between alternative haplotypes have been identified[31–34] on linkage groups (LGs) 1, 2, 7, and 12 of the gadMor2 reference genome assembly[35]. While the functional consequences of this differentiation remain largely unknown, the classification of the four regions as supergenes[36–39] is supported by the associations of the alternative haplotypes with different lifestyles [33, 36, 39] and environments[32, 37, 38, 40, 41]. One of the strongest of these associations is found between the alternative haplotypes on LG 1 and migratory and stationary Atlantic cod ecotypes[33, 38]. In the Northeast Atlantic, these ecotypes co-occur during the spawning season in March and April along the Norwegian coast, but are separated throughout the rest of the year, which the migratory ecotype — the Northeast Arctic cod (NEAC) — spends in its native habitat in the Barents Sea[42]. The association between the alternative haplotypes and the two ecotypes is based on contrasting frequencies of these haplotypes, as almost all stationary individuals are homozygous for one of the haplotypes and migratory individuals are either heterozygous or homozygous for the other haplotype[33, 43, 44]. This pattern of contrasting frequencies is repeated between migratory and stationary Atlantic cod ecotypes in the Northwest Atlantic and on Iceland (perhaps less clearly in the latter case)[45–47], corroborating the presumed functional link between haplotypes and ecotypes. A number of genes from within the differentiated region on LG 1 have been proposed as candidate genes under selection, including the *Ca6* gene that might play a role in adaptations to feeding at greater depths in the migratory ecotype; however, the targets of selection remain difficult to identify reliably due to the tight linkage across the entire region with close to 800 genes[36].

Similar to the different frequencies of LG 1 haplotypes between migratory and stationary ecotypes, one of the two alternative haplotypes on LG 2 is far more frequent in Atlantic cod from the Baltic sea compared to the nearest North Atlantic populations and has been suggested to carry genes adapted to the low salinity of the Baltic Sea[32, 41]. The alternative haplotypes on LGs 7 and 12 also differ in their frequencies among Atlantic cod populations, with one of the two haplotypes in each case being nearly absent in the Irish and Celtic Sea, possibly in relation to adaptation to higher temperatures during the spawning season in these populations[47–49], which are among the southernmost populations in the Northeast Atlantic. Similar geographic distributions of haplotypes from the four supergenes have been interpreted as evidence for interchromosomal linkage among these haplotypes[31]; however, to what extent these similarities are the result of epistatic interactions, adaptation to the same environment, or merely drift, remains uncertain[47].

Chromosomal inversions have long been suspected[45] to be the cause of recombination suppression in the four supergenes in Atlantic cod, but only recently confirmed, first for the supergene on LG 1 with the help of detailed linkage maps for that linkage group (which revealed that the supergene is formed not just of one but two adjacent inversions)[36], and then for the three other supergenes through comparison of long-read-based genome assemblies[48]. By mapping contigs from outgroup species to two different assemblies for LG 1 — one for a migratory and one for a stationary individual — it was also determined that it is the stationary ecotype that primarily carries the ancestral, noninverted haplotype on LG 1; however, this type of analysis has so far not been performed for the supergenes on LGs 2, 7, and 12.

The age of the supergene on LG 1 was estimated to be around 1.6 million years (Myr) based on genetic distances between the two alternative haplotypes and homologous sequences of Greenland cod (*Gadus ogac*), in combination with an assumed divergence time between Greenland and Atlantic cod of 3.5 million years ago (Ma)[36]. This assumed divergence time, however, may be unreliable, as it was based on a mitochondrial substitution rate[50, 51] that was originally calculated for Caribbean fishes[52], and because the same divergence appears only about half as old in another recent study[53]. For the supergenes on LGs 2, 7, and 12, no age estimates have yet been published. Due to these uncertainties, conclusions about a possible joint origin of all four supergenes have necessarily remained highly speculative. The role of introgression in the origin of the Atlantic cod's supergenes has so far also been uncertain: While introgression among codfishes (subfamily Gadinae) has already been investigated in one former study based on genome-scale sequence data[54], the results of that study were affected by mislabeled sequences, reference bias, and incorrect application of statistical tests (Supplementary Note 1), and thus remain inconclusive regarding the occurrence of introgression.

Here, we investigate the origin and the fate of supergenes in Atlantic cod as follows: We generate a new long-read-based genome assembly for a stationary Atlantic cod individual from northern Norway as a complement to the existing gadMor2 assembly[35], which represents a migratory individual also from northern Norway. Importantly, these two assemblies carry alternative haplotypes at each of the four supergenes. Through comparison of the two assemblies with each other and with an outgroup assembly, we corroborate that inversions are the cause of recombination suppression for each supergene, pinpoint the chromosomal boundaries of supergenes, and identify inverted and noninverted haplotypes. Using Bayesian time-calibrated phylogenetic analyses of newly generated and previously available genomic data, we reveal separate times of origin for the four supergenes and identify traces of introgression among closely related codfishes. Through demographic analyses, we find signatures of past bottlenecks associated with the origin of supergenes, and by applying D-statistics and sliding-window phylogenetic inference, we detect the occurrence of genetic exchange between haplotypes, both through gene conversion and double crossovers. Our results suggest that the long-term existence of supergenes may depend on the interaction of local adaptation preventing the fixation of one haplotype, genetic exchange between haplotypes countering mutation load, and selection acting on exchanged sequences to maintain the separation of the two haplotypes.

Results

A genome assembly for a stationary *Gadus morhua* individual. To allow a comparison of genome architecture between migratory and stationary *Gadus morhua*, we performed PacBio and Illumina sequencing for a stationary *Gadus morhua* indivdual sampled in northern Norway, at the Lofoten islands (Fig. 1a). The resulting genome assembly (gadMor_Stat), produced with Celera Assembler[55] and improved with Pilon[56], had a size of 565,431,517 bp composed of a total of 6,961 contigs with an N50 length of 121,508 bp. When aligned to the gadMor2 genome assembly[35] (representing a migratory individual), the gadMor_Stat assembly was highly similar on almost all gadMor2 linkage groups, with a genetic distance of 0.0040–0.0053 between the two assemblies. The exception to this were the four supergenes on LGs 1, 2, 7, and 12, that all showed an elevated genetic distance of 0.0066–0.0129 between the two assemblies, confirming that — as suggested in tests based on preliminary sequencing data — the sequenced stationary individual carried at all four supergenes the haplotype that is alternative to the one in the gadMor2 assembly (Supplementary Table 1). To determine the chromosomal boundaries of the regions of tight linkage associated with the supergenes, we investigated linkage disequilibrium (LD) on LGs 1, 2, 7, and

12 with a dataset of single-nucleotide polymorphisms (SNPs) for 100 *Gadus morhua* individuals. By quantifying the strength of linkage per SNP as the sum of the distances (in bp) with which the SNP is strongly linked, we identified sharp declines of linkage marking the boundaries of all four supergenes (Fig. 1b, Table 1), as expected under the assumption of large-scale chromosomal inversions[57].

The presence of large inversions on each of the four linkage groups was further supported by alignments of contigs from the gadMor_Stat assembly to the gadMor2 assembly, as we identified several contigs with split alignments, of which one part mapped unambiguously near the beginning and another mapped near the end of a supergene (Supplementary Table 2). The positions of split contig alignments allowed us to pinpoint the inversion breakpoints on the four linkage groups with varying precision (Table 1). The most informative alignments were those near the beginnings of the supergenes on LGs 1 and 7, which in both cases placed the breakpoints within a window of approximately 2 kbp. As also reported for inversions in *Drosophila*[28], this precise placement of the inversion breakpoints revealed that they do not match the positions of LD onset exactly, but that they were located up to 45 kbp inside of the region of tight linkage (Fig. 1b; Table 1). None of the four inversion regions included centromeres[48]; thus, all of them seemed to contain paracentric inversions, which, in contrast to pericentric inversions, may not decrease the fitness of heterozygotes[58].

To determine which of the two genomes carries the inversion in each case, we also aligned contigs from the long-read-based genome assembly of *Melanogrammus aeglefinus* (melAeg)[60], a closely related outgroup within the subfamily Gadinae, to the gadMor2 assembly. Split contig alignments were again identified mapping near the boundaries of the supergenes on LGs 1 and 7, indicating that for these supergenes, it is the gadMor2 genome that carries the inversions. In contrast, a single contig of the melAeg assembly was clearly colinear to the gadMor2 assembly in a region that extended about 150 kbp in both directions from one of the ends of the supergene on LG 2 (Fig. 1b, Supplementary Table 2), indicating that the inversion on LG 2 is carried not by the gadMor2 genome but by the gadMor_Stat genome instead. For the supergene on LG 12, on the other hand, no informative alignments between the melAeg assembly and the gadMor2 assembly were found; thus, our contig-mapping approach did not allow us to determine which of the two *Gadus morhua* genomes carries the inversion on this linkage group (however, our subsequent demographic analyses suggested that it is the gadMor_Stat genome that carries the inversion on LG 12; see below).

Rapid divergences and introgression among codfishes. To establish the phylogenetic context within which the inversions arose in *Gadus morhua*, and to test whether any of the four supergenes originated from introgression[4, 61], we performed Bayesian phylogenomic analyses at two different levels: In a first analysis, we used a set of alignments for 91 genes, representing all gadMor2 linkage groups, (total length: 106,566 bp) to infer divergence times among fishes of the subfamily Gadinae (Supplementary Table 3). Applying the multi-species coalescent model with StarBEAST2[62] reduced the age estimates for the group compared to other recent phylogenomic studies that did not account for incomplete lineage sorting[63, 64]. The StarBEAST2 analysis placed the origin of crown Gadinae at around 18.0 Ma (95% highest posterior density interval, HPD: 20.6–15.4 Ma)

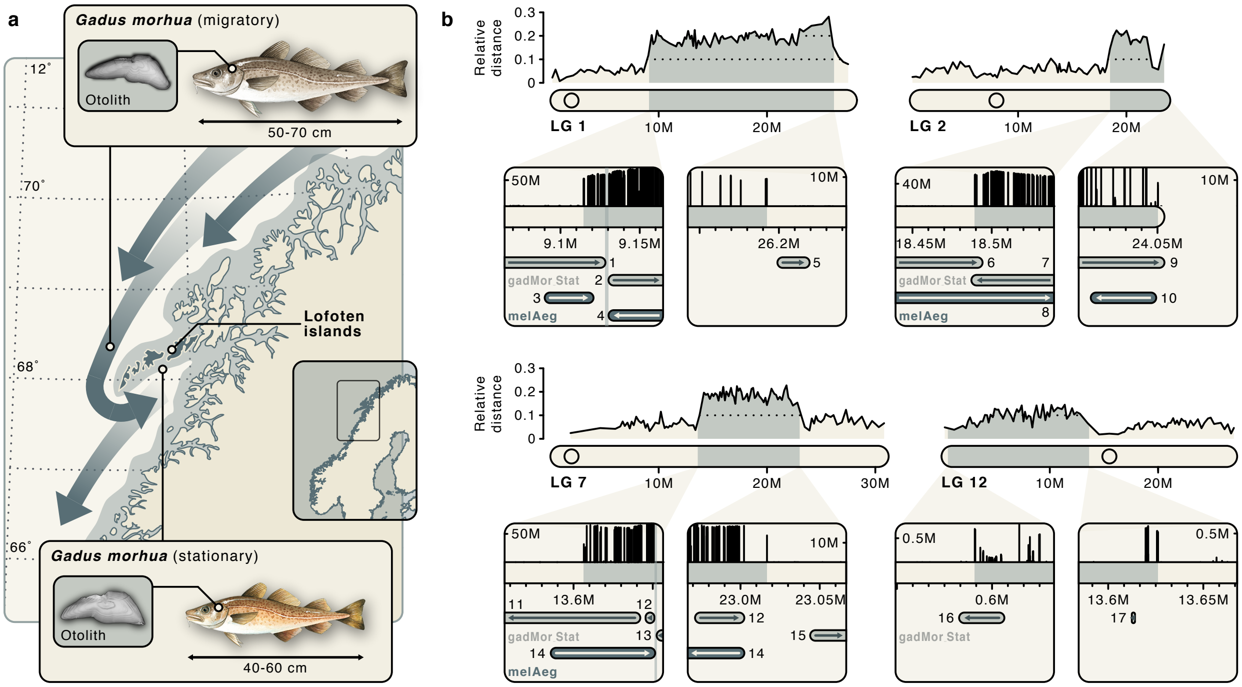


Fig. 1 Four supergenes associated with large chromosomal inversions in *Gadus morhua*. **a** Migratory and stationary *Gadus morhua* seasonally co-occur along the coast of northern Norway and differ in total length and otolith measurements[42, 59]. The distribution of stationary *Gadus morhua* is shaded in gray whereas the seasonal movements of migratory *Gadus morhua* are indicated with dark gray arrows. While the gadMor2 genome assembly[35] comes from a migratory individual, the gadMor_Stat assembly presented here comes from a stationary individual. Both of these individuals were sampled at the Lofoten islands. **b** Pairwise genetic distance between the gadMor2 and gadMor_Stat assemblies, relative to the genetic distance to an assembly for *Melanogrammus aeglefinus* (melAeg)[60] in a three-way whole-genome alignment. The alignment coordinates are according to the gadMor2 assembly. The four LGs 1, 2, 7, and 12 are shown as rounded horizontal bars, on which circles indicate the approximate centromere positions [48]. Supergene regions are shaded in gray, and the beginning and end of each of these regions are shown in more detail in the insets below each linkage group. Each of these insets focuses on a section of 100 kbp around a supergene's beginning or end. Shown in black above the bar representing that section is a per-SNP measure of linkage disequilibrium (LD), based on which the gray shading on the bar illustrates the beginning or the end of high LD. Drawn below the scale bar are contigs of the gadMor_Stat and melAeg assemblies, in light gray and dark gray, respectively, that align well to the shown sections. The arrows indicate the alignment orientation of contigs (forward or reverse complement), and contigs are labelled with numbers as in Supplementary Table 2. In the first insets for LGs 1 and 7, vertical bars indicate inferred inversion breakpoints, which are found up to 45 kbp (Table 1) after the onset of high LD. Fish drawings by Alexandra Viertler; otolith images by Côme Denechaud.

and revealed a rapidly radiating clade comprised of the genera *Pollachius*, *Melanogrammus*, *Merlangius*, *Boreogadus*, *Arctogadus*, and *Gadus* that occurs exclusively on the Northern hemisphere and began to diversify around 8.6 Ma (95% HPD: 9.9–7.2 Ma) (Supplementary Figure 1). We then refined the estimates for this clade with a second analysis that focused on the divergences among *Boreogadus*, *Arctogadus* (Fig. 2a; Supplementary Table 4), and *Gadus* and used 109 alignments (total length: 383,727 bp) sampled from across the genome. In this analysis, we co-estimated and accounted for possible introgression among species by applying the isolation-with-migration model as implemented in the AIM add-on package for BEAST 2[65, 66]. We found strong support for two topologies in which *Arctogadus* is either placed as the sister taxon to *Gadus* (Bayesian posterior probability, BPP: 0.763) or to *Boreogadus* (BPP: 0.234), with introgression supported (by Bayes factors greater than 10) between *Boreogadus* and *Arctogadus* in both cases and additionally from

Table 1 Tight linkage and chromosomal inversions in supergene regions in *Gadus morhua*.

LG	Beginning of high-LD	End of high-LD	Beginning of inversion region	End of inversion region
1	9,114,741[1]	26,192,489	9,128,372–9,130,274[2]	~26,100,000[3]
2	18,489,307	24,050,282	~18,490,000[4]	24,054,399–24,054,406[5]
7	13,606,502	23,016,726	13,651,003–13,652,432	23,002,424–23,043,967
12	589,105	13,631,347	607,782–662,878	13,386,293–13,614,908

[1] Coordinates refer to the gadMor2 assembly [35].

[2] Unless otherwise specified, boundaries of inversion regions were determined based on contig alignments (Supplementary Table 2).

[3] Comparison with the gadMor3 assembly [48] (Supplementary Figure 8) suggests that the actual end of the inversion region region is misplaced in the gadMor2 assembly, between positions 18,890,477 and 18,900,044, and that the region from position ~16,800,000 and ~18,900,000 in the gadMor2 assembly is instead located after the current position ~26,100,000.

[4] Due to repetitive sequences at the beginning of the inversion region, contigs mapping inside and outside of the region overlap between positions 18,487,151 and 18,494,225.

[5] This is the end of the linkage group in the gadMor2 assembly; however, comparison with the gadMor3 assembly suggests that the region from position ~22,600,000 to ~23,700,000 in the gadMor2 assembly is incorrectly placed and instead located at the end of the linkage group.

the common ancestor of the genus *Gadus* to *Arctogadus* in the latter case (Fig. 2b, Supplementary Figure 2). While being inconclusive about the phylogenetic position of *Arctogadus*, these results indicate that the genus is genetically more similar to both *Gadus* and *Boreogadus* than can be explained by a bifurcating tree without introgression. Regardless of the position of *Arctogadus*, our analysis with the isolation-with-migration model supported an age of around 4 Ma for the clade comprising the three genera (mean estimates: 3.81 Ma and 3.99 Ma; 95% HPD: 4.44–3.19 Ma and 4.56–3.33 Ma; Fig. 2b).

Introgression among the three genera was further supported by Patterson's D-statistic[67, 68]. Using Dsuite[69], we calculated for all possible species trios two versions of this statistic from a set of 19,035,318 SNPs: D_{fix}, for which taxa in the trio are arranged according to a provided input tree, and D_{BBAA}, for which taxa are arranged so that the number of sites with the "BBAA" pattern is maximized. The strongest signals of introgression were found once again between *Boreogadus* and *Arctogadus*, for example in a trio together with *Gadus morhua* (Fig. 2c): In this trio, 170,613 sites supported a sister-group relationship between *Arctogadus* and *Gadus morhua*, 280,258 sites supported a sister-group relationship between *Boreogadus* and *Arctogadus*, and 131,776 sites supported a sister-group relationship between *Boreogadus* and *Gadus morhua*, resulting in a significant D-statistic of $D_{\text{BBAA}} = 0.128$ and $D_{\text{fix}} = 0.360$ ($p < 10^{-10}$ in both cases; Supplementary Tables 5 and 6). Within the genus *Gadus*, *theD*-statistic additionally provided strong support for introgression between the geographically co-occurring *G. ogac* and *G. morhua* in a trio with *G. macrocephalus*: In this trio, the sister-group relationship between *G. ogac* and *G. macrocephalus* was unambiguously supported by 980,086 sites but while *G. ogac* shared 14,168 sites with *G. morhua*, *G. macrocephalus* and *G. morhua* only shared 7,797 sites, resulting in a D-statistic of $D_{\text{fix}} = D_{\text{BBAA}} = 0.283$ ($p < 10^{-10}$; Supplementary Tables 5 and 6). In both cases, signals of introgression appeared to be largely uniform throughout the genome (supergene regions were excluded from this analysis), and were not affected by the choice of genome representing *Gadus morhua* (Fig. 2d). The occurrence

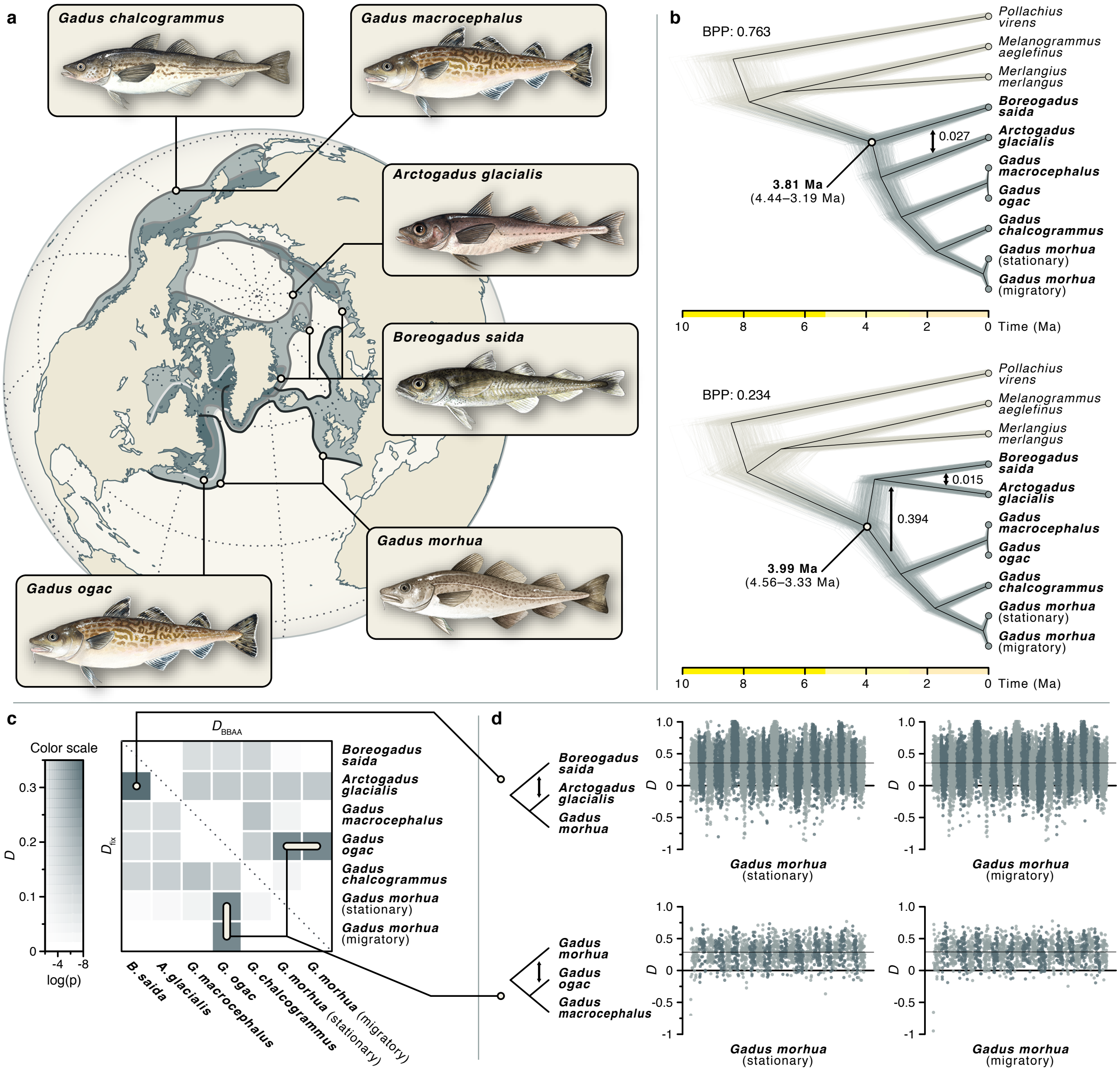


Fig. 2 Divergence times and introgression among Gadinae. **a** Distribution ranges of sampled species of the genera *Gadus*, *Arctogadus*, and *Boreogadus*. Partially overlapping distribution ranges are shown in dark gray. **b** Species tree of the six species and three outgroups (in beige; *P. virens*, *M. aeglefinus*, and *M. merlangus*), estimated under the isolation-with-migration model from 109 alignments with a total length 383,727 bp. The Bayesian analysis assigned 99.7% of the posterior probability to two tree topologies that differ in the position of *Arctogadus glacialis* and were supported with Bayesian posterior probabilities (BPP) of 0.763 and 0.234, respectively. Rates of introgression estimated in the Bayesian analysis are marked with arrows. Thin gray and beige lines show individual trees sampled from the posterior distribution; the black line indicates the maximum-clade-credibility summary tree, separately calculated for each of the two configurations. Of *G. morhua*, both migratory and stationary individuals were included. **c** Pairwise introgression among species of the genera *Gadus*, *Arctogadus*, and *Boreogadus*. Introgression was quantified with the *D*-statistic. The heatmap shows two versions of the *D*-statistic, D_{fix} and D_{BBAA}, below and above the diagonal, respectively. **d** Introgression across the genome. The *D*-statistic is shown for sliding windows in comparisons of three species. The top and bottom rows show support for introgression between *B. saida* and *A. glacialis* and between *G. morhua* and *G. ogac*, respectively. Results are shown separately for the stationary and migratory *G. morhua* genomes. The mean *D*-statistic across the genome is marked with a thin solid line. Fish drawings by Alexandra Viertler.

of introgression between *Boreogadus* and *Arctogadus* was corroborated by a tree-based equivalent to Patterson's D-statistic that does not rely on the molecular-clock assumption, the D_{tree}-statistic of Ronco et al.[70], and by genealogy interrogation[71, 72], but introgression between *G. ogac* and *G. morhua* did not receive this additional support, perhaps because introgressed regions were too short to affect tree topologies (Supplementary Figure 3).

Recent divergences among *Gadus morhua* populations. We performed phylogenomic analyses for individuals from eight *Gadus morhua* populations covering the species' distribution in the North Atlantic (Fig. 3a; Supplementary Table 7). In addition to the individuals used for the gadMor2 and gadMor_Stat assemblies, we selected from these populations 22 individuals for which preliminary analyses had shown that each of them carried, at each of the four supergene regions, two copies of the same haplotypes (i.e., they were homokaryotypic). For the four sampling localities Newfoundland, Iceland, Lofoten, and Møre, we discriminated between "migratory" and "stationary"

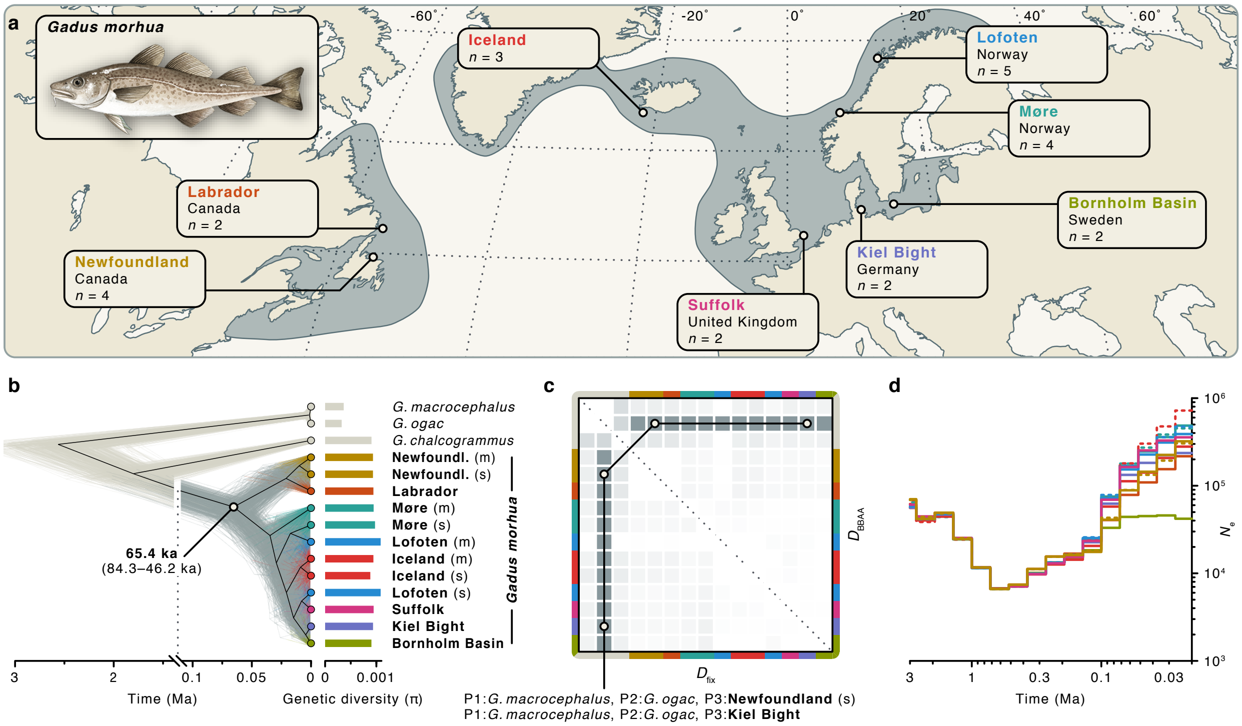


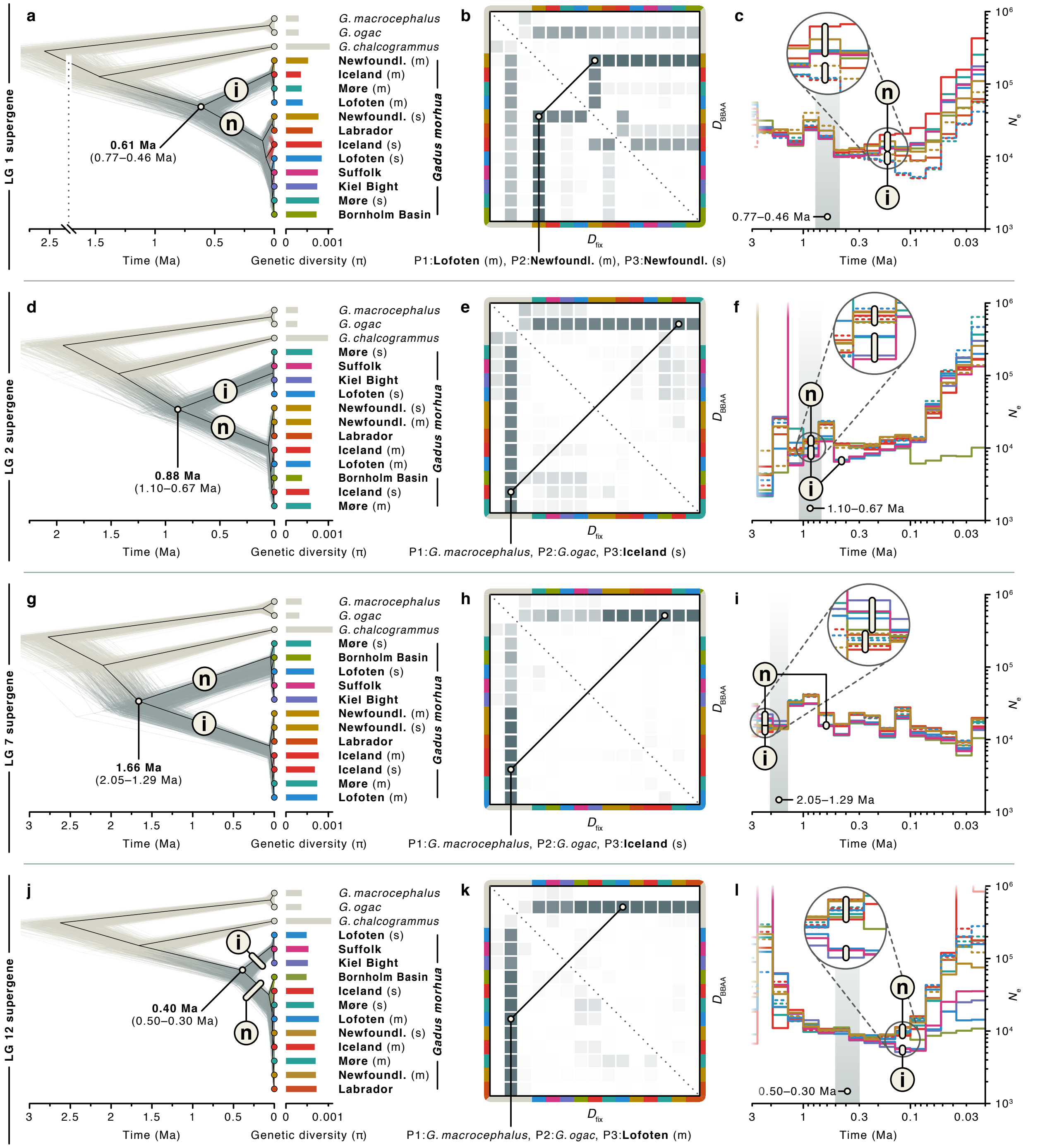
Fig. 3 Divergence times, demography, and introgression among *Gadus morhua* populations. **a** Geographic distribution of *Gadus morhua* in the North Atlantic and sampling locations for analyses of population divergence times, demography, and introgression. **b**, Tree of *Gadus morhua* populations and three outgroups (in beige; *G. macrocephalus*, *G. ogac*, and *G. chalcogrammus*), inferred under the multi-species coalescent model from 1,000 SNPs sampled across the genome (excluding inversion regions). Thin gray and beige lines show individual trees sampled from the posterior distribution; the black line indicates the maximum-clade-credibility summary tree. Estimates of the genetic diversity (π) per population are indicated by bars to the right of the tips of the tree. **c** Pairwise introgression among *Gadus morhua* populations and outgroup species. Two versions of the D-statistic, D_{fix} and D_{BBAA}, are shown above and below the diagonal, respectively. Color codes on the axes indicate populations, and heatmap coloring is as in Fig. 2c. The two trios with the strongest signals, supporting introgression between *G. ogac* and both the Labrador and the migratory Lofoten *G. morhua* population with $D_{\mathrm{BBAA}} = D_{\mathrm{fix}} = 0.201$ are marked. **d** Population sizes (N_{e}) over time in *Gadus morhua* populations, estimated with Relate. For the Newfoundland, Møre, Iceland, and Lofoten populations, migratory (m) and stationary (s) individuals were analyzed separately; dashed lines are used for migratory populations.

individuals based on whether they carried the same supergene haplotype on LG 1 as the gadMor2 genome or the same as the gadMor_Stat genome, respectively. For the individuals from Lofoten and Møre, this classification could be confirmed by an analysis of their otoliths[59], but otolith data was not available for the individuals from the other sampling localities. Based on a dataset of 20,402,423 genome-wide biallelic SNPs, we estimated relationships and divergence times among *Gadus morhua* populations under the multi-species coalescent model with SNAPP[73, 74], first only with data from outside of the supergene regions. In line with previous studies based on SNP arrays[31, 38, 75], we found the primary divergence within *Gadus morhua* to separate the populations of the Northwest Atlantic from those of the Northeast Atlantic (including Iceland). We estimated these groups to have diverged around 65.4 ka (95% HPD: 84.3–46.2 ka) but acknowledge that these results may underestimate the true divergence time because the applied model does not account for possible gene flow after divergence (Fig. 3b, Supplementary Figure 4).

The genetic diversity, quantified by π[76], was comparable among the populations of both groups, ranging from 8.82×10^{-4} to 1.084×10^{-3} (Supplementary Table 8). Applied to the set of 20,402,423 SNPs, Patterson's D-statistic corroborated the occurrence of introgression between *Gadus ogac* and *Gadus morhua* and showed that the signals of introgression are almost uniform among all populations. One of the strongest signals was found with the trio including *Gadus macrocephalus*, *G. ogac*, and the stationary Newfoundland *G. morhua* population, for which 1,409,330 sites supported the sister-group relationship between *G. macrocephalus* and *G. ogac* and 16,543 sites were shared between *G. ogac* and the stationary Newfoundland population, but only 9,955 sites were shared between *G. macrocephalus* and the stationary Newfoundland population, resulting in a D-statistic of $D_{\text{fix}} = D_{\text{BBAA}} = 0.249$ ($p < 10^{-10}$; Supplementary Tables 9 and 10). Estimating changes in the population size (N_{e}) over time for *Gadus morhua* with Relate[77] revealed a Pleistocene bottleneck, lasting from around 0.7 Ma to 0.3 Ma, during which the population size of the common ancestor of all populations decreased from around 50,000 to 7,000 diploid individuals. The subsequent increase in population sizes occurred in parallel with the diversification of *Gadus morhua* populations and was experienced by all of them to a similar degree, albeit slightly less so by the Northwest Atlantic populations and least by the Bornholm Basin population of the Baltic Sea (Fig. 3d).

Different age estimates for supergenes in *Gadus morhua*. To infer the ages of the supergenes on LGs 1, 2, 7, and 12, we applied SNAPP analyses to SNPs from each supergene separately,

Fig. 4 (next page) Divergence times, demography, and gene flow among *Gadus morhua* populations within supergene regions. **a**,**d**,**g**,**j** Trees of *Gadus morhua* populations and three outgroups (in beige; *G. macrocephalus*, *G. ogac*, and *G. chalcogrammus*), inferred under the multi-species coalescent model from 1,000 SNPs sampled from within the supergene regions on LGs 1 (**a**), 2 (**d**), 7 (**g**), and 12 (**j**). Thin gray and beige lines show individual trees sampled from the posterior distribution; the black line indicates the maximum-clade-credibility summary tree. Within *G. morhua*, inverted and noninverted supergene haplotypes are marked with labels "i" and "n", respectively. Estimates of the genetic diversity (π) per population within supergene regions are indicated by bars to the right of the tips of the tree. **b**,**e**,**h**,**k** Pairwise signals of past gene flow among *Gadus morhua* populations and outgroup species within the supergene regions on LGs 1 (**b**), 2 (**e**), 7 (**h**), and 12 (**k**). Two versions of the D-statistic, D_{fix} and D_{BBAA}, are shown above and below the diagonal, respectively. Color codes on the axes indicate populations, ordered as in **a**,**d**,**g**,**j**. The trios with the strongest signals of introgression are labelled. **c**,**f**,**i**,**l** Population sizes (N_{e}) over time in *Gadus morhua* populations for the supergene regions on LGs 1 (**c**), 2 (**f**), 7 (**i**), and 12 (**l**). For the Newfoundland, Møre, Iceland, and Lofoten populations, migratory (m) and stationary (s) individuals were analyzed separately; dashed lines are used for migratory populations. The gray regions indicate the confidence intervals for the inferred age of the split between the two haplotypes (from **a**,**d**,**g**,**j**).



extracted from the dataset of 20,402,423 biallelic SNPs. For each of the four supergenes, we recovered a deep divergence separating inverted and noninverted haplotypes; however, the age estimates for this divergence differed widely among the four supergenes, with mean age estimates of 0.61 Ma (95% HPD: 0.77–0.46 Ma) for the supergene on LG 1 (Fig. 4a), 0.88 Ma (95% HPD: 1.10–0.67 Ma) for the supergene on LG 2 (Fig. 4d), 1.66 Ma (95% HPD: 2.05–1.29 Ma) for the supergene on LG 7 (Fig. 4g), and 0.40 Ma (95% HPD: 0.50–0.30 Ma) for the supergene on LG 12 (Fig. 4j; Supplementary Figure 5). The groups sharing the same haplotype also differed among the four supergenes, even though the migratory individuals from Newfoundland, Iceland, Lofoten, and Møre always shared the same haplotype. The genetic diversity was on average lower for the inverted haplotypes on LGs 1 and 12, but increased compared to the noninverted haplotype on LGs 2 and 7 (Figs. 4a,d,g,j, Supplementary Table 8).

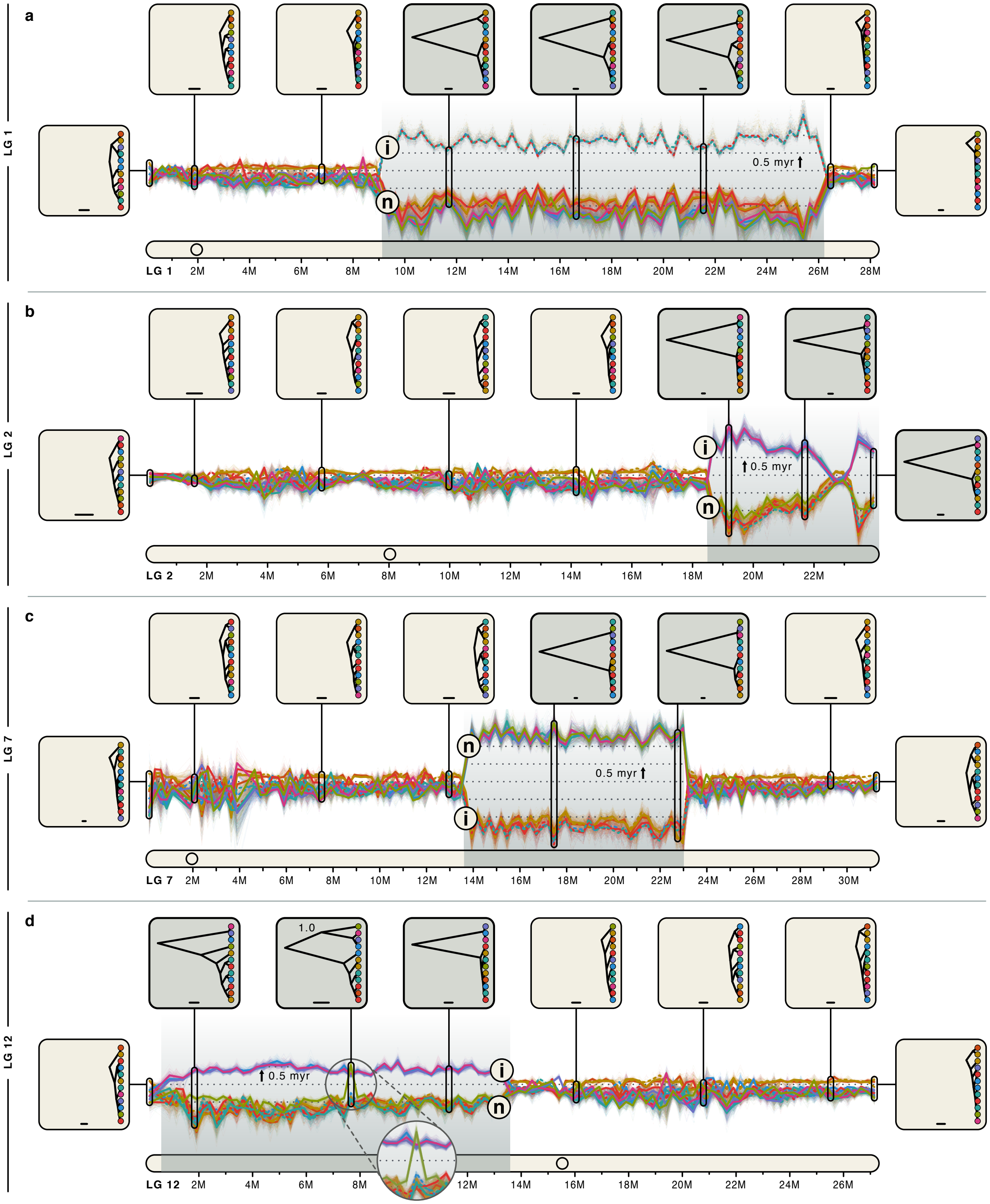
Gene flow between inverted and noninverted haplotypes via gene conversion. Like for the genome-wide estimate of the divergence time between Northwest and Northeast Atlantic populations, the ages of the separation between the two alternative haplotypes per supergene could be underestimated if gene flow has occurred after their divergence. However, given that single-crossover recombination is suppressed between inverted and noninverted haplotypes, gene flow between the two haplotypes can only occur through gene conversion or double-crossover events[29]. To shed light on the frequency of these two processes within the four supergene regions, we calculated the D-statistic for the sets of supergene-specific SNPs. The D-statistic supported gene flow between inverted and noninverted haplotypes particularly for the supergene on LG 1, in trios that included either the migratory or stationary Newfoundland population (Fig. 4b). When both of these populations were placed in the same trio together with the migratory Lofoten population, 14,466 sites supported a sister-group relationship between the two migratory populations, in agreement with the population tree inferred with SNAPP (Fig. 4a). However, with 1,041 sites shared between the migratory and stationary Newfoundland populations but only 465 sites shared between the migratory Lofoten population and the stationary Newfoundland population, the D-statistic was $D_{\mathrm{fix}} = D_{\mathrm{BBAA}} = 0.383$ and strongly supported gene flow — and thus gene conversion or double-crossover events — between the two geographically co-occurring migratory and stationary Newfoundland populations ($p < 10^{-10}$; Supplementary Tables 11 and 12). To test whether gene conversion could be the cause of this gene flow occurring between the Newfoundland populations, we tested for the GC bias expected from gene conversion[24], comparing GC-content of sites shared between the two Newfoundland populations ("ABBA" sites) to that of sites shared between the migratory Newfoundland population and other migratory populations ("BBAA" sites). The mean GC-content of the former, 0.482, is indeed significantly higher than that of the latter, 0.472 (one-sided t-test; $t = -4.74$, df $= 27833$, $p < 10^{-5}$), supporting gene conversion as the agent of gene flow between inverted and noninverted haplotypes in the Newfoundland populations. For the supergenes on LGs 2, 7, and 12, the D-statistic again — as for the genome-wide SNP data — primarily supported introgression between *Gadus ogac* and *Gadus morhua* populations, which was largely uniform among populations (Fig. 4e,h,k) and reached values between 0.333 and 0.376 (Supplementary Tables 13-18). For LG 7, this introgression was found to affect mostly the inverted haplotype, which might be related to the physical separation between *Gadus ogac* and the populations represented by the noninverted haplotype, as all of these occur in the Northeast Atlantic.

Demographic analyses recover signatures of bottlenecks following inversions. It is likely that each of the inversions associated with the four supergenes in *Gadus morhua* originated a single time, in a single individual, from which all of the current carriers of the inversion descended. The supergene origin should therefore have been equivalent to an extreme bottleneck event during which the population size was reduced to a single sequence, but which affected only the inversion region, and only the carriers of the inversion. We might thus expect to see signatures of this extreme bottleneck event in analyses of population size over time, in the form of differences in ancestral population sizes between inverted and noninverted haplotypes that date to the time of supergene origin.

To verify that signatures of extreme bottlenecks are detectable in descending genomes even after long periods of time, we first performed a series of power analyses based on coalescent simulations (Supplementary Note 2). After confirming that the program Relate is in principle able to pick up such signals, we tested for the presence of bottleneck signatures associated with inversion events by performing demographic analyses with Relate separately for sets of SNPs from each of the four supergene regions. As expected due to the smaller amount of input data (1–3% compared to the genome-wide SNP data), these analyses produced estimates that were less clear (Fig. 4c,f,i,l) than those obtained with genome-wide SNP data (Fig. 3d). Nevertheless, the supergene-specific demographic analyses supported a weak differentiation between the population sizes of inverted and noninverted haplotypes, with a temporary reduction of population sizes for the inverted haplotype that coincided with the estimated ages of the supergenes on LGs 2 and 7 (Fig. 4f,i). For the supergenes on LGs 1 and 12, no differentiation in the population sizes coinciding with the supergene ages was evident, but for the supergene on LG 1, it was the population size of the inverted haplotype that was reduced compared to the noninverted haplotype in the first time period that showed such a differentiation following the supergene origin (Fig. 4c). While our contig-mapping approach had not allowed us to infer which of the two haplotypes of the LG 12 supergene was inverted, the reduced inferred population size of the haplotype carried by the individuals from the Suffolk, Kiel Bight, and stationary Lofoten populations, in a time interval following the supergene origin, suggested that this haplotype might be the inverted one on that linkage group. The ancestral population sizes estimated from the supergene region on LG 7 further highlighted that estimates of the current genetic diversity may not always be useful for the identification of inverted haplotypes (Fig. 4i), as the population-size estimates for the noninverted haplotype were higher than those for the inverted haplotype at the time of supergene origin, but lower for most of the subsequent time towards the present (Fig. 4i).

Divergence profiles reveal double crossovers. To explore whether divergence times between the two haplotypes per supergene are homogeneous across the supergene region, we repeated divergence-time inference with SNAPP in sliding windows of 250 kbp along all linkage groups. We expected that if any gene flow between inverted and noninverted haplotypes should proceed via double crossovers, its effect should be less pronounced near the inversion breakpoints at the

Fig. 5 (next page) Divergence profiles for linkage groups with supergenes. **a**–**d** Illustration of between-population divergence times along LGs 1 (**a**), 2 (**b**), 7 (**c**), and 12 (**d**), estimated with SNAPP from SNPs in sliding windows. Supergene regions are indicated by gray backgrounds. Along the vertical axis, the distance between two adjacent lines shows the time by which the corresponding populations have been separated on the ladderized population tree for a given window. Examples of the population tree are shown in insets for eight selected windows. The scale bar in these insets indicates the branch length equivalent to 50,000 years. The node label in one inset in (**d**) indicates the support for the grouping of the Bornholm Basin population with three populations representing the inverted haplotype (BPP: 1.0).



356 boundaries of the supergenes and stronger towards their centers, which could generate U-shaped divergence profiles for supergene regions[22, 29, 30]. Contrary to this expectation, the divergence

profiles were relatively homogeneous from beginning to end, particularly for the supergenes on LGs 1, 7, and 12 (Fig. 5a,c,d; Supplementary Figure 7), suggesting either that double crossovers are rare within these supergenes, or that sequences exchanged through double crossovers are frequently purged from the recipient haplotypes. As the supergene on LG 1 is known to include not one but two adjacent inversions of roughly similar size[36], our results also suggested a similar age and possibly a joint origin for both of these inversions. The divergence profile for the supergene on LG 2 appeared consistent with the expectation of a U-shaped pattern; however, comparison of the gadMor2 assembly with the recently released gadMor3 assembly[48] showed that the end of this linkage group may be misassembled in gadMor2, and the region of low divergence appearing within the supergene (around position 22.5–23.0 Mbp) may in fact lie outside of it (Supplementary Figure 8). However, a closer look at the divergence profile for LG 12 revealed a single window within the supergene in which the otherwise clear separation between the groups carrying the alternative haplotypes was interrupted: Unlike in all other windows within this supergene, the Bornholm Basin population grouped (BPP: 1.0) with the three populations representing the inverted haplotype (Suffolk, Kiel Bight, and stationary Lofoten; see Fig. 4j) in the window for positions 7.50–7.75 Mbp (Fig. 5d). To investigate the genotypes of the two sampled Bornholm Basin individuals within this region in more detail, we identified 219 haplotype-informative sites between positions 7–8 Mbp on LG 12, and found that these individuals were both heterozygous at these sites, for a region of ~275 kbp between positions 7,478,537 bp and 7,752,994 bp (Fig. 6). The two individuals from the Bornholm Basin population thus carried a long sequence from the inverted haplotype even though they were otherwise clearly associated with the noninverted haplotype. As the length of this introduced sequence was far longer than the 50–1,000 bp expected to be copied per gene-conversion event[23, 24], it strongly supports a contribution of double crossover to gene flow between the two haplotypes of the LG 12 supergene.

Interestingly, the composition of the sequence introduced through double crossover indicated that selection may have played a role in maintaining it within the Bornholm Basin population. The region covered by the introduced sequence contains 24 predicted genes (Supplementary Table 20), amongst them three yolk precursor genes (vitellogenin) whose gene ontology classification "lipid transporter activity" was identified as being significantly enriched within this region at false discovery rate (FDR) 0.05 (Fisher's exact test; Supplementary Table 21). Due to their role in lipid transport, these vitellogenin genes are assumed to contribute to the proper hydration of spawned eggs, and, as a result of that, to the maintenance of neutral egg buoyancy[79, 80]. However, in contrast to the fully marine environments of the open North Atlantic, the almost entirely land-enclosed brackish Baltic Sea has a severely reduced salinity, and thus requires adaptations in the hydration of eggs, so that they remain neutrally buoyant at ~14 ppt (compared to ~35 ppt in the North Atlantic) and do not sink to anoxic layers[81–83]. The sequences of the three vitellogenin genes in fact differed strongly between the inverted and noninverted haplotypes, with a reading-frame shift in one of them that truncates its coding sequence from 2,721 bp to 1,104 bp, and ten amino-acid changes in a second of the three genes (Supplementary Table 22). Thus, the three vitellogenin genes may be under selection in *Gadus morhua* from the Baltic Sea[32], increasing the frequency of the introduced sequence within the Bornholm Basin population.

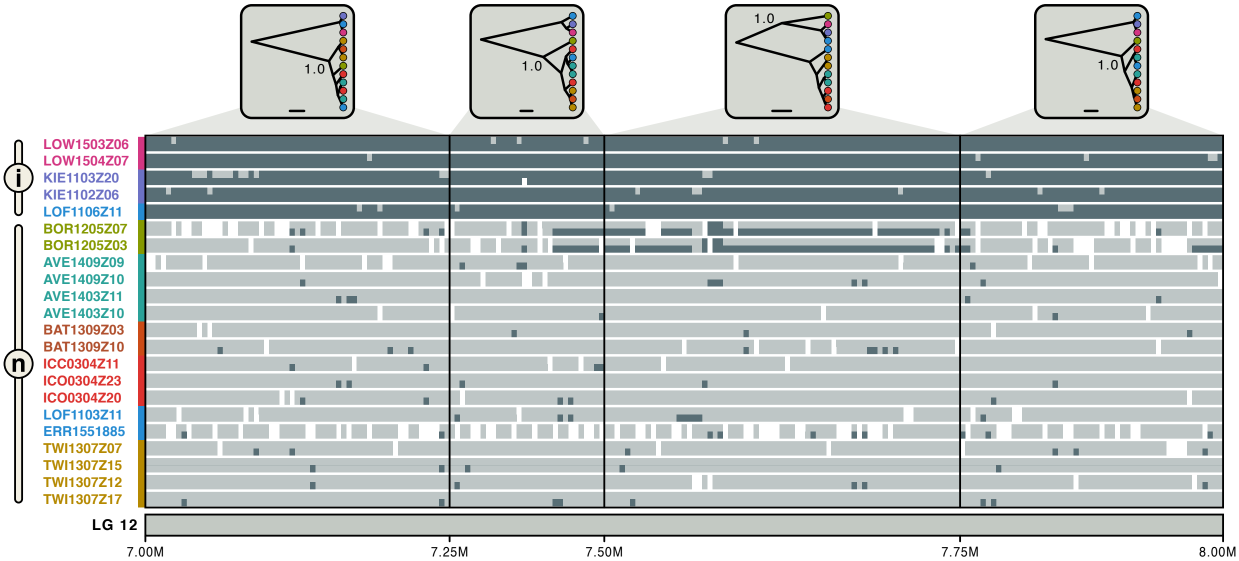


Fig. 6 Ancestry painting for part of the supergene on LG 12. The ancestry painting[72, 78] shows genotypes at 219 haplotype-informative sites between positions 7 and 8 Mbp on LG 12, within the supergene on that linkage group. For each of 22 *Gadus morhua* individuals, homozygous genotypes are shown in dark or light gray while heterozygous genotypes are illustrated with a light gray top half and a dark gray bottom half; white color indicates missing genotypes. As haplotype-informative sites, we selected those that have less than 10% missing data and strongly constrasting allele frequencies (≥ 0.9 in one group and ≤ 0.1 in the other) between the group carrying the inverted haplotype, composed of individuals from Suffolk, Kiel Bight, and stationary Lofoten, and the group carrying the noninverted haplotype, composed of individuals from the Møre, Labrador, Reykjavík, migratory Lofoten, and Newfoundland populations. The four insets at the top show population trees inferred with SNAPP; the node labels in these insets indicate Bayesian support for the grouping of the Bornholm Basin population with either the inverted or noninverted haplotype.

Discussion

Through comparison of long-read-based genome assemblies for migratory and stationary Atlantic cod individuals, we corroborated earlier conclusions that chromosomal inversions underlie all of the four supergenes in Atlantic cod[48]. The inversion breakpoints do not coincide exactly with the boundaries of supergenes, but lie up to 45 kbp inside of them, in agreement with findings reported for *Drosophila* that suggested that recombination suppression can extend beyond inversion breakpoints [28]. By also comparing the genome assemblies for Atlantic cod with an assembly of the closely related haddock (*Melanogrammus aeglefinus*)[60], we were further able to identify the gadMor2 assembly — representing a migratory Atlantic cod individual — as the carrier of the inverted haplotypes of the supergenes on LG 1 and 7, but not of that on LG 2. In addition, our demographic analyses indicated that the gadMor2 assembly might also carry the noninverted haplotype on LG 12. The inverted haplotypes were not consistently characterized by lower genetic diversity than noninverted haplotypes, contrary to assumptions that were used in previous studies to distinguish between them[33, 34]. As suggested by our demographic analyses (Fig. 4i) and simulations (Supplementary Table 19), this contrast can be explained by an ability of inverted haplotypes to recover from the initial bottleneck. While this recovery may require substantial frequencies of the noninverted haplotype and the passing of sufficient time since the inversion origin to allow the accumulation of new mutations[17], both of these requirements may be met for the

Atlantic cod supergenes. It is likely that local adaptation has played a role in maintaining both haplotypes per supergene at relatively high frequencies; however, the way in which the haplotypes convey local adaptation in Atlantic cod remains largely unknown.

According to our time-calibrated phylogenomic analyses, the four supergenes in Atlantic cod originated between ~0.40 and ~1.66 Ma. These dates could potentially be underestimates due to gene flow between the inverted and noninverted haplotypes that could not be accounted for in our age estimation. Nevertheless, we consider these age estimates as strong evidence for separate origins of all four supergenes, due to the wide range between these age estimates, the homogeneity in sliding-window age estimates from beginning to end of each supergene, and the support for demographic bottlenecks coinciding with the inferred age estimates (Figs. 4,5). Our age estimates thus indicate that all four supergenes evolved separately, after the Atlantic cod's divergence from the walleye pollock (*Gadus chalcogrammus*), and before the divergence of all extant Atlantic cod populations.

Introgression from other codfish species does not seem to have played a role in the origin of the four supergenes. However, we found strong signals of introgression between Atlantic cod and Greenland cod (*Gadus ogac*), as Atlantic cod unambigously shares more alleles with Greenland cod than with the Greenland cod's sister species, the Pacific cod (*Gadus macrocephalus*), despite the very recent divergence between Greenland and Pacific cod (~40 ka). This genetic similarity between Atlantic cod and Greenland cod does not appear to be an artifact resulting from reference bias and is best explained by gene flow into Greenland cod, from an Atlantic cod population co-occurring in the Northwest Atlantic (Supplementary Note 3). Besides the signals for introgression into Greenland cod, our results strongly support gene flow between the two haplotypes of the supergene on LG 1 (Fig. 4b): between all carriers of the noninverted haplotype and the migratory individuals from Newfoundland (which carry the inverted haplotype) and between all carriers of the inverted haplotype and the stationary individuals from Newfoundland (which carry the noninverted haplotype). Due to elevated GC-content of sites shared between the haplotypes, we interpret these signals of gene flow as evidence for gene conversion occurring at Newfoundland. It remains unclear why the same gene conversion does not seem to occur at other locations where the two haplotypes co-occur (e.g. in northern Norway). However, it could be speculated that this difference in the occurrence of gene conversion reflects different selection pressures on both sides of the Atlantic, with selection purging sequences exchanged between haplotypes in the Northeast Atlantic, but not in the Northwest Atlantic. As a second mechanism allowing gene flow between noninverted and inverted haplotypes, our results demonstrated the occurrence of double crossover, between the two haplotypes of the LG 12 supergene. Like for the inferred gene conversion, selection may have played a role in the maintenance of the exchanged sequence in the recipient haplotype — as suggested by its genetic content that includes three differentiated vitellogenin genes.

The presence of four long (4–17 Mbp) and old (0.40–1.66 Ma) inversion-based supergenes in Atlantic cod adds to recent findings of inversions of similar size and/or age in butterflies[12], ants[2], birds[4], lampreys[84], and *Drosophila*[30]. For non-model organisms, these findings are largely owed to the development of long-read-based genome sequencing within the last decade and may become even more common as long-read-based sequencing is applied to more and more species. These findings are, however, in contrast to earlier expectations based on theoretical work and empirical studies on selected model organisms: Only about 20 years ago, the available evidence indicated that inversions (and thus inversion-based supergenes) would be "generally not ancient"[22], with maximum ages on the order of N_e generations (which would correspond to ~100 ka in Atlantic cod), because they would either decay too rapidly due to the accumulation of mutation load, or erode if gene flow occurs through gene conversion and double crossover[22, 29, 85]. While we observed evidence for both gene conversion and double crossover in Atlantic cod, we did not find signs of supergene decay or erosion: Decay could be indicated by high repeat content or mutation load[21], but neither were increased within the four supergenes (Supplementary Figure 9). And supergene erosion — at least when resulting from double crossovers — would be expected to produce U-shaped divergence profiles[22, 29, 30], but no such profiles were found for the Atlantic cod supergenes (Fig. 5). These observations mirror those made recently by Yan et al.[2] for a supergene in fire ants, and we thus concur with their conclusion that "low levels of recombination and/or gene conversion may play an underappreciated role in preventing rapid degeneration of supergenes". But, since our results also indicated that selection may have acted on sequences exchanged between supergene haplotypes, we further suggest that — just like in interbreeding species that maintain stable species boundaries despite frequent hybridization[72] — selective purging of introduced sequences may also be important for the fate of supergenes, by maintaining the rate of gene flow between haplotypes at exactly the right balance, between too little of it and the consequential decay, and too much of it and the resulting supergene erosion.

Methods

Construction of the gadMor_Stat genome assembly. We performed high-coverage genome sequencing for a stationary *Gadus morhua* individual (LOF1106Z11) sampled at the Lofoten islands in northern Norway, for which preliminary investigations had suggested that it carried, homozygously on each of the four LGs 1, 2, 7, and 12, a supergene haplotype that was complementary to the one in the gadMor2 genome[35], which represents a migratory individual from the Lofoten islands. We used the Pacific Biosciences RS II platform, operated by the Norwegian Sequencing Centre (NSC; www.sequencing.uio.no), to generate 2.4 million PacBio SMRT reads with a total volume of 12.5 Gbp, approximately equivalent to an 18× coverage of the *Gadus morhua* genome. The PacBio SMRT reads were assembled with Celera Assembler v.8.3rc2[55], adjusting the following settings according to the nature of the PacBio reads (all others were left at their defaults): merSize =16, merThreshold=0, merDistinct=0.9995, merTotal=0.995, ovlErrorRate=0.40, ovlMinLen=500, utgGraphErrorRate=0.300, utgGraphErrorLimit=32.5, utgMergeErrorRate=0.35, utgMergeErrorLimit=40, utgBubblePopping=1, utgErrorRate=0.40, utgErrorLimit=25, cgwErrorRate=0.40, cnsErrorRate=0.40. The consensus sequence of the assembly was polished with Quiver v.0.9.0[86] and refined with Illumina reads sequenced for the same individual (see below). A total volume of 6.2 Gbp of Illumina reads were mapped to the assembly with BWA MEM v.0.7.12-r1039[87] and sorted and indexed with SAMtools v.1.10[88, 89]. Subsequently, Pilon v.1.16[56] was applied to recall consensus.

Whole-genome sequencing and population-level variant calling. *Gadus morhua* individuals sampled in Canada, Iceland, the United Kingdom, Germany, Sweden, and Norway (Fig. 3; Supplementary Table 7) were subjected to medium-coverage whole-genome Illumina sequencing. DNA extraction, library preparation, and sequencing were performed at the NSC; using the Illumina Truseq DNA PCR-free kit for DNA extraction and an Illumina HiSeq 2500 instrument with V4 chemistry for paired-end (2 × 125 bp) sequencing. Sequencing reads were mapped to the gadMor2 assembly for *Gadus morhua*[35] with BWA MEM v.0.7.17 and sorted and indexed with SAMtools v.1.9; read duplicates were marked and read groups were added with Picard tools v.2.18.27 (http://broadinstitute.github.io/picard). Variant calling was performed with GATK's v.4.1.2.0[90, 91] HaplotypeCaller and GenotypeGVCFs tools, followed by indexing with BCFtools v.1.9[89].

Delimiting high-LD regions associated with inversions. As chromosomal inversions locally suppress recombination between individuals carrying the inversion and those that do not, we used patterns of linkage disequilibrium (LD) to guide the delimitation of inversion regions for each of the four supergenes[32, 36, 48]. To maximize the signal of LD generated by the inversions, we selected 100 *Gadus morhua* individuals so that for each of the four supergenes, 50 individuals carried two copies of one of the two alternative supergene haplotypes, and the other 50 individuals carried two copies of the other. Variant calls of the 100 individuals were filtered with BCFtools, excluding all indels and multi-nucleotide polymorphisms and setting all genotypes with a Phred-scaled quality below 20, a read depth below 3, or a read depth above 80 to missing. Genotypes with more than 80% missing data or a minor allele count below 20 were then removed from the dataset with VCFtools v.0.1.14[92]. Linkage among single-nucleotide polymorphisms (SNPs) spaced less than 250,000 bp from each other was calculated with PLINK v.1.90b3b[93]. The strength of short- to mid-range linkage for each SNP was then quantified as the sum of the distances (in bp) between that SNP and all other SNPs with which it was found to be linked with $R^2 > 0.8$. We found this measure to illustrate well the sharp decline of linkage at the boundaries of the four supergenes (Fig. 1).

Contig mapping. To confirm the presence of chromosomal inversions within the four supergenes on LGs 1, 2, 7, and 12 of the *Gadus morhua* genome, we aligned contigs of the gadMor_Stat assembly to the gadMor2 assembly by using BLASTN v.2.2.29[94] searches with an e-value threshold of 10^{-10}, a match reward of 1, and mismatch, gap opening, and gap extension penalties of 2, 2, and 1, respectively. Matches were plotted and visually analyzed for contigs of the gadMor_Stat assembly that either span the boundaries of the four supergene regions or map partially close to both boundaries of one such region. We considered the latter to support the presence of a chromosomal inversion if one of two parts of a contig mapped just inside of one boundary and the other part mapped just outside of the other boundary, and if the two parts had opposite orientation; in contrast, an observation of contigs clearly spanning one of the boundaries would reject the assumption of an inversion. To further assess which of the two *Gadus morhua* genomes (gadMor2 or gadMor_Stat) carries the inversion at each of the four regions, we also aligned contigs of the genome assembly for *Melanogrammus aeglefinus* (melAeg)[60] to the gadMor2 assembly.

Threeway whole-genome alignment. We generated whole-genome alignments of the gadMor2, gadMor_Stat, and melAeg assemblies using three different approaches. First, we visually inspected the plots of BLASTN matches (see above), determined the order and orientation of all gadMor_Stat and melAeg contigs unambiguously mapping to the gadMor2 assembly, and then combined these contigs (or their reverse complement, depending on orientation) into a single FASTA file per species and gadMor2 linkage group. For each linkage group, pairwise alignments between the file with contigs from the gadMor_Stat assembly and the gadMor2 assembly, and between the file with contigs from the melAeg assembly and the gadMor2 assembly, were then produced with the program MASA-CUDAlign v.3.9.1.1024[95]. Second, we used the program LASTZ v.1.0.4[96] to align both the gadMor_Stat assembly and the melAeg assembly to the gadMor2 assembly, after masking repetitive regions in all three assemblies with RepeatMasker v.1.0.8 (http://www.repeatmasker.org). Per linkage group, the pairwise alignments generated with MASA-CUDAlign and LASTZ were then merged into a single alignment, which was refined with MAFFT v.7.300[97] within sliding windows of 1,000 bp. Third, Illumina sequencing reads of the individuals used for the three assemblies were mapped to the gadMor2 assembly with BWA MEM, followed by sorting and indexing with SAMtools. The resulting files were converted to FASTA format using a combination of SAMtools mpileup, BCFtools, and seqtk (https://github.com/lh3/seqtk) commands. Finally, we generated a conservative threeway whole-genome alignment by comparing the three different types of alignments and setting all sites to missing at which one or more of the three alignment types differed. Alignment sites that opened gaps in the gadMor2 sequence were deleted so that the resulting strict consensus alignment retained the coordinate system of the gadMor2 assembly.

Based on the threeway whole-genome alignment, we calculated the genetic distance between the gadMor_Stat and gadMor2 assemblies, relative to the genetic distance between the melAeg and gadMor2 assemblies, in sliding windows of 100,000 bp. We also used the threeway whole-genome alignment to generate a mask of unreliable alignment sites, including all sites that had been set to missing in the alignment.

Estimating divergence times of Gadinae. We estimated the divergence times among species of the subfamily Gadinae with a phylogenomic approach on the basis of published age estimates for two divergence events. The phylogenomic dataset used for these analyses comprised genome assemblies for eight Gadinae species released by Malmstrøm et al.[63], a genome assembly for the most closely related outgroup *Brosme brosme*[63], the gadMor2 assembly for *Gadus morhua*, and sets of unassembled Illumina reads for *Gadus macrocephalus* and *Gadus ogac*[54] (Supplementary Table 3). Aiming to identify sequences orthologous to 3,061 exon markers used in a recent phylogenomic analysis of teleost relationships by Roth et al.[98], we first performed targeted assembly of these markers from the sets of Illumina reads for *Gadus macrocephalus* and *Gadus ogac*. Targeted assembly was conducted with Kollector v.1.0.1[99], using marker sequences of *Gadus morhua* from Roth et al.[98] as queries. From the set of whole-genome and targeted assemblies, candidate orthologs to the 3,061 exon markers used by Roth et al.[98] were then identified through TBLASTN searches, using sequences of *Danio rerio* as queries as in the earlier study. The identified sequences were aligned with MAFFT and filtered to exclude potentially remaining paralogous sequences and misaligned regions: We removed all sequences with TBLASTN bitscore values below 0.9× the highest bitscore value and all sequences that had dN/dS values greater than 0.3 in comparison to the *Danio rerio* queries, we removed codons from the alignment for which BMGE v.1.1[100] determined a gap rate greater than 0.2 or an entropy-like score greater than 0.5, and we excluded exon alignments with a length shorter than 150 bp, more than two missing sequences, or a GC-content standard deviation greater than 0.04. We then grouped exon alignments by gene and excluded all genes that 1) were represented by less than three exons, 2) had one or more completely missing sequences, 3) were supported by a mean RAxML v.8.2.4[101] bootstrap value lower than 0.65, 4) were located within the four supergene regions, 5) exhibited significant exon tree discordance according to an analysis with Concaterpillar v.1.7.2[102], or 6) had a gene tree with non-clock-like evolution (mean estimate for coefficient of variation greater than 0.5 or 95% highest-posterior-density interval including 1.0) according to a relaxed-clock analysis with BEAST 2[66]. Finally, concatenated exon alignments per gene were inspected by eye, and six genes were removed due to remaining possible misalignment. The filtered dataset included alignments for 91 genes with a total alignment length of 106,566 bp and a completeness of 92.8%.

We inferred the species tree of Gadinae with StarBEAST2[62, 66] under the multi-species coalescent model, assuming a strict clock, constant population sizes, and the birth-death tree model[103], and averaging over substitution models with the bModelTest package[104] for BEAST 2. For time calibration, we placed lognormal prior distributions on the age of the divergence of the outgroup *Brosme brosme* from Gadinae (mean in real space: 32.325; standard deviation: 0.10) and on the crown age of Gadinae (mean in real space: 18.1358; standard deviation: 0.28); in both cases, the distribution parameters were chosen to approximate previous phylogenomic age estimates for these two divergence events[105]. We performed five replicate StarBEAST2 analyses, each with a length of one billion Markov-chain Monte Carlo (MCMC) iterations. After merging replicate posterior distributions, effective sample sizes (ESS) for all model parameters were greater than 1,000, indicating full stationarity and convergence of MCMC chains. We then used TreeAnnotator from the BEAST 2 package to summarize the posterior tree distribution in the form of a maximum-clade-credibility (MCC) consensus tree with Bayesian posterior probabilities as node support[106].

Estimating divergence times and introgression among species of the genera *Gadus*, *Arctogadus*, and *Boreogadus*. To further investigate divergence times and introgression among species of the closely related genera *Gadus*, *Arctogadus*, and *Boreogadus*, we used a second phylogenomic dataset based on read mapping to the gadMor2 assembly. This dataset included Illumina read data for all four species of the genus *Gadus* (*G. morhua*, *G. chalcogrammus*, *G. macrocephalus*, and *G. ogac*)[54, 63], *Arctogadus glacialis*[63], and *Boreogadus saida*[63], as well as *Merlangius merlangius*, *Melanogrammus aeglefinus*, and *Pollachius virens*[60, 63], which we here considered outgroups (Supplementary Table 4). Read data from both a stationary and a migratory indivual (both sampled at the Lofoten islands) were used to represent *Gadus morhua*, to assess if one of the two received weaker or stronger introgression from other taxa. Mapping, read sorting, and indexing were again performed with BWA MEM and SAMtools, and variant calling was again performed with GATK's HaplotypeCaller and GenotypeGVCFs tools as described above except that we now also exported invariant sites to the output file. To limit the dataset to the most reliably mapping genomic regions, we applied the mask of unreliable sites generated from the threeway whole-genome alignment (see above), resulting in set of 19,035,318 SNPs. We then extracted alignments from GATK's output files for each non-overlapping window of 5,000 bp for which no more than 4,000 sites were masked, setting all genotypes with a Phred-scaled likelihood below 20 to missing. Alignments were not extracted from the four supergene regions and those windows with less than 100 variable sites were ignored. As we did not model recombination within alignments in our phylogenomic inference, the most suitable alignments for the inference were those with weak signals of recombination. Therefore, we calculated the number of hemiplasies per alignment by comparing the number of variable sites with the parsimony score, estimated with PAUP*[107], and excluded all alignments that had more than ten hemiplasies. Finally, we again removed all alignment sites for which BMGE determined a gap rate greater than 0.2 or an entropy-like score greater than 0.5. The resulting filtered dataset was composed of 109 alignments with a total length of 383,727 bp and a completeness of 91.0%.

We estimated the species tree and introgression among *Gadus*, *Arctogadus*, and *Boreogadus* under the isolation-with-migration model implemented in the AIM package[65] for BEAST 2. The inference assumed a strict clock, constant population sizes, the pure-birth tree model[108], and the HKY[109] substitution model with gamma-distributed rate variation among sites[110]. We time-calibrated the species tree with a single lognormal prior distribution on the divergence of *Pollachius virens* from all other taxa of the dataset (mean in real space: 8.56; standard deviation: 0.08), constraining the age of this divergence event according to the results of the analysis of divergence times of Gadinae (see above; Supplementary Figure 1). We performed ten replicate analyses that each had a length of five billion MCMC iterations, resulting in ESS values greater than 400 for all model parameters. The posterior tree distribution was subdivided according to tree topology and inferred gene flow and we produced separate MCC consensus trees for each of the tree subsets.

To further test for introgression among *Gadus*, *Arctogadus*, and *Boreogadus*, we calculated Patterson's D-statistic from the masked dataset for all possible species trios (with *Pollachius virens* fixed as outgroup) using the "Dtrios" function of Dsuite v.0.1.r3[69]. For the calculation of the D-statistic, species trios were sorted in two ways; with a topology fixed according to the species tree inferred under the isolation-with-migration model (D_{fix}), and so that the number of "BBAA" patterns was greater than those of "ABBA" and "BABA" patterns (D_{BBAA}). The significance of the statistic was assessed through block-jackknifing with 20 blocks of equal size. For the trios with the most significant signals of introgression, we further used the "Dinvestigate" function of Dsuite to calculate the D-statistic within sliding windows of 50 SNPs, overlapping by 25 SNPs.

To corroborate the introgression patterns inferred with Dsuite, we performed two analyses based on comparisons of the frequencies of trio topologies in maximum-likelihood phylogenies. Alignments for these analyses were selected as for the species-tree inference under the isolation-with-migration model, except that up to 20 hemiplasies were allowed per alignment. The resulting set of 851 alignments had a total length of 3,052,697 bp and a completeness of 91.0%. From each of these alignments, a maximum-likelihood phylogeny was inferred with IQ-TREE v.1.6.8[111] with a substitution model selected through IQ-TREE's standard model selection. Branches with a length below 0.001 were collapsed into polytomies. Based on the inferred maximum-likelihood trees, we calculated, for all possible species trios, the D_{tree}-statistic of Ronco et al., a tree-based equivalent to Patterson's D-statistic in which the frequencies of pairs of sister taxa are counted in a set of trees instead of the frequencies of shared sites in a genome (a related measure was proposed by Huson et al.[112]): $D_{\mathrm{tree}} = (f_{\mathrm{2nd}} - f_{\mathrm{3rd}})/(f_{\mathrm{2nd}} + f_{\mathrm{3rd}})$, where for a given trio, f_{2nd} is the frequency of the second-most frequent pair of sisters and f_{3rd} is the frequency of the third-most frequent (thus, the least frequent) pair of sisters. As a second tree-based analysis of introgression, we applied genealogy interrogation[71], comparing the likelihoods of trees with alternative topological constraints for the same alignment, as in Barth et al.[72]. We tested two hypotheses of introgression with this method: 1) Introgression between *Arctogadus glacialis* and either *Boreogadus saida* or the group of the four species of the genus *Gadus*; and 2) introgression between *Gadus ogac* and the sister species *Gadus chalcogrammus* and *Gadus morhua*.

Estimating divergence times, demography, and gene flow among *Gadus morhua* populations. To investigate divergence times among *Gadus morhua* populations, we applied phylogenetic analyses to the dataset based on whole-genome sequencing and variant calling for 24 *Gadus morhua* individuals (Supplementary Table 7). This dataset included, now considered as outgroups, the same representatives of *Gadus chalcogrammus*, *G. macrocephalus*, *G. ogac*, *Arctogadus glacialis*, and *Boreogadus saida* as our analyses of divergence times and introgression among *Gadus*, *Arctogadus*, and *Boreogadus* (see above). "Migratory" and "stationary" *Gadus morhua* individuals from Newfoundland, Iceland, Lofoten, and Møre were used as separate groups in these analyses. Subsequent to mapping with BWA MEM and variant calling with GATK's HaplotypeCaller and GenotypeGVCFs tools, we filtered the called variants with BCFtools to include only sites for which the Phred-scaled p value for Fisher's exact test was smaller than 20, the quality score normalized by read depth was greater than 2, the root-mean-square mapping quality was greater than 20, the overall read depth across all individuals was between the 10 and 90% quantiles, and the inbreeding coefficient was greater than -0.5. We further excluded sites if their Mann-Whitney-Wilcoxon rank-sum test statistic was smaller than -0.5 either for site position bias within reads or for mapping quality bias between reference and alternative alleles. After normalizing indels with BCFtools, SNPs in proximity to indels were discarded with a filter that took into account the length of the indel: SNPs were removed within 10 bp of indels that were 5 bp or longer, but only within 5, 3, or 2 bp if the indel was 3–4, 2, or 1 bp long, respectively. After applying this filter, all indels were removed from the dataset. For the remaining SNPs, genotypes with a read depth below 4 or a genotype quality below 20 were set to missing. Finally, we excluded all sites that were no longer variable or had more than two different alleles; the filtered dataset then contained 20,402,423 biallelic SNPs.

We inferred the divergence times among *Gadus morhua* populations from the SNP data under the multi-species coalescent model with the SNAPP add-on package for BEAST 2. Due to the high computational demand of SNAPP, we performed this analysis only with a further reduced set of 1,000 SNPs, randomly selected from all biallelic SNPs that were without missing genotypes and located outside of the supergene regions. The input files for SNAPP were prepared with the script snapp_prep.rb[74], implementing a strict-clock model and a pure-birth tree model. The tree of *Gadus morhua* populations and outgroup species was time-calibrated with a single lognormal prior distribution (mean in real space: 3.83; standard deviation: 0.093) that constrained the root age of the tree according to the results of the analysis of divergence times and introgression among *Gadus*, *Arctogadus*, and *Boreogadus* (see above; Fig. 2b, Supplementary Figure 2A). We performed three replicate SNAPP analyses, each with a length of 400,000 MCMC iterations, resulting in ESS values that were all greater than 400. The posterior tree distribution was again summarized as a MCC consensus tree.

Gene flow among *Gadus morhua* populations and outgroup species was investigated with Dsuite from all biallelic SNPs that were without missing genotypes and located outside of the four supergene regions; there were 408,574 of these. The gene flow analyses were performed with Dsuite's "Dtrios" function as described above.

Population sizes over time were estimated for all sampled *Gadus morhua* populations with Relate v.1.1.2[77]. To maximize the number of suitable SNPs for this analysis, we excluded all outgroups except the sister species, *Gadus chalcogrammus*, and repeated variant calling and filtering with the same settings as before. After applying a mask to exclude all variants from repetitive regions in the gadMor2 assembly (784,488 bp in total)[35], 10,872,496 biallelic SNPs remained and were phased with BEAGLE v.5.1[113], setting the population size assumed by BEAGLE to 10,000. We excluded all sites that were heterozygous in the *Gadus chalcogrammus* individual and then reconstructed an "ancestral" genome sequence from the gadMor2 assembly and the called variants for *G. chalcogrammus*. Following this reconstruction, we removed *G. chalcogrammus* from the set of SNPs and excluded all sites that had become monomorphic after this removal, leaving 7,101,144 SNPs that were biallelic among the sampled *Gadus morhua* individuals. In addition to the "ancestral" genome sequence and the set of biallelic SNPs, we prepared a mask for the Relate analysis, covering all sites that were also included in the mask for repetitive regions, all sites that would have been excluded from variant calling due to proximity to indels (see above), and all sites that were ignored in the reconstruction of the "ancestral" sequence due to heterozygous genotype calls for the *G. chalcogrammus* individual.

As Relate further requires an estimate of the mutation rate, we calculated this rate for the filtered set of SNPs as the mean pairwise genetic distance between *Gadus morhua* individuals from the Northwest Atlantic (thus, from the populations Newfoundland and Labrador) and those from the Northeast Atlantic (thus, from all other populations), divided by 2 times the expected coalescence time between the two groups and the genome size. We excluded the four linkage groups carrying supergenes from this calculation. The expected coalescence time was calculated as the divergence time between the two groups, which was estimated in the analysis with SNAPP as 65,400 years (Fig. 3), plus the expected time to coalescence within the common ancestor, which is the product of the generation time and the diploid population size under the assumption of a panmictic ancestral population. With an assumed generation time of 10 years[114] and a population size of 57,400, as estimated in the SNAPP analysis, the expected time to coalescence within the common ancestor is 574,000 years, and the total expected coalescence time was thus set to $65,400 + 574,000 = 639,400$ years. As the mean pairwise genetic distance between the individuals of the two groups was 878,704.31 and the size of the gadMor2 assembly without LGs 1, 2, 7, and 12, and excluding masked sites, is 419,183,531 bp, the calculated mutation rate was $\mu = 878,704.31/(2 \times 639,400 \times$

$419,183,531) = 1.64 \times 10^{-9}$ per bp and per year, or 1.64×10^{-8} per bp and per generation. Because the genetic distance was calculated from the filtered set of SNPs, this rate is likely to underestimate the true mutation rate of *Gadus morhua*; however, because the same filtered set of SNPs was used as input for Relate, this rate is applicable in our inference of population sizes over time. The input file was converted from variant call format to haplotype format using RelateFileFormats with the flag "–mode ConvertFromVcf". The script PrepareInputFiles.sh was used to flip genotypes according to the reconstructed "ancestral" genome sequence and to adjust distances between SNPs using the mask prepared for this analysis. Relate was first run to infer genome-wide genealogies and mutations assuming the above calculated mutation rate of 1.64×10^{-8} per bp and per generation and a diploid effective population size of 50,000. This was followed by an estimation of the population sizes over time by running the script EstimatePopulationSize.sh for five iterations, applying the same mutation rate and setting the threshold to remove uninformative trees to 0.5. The tools and scripts RelateFileFormats, PrepareInputFiles.sh, and EstimatePopulationSize.sh are all distributed with Relate.

Estimating divergence times, demography, and gene flow specific to supergenes. The analyses of divergence times, demography, and gene flow among *Gadus morhua* populations were repeated separately with SNPs from each of the four supergene regions on LGs 1, 2, 7, and 12. While the SNAPP analyses for these regions were again performed with reduced subsets of 1,000 SNPs per region, the data subsets used in analyses of gene flow with Dsuite comprised 11,474, 3,123, 10,4121, and 10,339 biallelic SNPs, and those used in the analyses of demography with Relate comprised 211,057, 71,046, 130,918, and 130,620 biallelic SNPs, respectively. The mutation rate used as input for these Relate analyses was identical to the one used for the analysis with genome-wide SNPs.

Estimating population divergence times across the genome. In addition to the genome-wide and supergene-specific SNAPP analyses that used biallelic SNPs from the entire genome or the entire length of supergene regions, we also performed sliding-window SNAPP analyses across all linkage groups to quantify differences in population divergence times across the genome. Our motivation for these analyses was primarily to assess whether or not divergence times were homogeneous over the lengths of supergenes, as differences in these divergence times within supergenes could be informative both about the presence of separate inversion within these regions and about their erosion processes. Additionally, we expected that these analyses could reveal further putative inversions elsewhere in the genome if they should exist.

From the set of 20,402,423 biallelic SNPs, we extracted subsets of SNPs for each non-overlapping window of a length of 250,000 bp, with a minimum distance between SNPs of 50 bp. We discarded windows with less than 500 remaining biallelic SNPs and used a maximum of 1,000 biallelic SNPs per window; these were selected at random if more biallelic SNPs were available per window. Input files for SNAPP were then prepared as for the genome-wide and supergene-specific SNAPP analyses. Per window, we performed two replicate SNAPP analyses with an initial length of 100,000 MCMC iterations, and these analyses were resumed up to a maximum of 500,000 MCMC iterations as long as the lowest ESS value was below 100. Windows with less than 300 sufficiently complete SNPs for SNAPP analyses, with an ESS value below 100 after the maximum number of MCMC iterations, or with a mean BPP node support value below 0.5 were discarded after the analysis. Per remaining window, posterior tree distributions from the two replicate analyses were combined and summarized in the form of MCC consensus trees. Additionally, a random sample of 100 trees was drawn from each combined posterior distribution.

Instead of showing all resulting trees, we developed a type of plot that shows, without loss of phylogenetic information, the divergence times stacked upon each other on a single axis, which allowed us to illustrate these divergence times efficiently across linkage groups. For this plot, all trees were first ladderized, outgroups were pruned, and the divergence times between each pair of populations adjacent to each other on the ladderized trees were extracted. Per window, the order of populations on the ladderized tree, together with the extracted divergence times between them, was used to define the positions of points on the vertical axis of the plot, so that each point represents a population and their vertical distances indicate the divergence times between populations that are next to each other on the ladderized tree. The positions of window on the linkage group were used to place these dots on the horizontal axis of the plot, and all dots representing the same population were connected by lines to produce the complete plot of divergence times across linkage groups.

GO term analysis. Gene ontology (GO) term enrichment tests were performed for 24 genes located within the region of the supergene on LG 12 with evidence for double crossover (positions 7,478,537 bp to 7,752,994 bp), against a reference set containing all 14,060 predicted genes associated with GO terms in the gadMor2 assembly[35]. Significant enrichment in biological processes, molecular functions, or cellular components was tested for using the topGO package v.2.42[115] for R v.4.0.2. We applied Fisher's exact test with the algorithm "weight01" and adjusted for multiple testing, considering values with a false discovery rate (FDR) below 0.05 as significant.

Code availability

Code for computational analyses is available from Github (http://github.com/mmatschiner/supergenes).

Data availability

The gadMor_Stat assembly (ENA accession ERZ1743403) and read data for all *Gadus morhua* specimens listed in Supplementary Table 7 are deposited on ENA with project number PRJEB43149. Alignment files, SNP datasets in PED and VCF format, and input and output of phylogenetic analyses are available from Zenodo (doi: 10.5281/zenodo.4560275).

Acknowledgements

We thank M. Malmstrøm, P. Berg, and D. Righton for help with fieldwork, and M. Skage, S. Kollias, M. S. Hansen, and A. Tooming-Klunderud from the Norwegian Sequencing Centre (NSC; https://www.sequencing.uio.no) for sequencing and processing of samples. PacBio and Illumina library creation and high-throughput sequencing were carried out at NSC, University of Oslo, Norway. All computational analyses were performed on the Abel and Saga supercomputing clusters (Norwegian metacenter for High Performance Computing and the University of Oslo) operated by the Research Computing Services group at USIT, the University of Oslo IT-department. We also thank Alexandra Viertler for drawings of codfishes and Côme Denechaud for providing otolith images.

Author contributions

M.M., K.S.J., and S.J. conceived this study. M.M. performed most analyses. J.M.I.B. contributed demographic analyses and the GO term analysis, O.K.T. produced the gadMor_Stat assembly, and B.S. performed variant calling. H.T.B. and M.S.O.B. contributed to the organization of the study, and K.S.J. and S.J. arranged whole-genome sequencing. C.P. and I.B. provided samples for sequencing. M.M. wrote the manuscript, with individual sections contributed by J.M.I.B. and O.K.T. All authors provided feedback and approved the final version of the manuscript.

References

1. Joron, M. *et al.* Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477,** 203–206 (2011).
2. Yan, Z. *et al.* Evolution of a supergene that regulates a trans-species social polymorphism. *Nat. Ecol. Evol.* **4,** 210–249 (2020).
3. Lamichhaney, S. *et al.* Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax). Nat. Genet.* **48,** 84–88 (2016).
4. Tuttle, E. M. *et al.* Divergence and functional degradation of a sex chromosome-like supergene. *Curr. Biol.* **26,** 344–350 (2016).
5. Li, J. *et al.* Genetic architecture and evolution of the S locus supergene in *Primula vulgaris. Nat. Plants* **2,** 16188 (2016).
6. Thompson, M. J. & Jiggins, C. D. Supergenes and their role in evolution. *Heredity* **113,** 1–8 (2014).
7. Schwander, T., Libbrecht, R. & Keller, L. Supergenes and complex phenotypes. *Curr. Biol.* **24,** R288–R294 (2014).

8. Tigano, A. & Friesen, V. L. Genomics of local adaptation with gene flow. *Mol. Ecol.* **25,** 2144–2164 (2016).

9. Gutiérrez-Valencia, J., Hughes, W., Berdan, E. L. & Slotte, T. The genomic architecture and evolutionary fates of supergenes. *arXiv.* arXiv:2012.11508 (2020).

10. Fisher, R. A. *The genetical theory of natural selection* (Clarendon Press, Oxford, UK, 1930).

11. Kirkpatrick, M. Chromosome inversions, local adaptation and speciation. *Genetics* **173,** 419–434 (2006).

12. Jay, P. *et al.* Supergene evolution triggered by the introgression of a chromosomal inversion. *Curr. Biol.* **28,** 1839–1845.e3 (2018).

13. Jay, P., Aubier, T. G. & Joron, M. Admixture can readily lead to the formation of supergenes. *bioRxiv.* doi:10.1101/2020.11.19.389577 (2020).

14. Dobzhansky, T. & Epling, C. The suppression of crossing over in inversion heterozygotes of *Drosophila pseudoobscura. Proc. Natl. Acad. Sci. U.S.A.* **34,** 137–141 (1948).

15. Sturtevant, A. H. & Beadle, G. W. The relations of inversions in the X chromosome of *Drosophila melanogaster* to crossing over and disjunction. *Genetics* **21,** 554–604 (1936).

16. Anton, E., Blanco, J., Egozcue, J. & Vidal, F. Sperm studies in heterozygote inversion carriers: a review. *Cytogenet. Genome Res.* **111,** 297–304 (2005).

17. Navarro, A., Barbadilla, A. & Ruiz, A. Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in *Drosophila. Genetics* **155,** 685–698 (2000).

18. Faria, R., Johannesson, K., Butlin, R. K. & Westram, A. M. Evolving inversions. *Trends Ecol. Evol.* **34,** 239–248 (2019).

19. Berdan, E. L., Blanckaert, A., Butlin, R. K. & Bank, C. Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *bioRxiv.* doi:10.1101/606012 (2020).

20. Bachtrog, D. A dynamic view of sex chromosome evolution. *Curr. Opin. Genet. Dev.* **16,** 578–585 (2006).

21. Blaser, O., Grossen, C., Neuenschwander, S. & Perrin, N. Sex-chromosome turnovers induced by deleterious mutation load. *Evolution* **67,** 635–645 (2012).

22. Andolfatto, P., Depaulis, F. & Navarro, A. Inversion polymorphisms and nucleotide variability in *Drosophila. Genet. Res.* **77,** 1–8 (2001).

23. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat. Genet.* **36,** 151–156 (2004).

24. Williams, A. L. *et al.* Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLIFE* **4,** e04637 (2015).

25. Chovnick, A. Gene conversion and transfer of genetic information within the inverted region of inversion heterozygotes. *Genetics* **75,** 123–131 (1973).

26. Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C. & Patrinos, G. P. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet.* **8,** 762–775 (2007).

27. Korunes, K. L. & Noor, M. A. F. Pervasive gene conversion in chromosomal inversion heterozygotes. *Mol. Ecol.* **28,** 1302–1315 (2019).

28. Stevison, L. S., Hoehn, K. B. & Noor, M. A. F. Effects of inversions on within- and between-species recombination and divergence. *Genome Biol. Evol.* **3,** 830–841 (2011).

29. Navarro, A., Betrán, E., Barbadilla, A. & Ruiz, A. Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics* **146,** 695–709 (1997).

30. Reis, M., Vieira, C. P., Lata, R., Posnien, N. & Vieira, J. Origin and consequences of chromosomal inversions in the *virilis* group of *Drosophila. Genome Biol. Evol.* **10,** 3152–3166 (2018).

31. Bradbury, I. R. *et al.* Long distance linkage disequilibrium and limited hybridization suggest cryptic speciation in Atlantic Cod. *PLOS ONE* **9,** e106380 (2014).

32. Berg, P. R. *et al.* Adaptation to low salinity promotes genomic divergence in Atlantic cod (*Gadus morhua* L.) *Genome Biol. Evol.* **7,** 1644–1663 (2015).

33. Berg, P. R. *et al.* Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Sci. Rep.* **6,** 23246 (2016).

34. Sodeland, M. *et al.* "Islands of divergence" in the Atlantic cod genome represent polymorphic chromosomal rearrangements. *Genome Biol. Evol.* **8,** 1012–1022 (2016).

35. Tørresen, O. K. *et al.* An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics* **18,** 95 (2017).

36. Kirubakaran, T. G. *et al.* Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Mol. Ecol.* **25,** 2130–2143 (2016).

37. Barney, B. T., Munkholm, C., Walt, D. R. & Palumbi, S. R. Highly localized divergence within supergenes in Atlantic cod (*Gadus morhua*) within the Gulf of Maine. *BMC Genomics* **18,** 271 (2017).

38. Berg, P. R. *et al.* Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity* **119,** 418–428 (2017).

39. Kess, T. *et al.* A migration-associated supergene reveals loss of biocomplexity in Atlantic cod. *Sci. Adv.* **5,** eaav2461 (2019).

40. Barth, J. M. I. *et al.* Genome architecture enables local adaptation of Atlantic cod despite high connectivity. *Mol. Ecol.* **26,** 4452–4466 (2017).

41. Barth, J. M. I. *et al.* Disentangling structural genomic and behavioural barriers in a sea of connectivity. *Mol. Ecol.* **87,** 449 (2019).

42. Berg, E. & Albert, O. T. Cod in fjords and coastal waters of North Norway: distribution and variation in length and maturity at age. *ICES J. Mar. Sci.* **60,** 787–797 (2003).

43. Case, R., Hutchinson, W. F., Hauser, L., Van Oosterhout, C. & Carvalho, G. R. Macro- and micro-geographic variation in pantophysin (PanI) allele frequencies in NE Atlantic cod *Gadus morhua. Mar. Ecol. Prog. Ser.* **301,** 267–278 (2005).

44. Star, B. *et al.* Ancient DNA reveals the Arctic origin of Viking Age cod from Haithabu, Germany. *Proc. Natl. Acad. Sci. U.S.A.* **114,** 9152–9157 (2017).

45. Hemmer-Hansen, J. *et al.* A genomic island linked to ecotype divergence in Atlantic cod. *Mol. Ecol.* **22,** 2653–2667 (2013).

46. Sinclair-Waters, M. *et al.* Ancient chromosomal rearrangement associated with local adaptation of a postglacially colonized population of Atlantic Cod in the northwest Atlantic. *Mol. Ecol.* **27,** 339–351 (2017).

47. Kess, T. *et al.* Modular chromosome rearrangements reveal parallel and nonparallel adaptation in a marine fish. *Ecol. Evol.* **10,** 638–653 (2020).

48. Kirubakaran, T. G. *et al.* A nanopore based chromosome-level assembly representing Atlantic cod from the Celtic Sea. *G3,* g3.401423.2020 (2020).

49. Johansen, T. *et al.* Genomic analysis reveals neutral and adaptive patterns that challenge the current management regime for East Atlantic cod *Gadus morhua* L. *Evol. Appl.* **13,** 2673–2688 (2020).

50. Carr, S. M., Kivlichan, D. S., Pepin, P. & Crutcher, D. C. Molecular systematics of gadid fishes: implications for the biogeographic origins of Pacific species. *Can. J. Zool.* **77,** 19–26 (1999).

51. Coulson, M. W., Marshall, H. D., Pepin, P. & Carr, S. M. Mitochondrial genomics of gadine fishes: implications for taxonomy and biogeographic origins from whole-genome data sets. *Genome* **49,** 1115–1130 (2006).

52. Bermingham, E., McCafferty, S. S. & Martin, A. P. in *Molecular Systematics of Fishes* (eds Kocher, T. D. & Stepien, C. A.) 113–128 (Academic Press, San Diego, USA, 1997).

53. Owens, H. L. Evolution of codfishes (Teleostei: Gadinae) in geographical and ecological space: evidence that physiological limits drove diversification of subarctic fishes. *J. Biogeogr.* **42,** 1091–1102 (2015).

54. Árnason, E. & Halldórsdóttir, K. Codweb: Whole-genome sequencing uncovers extensive reticulations fueling adaptation among Atlantic, Arctic, and Pacific gadids. *Sci. Adv.* **5,** eaat8788 (2019).

55. Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24,** 2818–2824 (2008).

56. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9,** e112963 (2014).

57. Sturtevant, A. H. A case of rearrangement of genes in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.* **7,** 235–237 (1921).

58. Kirkpatrick, M. How and why chromosome inversions evolve. *PLOS Biol.* **8,** e1000501 (2010).

59. Stransky, C. *et al.* Separation of Norwegian coastal cod and Northeast Arctic cod by outer otolith shape analysis. *Fish. Res.* **90,** 26–35 (2008).

60. Tørresen, O. K. *et al.* Genomic architecture of haddock (*Melanogrammus aeglefinus*) shows expansions of innate immune genes and short tandem repeats. *BMC Genomics* **19,** 240 (2018).

61. Edelman, N. B. *et al.* Genomic architecture and introgression shape a butterfly radiation. *Science* **366,** 594–599 (2019).

62. Ogilvie, H. A., Bouckaert, R. R. & Drummond, A. J. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Mol. Biol. Evol.* **34,** 2101–2114 (2017).

63. Malmstrøm, M. *et al.* Evolution of the immune system influences speciation rates in teleost fishes. *Nat. Genet.* **48,** 1204–1210 (2016).

64. Hughes, L. C. *et al.* Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proc. Natl. Acad. Sci. U.S.A.* **5,** 201719358 (2018).

65. Müller, N. F., Ogilvie, H. A., Zhang, C., Drummond, A. & Stadler, T. Inference of species histories in the presence of gene flow. *bioRxiv.* doi:10.1101/348391 (2018).

66. Bouckaert, R. R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15,** e1006650 (2019).

67. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).

68. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28,** 2239–2252 (2011).

69. Malinsky, M., Matschiner, M. & Svardal, H. Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Mol. Ecol. Resour.* **19,** 1655 (2020).

70. Ronco, F. *et al.* Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature* **589,** 76–81 (2021).

71. Arcila, D. *et al.* Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nat. Ecol. Evol.* **1,** 1–10 (2017).

72. Barth, J. M. I. *et al.* Stable species boundaries despite ten million years of hybridization in tropical eels. *Nat. Commun.* **11,** 1433 (2020).

73. Bryant, D., Bouckaert, R. R., Felsenstein, J., Rosenberg, N. A. & RoyChoudhury, A. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* **29,** 1917–1932 (2012).

74. Stange, M., Sánchez-Villagra, M. R., Salzburger, W. & Matschiner, M. Bayesian divergence-time estimation with genome-wide SNP data of sea catfishes (Ariidae) supports Miocene closure of the Panamanian Isthmus. *Syst. Biol.* **67,** 681–699 (2018).

75. Bradbury, I. R. *et al.* Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proc. R. Soc. Lond. B* **277,** 3725–3734 (2010).

76. Ruegg, K., Anderson, E. C., Boone, J., Pouls, J. & Smith, T. B. A role for migration-linked genes and genomic islands in divergence of a songbird. *Mol. Ecol.* **23,** 4757–4769 (2014).

77. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* 1321–1329 (2019).

78. Runemark, A. *et al.* Variation and constraints in hybrid genome formation. *Nat. Ecol. Evol.* **2,** 549–556 (2018).

79. Thorsen, A., Kjesbu, O. S., Fyhn, H. J. & Solemidal, P. Physiological mechanisms of buoyancy in eggs from brackish water cod. *J. Fish Biol.* **48,** 457–477 (1996).

80. Finn, R. N. & Fyhn, H. J. Requirement for amino acids in ontogeny of fish. *Aquac. Res.* **41,** 684–716 (2010).

81. Johannesson, K., Smolarz, K., Grahn, M. & André, C. The future of Baltic Sea populations: local extinction or evolutionary rescue? *AMBIO* **40,** 179–190 (2011).

82. Nissling, A., Kryvi, H. & Vallin, L. Variation in egg buoyancy of Baltic cod *Gadus morhua* and its implications for egg survival in prevailing conditions in the Baltic Sea. *Mar. Ecol. Prog. Ser.* **110,** 67–74 (1994).

83. Nissling, A. & Westin, L. Salinity requirements for successful spawning of Baltic and Belt Sea cod and the potential for cod stock interactions in the Baltic Sea. *Mar. Ecol. Prog. Ser.* **152,** 261–271 (1997).

84. Hess, J. E. *et al.* Genomic islands of divergence infer a phenotypic landscape in Pacific lamprey. *Molecular Ecology* **29,** 3841–3856 (2020).

85. Schaeffer, S. W. & Anderson, W. W. Mechanisms of genetic exchange within the chromosomal inversions of *Drosophila pseudoobscura. Genetics* **171,** 1729–1739 (2005).

86. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10,** 563–569 (2013).

87. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

88. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

89. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

90. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

91. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv.* doi:10.1101/201178 (2018).

92. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

93. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

94. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

95. Sandes, E. F. d. O., Miranda, G., Melo, A. C. M. A. d., Martorell, X. & Ayguade, E. CUDAlign 3.0: Parallel Biological Sequence Comparison in Large GPU Clusters. *2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid),* 160–169 (2014).

96. Harris, R. S. *Improved pairwise alignment of genomic DNA.* PhD thesis (Pennsylvania State University, 2007).

97. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).

98. Roth, O. *et al.* Evolution of male pregnancy associated with remodeling of canonical vertebrate immunity in seahorses and pipefishes. *Proc. Natl. Acad. Sci. U.S.A.* **117,** 9431–9439 (2020).

99. Kucuk, E. *et al.* Kollector: transcript-informed, targeted de novo assembly of gene loci. *Bioinformatics* **33,** 1782–1788 (2017).

100. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10,** 210 (2010).

101. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).

102. Leigh, J. W., Susko, E., Baumgartner, M. & Roger, A. J. Testing congruence in phylogenomic analysis. *Syst. Biol.* **57,** 104–115 (2008).

103. Gernhard, T. The conditioned reconstructed process. *J. Theor. Biol.* **253,** 769–778 (2008).

104. Bouckaert, R. R. & Drummond, A. J. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol. Biol.* **17,** 42 (2017).

105. Musilova, Z. *et al.* Vision using multiple distinct rod opsins in deep-sea fishes. *Science* **364,** 588–592 (2019).

106. Heled, J. & Bouckaert, R. R. Looking for trees in the forest: summary tree from posterior samples. *BMC Evol. Biol.* **13,** 221 (2013).

107. Swofford, D. L. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4. (2003).

108. Yule, G. U. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Phil. Trans. R. Soc. B* **213,** 21–87 (1925).

109. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22,** 160–174 (1985).

110. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39,** 306–314 (1994).

111. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32,** 268–274 (2015).

112. Huson, D. H., Klöpper, T., Lockhart, P. J. & Steel, M. A. in *Research in Computational Molecular Biology. RECOMB 2005. Lecture Notes in Computer Science* (eds Miyano, S. *et al.*) 233–249 (Springer, Berlin, Heidelberg, 2005).

113. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).

114. Smedbol, R. K., Shelton, P. A., Fréchet, A. & Chouinard, G. A. Review of population structure, distribution and abundance of cod (*Gadus morhua*) in Atlantic Canada in a species-at-risk context. *Canadian Science Advisory Secretariat Research document 2002/082,* 1–134 (2002).

115. Alexa, A. & J, R. topGO: Enrichment Analysis for Gene Ontology. R package version 2.42.0. (2020).