

Article

Origin and fate of supergenes in Atlantic cod

Michael Matschiner^{1,2,3*}, Julia Maria Isis Barth⁴, Ole Kristian Tørresen¹, Bastiaan Star¹, Helle Tessand Baalsrud¹, Marine Servane Ono Briec¹, Christophe Pampoulie⁵, Ian Bradbury⁶, Kjetill Sigurd Jakobsen¹, Sissel Jentoft^{1*}

Addresses:

¹Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway.

²Department of Palaeontology and Museum, University of Zurich, Zurich, Switzerland.

³Current address: Natural History Museum, University of Oslo, Oslo, Norway.

⁴Zoological Institute, Department of Environmental Sciences, University of Basel, Basel, Switzerland.

⁵Marine and Freshwater Research Institute, Hafnarfjörður, Iceland.

⁶Fisheries and Oceans Canada, St. John's, Canada.

*Corresponding authors: E-mail: michael.matschiner@nhm.uio.no, sissel.jentoft@ibv.uio.no

1 Abstract

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

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

18 Introduction

19 Some of the most spectacular phenotypical variation within species, such as mimicry patterns in but-
20 terflies[1], social organization in ants[2], plumage morphs in birds[3, 4], and floral types in plants[5],
21 is encoded by supergenes — tightly linked sets of coadapted genes that are inherited as a single
22 Mendelian locus[6–9]. Even though supergenes have now been known for nearly a century[10], their
23 origin remains a challenging question[9], because it requires that beneficially interacting mutations
24 occur in at least two genes together with a reduction of recombination between these genes[7]. As
25 a scenario in which these requirements can be met, recent research has pointed to chromosomal in-
26 versions arising in incompletely separated groups: either locally adapted populations that exchange
27 migrants, or species that receive genetic material — introgression — through hybridization[11–13].
28 In these systems, the beneficial interaction between mutations in different genes can come from
29 their joint adaptation to the same environment, and these mutations can become linked if they are
30 captured by the same inversion[6–9, 14]. This linkage between mutations within inversions is the
31 result of a loop formation that occurs when chromosomes with the inverted haplotype pair with
32 chromosomes without it during meiosis. If a single crossover occurs between the two chromosomes
33 within the loop region, the recombinant chromosomes are affected by both duplications and dele-
34 tions and therefore unbalanced. The gametes carrying these unbalanced chromosomes are usually
35 lethal, and thus do not contribute to the next generation[15, 16]. When viewed backwards in time,
36 the recombination rate between inverted and noninverted haplotypes therefore appears reduced. On
37 the other hand, crossovers among two inverted haplotypes do not affect the viability of gametes, so

38 that their recombination rate remains unchanged. Since most inversions originate just once, in a
39 single individual, the number of individuals in which inverted haplotypes can successfully recombine
40 is initially very low, increasing only as the inversion becomes more frequent in the species. The
41 origin of a supergene is therefore expected to be equivalent to a severe bottleneck (down to a single
42 sequence) that affects part of the genome (the inversion region) in a part of the species (the carriers
43 of the inversion)[17].

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

72 In Atlantic cod (*Gadus morhua*), large genomic regions with tight linkage over 4–17 Megabase-
73 pairs (Mbp) and strong differentiation between alternative haplotypes have been identified[31–34]
74 on linkage groups (LGs) 1, 2, 7, and 12 of the gadMor2 reference genome assembly[35]. While the
75 functional consequences of this differentiation remain largely unknown, the classification of the four
76 regions as supergenes[36–39] is supported by the associations of the alternative haplotypes with
77 different lifestyles [33, 36, 39] and environments[32, 37, 38, 40, 41]. One of the strongest of these
78 associations is found between the alternative haplotypes on LG 1 and migratory and stationary At-

79 lantic cod ecotypes[33, 38]. In the Northeast Atlantic, these ecotypes co-occur during the spawning
80 season in March and April along the Norwegian coast, but are separated throughout the rest of
81 the year, which the migratory ecotype — the Northeast Arctic cod (NEAC) — spends in its native
82 habitat in the Barents Sea[42]. The association between the alternative haplotypes and the two
83 ecotypes is based on contrasting frequencies of these haplotypes, as almost all stationary individu-
84 als are homozygous for one of the haplotypes and migratory individuals are either heterozygous or
85 homozygous for the other haplotype[33, 43, 44]. This pattern of contrasting frequencies is repeated
86 between migratory and stationary Atlantic cod ecotypes in the Northwest Atlantic and on Iceland
87 (perhaps less clearly in the latter case)[45–47], corroborating the presumed functional link between
88 haplotypes and ecotypes. A number of genes from within the differentiated region on LG 1 have
89 been proposed as candidate genes under selection, including the *Ca6* gene that might play a role in
90 adaptations to feeding at greater depths in the migratory ecotype; however, the targets of selection
91 remain difficult to identify reliably due to the tight linkage across the entire region with close to
92 800 genes[36].

93 Similar to the different frequencies of LG 1 haplotypes between migratory and stationary eco-
94 types, one of the two alternative haplotypes on LG 2 is far more frequent in Atlantic cod from the
95 Baltic sea compared to the nearest North Atlantic populations and has been suggested to carry
96 genes adapted to the low salinity of the Baltic Sea[32, 41]. The alternative haplotypes on LGs 7 and
97 12 also differ in their frequencies among Atlantic cod populations, with one of the two haplotypes in
98 each case being nearly absent in the Irish and Celtic Sea, possibly in relation to adaptation to higher
99 temperatures during the spawning season in these populations[47–49], which are among the south-
100 ernmost populations in the Northeast Atlantic. Similar geographic distributions of haplotypes from
101 the four supergenes have been interpreted as evidence for interchromosomal linkage among these
102 haplotypes[31]; however, to what extent these similarities are the result of epistatic interactions,
103 adaptation to the same environment, or merely drift, remains uncertain[47].

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

113 The age of the supergene on LG 1 was estimated to be around 1.6 million years (Myr) based on
114 genetic distances between the two alternative haplotypes and homologous sequences of Greenland
115 cod (*Gadus ogac*), in combination with an assumed divergence time between Greenland and Atlantic
116 cod of 3.5 million years ago (Ma)[36]. This assumed divergence time, however, may be unreliable,
117 as it was based on a mitochondrial substitution rate[50, 51] that was originally calculated for
118 Caribbean fishes[52], and because the same divergence appears only about half as old in another

119 recent study[53]. For the supergenes on LGs 2, 7, and 12, no age estimates have yet been published.
120 Due to these uncertainties, conclusions about a possible joint origin of all four supergenes have
121 necessarily remained highly speculative. The role of introgression in the origin of the Atlantic cod's
122 supergenes has so far also been uncertain: While introgression among codfishes (subfamily Gadinae)
123 has already been investigated in one former study based on genome-scale sequence data[54], the
124 results of that study were affected by mislabeled sequences, reference bias, and incorrect application
125 of statistical tests (Supplementary Note 1), and thus remain inconclusive regarding the occurrence
126 of introgression.

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

143 Results

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

158 12 with a dataset of single-nucleotide polymorphisms (SNPs) for 100 *Gadus morhua* individuals.
159 By quantifying the strength of linkage per SNP as the sum of the distances (in bp) with which
160 the SNP is strongly linked, we identified sharp declines of linkage marking the boundaries of all
161 four supergenes (Fig. 1b, Table 1), as expected under the assumption of large-scale chromosomal
162 inversions[57].

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

176 To determine which of the two genomes carries the inversion in each case, we also aligned contigs
177 from the long-read-based genome assembly of *Melanogrammus aeglefinus* (melAeg)[60], a closely
178 related outgroup within the subfamily Gadinae, to the gadMor2 assembly. Split contig alignments
179 were again identified mapping near the boundaries of the supergenes on LGs 1 and 7, indicating
180 that for these supergenes, it is the gadMor2 genome that carries the inversions. In contrast, a
181 single contig of the melAeg assembly was clearly colinear to the gadMor2 assembly in a region that
182 extended about 150 kbp in both directions from one of the ends of the supergene on LG 2 (Fig.
183 1b, Supplementary Table 2), indicating that the inversion on LG 2 is carried not by the gadMor2
184 genome but by the gadMor_Stat genome instead. For the supergene on LG 12, on the other hand,
185 no informative alignments between the melAeg assembly and the gadMor2 assembly were found;
186 thus, our contig-mapping approach did not allow us to determine which of the two *Gadus morhua*
187 genomes carries the inversion on this linkage group (however, our subsequent demographic analyses
188 suggested that it is the gadMor_Stat genome that carries the inversion on LG 12; see below).

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

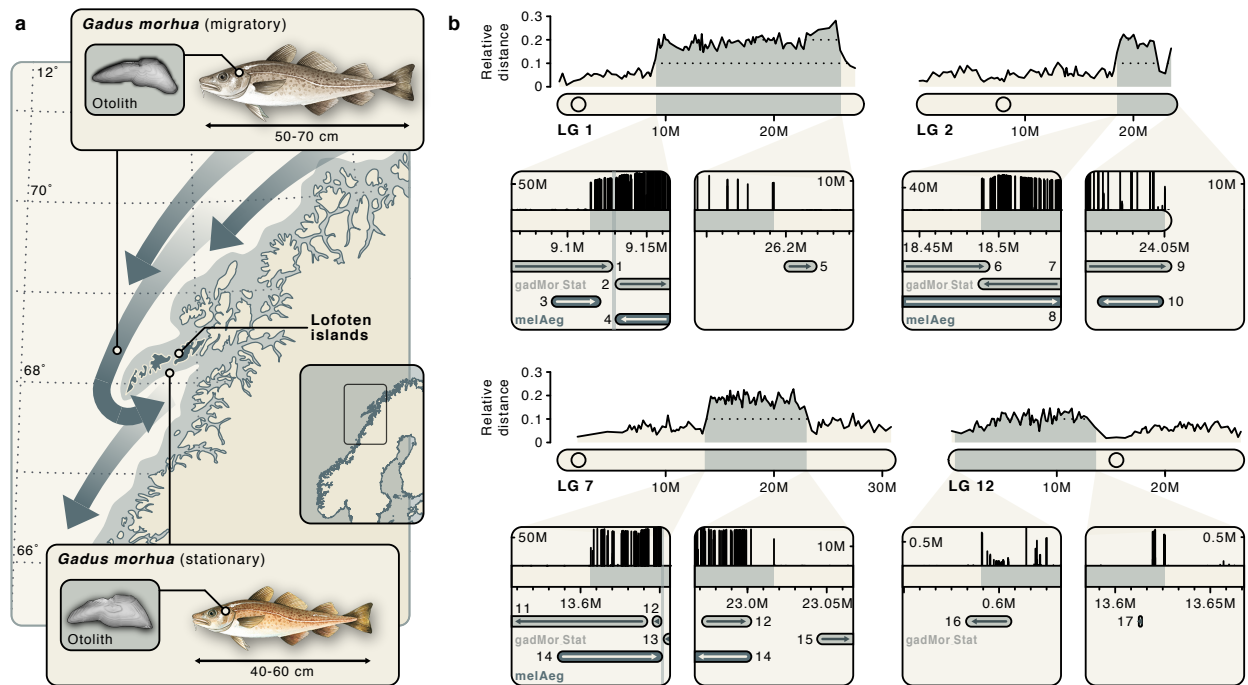


Fig. 1 Four supergenes associated with large chromosomal inversions in *Gadus morhua*. **a** Migratory and stationary *Gadus morhua* seasonally co-occur along the coast of northern Norway and differ in total length and otolith measurements [42, 59]. The distribution of stationary *Gadus morhua* is shaded 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.

198 and revealed a rapidly radiating clade comprised of the genera *Pollachius*, *Melanogrammus*, *Mer-*
 199 *langius*, *Boreogadus*, *Arctogadus*, and *Gadus* that occurs exclusively on the Northern hemisphere
 200 and began to diversify around 8.6 Ma (95% HPD: 9.9–7.2 Ma) (Supplementary Figure 1). We then
 201 refined the estimates for this clade with a second analysis that focused on the divergences among
 202 *Boreogadus*, *Arctogadus* (Fig. 2a; Supplementary Table 4), and *Gadus* and used 109 alignments
 203 (total length: 383,727 bp) sampled from across the genome. In this analysis, we co-estimated and
 204 accounted for possible introgression among species by applying the isolation-with-migration model
 205 as implemented in the AIM add-on package for BEAST 2 [65, 66]. We found strong support for two
 206 topologies in which *Arctogadus* is either placed as the sister taxon to *Gadus* (Bayesian posterior
 207 probability, BPP: 0.763) or to *Boreogadus* (BPP: 0.234), with introgression supported (by Bayes
 208 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

¹Coordinates refer to the gadMor2 assembly [35].

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

³Comparison 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.

⁴Due 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.

⁵This 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.

209 the common ancestor of the genus *Gadus* to *Arctogadus* in the latter case (Fig. 2b, Supplementary
210 Figure 2). While being inconclusive about the phylogenetic position of *Arctogadus*, these results
211 indicate that the genus is genetically more similar to both *Gadus* and *Boreogadus* than can be
212 explained by a bifurcating tree without introgression. Regardless of the position of *Arctogadus*, our
213 analysis with the isolation-with-migration model supported an age of around 4 Ma for the clade
214 comprising the three genera (mean estimates: 3.81 Ma and 3.99 Ma; 95% HPD: 4.44–3.19 Ma and
215 4.56–3.33 Ma; Fig. 2b).

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

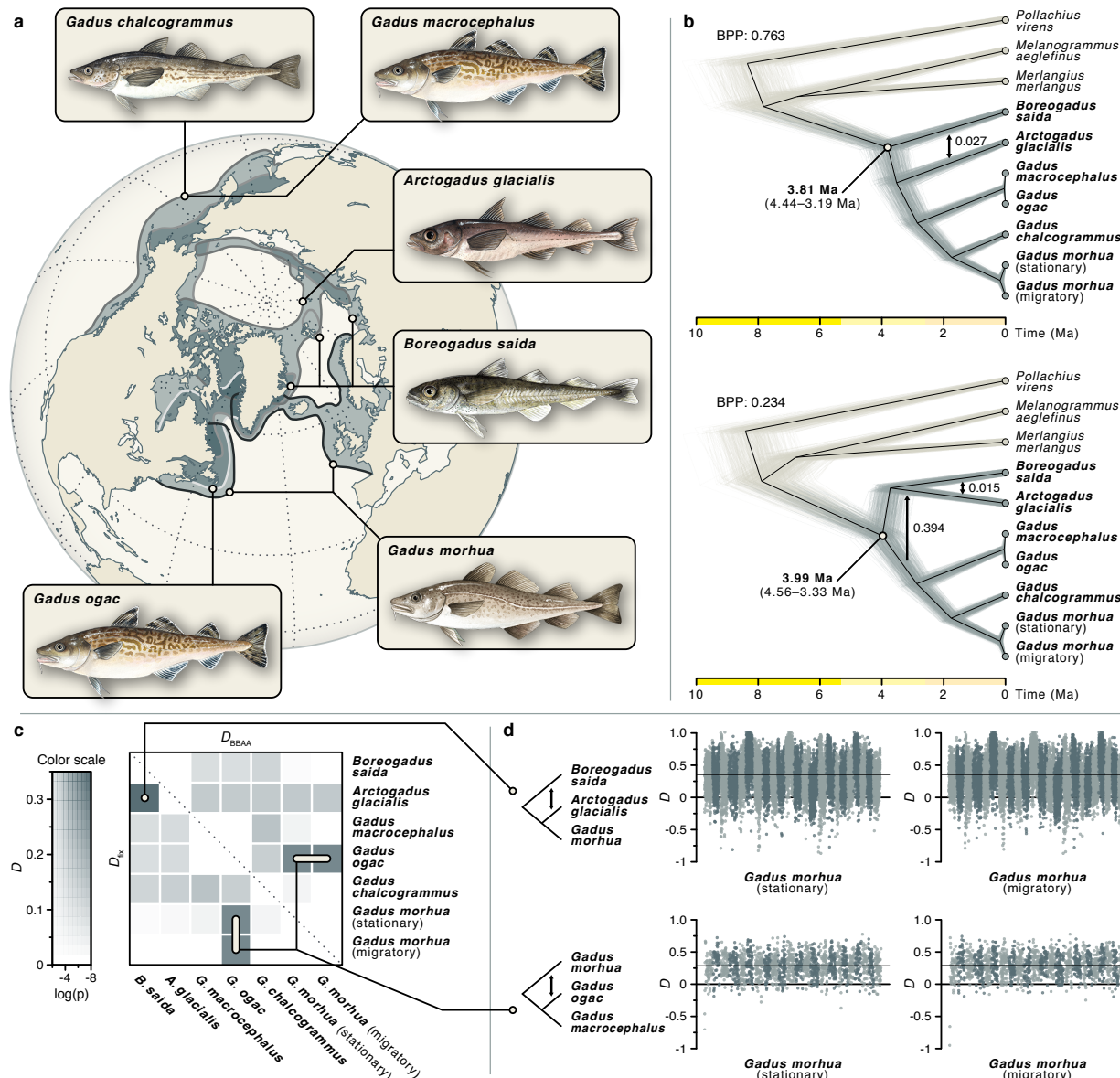


Fig. 2 Divergence times and introgression among Gadinae. **a** Distribution ranges of sampled species of the genera *Gadus*, *Arctogadus*, and *Boreogadus*. Partially overlapping distribution ranges are shown in dark gray. **b** Species tree of the six species and three outgroups (in beige; *P. virens*, *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_{BBA} , 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.

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

239 **Recent divergences among *Gadus morhua* populations.** We performed phylogenomic anal-
 240 yses for individuals from eight *Gadus morhua* populations covering the species' distribution in the
 241 North Atlantic (Fig. 3a; Supplementary Table 7). In addition to the individuals used for the gad-
 242 Mor2 and gadMor.Stat assemblies, we selected from these populations 22 individuals for which
 243 preliminary analyses had shown that each of them carried, at each of the four supergene regions,
 244 two copies of the same haplotypes (i.e., they were homokaryotypic). For the four sampling localities
 245 Newfoundland, Iceland, Lofoten, and Møre, we discriminated between "migratory" and "stationary"

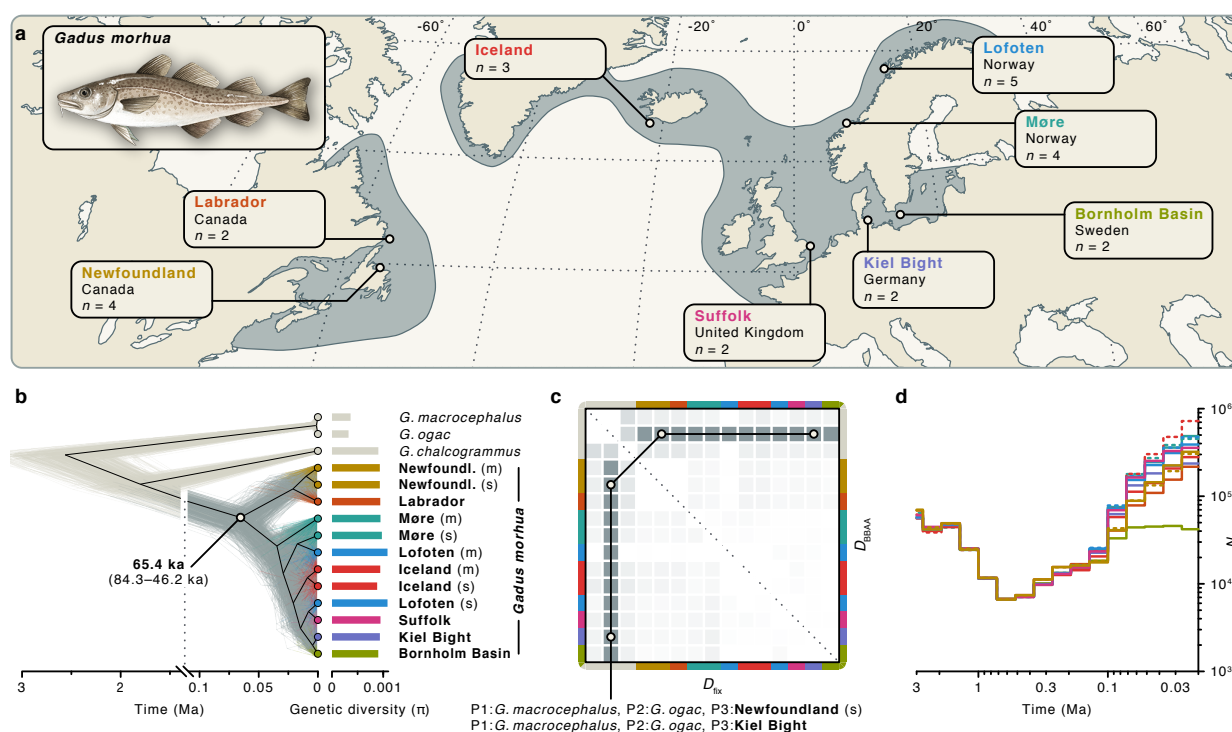


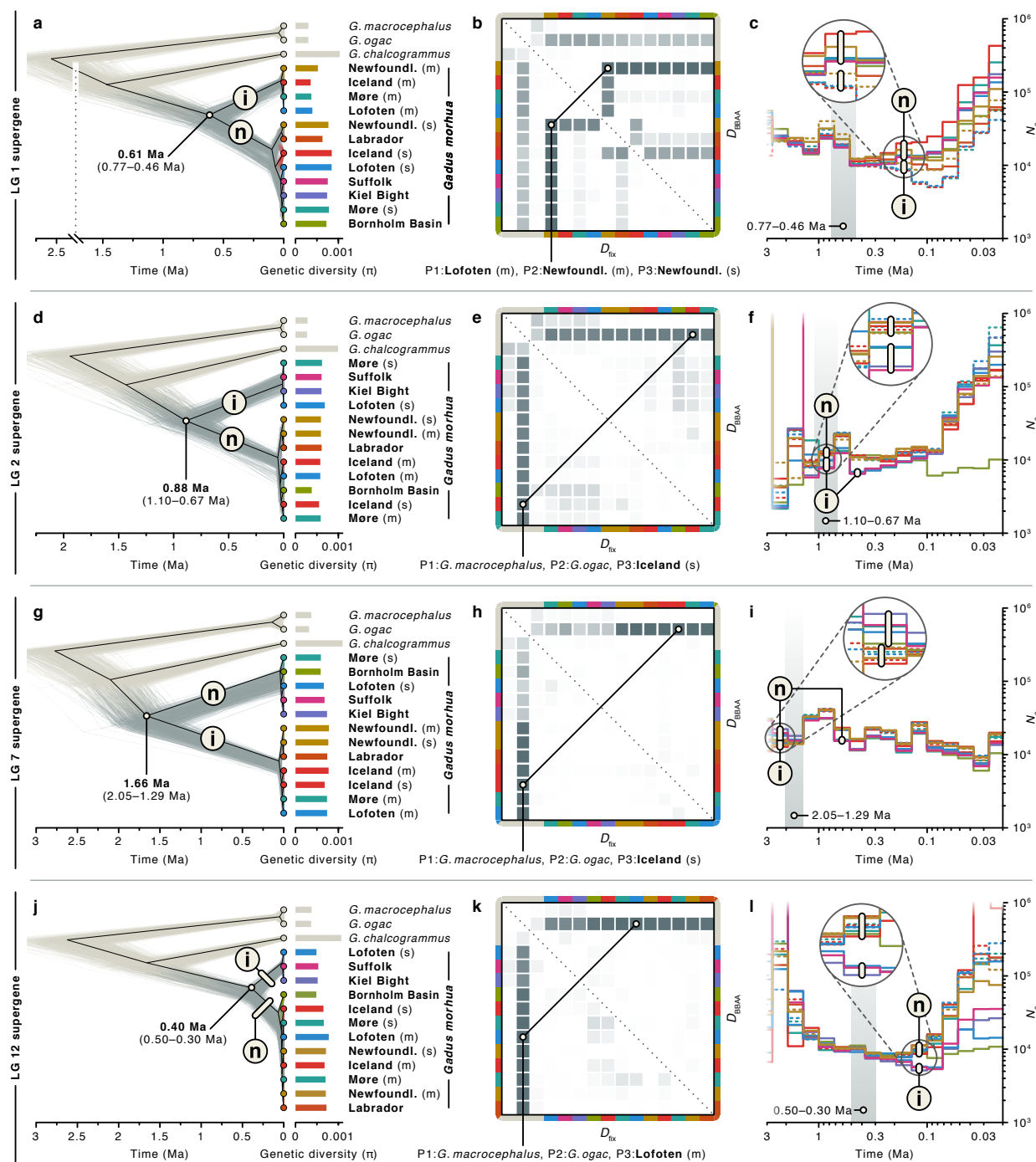
Fig. 3 Divergence times, demography, and introgression among *Gadus morhua* populations. **a** Geographic distribution of *Gadus morhua* in the North Atlantic and sampling locations for analyses of population divergence times, demography, and introgression. **b**, Tree of *Gadus morhua* populations and 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 $D_{\text{BBAA}} = D_{\text{fix}} = 0.201$ 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.

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

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

274 **Different age estimates for supergenes in *Gadus morhua*.** To infer the ages of the super-
275 genes 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_{BBA} , 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**).



276 extracted from the dataset of 20,402,423 biallelic SNPs. For each of the four supergenes, we recov-
 277 ered a deep divergence separating inverted and noninverted haplotypes; however, the age estimates
 278 for this divergence differed widely among the four supergenes, with mean age estimates of 0.61 Ma
 279 (95% HPD: 0.77–0.46 Ma) for the supergene on LG 1 (Fig. 4a), 0.88 Ma (95% HPD: 1.10–0.67
 280 Ma) for the supergene on LG 2 (Fig. 4d), 1.66 Ma (95% HPD: 2.05–1.29 Ma) for the supergene

281 on LG 7 (Fig. 4g), and 0.40 Ma (95% HPD: 0.50–0.30 Ma) for the supergene on LG 12 (Fig. 4j;
282 Supplementary Figure 5). The groups sharing the same haplotype also differed among the four
283 supergenes, even though the migratory individuals from Newfoundland, Iceland, Lofoten, and Møre
284 always shared the same haplotype. The genetic diversity was on average lower for the inverted
285 haplotypes on LGs 1 and 12, but increased compared to the noninverted haplotype on LGs 2 and
286 7 (Figs. 4a,d,g,j, Supplementary Table 8).

287 **Gene flow between inverted and noninverted haplotypes via gene conversion.** Like for
288 the genome-wide estimate of the divergence time between Northwest and Northeast Atlantic pop-
289 ulations, the ages of the separation between the two alternative haplotypes per supergene could
290 be underestimated if gene flow has occurred after their divergence. However, given that single-
291 crossover recombination is suppressed between inverted and noninverted haplotypes, gene flow be-
292 tween the two haplotypes can only occur through gene conversion or double-crossover events[29].
293 To shed light on the frequency of these two processes within the four supergene regions, we cal-
294 culated the D -statistic for the sets of supergene-specific SNPs. The D -statistic supported gene
295 flow between inverted and noninverted haplotypes particularly for the supergene on LG 1, in trios
296 that included either the migratory or stationary Newfoundland population (Fig. 4b). When both
297 of these populations were placed in the same trio together with the migratory Lofoten popula-
298 tion, 14,466 sites supported a sister-group relationship between the two migratory populations, in
299 agreement with the population tree inferred with SNAPP (Fig. 4a). However, with 1,041 sites
300 shared between the migratory and stationary Newfoundland populations but only 465 sites shared
301 between the migratory Lofoten population and the stationary Newfoundland population, the D -
302 statistic was $D_{\text{fix}} = D_{\text{BBAA}} = 0.383$ and strongly supported gene flow — and thus gene conversion
303 or double-crossover events — between the two geographically co-occurring migratory and station-
304 ary Newfoundland populations ($p < 10^{-10}$; Supplementary Tables 11 and 12). To test whether
305 gene conversion could be the cause of this gene flow occurring between the Newfoundland popula-
306 tions, we tested for the GC bias expected from gene conversion[24], comparing GC-content of sites
307 shared between the two Newfoundland populations (“ABBA” sites) to that of sites shared between
308 the migratory Newfoundland population and other migratory populations (“BBAA” sites). The
309 mean GC-content of the former, 0.482, is indeed significantly higher than that of the latter, 0.472
310 ($p < 10^{-5}$; one-sided t-test), supporting gene conversion as the agent of gene flow between inverted
311 and noninverted haplotypes in the Newfoundland populations. For the supergenes on LGs 2, 7, and
312 12, the D -statistic again — as for the genome-wide SNP data — primarily supported introgression
313 between *Gadus ogac* and *Gadus morhua* populations, which was largely uniform among populations
314 (Fig. 4e,h,k) and reached values between 0.333 and 0.376 (Supplementary Tables 13-18). For LG
315 7, this introgression was found to affect mostly the inverted haplotype, which might be related to
316 the physical separation between *Gadus ogac* and the populations represented by the noninverted
317 haplotype, as all of these occur in the Northeast Atlantic.

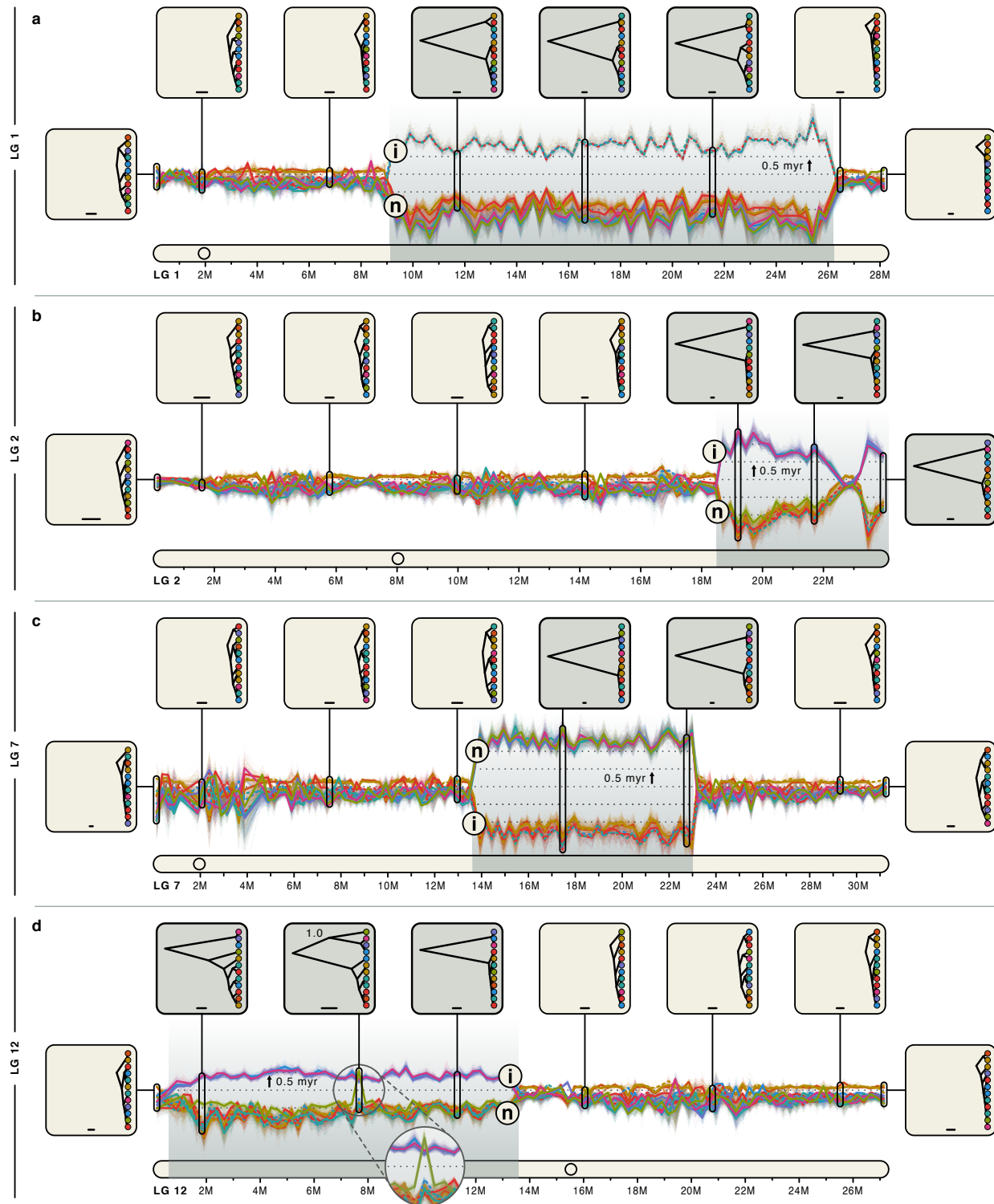
318 **Demographic analyses recover signatures of bottlenecks following inversions.** It is likely
319 that each of the inversions associated with the four supergenes in *Gadus morhua* originated a single
320 time, in a single individual, from which all of the current carriers of the inversion descended. The
321 supergene origin should therefore have been equivalent to an extreme bottleneck event during which

322 the population size was reduced to a single sequence, but which affected only the inversion region,
323 and only the carriers of the inversion. We might thus expect to see signatures of this extreme
324 bottleneck event in analyses of population size over time, in the form of differences in ancestral
325 population sizes between inverted and noninverted haplotypes that date to the time of supergene
326 origin.

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

351 **Divergence profiles reveal double crossovers.** To explore whether divergence times between
352 the two haplotypes per supergene are homogeneous across the supergene region, we repeated
353 divergence-time inference with SNAPP in sliding windows of 250 kbp along all linkage groups.
354 We expected that if any gene flow between inverted and noninverted haplotypes should proceed
355 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
357 divergence profiles for supergene regions[22, 29, 30]. Contrary to this expectation, the divergence

358 profiles were relatively homogeneous from beginning to end, particularly for the supergenes on LGs
359 1, 7, and 12 (Fig. 5a,c,d; Supplementary Figure 7), suggesting either that double crossovers are
360 rare within these supergenes, or that sequences exchanged through double crossovers are frequently
361 purged from the recipient haplotypes. As the supergene on LG 1 is known to include not one but two
362 adjacent inversions of roughly similar size[36], our results also suggested a similar age and possibly
363 a joint origin for both of these inversions. The divergence profile for the supergene on LG 2 ap-
364 peared consistent with the expectation of a U-shaped pattern; however, comparison of the gadMor2
365 assembly with the recently released gadMor3 assembly[48] showed that the end of this linkage group
366 may be misassembled in gadMor2, and the region of low divergence appearing within the supergene
367 (around position 22.5–23.0 Mbp) may in fact lie outside of it (Supplementary Figure 8). However,
368 a closer look at the divergence profile for LG 12 revealed a single window within the supergene in
369 which the otherwise clear separation between the groups carrying the alternative haplotypes was
370 interrupted: Unlike in all other windows within this supergene, the Bornholm Basin population
371 grouped (BPP: 1.0) with the three populations representing the inverted haplotype (Suffolk, Kiel
372 Bight, and stationary Lofoten; see Fig. 4j) in the window for positions 7.50–7.75 Mbp (Fig. 5d).
373 To investigate the genotypes of the two sampled Bornholm Basin individuals within this region in
374 more detail, we identified 219 haplotype-informative sites between positions 7–8 Mbp on LG 12,
375 and found that these individuals were both heterozygous at these sites, for a region of ~275 kbp
376 between positions 7,478,537 bp and 7,752,994 bp (Fig. 6). The two individuals from the Bornholm
377 Basin population thus carried a long sequence from the inverted haplotype even though they were
378 otherwise clearly associated with the noninverted haplotype. As the length of this introduced se-
379 quence was far longer than the 50–1,000 bp expected to be copied per gene-conversion event[23, 24],
380 it strongly supports a contribution of double crossover to gene flow between the two haplotypes of
381 the LG 12 supergene.

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

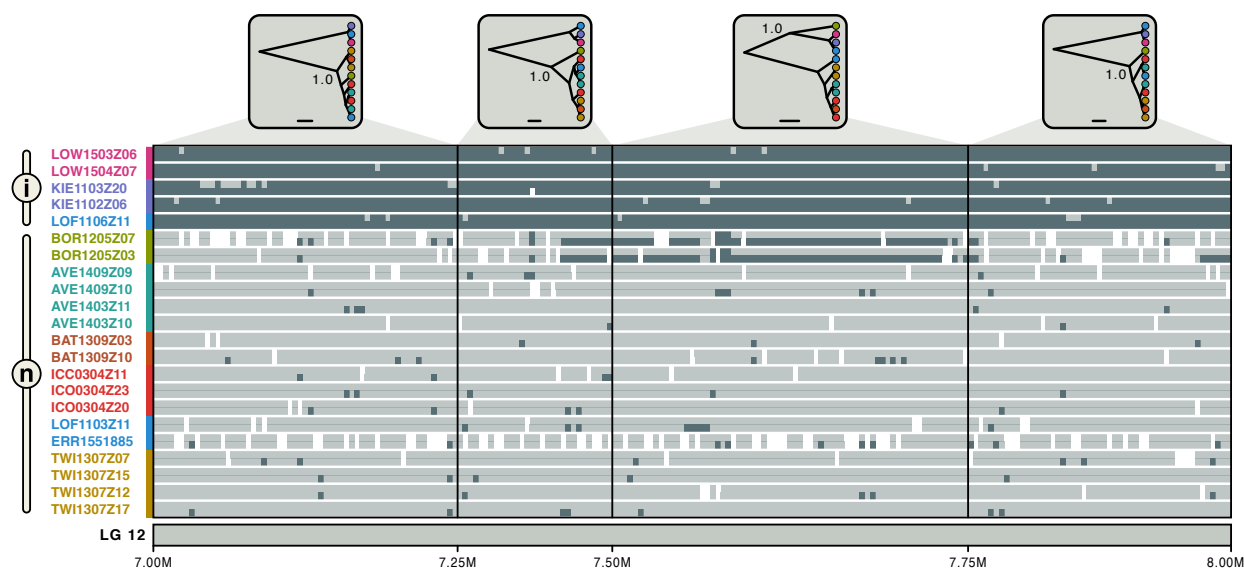


Fig. 6 Ancestry painting for part of the supergene on LG 12. The ancestry painting [72, 78] shows genotypes at 219 haplotype-informative sites between positions 7 and 8 Mbp on LG 12, within the supergene on that linkage group. For each of 22 *Gadus morhua* individuals, homozygous genotypes are shown 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 contrasting 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.

399 Discussion

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

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

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

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

453 The presence of four long (4–17 Mbp) and old (0.40–1.66 Ma) inversion-based supergenes in
454 Atlantic cod adds to recent findings of inversions of similar size and/or age in butterflies[12], ants[2],
455 birds[4], lampreys[84], and *Drosophila*[30]. For non-model organisms, these findings are largely owed
456 to the development of long-read-based genome sequencing within the last decade and may become

457 even more common as long-read-based sequencing is applied to more and more species. These
458 findings are, however, in contrast to earlier expectations based on theoretical work and empirical
459 studies on selected model organisms: Only about 20 years ago, the available evidence indicated
460 that inversions (and thus inversion-based supergenes) would be “generally not ancient” [22], with
461 maximum ages on the order of N_e generations (which would correspond to ~ 100 ka in Atlantic cod),
462 because they would either decay too rapidly due to the accumulation of mutation load, or erode
463 if gene flow occurs through gene conversion and double crossover [22, 29, 85]. While we observed
464 evidence for both gene conversion and double crossover in Atlantic cod, we did not find signs of
465 supergene decay or erosion: Decay could be indicated by high repeat content or mutation load [21],
466 but neither were increased within the four supergenes (Supplementary Figure 9). And supergene
467 erosion — at least when resulting from double crossovers — would be expected to produce U-shaped
468 divergence profiles [22, 29, 30], but no such profiles were found for the Atlantic cod supergenes (Fig.
469 5). These observations mirror those made recently by Yan et al. [2] for a supergene in fire ants,
470 and we thus concur with their conclusion that “low levels of recombination and/or gene conversion
471 may play an underappreciated role in preventing rapid degeneration of supergenes”. But, since our
472 results also indicated that selection may have acted on sequences exchanged between supergene
473 haplotypes, we further suggest that — just like in interbreeding species that maintain stable species
474 boundaries despite frequent hybridization [72] — selective purging of introduced sequences may also
475 be important for the fate of supergenes, by maintaining the rate of gene flow between haplotypes
476 at exactly the right balance, between too little of it and the consequential decay, and too much of
477 it and the resulting supergene erosion.

478 **Methods**

479 **Construction of the gadMor_Stat genome assembly.** We performed high-coverage genome
480 sequencing for a stationary *Gadus morhua* individual (LOF1106Z11) sampled at the Lofoten islands
481 in northern Norway, for which preliminary investigations had suggested that it carried, homozy-
482 gously on each of the four LGs 1, 2, 7, and 12, a supergene haplotype that was complementary
483 to the one in the gadMor2 genome [35], which represents a migratory individual from the Lofoten
484 islands. We used the Pacific Biosciences RS II platform, operated by the Norwegian Sequencing
485 Centre (NSC; www.sequencing.uio.no), to generate 2.4 million PacBio SMRT reads with a total
486 volume of 12.5 Gbp, approximately equivalent to an $18\times$ coverage of the *Gadus morhua* genome.
487 The PacBio SMRT reads were assembled with Celera Assembler v.8.3rc2 [55], adjusting the following
488 settings according to the nature of the PacBio reads (all others were left at their defaults): merSize
489 =16, merThreshold=0, merDistinct=0.9995, merTotal=0.995, ovErrorRate=0.40, ovMinLen=500,
490 utgGraphErrorRate=0.300, utgGraphErrorLimit=32.5, utgMergeErrorRate=0.35, utgMergeError-
491 Limit=40, utgBubblePopping=1, utgErrorRate=0.40, utgErrorLimit=25, cgwErrorRate=0.40, cns-
492 ErrorRate=0.40. The consensus sequence of the assembly was polished with Quiver v.0.9.0 [86] and
493 refined with Illumina reads sequenced for the same individual (see below). A total volume of 6.2
494 Gbp of Illumina reads were mapped to the assembly with BWA MEM v.0.7.12-r1039 [87] and sorted

495 and indexed with SAMtools v.1.10[88, 89]. Subsequently, Pilon v.1.16[56] was applied to recall
496 consensus.

497 **Whole-genome sequencing and population-level variant calling.** *Gadus morhua* individuals
498 sampled in Canada, Iceland, the United Kingdom, Germany, Sweden, and Norway (Fig. 3; Sup-
499 plementary Table 7) were subjected to medium-coverage whole-genome Illumina sequencing. DNA
500 extraction, library preparation, and sequencing were performed at the NSC; using the Illumina
501 Truseq DNA PCR-free kit for DNA extraction and an Illumina HiSeq 2500 instrument with V4
502 chemistry for paired-end (2×125 bp) sequencing. Sequencing reads were mapped to the gadMor2
503 assembly for *Gadus morhua*[35] with BWA MEM v.0.7.17 and sorted and indexed with SAM-
504 tools v.1.9; read duplicates were marked and read groups were added with Picard tools v.2.18.27
505 (<http://broadinstitute.github.io/picard>). Variant calling was performed with GATK's v.4.1.2.0[90,
506 91] HaplotypeCaller and GenotypeGVCFs tools, followed by indexing with BCFtools v.1.9[89].

507 **Delimiting high-LD regions associated with inversions.** As chromosomal inversions locally
508 suppress recombination between individuals carrying the inversion and those that do not, we used
509 patterns of linkage disequilibrium (LD) to guide the delimitation of inversion regions for each of the
510 four supergenes[32, 36, 48]. To maximize the signal of LD generated by the inversions, we selected
511 100 *Gadus morhua* individuals so that for each of the four supergenes, 50 individuals carried two
512 copies of one of the two alternative supergene haplotypes, and the other 50 individuals carried two
513 copies of the other. Variant calls of the 100 individuals were filtered with BCFtools, excluding all
514 indels and multi-nucleotide polymorphisms and setting all genotypes with a Phred-scaled quality
515 below 20, a read depth below 3, or a read depth above 80 to missing. Genotypes with more than 80%
516 missing data or a minor allele count below 20 were then removed from the dataset with VCFtools
517 v.0.1.14[92]. Linkage among single-nucleotide polymorphisms (SNPs) spaced less than 250,000 bp
518 from each other was calculated with PLINK v.1.90b3b[93]. The strength of short- to mid-range
519 linkage for each SNP was then quantified as the sum of the distances (in bp) between that SNP
520 and all other SNPs with which it was found to be linked with $R^2 > 0.8$. We found this measure to
521 illustrate well the sharp decline of linkage at the boundaries of the four supergenes (Fig. 1).

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

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

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

561 **Estimating divergence times of Gadinae.** We estimated the divergence times among species
562 of the subfamily Gadinae with a phylogenomic approach on the basis of published age estimates for
563 two divergence events. The phylogenomic dataset used for these analyses comprised genome assem-
564 blies for eight Gadinae species released by Malmstrøm et al.[63], a genome assembly for the most
565 closely related outgroup *Brosme brosme*[63], the gadMor2 assembly for *Gadus morhua*, and sets of
566 unassembled Illumina reads for *Gadus macrocephalus* and *Gadus ogac*[54] (Supplementary Table
567 3). Aiming to identify sequences orthologous to 3,061 exon markers used in a recent phylogenomic
568 analysis of teleost relationships by Roth et al.[98], we first performed targeted assembly of these
569 markers from the sets of Illumina reads for *Gadus macrocephalus* and *Gadus ogac*. Targeted assem-
570 bly was conducted with Kollector v.1.0.1[99], using marker sequences of *Gadus morhua* from Roth et
571 al.[98] as queries. From the set of whole-genome and targeted assemblies, candidate orthologs to the
572 3,061 exon markers used by Roth et al.[98] were then identified through TBLASTN searches, using
573 sequences of *Danio rerio* as queries as in the earlier study. The identified sequences were aligned
574 with MAFFT and filtered to exclude potentially remaining paralogous sequences and misaligned
575 regions: We removed all sequences with TBLASTN bitscore values below $0.9 \times$ the highest bitscore

576 value and all sequences that had dN/dS values greater than 0.3 in comparison to the *Danio rerio*
577 queries, we removed codons from the alignment for which BMGE v.1.1[100] determined a gap rate
578 greater than 0.2 or an entropy-like score greater than 0.5, and we excluded exon alignments with a
579 length shorter than 150 bp, more than two missing sequences, or a GC-content standard deviation
580 greater than 0.04. We then grouped exon alignments by gene and excluded all genes that 1) were
581 represented by less than three exons, 2) had one or more completely missing sequences, 3) were
582 supported by a mean RAxML v.8.2.4[101] bootstrap value lower than 0.65, 4) were located within
583 the four supergene regions, 5) exhibited significant exon tree discordance according to an analysis
584 with Concatenator v.1.7.2[102], or 6) had a gene tree with non-clock-like evolution (mean estimate
585 for coefficient of variation greater than 0.5 or 95% highest-posterior-density interval including 1.0)
586 according to a relaxed-clock analysis with BEAST 2[66]. Finally, concatenated exon alignments per
587 gene were inspected by eye, and six genes were removed due to remaining possible misalignment.
588 The filtered dataset included alignments for 91 genes with a total alignment length of 106,566 bp
589 and a completeness of 92.8%.

590 We inferred the species tree of Gadinae with StarBEAST2[62, 66] under the multi-species coales-
591 cent model, assuming a strict clock, constant population sizes, and the birth-death tree model[103],
592 and averaging over substitution models with the bModelTest package[104] for BEAST 2. For time
593 calibration, we placed lognormal prior distributions on the age of the divergence of the outgroup
594 *Brosme brosme* from Gadinae (mean in real space: 32.325; standard deviation: 0.10) and on the
595 crown age of Gadinae (mean in real space: 18.1358; standard deviation: 0.28); in both cases, the
596 distribution parameters were chosen to approximate previous phylogenomic age estimates for these
597 two divergence events[105]. We performed five replicate StarBEAST2 analyses, each with a length
598 of one billion Markov-chain Monte Carlo (MCMC) iterations. After merging replicate posterior dis-
599 tributions, effective sample sizes (ESS) for all model parameters were greater than 1,000, indicating
600 full stationarity and convergence of MCMC chains. We then used TreeAnnotator from the BEAST
601 2 package to summarize the posterior tree distribution in the form of a maximum-clade-credibility
602 (MCC) consensus tree with Bayesian posterior probabilities as node support[106].

603 **Estimating divergence times and introgression among species of the genera *Gadus*,**
604 ***Arctogadus*, and *Boreogadus*.** To further investigate divergence times and introgression among
605 species of the closely related genera *Gadus*, *Arctogadus*, and *Boreogadus*, we used a second phyloge-
606 nomic dataset based on read mapping to the gadMor2 assembly. This dataset included Illumina
607 read data for all four species of the genus *Gadus* (*G. morhua*, *G. chalcogrammus*, *G. macrocephalus*,
608 and *G. ogac*)[54, 63], *Arctogadus glacialis*[63], and *Boreogadus saida*[63], as well as *Merlangius*
609 *merlangius*, *Melanogrammus aeglefinus*, and *Pollachius virens*[60, 63], which we here considered
610 outgroups (Supplementary Table 4). Read data from both a stationary and a migratory individual
611 (both sampled at the Lofoten islands) were used to represent *Gadus morhua*, to assess if one of the
612 two received weaker or stronger introgression from other taxa. Mapping, read sorting, and indexing
613 were again performed with BWA MEM and SAMtools, and variant calling was again performed
614 with GATK's HaplotypeCaller and GenotypeGVCFs tools as described above except that we now
615 also exported invariant sites to the output file. To limit the dataset to the most reliably mapping ge-
616 nomic regions, we applied the mask of unreliable sites generated from the threeway whole-genome

617 alignment (see above), resulting in set of 19,035,318 SNPs. We then extracted alignments from
618 GATK’s output files for each non-overlapping window of 5,000 bp for which no more than 4,000
619 sites were masked, setting all genotypes with a Phred-scaled likelihood below 20 to missing. Align-
620 ments were not extracted from the four supergene regions and those windows with less than 100
621 variable sites were ignored. As we did not model recombination within alignments in our phyloge-
622 nomic inference, the most suitable alignments for the inference were those with weak signals of
623 recombination. Therefore, we calculated the number of hemiplasies per alignment by comparing
624 the number of variable sites with the parsimony score, estimated with PAUP*[107], and excluded
625 all alignments that had more than ten hemiplasies. Finally, we again removed all alignment sites
626 for which BMGE determined a gap rate greater than 0.2 or an entropy-like score greater than 0.5.
627 The resulting filtered dataset was composed of 109 alignments with a total length of 383,727 bp
628 and a completeness of 91.0%.

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

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

649 To corroborate the introgression patterns inferred with Dsuite, we performed two analyses based
650 on comparisons of the frequencies of trio topologies in maximum-likelihood phylogenies. Alignments
651 for these analyses were selected as for the species-tree inference under the isolation-with-migration
652 model, except that up to 20 hemiplasies were allowed per alignment. The resulting set of 851
653 alignments had a total length of 3,052,697 bp and a completeness of 91.0%. From each of these
654 alignments, a maximum-likelihood phylogeny was inferred with IQ-TREE v.1.6.8[111] with a sub-
655 stitution model selected through IQ-TREE’s standard model selection. Branches with a length
656 below 0.001 were collapsed into polytomies. Based on the inferred maximum-likelihood trees, we

657 calculated, for all possible species trios, the D_{tree} -statistic of Ronco et al., a tree-based equivalent to
658 Patterson’s D -statistic in which the frequencies of pairs of sister taxa are counted in a set of trees
659 instead of the frequencies of shared sites in a genome (a related measure was proposed by Huson
660 et al.[112]): $D_{\text{tree}} = (f_{2\text{nd}} - f_{3\text{rd}})/(f_{2\text{nd}} + f_{3\text{rd}})$, where for a given trio, $f_{2\text{nd}}$ is the frequency of the
661 second-most frequent pair of sisters and $f_{3\text{rd}}$ is the frequency of the third-most frequent (thus, the
662 least frequent) pair of sisters. As a second tree-based analysis of introgression, we applied genealogy
663 interrogation[71], comparing the likelihoods of trees with alternative topological constraints for the
664 same alignment, as in Barth et al.[72]. We tested two hypotheses of introgression with this method:
665 1) Introgression between *Arctogadus glacialis* and either *Boreogadus saida* or the group of the four
666 species of the genus *Gadus*; and 2) introgression between *Gadus ogac* and the sister species *Gadus*
667 *chalcogrammus* and *Gadus morhua*.

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

690 We inferred the divergence times among *Gadus morhua* populations from the SNP data under
691 the multi-species coalescent model with the SNAPP add-on package for BEAST 2. Due to the
692 high computational demand of SNAPP, we performed this analysis only with a further reduced
693 set of 1,000 SNPs, randomly selected from all biallelic SNPs that were without missing genotypes
694 and located outside of the supergene regions. The input files for SNAPP were prepared with the
695 script `snapp_prep.rb`[74], implementing a strict-clock model and a pure-birth tree model. The tree
696 of *Gadus morhua* populations and outgroup species was time-calibrated with a single lognormal
697 prior distribution (mean in real space: 3.83; standard deviation: 0.093) that constrained the root

698 age of the tree according to the results of the analysis of divergence times and introgression among
699 *Gadus*, *Arctogadus*, and *Boreogadus* (see above; Fig. 2b, Supplementary Figure 2A). We performed
700 three replicate SNAPP analyses, each with a length of 400,000 MCMC iterations, resulting in ESS
701 values that were all greater than 400. The posterior tree distribution was again summarized as a
702 MCC consensus tree.

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

707 Population sizes over time were estimated for all sampled *Gadus morhua* populations with
708 Relate v.1.1.2[77]. To maximize the number of suitable SNPs for this analysis, we excluded all
709 outgroups except the sister species, *Gadus chalcogrammus*, and repeated variant calling and filtering
710 with the same settings as before. After applying a mask to exclude all variants from repetitive
711 regions in the gadMor2 assembly (784,488 bp in total)[35], 10,872,496 biallelic SNPs remained
712 and were phased with BEAGLE v.5.1[113], setting the population size assumed by BEAGLE to
713 10,000. We excluded all sites that were heterozygous in the *Gadus chalcogrammus* individual and
714 then reconstructed an “ancestral” genome sequence from the gadMor2 assembly and the called
715 variants for *G. chalcogrammus*. Following this reconstruction, we removed *G. chalcogrammus* from
716 the set of SNPs and excluded all sites that had become monomorphic after this removal, leaving
717 7,101,144 SNPs that were biallelic among the sampled *Gadus morhua* individuals. In addition to
718 the “ancestral” genome sequence and the set of biallelic SNPs, we prepared a mask for the Relate
719 analysis, covering all sites that were also included in the mask for repetitive regions, all sites that
720 would have been excluded from variant calling due to proximity to indels (see above), and all sites
721 that were ignored in the reconstruction of the “ancestral” sequence due to heterozygous genotype
722 calls for the *G. chalcogrammus* individual.

723 As Relate further requires an estimate of the mutation rate, we calculated this rate for the
724 filtered set of SNPs as the mean pairwise genetic distance between *Gadus morhua* individuals from
725 the Northwest Atlantic (thus, from the populations Newfoundland and Labrador) and those from the
726 Northeast Atlantic (thus, from all other populations), divided by 2 times the expected coalescence
727 time between the two groups and the genome size. We excluded the four linkage groups carrying
728 supergenes from this calculation. The expected coalescence time was calculated as the divergence
729 time between the two groups, which was estimated in the analysis with SNAPP as 65,400 years
730 (Fig. 3), plus the expected time to coalescence within the common ancestor, which is the product of
731 the generation time and the diploid population size under the assumption of a panmictic ancestral
732 population. With an assumed generation time of 10 years[114] and a population size of 57,400, as
733 estimated in the SNAPP analysis, the expected time to coalescence within the common ancestor
734 is 574,000 years, and the total expected coalescence time was thus set to $65,400 + 574,000 =$
735 $639,400$ years. As the mean pairwise genetic distance between the individuals of the two groups
736 was 878,704.31 and the size of the gadMor2 assembly without LGs 1, 2, 7, and 12, and excluding
737 masked sites, is 419,183,531 bp, the calculated mutation rate was $\mu = 878,704.31 / (2 \times 639,400 \times$

738 $419,183,531) = 1.64 \times 10^{-9}$ per bp and per year, or 1.64×10^{-8} per bp and per generation. Because
739 the genetic distance was calculated from the filtered set of SNPs, this rate is likely to underestimate
740 the true mutation rate of *Gadus morhua*; however, because the same filtered set of SNPs was used
741 as input for Relate, this rate is applicable in our inference of population sizes over time. The
742 input file was converted from variant call format to haplotype format using RelateFileFormats with
743 the flag “-mode ConvertFromVcf”. The script PrepareInputFiles.sh was used to flip genotypes
744 according to the reconstructed “ancestral” genome sequence and to adjust distances between SNPs
745 using the mask prepared for this analysis. Relate was first run to infer genome-wide genealogies and
746 mutations assuming the above calculated mutation rate of 1.64×10^{-8} per bp and per generation and
747 a diploid effective population size of 50,000. This was followed by an estimation of the population
748 sizes over time by running the script EstimatePopulationSize.sh for five iterations, applying the
749 same mutation rate and setting the threshold to remove uninformative trees to 0.5. The tools and
750 scripts RelateFileFormats, PrepareInputFiles.sh, and EstimatePopulationSize.sh are all distributed
751 with Relate.

752 **Estimating divergence times, demography, and gene flow specific to supergenes.** The
753 analyses of divergence times, demography, and gene flow among *Gadus morhua* populations were
754 repeated separately with SNPs from each of the four supergene regions on LGs 1, 2, 7, and 12. While
755 the SNAPP analyses for these regions were again performed with reduced subsets of 1,000 SNPs per
756 region, the data subsets used in analyses of gene flow with Dsuite comprised 11,474, 3,123, 10,4121,
757 and 10,339 biallelic SNPs, and those used in the analyses of demography with Relate comprised
758 211,057, 71,046, 130,918, and 130,620 biallelic SNPs, respectively. The mutation rate used as input
759 for these Relate analyses was identical to the one used for the analysis with genome-wide SNPs.

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

769 From the set of 20,402,423 biallelic SNPs, we extracted subsets of SNPs for each non-overlapping
770 window of a length of 250,000 bp, with a minimum distance between SNPs of 50 bp. We discarded
771 windows with less than 500 remaining biallelic SNPs and used a maximum of 1,000 biallelic SNPs
772 per window; these were selected at random if more biallelic SNPs were available per window. Input
773 files for SNAPP were then prepared as for the genome-wide and supergene-specific SNAPP analyses.
774 Per window, we performed two replicate SNAPP analyses with an initial length of 100,000 MCMC
775 iterations, and these analyses were resumed up to a maximum of 500,000 MCMC iterations as long
776 as the lowest ESS value was below 100. Windows with less than 300 sufficiently complete SNPs for
777 SNAPP analyses, with an ESS value below 100 after the maximum number of MCMC iterations,

778 or with a mean BPP node support value below 0.5 were discarded after the analysis. Per remaining
779 window, posterior tree distributions from the two replicate analyses were combined and summarized
780 in the form of MCC consensus trees. Additionally, a random sample of 100 trees was drawn from
781 each combined posterior distribution.

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

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

800 **Code availability**

801 Code for computational analyses is available from Github (<http://github.com/mmatschiner/super->
802 [genes](http://github.com/mmatschiner/super-genes)).

803 **Data availability**

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

808 Acknowledgements

809 We thank M. Malmstrøm, P. Berg, and D. Righton for help with fieldwork, and M. Skage, S.
810 Kollias, M. S. Hansen, and A. Tooming-Klunderud from the Norwegian Sequencing Centre (NSC;
811 <https://www.sequencing.uio.no>) for sequencing and processing of samples. PacBio and Illumina
812 library creation and high-throughput sequencing were carried out at NSC, University of Oslo, Nor-
813 way. All computational analyses were performed on the Abel and Saga supercomputing clusters
814 (Norwegian metacenter for High Performance Computing and the University of Oslo) operated
815 by the Research Computing Services group at USIT, the University of Oslo IT-department. We
816 also thank Alexandra Viertler for drawings of codfishes and Côme Denechaud for providing otolith
817 images.

818 Author contributions

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

825 References

- 826 1. Joron, M. *et al.* Chromosomal rearrangements maintain a polymorphic supergene controlling
827 butterfly mimicry. *Nature* **477**, 203–206 (2011).
- 828 2. Yan, Z. *et al.* Evolution of a supergene that regulates a trans-species social polymorphism.
829 *Nat. Ecol. Evol.* **4**, 210–249 (2020).
- 830 3. Lamichhaney, S. *et al.* Structural genomic changes underlie alternative reproductive strategies
831 in the ruff (*Philomachus pugnax*). *Nat. Genet.* **48**, 84–88 (2016).
- 832 4. Tuttle, E. M. *et al.* Divergence and functional degradation of a sex chromosome-like supergene.
833 *Curr. Biol.* **26**, 344–350 (2016).
- 834 5. Li, J. *et al.* Genetic architecture and evolution of the S locus supergene in *Primula vulgaris*.
835 *Nat. Plants* **2**, 16188 (2016).
- 836 6. Thompson, M. J. & Jiggins, C. D. Supergenes and their role in evolution. *Heredity* **113**, 1–8
837 (2014).
- 838 7. Schwander, T., Libbrecht, R. & Keller, L. Supergenes and complex phenotypes. *Curr. Biol.*
839 **24**, R288–R294 (2014).

- 840 8. Tigano, A. & Friesen, V. L. Genomics of local adaptation with gene flow. *Mol. Ecol.* **25**,
841 2144–2164 (2016).
- 842 9. Gutiérrez-Valencia, J., Hughes, W., Berdan, E. L. & Slotte, T. The genomic architecture and
843 evolutionary fates of supergenes. *arXiv*. arXiv:2012.11508 (2020).
- 844 10. Fisher, R. A. *The genetical theory of natural selection* (Clarendon Press, Oxford, UK, 1930).
- 845 11. Kirkpatrick, M. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–
846 434 (2006).
- 847 12. Jay, P. *et al.* Supergene evolution triggered by the introgression of a chromosomal inversion.
848 *Curr. Biol.* **28**, 1839–1845.e3 (2018).
- 849 13. Jay, P., Aubier, T. G. & Joron, M. Admixture can readily lead to the formation of supergenes.
850 *bioRxiv*. doi:10.1101/2020.11.19.389577 (2020).
- 851 14. Dobzhansky, T. & Epling, C. The suppression of crossing over in inversion heterozygotes of
852 *Drosophila pseudoobscura*. *Proc. Natl. Acad. Sci. U.S.A.* **34**, 137–141 (1948).
- 853 15. Sturtevant, A. H. & Beadle, G. W. The relations of inversions in the X chromosome of
854 *Drosophila melanogaster* to crossing over and disjunction. *Genetics* **21**, 554–604 (1936).
- 855 16. Anton, E., Blanco, J., Egozcue, J. & Vidal, F. Sperm studies in heterozygote inversion carriers:
856 a review. *Cytogenet. Genome Res.* **111**, 297–304 (2005).
- 857 17. Navarro, A., Barbadilla, A. & Ruiz, A. Effect of inversion polymorphism on the neutral
858 nucleotide variability of linked chromosomal regions in *Drosophila*. *Genetics* **155**, 685–698
859 (2000).
- 860 18. Faria, R., Johannesson, K., Butlin, R. K. & Westram, A. M. Evolving inversions. *Trends Ecol.*
861 *Evol.* **34**, 239–248 (2019).
- 862 19. Berdan, E. L., Blanckaert, A., Butlin, R. K. & Bank, C. Deleterious mutation accumulation
863 and the long-term fate of chromosomal inversions. *bioRxiv*. doi:10.1101/606012 (2020).
- 864 20. Bachtrog, D. A dynamic view of sex chromosome evolution. *Curr. Opin. Genet. Dev.* **16**,
865 578–585 (2006).
- 866 21. Blaser, O., Grossen, C., Neuenschwander, S. & Perrin, N. Sex-chromosome turnovers induced
867 by deleterious mutation load. *Evolution* **67**, 635–645 (2012).
- 868 22. Andolfatto, P., Depaulis, F. & Navarro, A. Inversion polymorphisms and nucleotide variability
869 in *Drosophila*. *Genet. Res.* **77**, 1–8 (2001).
- 870 23. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion activity in human
871 meiotic crossover hot spots. *Nat. Genet.* **36**, 151–156 (2004).
- 872 24. Williams, A. L. *et al.* Non-crossover gene conversions show strong GC bias and unexpected
873 clustering in humans. *eLIFE* **4**, e04637 (2015).
- 874 25. Chovnick, A. Gene conversion and transfer of genetic information within the inverted region
875 of inversion heterozygotes. *Genetics* **75**, 123–131 (1973).

- 876 26. Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C. & Patrinos, G. P. Gene conversion:
877 mechanisms, evolution and human disease. *Nat. Rev. Genet.* **8**, 762–775 (2007).
- 878 27. Korunes, K. L. & Noor, M. A. F. Pervasive gene conversion in chromosomal inversion het-
879 erozygotes. *Mol. Ecol.* **28**, 1302–1315 (2019).
- 880 28. Stevison, L. S., Hoehn, K. B. & Noor, M. A. F. Effects of inversions on within- and between-
881 species recombination and divergence. *Genome Biol. Evol.* **3**, 830–841 (2011).
- 882 29. Navarro, A., Betrán, E., Barbadilla, A. & Ruiz, A. Recombination and gene flux caused
883 by gene conversion and crossing over in inversion heterokaryotypes. *Genetics* **146**, 695–709
884 (1997).
- 885 30. Reis, M., Vieira, C. P., Lata, R., Posnien, N. & Vieira, J. Origin and consequences of chro-
886 mosomal inversions in the *virilis* group of *Drosophila*. *Genome Biol. Evol.* **10**, 3152–3166
887 (2018).
- 888 31. Bradbury, I. R. *et al.* Long distance linkage disequilibrium and limited hybridization suggest
889 cryptic speciation in Atlantic Cod. *PLOS ONE* **9**, e106380 (2014).
- 890 32. Berg, P. R. *et al.* Adaptation to low salinity promotes genomic divergence in Atlantic cod
891 (*Gadus morhua* L.) *Genome Biol. Evol.* **7**, 1644–1663 (2015).
- 892 33. Berg, P. R. *et al.* Three chromosomal rearrangements promote genomic divergence between
893 migratory and stationary ecotypes of Atlantic cod. *Sci. Rep.* **6**, 23246 (2016).
- 894 34. Sodeland, M. *et al.* “Islands of divergence” in the Atlantic cod genome represent polymorphic
895 chromosomal rearrangements. *Genome Biol. Evol.* **8**, 1012–1022 (2016).
- 896 35. Tørresen, O. K. *et al.* An improved genome assembly uncovers prolific tandem repeats in
897 Atlantic cod. *BMC Genomics* **18**, 95 (2017).
- 898 36. Kirubakaran, T. G. *et al.* Two adjacent inversions maintain genomic differentiation between
899 migratory and stationary ecotypes of Atlantic cod. *Mol. Ecol.* **25**, 2130–2143 (2016).
- 900 37. Barney, B. T., Munkholm, C., Walt, D. R. & Palumbi, S. R. Highly localized divergence within
901 supergenes in Atlantic cod (*Gadus morhua*) within the Gulf of Maine. *BMC Genomics* **18**,
902 271 (2017).
- 903 38. Berg, P. R. *et al.* Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated
904 with large inversions. *Heredity* **119**, 418–428 (2017).
- 905 39. Kess, T. *et al.* A migration-associated supergene reveals loss of biocomplexity in Atlantic cod.
906 *Sci. Adv.* **5**, eaav2461 (2019).
- 907 40. Barth, J. M. I. *et al.* Genome architecture enables local adaptation of Atlantic cod despite
908 high connectivity. *Mol. Ecol.* **26**, 4452–4466 (2017).
- 909 41. Barth, J. M. I. *et al.* Disentangling structural genomic and behavioural barriers in a sea of
910 connectivity. *Mol. Ecol.* **87**, 449 (2019).
- 911 42. Berg, E. & Albert, O. T. Cod in fjords and coastal waters of North Norway: distribution and
912 variation in length and maturity at age. *ICES J. Mar. Sci.* **60**, 787–797 (2003).

-
- 913 43. Case, R., Hutchinson, W. F., Hauser, L., Van Oosterhout, C. & Carvalho, G. R. Macro- and
914 micro-geographic variation in pantophysin (PanI) allele frequencies in NE Atlantic cod *Gadus*
915 *morhua*. *Mar. Ecol. Prog. Ser.* **301**, 267–278 (2005).
- 916 44. Star, B. *et al.* Ancient DNA reveals the Arctic origin of Viking Age cod from Haithabu,
917 Germany. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 9152–9157 (2017).
- 918 45. Hemmer-Hansen, J. *et al.* A genomic island linked to ecotype divergence in Atlantic cod. *Mol.*
919 *Ecol.* **22**, 2653–2667 (2013).
- 920 46. Sinclair-Waters, M. *et al.* Ancient chromosomal rearrangement associated with local adapta-
921 tion of a postglacially colonized population of Atlantic Cod in the northwest Atlantic. *Mol.*
922 *Ecol.* **27**, 339–351 (2017).
- 923 47. Kess, T. *et al.* Modular chromosome rearrangements reveal parallel and nonparallel adapta-
924 tion in a marine fish. *Ecol. Evol.* **10**, 638–653 (2020).
- 925 48. Kirubakaran, T. G. *et al.* A nanopore based chromosome-level assembly representing Atlantic
926 cod from the Celtic Sea. *G3*, g3.401423.2020 (2020).
- 927 49. Johansen, T. *et al.* Genomic analysis reveals neutral and adaptive patterns that challenge the
928 current management regime for East Atlantic cod *Gadus morhua* L. *Evol. Appl.* **13**, 2673–
929 2688 (2020).
- 930 50. Carr, S. M., Kivlichan, D. S., Pepin, P. & Crutcher, D. C. Molecular systematics of gadid
931 fishes: implications for the biogeographic origins of Pacific species. *Can. J. Zool.* **77**, 19–26
932 (1999).
- 933 51. Coulson, M. W., Marshall, H. D., Pepin, P. & Carr, S. M. Mitochondrial genomics of gadine
934 fishes: implications for taxonomy and biogeographic origins from whole-genome data sets.
935 *Genome* **49**, 1115–1130 (2006).
- 936 52. Bermingham, E., McCafferty, S. S. & Martin, A. P. in *Molecular Systematics of Fishes* (eds
937 Kocher, T. D. & Stepien, C. A.) 113–128 (Academic Press, San Diego, USA, 1997).
- 938 53. Owens, H. L. Evolution of codfishes (Teleostei: Gadinae) in geographical and ecological space:
939 evidence that physiological limits drove diversification of subarctic fishes. *J. Biogeogr.* **42**,
940 1091–1102 (2015).
- 941 54. Árnason, E. & Halldórsdóttir, K. Codweb: Whole-genome sequencing uncovers extensive retic-
942 ulations fueling adaptation among Atlantic, Arctic, and Pacific gadids. *Sci. Adv.* **5**, eaat8788
943 (2019).
- 944 55. Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*
945 **24**, 2818–2824 (2008).
- 946 56. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection
947 and genome assembly improvement. *PLOS ONE* **9**, e112963 (2014).
- 948 57. Sturtevant, A. H. A case of rearrangement of genes in *Drosophila*. *Proc. Natl. Acad. Sci.*
949 *U.S.A.* **7**, 235–237 (1921).

-
- 950 58. Kirkpatrick, M. How and why chromosome inversions evolve. *PLOS Biol.* **8**, e1000501 (2010).
- 951 59. Stransky, C. *et al.* Separation of Norwegian coastal cod and Northeast Arctic cod by outer
952 otolith shape analysis. *Fish. Res.* **90**, 26–35 (2008).
- 953 60. Tørresen, O. K. *et al.* Genomic architecture of haddock (*Melanogrammus aeglefinus*) shows
954 expansions of innate immune genes and short tandem repeats. *BMC Genomics* **19**, 240 (2018).
- 955 61. Edelman, N. B. *et al.* Genomic architecture and introgression shape a butterfly radiation.
956 *Science* **366**, 594–599 (2019).
- 957 62. Ogilvie, H. A., Bouckaert, R. R. & Drummond, A. J. StarBEAST2 brings faster species tree
958 inference and accurate estimates of substitution rates. *Mol. Biol. Evol.* **34**, 2101–2114 (2017).
- 959 63. Malmstrøm, M. *et al.* Evolution of the immune system influences speciation rates in teleost
960 fishes. *Nat. Genet.* **48**, 1204–1210 (2016).
- 961 64. Hughes, L. C. *et al.* Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on
962 transcriptomic and genomic data. *Proc. Natl. Acad. Sci. U.S.A.* **5**, 201719358 (2018).
- 963 65. Müller, N. F., Ogilvie, H. A., Zhang, C., Drummond, A. & Stadler, T. Inference of species
964 histories in the presence of gene flow. *bioRxiv*. doi:10.1101/348391 (2018).
- 965 66. Bouckaert, R. R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary
966 analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).
- 967 67. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
- 968 68. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between
969 closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
- 970 69. Malinsky, M., Matschiner, M. & Svardal, H. Dsuite - Fast D-statistics and related admixture
971 evidence from VCF files. *Mol. Ecol. Resour.* **19**, 1655 (2020).
- 972 70. Ronco, F. *et al.* Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature*
973 **589**, 76–81 (2021).
- 974 71. Arcila, D. *et al.* Genome-wide interrogation advances resolution of recalcitrant groups in the
975 tree of life. *Nat. Ecol. Evol.* **1**, 1–10 (2017).
- 976 72. Barth, J. M. I. *et al.* Stable species boundaries despite ten million years of hybridization in
977 tropical eels. *Nat. Commun.* **11**, 1433 (2020).
- 978 73. Bryant, D., Bouckaert, R. R., Felsenstein, J., Rosenberg, N. A. & RoyChoudhury, A. Inferring
979 species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent
980 analysis. *Mol. Biol. Evol.* **29**, 1917–1932 (2012).
- 981 74. Stange, M., Sánchez-Villagra, M. R., Salzburger, W. & Matschiner, M. Bayesian divergence-
982 time estimation with genome-wide SNP data of sea catfishes (Ariidae) supports Miocene
983 closure of the Panamanian Isthmus. *Syst. Biol.* **67**, 681–699 (2018).
- 984 75. Bradbury, I. R. *et al.* Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic
985 Ocean in response to temperature. *Proc. R. Soc. Lond. B* **277**, 3725–3734 (2010).

-
- 986 76. Ruegg, K., Anderson, E. C., Boone, J., Pouls, J. & Smith, T. B. A role for migration-linked
987 genes and genomic islands in divergence of a songbird. *Mol. Ecol.* **23**, 4757–4769 (2014).
- 988 77. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation
989 for thousands of samples. *Nat. Genet.* 1321–1329 (2019).
- 990 78. Runemark, A. *et al.* Variation and constraints in hybrid genome formation. *Nat. Ecol. Evol.*
991 **2**, 549–556 (2018).
- 992 79. Thorsen, A., Kjesbu, O. S., Fyhn, H. J. & Solemidal, P. Physiological mechanisms of buoyancy
993 in eggs from brackish water cod. *J. Fish Biol.* **48**, 457–477 (1996).
- 994 80. Finn, R. N. & Fyhn, H. J. Requirement for amino acids in ontogeny of fish. *Aquac. Res.* **41**,
995 684–716 (2010).
- 996 81. Johannesson, K., Smolarz, K., Grahn, M. & André, C. The future of Baltic Sea populations:
997 local extinction or evolutionary rescue? *AMBIO* **40**, 179–190 (2011).
- 998 82. Nissling, A., Kryvi, H. & Vallin, L. Variation in egg buoyancy of Baltic cod *Gadus morhua*
999 and its implications for egg survival in prevailing conditions in the Baltic Sea. *Mar. Ecol.*
1000 *Prog. Ser.* **110**, 67–74 (1994).
- 1001 83. Nissling, A. & Westin, L. Salinity requirements for successful spawning of Baltic and Belt
1002 Sea cod and the potential for cod stock interactions in the Baltic Sea. *Mar. Ecol. Prog. Ser.*
1003 **152**, 261–271 (1997).
- 1004 84. Hess, J. E. *et al.* Genomic islands of divergence infer a phenotypic landscape in Pacific lam-
1005 prey. *Molecular Ecology* **29**, 3841–3856 (2020).
- 1006 85. Schaeffer, S. W. & Anderson, W. W. Mechanisms of genetic exchange within the chromosomal
1007 inversions of *Drosophila pseudoobscura*. *Genetics* **171**, 1729–1739 (2005).
- 1008 86. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT
1009 sequencing data. *Nat. Methods* **10**, 563–569 (2013).
- 1010 87. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform.
1011 *Bioinformatics* **26**, 589–595 (2010).
- 1012 88. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
1013 2079 (2009).
- 1014 89. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and
1015 population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–
1016 2993 (2011).
- 1017 90. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing
1018 next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
- 1019 91. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples.
1020 *bioRxiv*. doi:10.1101/201178 (2018).
- 1021 92. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158
1022 (2011).

-
- 1023 93. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage
1024 analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- 1025 94. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
1026 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- 1027 95. Sandes, E. F. d. O., Miranda, G., Melo, A. C. M. A. d., Martorell, X. & Ayguade, E. CUDAlign
1028 3.0: Parallel Biological Sequence Comparison in Large GPU Clusters. *2014 14th IEEE/ACM
1029 International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, 160–169 (2014).
- 1030 96. Harris, R. S. *Improved pairwise alignment of genomic DNA*. PhD thesis (Pennsylvania State
1031 University, 2007).
- 1032 97. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: im-
1033 provements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
- 1034 98. Roth, O. *et al.* Evolution of male pregnancy associated with remodeling of canonical verte-
1035 brate immunity in seahorses and pipefishes. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9431–9439
1036 (2020).
- 1037 99. Kucuk, E. *et al.* Kollector: transcript-informed, targeted de novo assembly of gene loci. *Bioin-
1038 formatics* **33**, 1782–1788 (2017).
- 1039 100. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new
1040 software for selection of phylogenetic informative regions from multiple sequence alignments.
1041 *BMC Evol. Biol.* **10**, 210 (2010).
- 1042 101. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
1043 phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
- 1044 102. Leigh, J. W., Susko, E., Baumgartner, M. & Roger, A. J. Testing congruence in phylogenomic
1045 analysis. *Syst. Biol.* **57**, 104–115 (2008).
- 1046 103. Gernhard, T. The conditioned reconstructed process. *J. Theor. Biol.* **253**, 769–778 (2008).
- 1047 104. Bouckaert, R. R. & Drummond, A. J. bModelTest: Bayesian phylogenetic site model averaging
1048 and model comparison. *BMC Evol. Biol.* **17**, 42 (2017).
- 1049 105. Musilova, Z. *et al.* Vision using multiple distinct rod opsins in deep-sea fishes. *Science* **364**,
1050 588–592 (2019).
- 1051 106. Heled, J. & Bouckaert, R. R. Looking for trees in the forest: summary tree from posterior
1052 samples. *BMC Evol. Biol.* **13**, 221 (2013).
- 1053 107. Swofford, D. L. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Ver-
1054 sion 4. (2003).
- 1055 108. Yule, G. U. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis,
1056 F.R.S. *Phil. Trans. R. Soc. B* **213**, 21–87 (1925).
- 1057 109. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular
1058 clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).

-
- 1059 110. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable
1060 rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).
- 1061 111. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective
1062 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**,
1063 268–274 (2015).
- 1064 112. Huson, D. H., Klöpper, T., Lockhart, P. J. & Steel, M. A. in *Research in Computational*
1065 *Molecular Biology. RECOMB 2005. Lecture Notes in Computer Science* (eds Miyano, S. *et*
1066 *al.*) 233–249 (Springer, Berlin, Heidelberg, 2005).
- 1067 113. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data
1068 inference for whole-genome association studies by use of localized haplotype clustering. *Am.*
1069 *J. Hum. Genet.* **81**, 1084–1097 (2007).
- 1070 114. Smedbol, R. K., Shelton, P. A., Fréchet, A. & Chouinard, G. A. Review of population struc-
1071 ture, distribution and abundance of cod (*Gadus morhua*) in Atlantic Canada in a species-
1072 at-risk context. *Canadian Science Advisory Secretariat Research document 2002/082*, 1–134
1073 (2002).
- 1074 115. Alexa, A. & J, R. topGO: Enrichment Analysis for Gene Ontology. R package version 2.42.0.
1075 (2020).