








# Shared ancestral polymorphisms and chromosomal rearrangements as potential drivers of local adaptation in a marine fish

Hugo Cayuela<sup>1\*</sup>  | Quentin Rougemont<sup>1\*</sup>  | Martin Laporte<sup>1</sup>  | Claire Mérot<sup>1</sup>  |  
Eric Normandeau<sup>1</sup> | Yann Dorant<sup>1</sup>  | Ole K. Tørresen<sup>2</sup> | Siv Nam Khang Hoff<sup>2</sup>  |  
Sissel Jentoft<sup>2</sup>  | Pascal Sirois<sup>3</sup> | Martin Castonguay<sup>4</sup> | Teunis Jansen<sup>5,6</sup> |  
Kim Praebel<sup>7</sup> | Marie Clément<sup>8,9</sup> | Louis Bernatchez<sup>1</sup>

<sup>1</sup>Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada

<sup>2</sup>Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway

<sup>3</sup>Département des sciences fondamentales, Université du Québec à Chicoutimi, Chicoutimi, QC, Canada

<sup>4</sup>Fisheries and Oceans Canada, Institut Maurice-Lamontagne, Mont-Joli, QC, Canada

<sup>5</sup>GINR-Greenland Institute of Natural Resources, Nuuk, Greenland

<sup>6</sup>DTU Aqua-National Institute of Aquatic Resources, Technical University of Denmark, Charlottenlund Castle, Charlottenlund, Denmark

<sup>7</sup>Norwegian College of Fishery Science, Faculty of Biosciences, Fisheries and Economics, UiT The Arctic University of Norway, Tromsø, Norway

<sup>8</sup>Center for Fisheries Ecosystems Research, Fisheries and Marine Institute of Memorial, University of Newfoundland, St. John's, NL, Canada

<sup>9</sup>Labrador Institute of Memorial University of Newfoundland, Happy Valley-Goose Bay, NL, Canada

## Correspondence

Hugo Cayuela, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada.  
Email: hugo.cayuela51@gmail.com

## Abstract

Gene flow has tremendous importance for local adaptation, by influencing the fate of de novo mutations, maintaining standing genetic variation and driving adaptive introgression. Furthermore, structural variation as chromosomal rearrangements may facilitate adaptation despite high gene flow. However, our understanding of the evolutionary mechanisms impeding or favouring local adaptation in the presence of gene flow is still limited to a restricted number of study systems. In this study, we examined how demographic history, shared ancestral polymorphism, and gene flow among glacial lineages contribute to local adaptation to sea conditions in a marine fish, the capelin (*Mallotus villosus*). We first assembled a 490-Mbp draft genome of *M. villosus* to map our RAD sequence reads. Then, we used a large data set of genome-wide single nucleotide polymorphisms (25,904 filtered SNPs) genotyped in 1,310 individuals collected from 31 spawning sites in the northwest Atlantic. We reconstructed the history of divergence among three glacial lineages and showed that they probably diverged from 3.8 to 1.8 million years ago and experienced secondary contacts. Within each lineage, our analyses provided evidence for large  $N_e$  and high gene flow among spawning sites. Within the Northwest Atlantic lineage, we detected a polymorphic chromosomal rearrangement leading to the occurrence of three haplogroups. Genotype–environment associations revealed molecular signatures of local adaptation to environmental conditions prevailing at spawning sites. Our study also suggests that both shared polymorphisms among lineages, resulting from standing genetic variation or introgression, and chromosomal rearrangements may contribute to local adaptation in the presence of high gene flow.

## KEYWORDS

Øaði, fish, inversion, joint Site Frequency Spectrum, *Mallotus villosus*, population genomics, RAD, speciation

\*Cayuela and Rougemont contributed equally to this work.

## 1 | INTRODUCTION

Local adaptation plays a critical role in organisms' response to the spatiotemporal variation of their environment (Blanquart, Kaltz, Nuismer, & Gandon, 2013; Kawecki & Ebert, 2004; Savolainen, Lascoux, & Merilä, 2013). A population is considered to be locally adapted when individuals have a higher fitness in their home environment than their counterparts from other populations experiencing different environmental conditions (Savolainen et al., 2013). Local adaptation relies on three forms of genetic variation: de novo mutations, standing genetic variation, and adaptive introgression (Savolainen et al., 2013; Tigano & Friesen, 2016). Classic population genetics theory states that de novo mutations are the raw source of genetic adaptation (Barton, 1998; Kaplan, Hudson, & Langley, 1989; Smith & Haigh, 1974) for which empirical evidence has been previously reported (e.g., Linnen, Kingsley, Jensen, & Hoekstra, 2009). Nevertheless, increasing evidence suggests that standing genetic variation plays a central role in adaptation (Bitter, Kapsenberg, Gattuso, & Pfister, 2019; Haenel, Roesti, Moser, MacColl, & Berner, 2019; Jones et al., 2012; Lai et al., 2019) and might allow faster adaptation to a new environment than de novo mutation (Barrett & Schluter, 2008). Finally, adaptive introgression may also contribute to local adaptation when new mutants or variants from standing genetic variation with beneficial effects on fitness are introduced in the population through interbreeding with related taxa (Hedrick, 2013; Oziolor et al., 2019; Racimo, Sankararaman, Nielsen, & Huerta-Sánchez, 2015).

Gene flow, which is determined by organisms' dispersal capacity and land/seascape resistance, have tremendous effects on local adaptation (Lenormand, 2002; Tigano & Friesen, 2016). When selection is spatially heterogeneous but temporally constant, gene flow may erode local adaptation by swamping local alleles or impose a fitness cost to dispersers (Blanquart & Gandon, 2011; Lenormand, 2002). In the case of de novo mutations, simulation-based studies have found that the interplay between selection  $s$  and the rate of gene flow  $m$  is the strongest determinant of whether a beneficial de novo mutation establishes in a population. A de novo mutation is most likely to establish when  $s$  is greater than  $m$  (Lenormand, 2002; Wright, 1931; Yeaman & Otto, 2011), or otherwise it will be swamped from the local genetic pool. By contrast, directed gene flow may favour local adaptation when organisms' dispersal decisions (i.e., context-dependent dispersal; Clobert, Le Galliard, Cote, Meylan, & Massot, 2009) are adjusted according to local fitness prospects, leading to habitat matching choice (Edelaar, Siepielski, & Clobert, 2008), a process that has been reported in various taxa (e.g., Camacho & Hendry, 2020; Jacob et al., 2017; Lowe & Addis, 2019). Furthermore, gene flow can contribute to increase and/or maintain standing genetic variation on which selection can act, thus increasing the potential for local adaptation (Monnahan, Colicchio, & Kelly, 2015; Prezeworski, Coop, & Wall, 2005; Tigano & Friesen, 2016). In the case of polygenic adaptation in particular, high heritable genetic variation may provide a pool of potential combinations of small-effect alleles determining the expression of fitness-related traits (Csilléry, Rodríguez-Verdugo,

Rellstab, & Guillaume, 2018; Hancock, Alkorta-Aranburu, Witonsky, & Di Rienzo, 2010; Pritchard & Di Rienzo, 2010). Lastly, gene flow may also favour the introduction of beneficial variants from related species (i.e., adaptive introgression), the chance of fixation of introgressed alleles depending on the hybridization frequency and fitness of  $F_1$  hybrids and backcrosses (Hedrick, 2013).

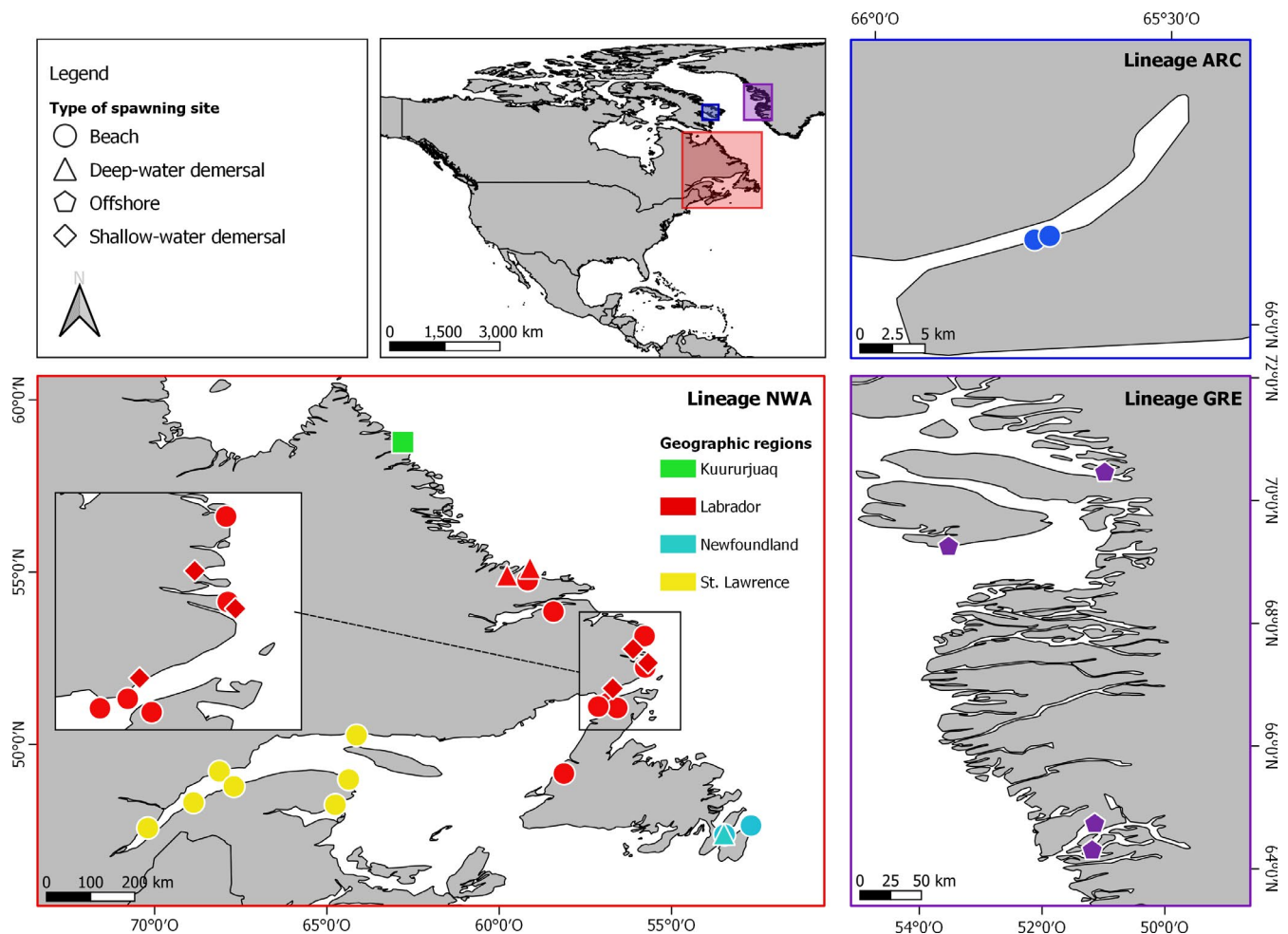
Three mechanisms allow maintenance of local adaptation despite the presence of gene flow: (a) linkage with an already diverged locus, (b) increased resistance to gene flow following secondary contact due to genetic incompatibilities, and (c) competition among genomic architectures including mechanisms that reduce or suppress recombination (Tigano & Friesen, 2016). In the last few years, molecular ecologists have paid increasing attention to the third mechanism and have shown that chromosomal rearrangements, which involve structural changes such as inversion, fusion, fission and translocation, may be a critical driver of local adaptation (Faria, Johannesson, Butlin, & Westram, 2019; Wellenreuther & Bernatchez, 2018; Wellenreuther, Mérot, Berdan, & Bernatchez, 2019). Low recombination within chromosomal rearrangements may lead to independent evolution of the affected genomic regions despite high gene flow in the rest of the genome (Faria & Navarro, 2010; Wellenreuther et al., 2019), which allows the expression of specialized phenotypes associated with local adaptation (Berg et al., 2017; Mérot et al., 2018; Wellband et al., 2019; Westram et al., 2018).

To date, empirical work focusing on the joint inference of local adaption and demographic history in the presence of high gene flow has been limited to a restricted number of study systems (e.g., in fishes: Barth et al., 2017; Le Moan, Bekkevold, & Hemmer-Hansen, 2019; Pettersson et al., 2019; Tine et al., 2014). Marine organisms are relevant biological models to address this issue because many species experience highly heterogeneous sea conditions potentially acting as selective agents, yet displaying very weak genetic differentiation due to large  $N_e$  and high connectivity associated with strong dispersal capacities (Bradbury, Laurel, Snelgrove, Bentzen, & Campana, 2008; Laporte et al., 2016; Palumbi, 1992; Selkoe et al., 2016; Xuereb, Kimber, Curtis, Bernatchez, & Fortin, 2018). Capelin (*Mallotus villosus*) is a key-forage fish species (Buren et al., 2014) that displays such characteristics typical of marine organisms. Using mitochondrial DNA (mtDNA), microsatellite and amplified fragment length polymorphism (AFLP) markers, previous studies have shown that capelin is characterized by high genetic diversity and weak local genetic structure, suggesting large  $N_e$  and high gene flow (Colbeck, Turgeon, Sirois, & Dodson, 2011; Præbel, Westgaard, Fevolden, & Christiansen, 2008). At a larger scale, three genetically divergent, parapatric glacial lineages have been reported in the North Atlantic and Arctic seas without apparent physical or geographical barriers separating them (Dodson, Tremblay, Colombani, Carscadden, & Lecomte, 2007). Within each lineage, capelin experiences spatio-temporally variable sea conditions that could be important selective drivers (Carscadden, Frank, & Leggett, 2001; Rose, 2005). Following the prenuptial migration, adults may reproduce at sites located in a continuum spanning from the intertidal (i.e., beach spawning sites) to the benthic zone from 1 to 280 m (i.e., demersal-spawning

sites) that differ strongly in terms of biotic and abiotic conditions (Nakashima & Wheeler, 2002). Environmental variation is especially high in the intertidal zone where the survival and/or development of embryos and larvae depends on temperature, salinity and trophic productivity (Frank & Leggett, 1981a, 1981b, 1982; Leggett, Frank, & Carscadden, 1984; Præbel, Christiansen, Kettunen-Præbel, & Fevolden, 2013; Purchase, 2018). The position along the gradient of water depth as well as the sea conditions prevailing in the intertidal zone are expected to exert strong selective pressures on local populations. This offers a suitable system to investigate connectivity and adaptation at different scales in a marine ecosystem, between and within lineages. Studies that have previously focused on inter- and intralocus genetic variation in the capelin were undertaken before the “genomic era,” which hampered the investigation of complex scenarios of divergence history and the search for molecular signals (including chromosomal rearrangements) associated with adaptation to sea conditions.

In this study, we examined how shared polymorphisms among the three aforementioned glacial lineages and chromosomal rearrangements underlie local adaptation to prevailing sea conditions

in capelin spawning sites. We used a large data set of genome-wide single nucleotide polymorphisms (SNPs) from 1,310 fish collected at 31 spawning sites throughout the Northwest Atlantic and Arctic waters. First, we confirmed the existence of the three glacial lineages (Northwest Atlantic lineage, NWA; Greenland lineage, GRE; and Arctic lineage, ARC) identified by Dodson et al. (2007) by analysing the pattern of genetic diversity and differentiation in the whole study area (Figure 1). Then, we inferred the demographic history of the three lineages using joint Site Frequency Spectrum (jSFS) for each pair of lineages. We took into account the confounding effect of barriers to gene flow, which affects the rate of migration (Barton & Bengtsson, 1986), and the confounding effect of linked selection, approximated as a reduction in local effective population size  $N_e$  (Charlesworth, Morgan, & Charlesworth, 1993). We then analysed in detail the genetic structure within the NWA lineage (Figure 1). In particular, we detected the molecular signature of putative chromosomal rearrangements (absent in the other glacial lineages). We next examined the molecular signature of local adaptation by identifying outlier loci putatively associated with the spawning site position along the water depth gradient (three



**FIGURE 1** Map of the study area. A total of 31 sampling locations were considered in the distribution range of three glacial lineages (NWA, ARC and GRE) of capelin (*Mallotus villosus*). In the NWA lineage distribution range, we sampled individuals according to three types of spawning sites: beach-spawning sites, shallow-water demersal sites, and deep-water demersal sites. In the ARC lineage, the two sampling sites are very close together (few kilometres) and cannot be distinguished on the map [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

categories of sites: beach spawning sites, demersal shallow-water sites and demersal deep-water sites) and environmental variables in beach spawning sites, including temperature and trophic production. We also evaluated the contribution of ancestral polymorphisms to local adaptation by examining an excess of shared polymorphisms among lineages at outlier loci associated with environmental factors. Finally, we explored the role of the chromosomal rearrangement on local adaptive variation by determining the proportion of candidate loci included in this genomic region.

## 2 | MATERIALS AND METHODS

### 2.1 | Genome sequencing and draft assembly

A first draft version of the capelin genome was constructed and used for the major purposes of mapping our RAD sequence reads. To estimate the chromosome-scale position of each SNP on the draft genome, we performed a synteny analysis with four other fish genomes. This allowed us to attribute each of the capelin contigs to one of the 24 orthologous chromosomes, that is to the chromosome with the longest total alignment across the four species. Capelin genome sequencing and assembly and synteny analysis are described in details in the Supporting Information methods section).

### 2.2 | Sampling area and genotyping-by-sequencing approach

#### 2.2.1 | Sampling area and molecular analyses

A total of 1,359 capelins were sampled from 31 sites in the Northwest Atlantic, both in Canadian and Greenland waters (Figure 1; Table S6), which were expected to include representatives from the three lineages according to a previous study using mtDNA and microsatellite genetic markers (Dodson et al., 2007). We sampled 25 spawning sites within the presumed NWA lineage including three demersal shallow-water sites, three demersal deep-water sites and 18 beach spawning sites. These sites were located in four geographical regions (Figure 1): Kuururjuaq, Labrador, Newfoundland and St. Lawrence. In parallel, we sampled two sites (beach-spawning fishes) within the presumed range of the ARC lineage and four offshore sampling sites in the range of the GRE lineage. The median sample size was 46.5 (range: 19–50) individuals per site. Fish were collected during the breeding period (captured by hand at beach spawning sites and using nets at demersal spawning sites) and sexed, and a piece of fin was preserved in RNA later.

Both DNA extractions and library preparation were performed following protocols fully described elsewhere (e.g., Moore et al., 2017; Rougemont et al., 2019). The methods are detailed in the Supporting Information methods section. A scheme presenting all analyses performed in our study can be found in Figure S13.

### 2.2.2 | DNA sequencing, genotyping and diversity analysis

Barcodes were removed using CUTADAPT (Martin, 2011) and trimmed to 80 bp, allowing for an error rate of 0.2. They were then demultiplexed using the “process\_radtags” module of STACKS version 1.44 (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013) and aligned to the capelin draft genome assembly using BWA-MEM (Li, 2013) with default parameters. Then, aligned reads were processed with STACKS version 1.44 for SNP calling and genotyping. The “pstacks” module was used with a minimum depth of three and up to three mismatches were allowed in the catalogue creation. We then ran the “populations” module to produce a vcf file that was further filtered using python ([https://github.com/enormandeau/stacks\\_workflow](https://github.com/enormandeau/stacks_workflow)) and bash scripts. SNPs were kept if they displayed a read depth greater than five and less than 70 (based on the mean distribution of read depth computed with VCFtools), the latter being set to control for paralogy and highly repetitive regions. Then, SNPs present in a least 70% in each sampling location and with heterozygosity lower than 0.60 were kept to further control for paralogues and Hardy–Weinberg disequilibrium (Dorant et al., 2019; Rougemont et al., 2019). The resulting vcf file comprised 550,724 SNPs spread over 27,442 loci and was further filtered prior to being used for demographic inferences with *daði* (see filtering details provided in the Supporting Information methods section). This vcf file was then subsampled to meet the different assumptions (in particular, the use of unlinked SNPs) of the models underlying the different population genetic analyses performed as described below. We therefore kept one single SNP per locus (the one with the highest minor allele frequency) and further removed SNPs with  $r^2$  values  $< .2$  using PLINK in sliding windows of 50 SNPs. Finally, we excluded any SNPs not present in at least 90% of the total data set, resulting in a final vcf of 25,904 high quality SNPs.

### 2.3 | Hierarchical patterns of population genetic structure

#### 2.3.1 | Genetic differentiation between glacial lineages

We documented the extent of genetic differentiation among lineages using the whole data set (i.e., the 31 sampling sites, Table 1). A principal component analysis (PCA) was performed to investigate the molecular variation among individuals using the R package *Adegenet* (Jombart, 2008). Then, we quantified ancestry and admixture proportion using the *snmf* function included in the R package LEA (Frichot & François, 2015), with K-values from 1 to 10. We estimated ancestry coefficients for all levels of K and obtained cross-validations using the cross-entropy criterion with 5% masked genotypes. Next, we quantified the level of genetic differentiation among sampling sites using the  $F_{ST}$  estimator (Weir & Cockerham, 1984) in the STAMPP R package along with 95% confidence intervals using 1,000 bootstraps (Pembleton, Cogan, & Forster, 2013).

**TABLE 1** Parameter estimates under the best demographic models for each of the three lineages

	NWA versus GRE	GRE versus ARC	NWA versus ARC
best	SC2N2m	SC2N2m	IM2N2m
$N_{\text{ref}}$	131,000 [11,000–250,000]	208,000 [72,760–343,000]	155,000 [102,000–208,000]
$N_{e1}$	828,000 [397,000–1,259,000]	592,000 [521,000–662,000]	1,329,000 [1,166,000–1,491,700]
$N_{e2}$	311,000 [162,000–460,000]	348,000 [239,590–456,000]	226,000 [211,000–241,000]
$m1$	0.000001 [0–0.000003]	0.0000003 [0–0.000001]	0.0000006 [0–0.000002]
$m2$	0.0000016 [0–0.000003]	0.0000002 [0–0.000003]	0.0000001 [0–0.0000003]
$m_{e1}$	0.000009 [0–0.00004]	0.000002 [0–0.00003]	0.0000006 [0–0.000004]
$m_{e2}$	0.00002 [0–0.00008]	0.000003 [0–0.00009]	0.000001 [0–0.000006]
$T_{\text{split}}$	1,785,000 [811,000–2,758,000]	2,700,000 [2,119,900–3,280,000]	3,760,000 [3,697,000–3,823,000]
$T_{\text{sc}}$	898,000 [0–1,807,000]	775,000 [732,000–818,000]	NA
$hrf$	0.10 [0.03–0.17]	0.24 [0.23–0.25]	0.12 [0.00–0.24]
$P$	0.53	0.57	0.01
$Q$	0.46	0.57	0.5

$N_{e1}$  and  $N_{e2}$ , effective population size of the compared pair,  $N_{\text{ref}}$  = effective population size of the ancestral (reference) population;  $m1 \leftarrow 2$  and  $m2 \leftarrow 1$ , migration from population 2 to population 1 and migration from population 1 to population 2.  $m_{e12}$  and  $m_{e21}$ , effective migration rate estimated in regions undergoing reduced introgression;  $T_{\text{split}}$ : time of split of the ancestral population in the two daughter species;  $T_{\text{sc}}$ , duration of secondary contact;  $P$ , proportion of the genome freely exchanged ( $1 - P$  provides the proportion of the genome non-neutrally exchanged);  $Q$ , proportion of the genome with a reduced  $N_e$  due to selection at linked sites ( $1 - Q$  provides the proportion of the genome neutrally exchanged);  $hrf$  = Hill–Robertson factor representing the reduction of  $N_e$  in the region ( $Q$ ) with reduced  $N_e$ . Values in brackets provide confidence intervals around the estimated parameters.

### 2.3.2 | Geographical and habitat-dependent genetic structure within lineages

A second PCA was executed for each of the three lineages separately to examine within-lineage genetic variation. Admixture analyses were conducted only on the NWA lineage to examine intralinear genetic structure. We did not analyse intralinear structure for ARC and GRE because of the limited number of sampling sites (two and four, respectively) and the absence of any genetic structure based on PCA results (Figure S4). Moreover, we quantified  $F_{ST}$  estimates among all sampling sites, and between demersal and beach spawning sites more specifically.

In addition, we tested for patterns of isolation-by-distance (IBD) within the NWA lineage separately for neutral SNPs and SNPs putatively under divergent selection using a procedure detailed in the Supporting Information methods section. IBD was tested using a linear model in which pairwise  $F_{ST}$  ( $F_{ST}/(1 - F_{ST})$ ) was included as the response variable and the Euclidean distance (z-score) was incorporated as an explanatory term. The significance of the effect of Euclidean distance was assessed using adjusted  $R^2$  and ANOVA with an  $F$ -test.

### 2.3.3 | Population structure as revealed by haplogroup distribution in the NWA lineage

Within the NWA lineage, we detected three genetic clusters using a PCA (hereafter called haplogroup1, haplogroup2 and

haplogroup3), which occurred in all sampling sites across the range of the NWA lineage but were absent in the other two lineages. This pattern was also supported by a clustering analysis performed for the NWA lineage alone (see Results). This signature may be due to a chromosomal rearrangement (i.e., three haplogroups with higher heterozygosity for the heterokaryote and lower heterozygosity for the two homokaryotes; Berg et al., 2017; Wellband et al., 2019). We thus performed the following analyses to confirm this hypothesis. We identified a set of SNPs mainly located in Chr2 and Chr9 associated with these haplogroups based on their loadings along the PC1 axis (Figures S4 and S5). We tested whether the heterozygosity of those SNPs differed between haplogroups using a beta regression model; heterozygosity was included in the model as the dependent variable and the haplogroup as discrete explanatory variable. Furthermore, for each sampling site, we tested for Hardy–Weinberg equilibrium when considering the haplogroups as different karyotypes of a putative rearrangement using a chi-square test and a ternary plot implemented in the R package HARDYWEINBERG (Graffelman, 2015).

Next, we examined if the assignment probability of individuals to the haplogroups depended on their sex. We built a logistic regression model for each haplogroup in which individual assignment to the haplogroup was coded as a binary response variable (0 = not assigned to the haplogroup; 1 = assigned) and sex as a discrete explanatory variable. Moreover, we investigated how the frequency of the haplogroup within sampling sites was affected by the spawning habitat (deep-water and shallow-water demersal



sites were merged into a single category to increase the power of the analysis). In addition, we examined how temperature and chlorophyll concentration in beach spawning sites influenced haplogroup frequency. The marine data layers were downloaded from Bio-ORACLE (<http://www.bio-oracle.org/>) and the two environmental variables were extracted using the R package *SDMPREDICTORS* (Bosch, Tyberghein, & De Clerck, 2017); we present the environmental data in Table S5. We used a linear model where the frequency of haplogroups 1 and 2 [ $(2 \times N_{\text{haplogroup1}} + N_{\text{haplogroup2}} / N_{\text{total}})$ ] was included as the response variable and spawning habitat, temperature and chlorophyll concentration as explanatory variables. Each explanatory variable was introduced separately in the model to avoid model over-fitting. The significance of explanatory variable effects was evaluated using adjusted  $R^2$  and ANOVA with an  $F$ -test.

## 2.4 | Demographic history of divergence in $\delta a\delta i$ and identification of interlineage outliers

### 2.4.1 | Divergence history of the three capelin lineages

We reconstructed the most likely demographic history using  $\delta a\delta i$  (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009). We compared four alternative models of historical divergence: (a) a model of Strict Isolation (SI), (b) a model of divergence with continuous gene flow or Isolation with Migration (IM), (c) a model of divergence with initial migration or Ancient Migration (AM) and (d) a model of Secondary Contact (SC) (fully described in Figure S3). These models incorporate the confounding effects of selection at linked sites locally affecting  $N_e$  and the effect of differential introgression (i.e., due to barriers to gene flow accumulated during population divergence; Barton & Bengtsson, 1986; Roux et al., 2014) affecting the rate of migration ( $m$ ) along the genome (Le Moan, Gagnaire, & Bonhomme, 2016; Rougemont et al., 2017; Tine et al., 2014). Model choice was performed using the Akaike information criterion (AIC),  $\Delta AIC$  and AIC weights. We attempted to estimate biological parameters assuming a generation time of 3.8 years (Dodson et al., 2007) and a standard mutation rate value of  $1e-8$  mutation/bp/generation. Furthermore, we performed additional analyses of the possible confounding effect of the two chromosomal blocks (on separate Chr2 and Chr9). The modelling approach is detailed in the Supporting Information methods section. Finally, to obtain an overview of the overall levels of divergence among the lineage we computed the net sequence divergence ( $D_a$ ) as well as level of absolute sequence divergence ( $D_{xy}$ ), the former being informative of the position of a pair of populations along the speciation continuum (Roux et al., 2016). We used the *vcf* haplotype file obtained from *STACKS* and computed these summary statistics using *MSCALC* (Roux et al., 2011) to account for missing data and integrate the length of each RAD locus.

### 2.4.2 | Coalescent simulations and identification of interlineage outliers

We used the demographic parameter estimates along with their standard deviation under each best model for each pair of lineages to perform 4,000,000 coalescent simulations using the coalescent simulator *ms* (Hudson, 2002). We assumed a strictly neutral model with homogeneous population size and homogeneous migration rate. For each SNP, we computed levels of expected heterozygosity ( $H_E$ ) as well as Weir & Cockerham's  $F_{ST}$  estimator between lineages. As proposed by Beaumont and Nichols (1996) we computed  $F_{ST}$  quantiles in heterozygosity intervals of 0.025. Unlike several methods that rely on simplistic demographic models, this approach use an explicitly fitted model and considers the uncertainty surrounding parameter estimates (Leroy et al., 2019). Loci departing from our neutral envelope (outside the 99.99th upper quantiles of the conditioned  $F_{ST}$  distribution) were inferred as candidate outliers. Finally, we used the sequences of the markers potentially under divergent selection between the three lineages to perform a BLAST search (Altschul, Gish, Miller, Myers, & Lipman, 1990). More specifically, we extracted a 1-kb-long sequence on each side of the SNPs of interest. Then, we performed blasts of those sequences against the NCBI NR database. To consider a SNP further, we required a BLAST length of at least 80 bp and an e-value of  $1e-10$ . Next, to identify nonsynonymous mutations, we blasted the *FASTA* sequences against the reference transcriptome. We kept hits with at least 90% similarity and a minimum amino acid sequence length alignment of 25. Both variants of each SNP were used with the BLAST results against the transcriptome and translation results were compared pairwise. For each locus, results were only kept if the lowest e-value hit for both variants was the same (i.e., same protein name and length).

## 2.5 | Searching for signals of local adaptation patterns within the NWA lineage

### 2.5.1 | Genotype–environment associations and local adaptation to spawning habitat

We examined the molecular signature of local adaptation to spawning sites within the NWA lineage only, because not enough sampling sites were available within the other two lineages. More specifically, we aimed to detect outlier loci associated with spawning habitats distributed along a water depth gradient: the beach spawning site located in the intertidal zone, the demersal spawning site in shallow water (from 2 to 5 m), and the demersal spawning site in deeper water (from 10 to 20 m). As recommended by Forester, Lasky, Wagner, and Urban (2018), we used a combination of latent factor mixed models (LFMMs; Frichot & François, 2015; Frichot, Schoville, Bouchard, & François, 2013) and partial redundancy analysis (pRDA; “rda” function in the R package *VEGAN*; Lasky et al., 2012; Forester et al., 2018) to detect candidate loci.

We ran LFMMs in the R package *LEA*. Models were parameterized using the  $K$ -value ( $K = 2$ ) obtained in our previous clustering analyses in *LEA* (Figure 3c). By doing so, we accounted for the genetic variation associated with the haplogroups in our analyses of genotype–environment associations. We ran each model 10 times with 5,000 iterations and a burnin of 2,500. The median of the squared  $z$ -scores was used to rank loci and to calculate a genomic inflation factor to evaluate model fit (François, Martins, Caye, 2016).

We used a pRDA to detect multilocus outliers that were associated with spawning habitats after controlling for the haplogroups (see Forester et al., 2018; Legendre & Legendre, 2012 for details). The standard deviation of marker scores was multiplied by 3.5 to establish the multilocus outlier detection  $p$ -value at .001 (Forester et al., 2018). We retained the outliers that were detected by both LFMMs and pRDA to reduce the number of false positives as recommended by Forester et al. (2018) and represented the overlap with a Venn diagram. For each outlier locus associated with spawning habitats, we performed enrichment tests as detailed in the Supporting Information methods section.

### 2.5.2 | Genotype–environment associations and local adaptation to environmental conditions in beach spawning sites

We investigated the molecular signature of adaptation to local environmental conditions across beach spawning sites within the NWA lineage (i.e., 19 sampling sites). We considered two environmental variables: sea surface temperature and chlorophyll concentration as a proxy for trophic productivity, which may have dramatic effects on embryonic and larval performances (Frank & Leggett, 1981a, 1981b, 1982; Leggett et al., 1984). The correlation between temperature and chlorophyll concentration ( $r = .47$ ) was relatively low (for a discussion on multicollinearity in ecological studies, see Dormann et al., 2013). For this reason, these two variables were simultaneously introduced in the pRDAs presented below. We determined candidate loci associated with the two variables using a combination of LFMM and pRDA. In LFMM, the variables were introduced one by one in the models to identify the candidate loci associated with each variable. We then used the same procedure described for spawning site analyses for *go\_enrichment* and *blasting*.

## 3 | RESULTS

### 3.1 | Draft genome assembly and statistics

The resulting draft genome assembly was 490 Mbp with a N50 contig size of 230 kbp. Running *busco* with the *Actinopterygii* data set on the assembly found 4,070 complete (89%), 209 fragmented (4.5%) and 305 missing genes (6.5%) out of a total of 4,584 genes searched. The genome assembly is available at <https://doi.org/10.6084/m9.figshare.9752558>.

#### 3.1.1 | Genetic differentiation between glacial lineages

The raw *vcf* file from *STACKS* contained 1,081,533 SNPs spread on 67,305 loci before filtration. After filtration, we kept a final set of 25,904 SNPs spread on 25,904 loci (see details in Methods). The mapping rate of our GBS data on the draft genome was 74%. The median number of reads per sample was 1,033,000 and the median depth was 21.

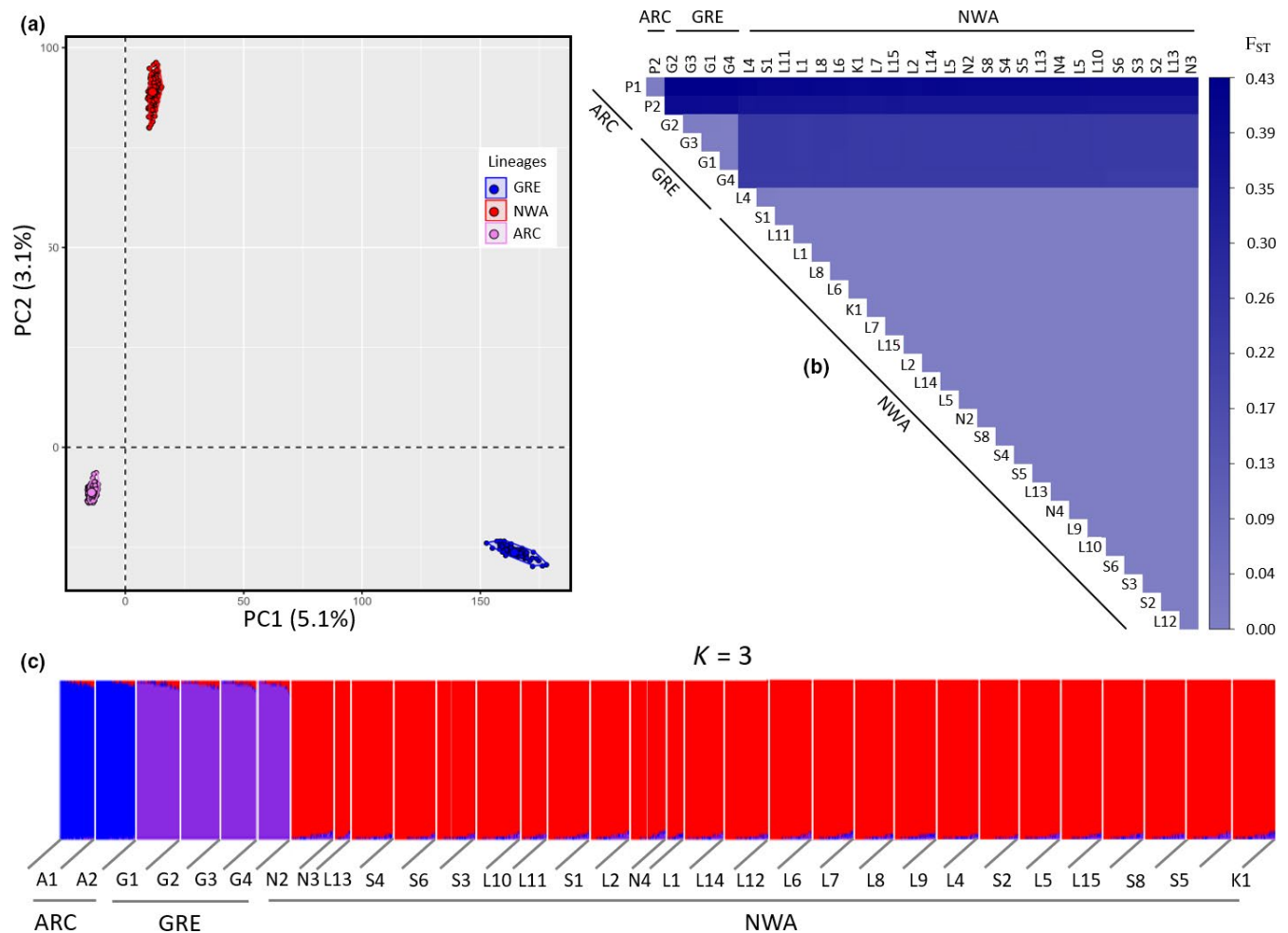
A high and heterogeneous level of differentiation was found along the genome for the comparison between lineages (Figure S1). The first and second axes of the PCA explained 8.2% of the genetic variation (Figure 2a) and they supported the pattern found with  $F_{ST}$  (see below). The ARC lineage was the most divergent among all three lineages. The clustering analysis also supported the existence of three lineages ( $K = 3$ ) with an absence of admixture between lineages (Figure 2c).

Measures of net sequence divergence ( $d_a$ ) which, unlike  $F_{ST}$ , are not affected by variation in the common ancestor, revealed a level of net divergence that was about twice as high between the ARC and the GRE lineages ( $d_a = 0.00146$ ) and between the ARC and NWA lineages ( $d_a = 0.00140$ ) compared to that between the GRE and NWA clades ( $d_a = 0.00067$ ). The same was true for  $D_{xy}$  values with values of 0.0032, 0.0041 and 0.0041 between NWA and GRE, ARC and NWA, and ARC and GRE, respectively. Moreover, the mean  $F_{ST}$  was 0.2435 between NWA and GRE, 0.3783 between NWA and ARC, and 0.4106 between GRE and ARC (Figure 2c), thus revealing a high level of genetic differentiation for a marine species (DeWoody & Avise, 2000).

#### 3.1.2 | Geographical and habitat-dependent genetic structure within lineages

At the intralocus level, pairwise  $F_{ST}$  values between sampling sites were three orders of magnitude lower than observed between the three lineages (Figure 2b). Mean  $F_{ST}$  was 0.0033 within the GRE lineage (ranging from 0.0021 to 0.0054) and 0.0044 in the ARC lineage (only two sampling sites in this lineage). Within the NWA lineage, mean  $F_{ST}$  was 0.0010 (ranging from 0 to 0.0090); 91% of the  $F_{ST}$  estimates had a  $p$ -value <0.01 and a 95% confidence interval (CI) that did not include 0 (Table S13). Within the NWA lineage, mean  $F_{ST}$  among shallow-water spawning sites and among deep-water demersal sites was 0.0019 and 0.0017, respectively, whereas it was 0.0029 among beach spawning sites. Mean  $F_{ST}$  between shallow-water demersal sites and beach sites and between deep-water demersal sites and beach sites was 0.0031 and 0.0027, respectively.

We then tested for IBD within the NWA lineage, using neutral loci and loci under selection (detailed methods are given in the Supporting Information methods section, and Figure S12). Combining *BAYESCAN* and *PCADAPT* revealed a total of 12,600 putatively neutral SNPs while each method revealed respectively 20 and 42 SNPs putatively under divergent selection (additional results are



**FIGURE 2** Genetic structure and differentiation among the three lineages (NWA, GRE and ARC) of capelin (*Mallotus villosus*). (a) Principal component analysis (axes 1 and 2) showing the genetic variation among the lineages. (b) Heatmap of  $F_{ST}$  between and within three lineages. (c) Clustering analysis performed using the LEA R package [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

given in the Supporting Information results section). We detected a weak signal of IBD using the two sets of markers (with SNPs under divergent selection:  $F = 5.38$ ,  $R^2 = 0.01$ ,  $p = .02$ ; with neutral SNPs:  $F = 3.89$ ,  $R^2 = 0.01$ ,  $p = .05$ ) (Figure S12).

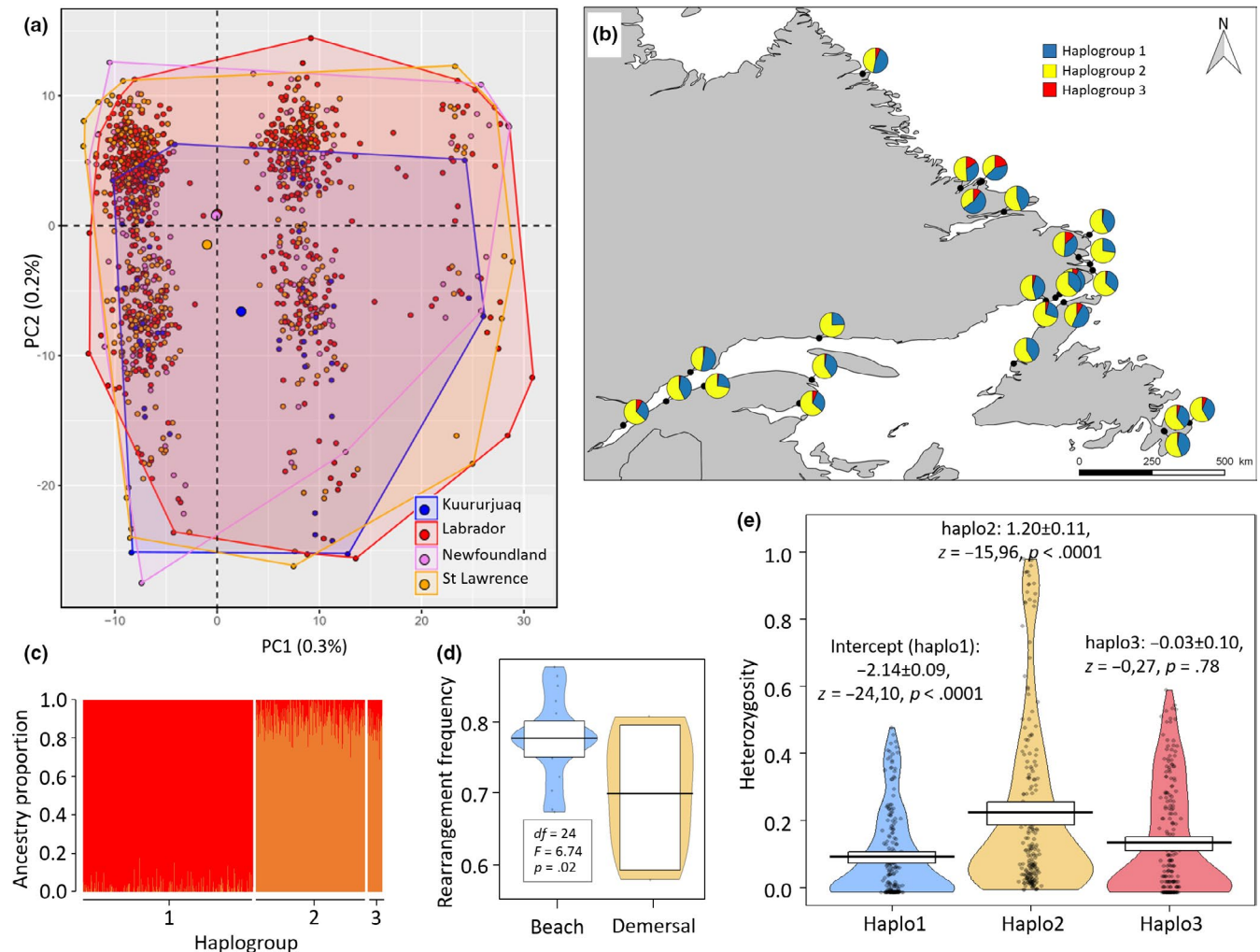
### 3.1.3 | A putative chromosomal rearrangement associated with population within the NWA lineage

The PCA performed for the GRE and ARC lineages did not resolve any genetic structure (Figure S4). By contrast, the PCA revealed pronounced genetic structure within the NWA lineage (Figure 3a), with three genetic clusters along PC1 (explaining 0.3% of the variation). The distribution of the clusters did not show any geographical pattern (Figure 3b, but see below). A set of 248 SNPs displayed the highest loading along the first axis. Mapping of these 248 SNPs revealed that 85% of them were preferentially located on ancestral chromosomes Chr2 ( $n = 74$  SNPs) and Chr9 ( $n = 139$  SNPs), whereas an average of two SNPs was mapped on the remaining chromosomes. The PCA performed for Chr2 and Chr9 separately showed

similar clustering patterns (Figures S10 and S11). Altogether, these observations raised the hypothesis that such genetic substructure may be due to a chromosomal inversion in Chr2 and Chr9 that could be fused in the capelin; alternatively, it could be two covarying chromosomal inversions or a polymorphic fusion. We thus considered the three clusters as three haplogroups. Based on the PCA, 57% of the individuals were assigned to haplogroup1, 38% to haplogroup2 and 5% to haplogroup3.

SNPs with the highest levels of differentiation between haplogroups were also predominantly located on Chr2 and Chr9 (Figure 4). The highest level of genetic differentiation was observed between haplogroups 1 and 3, with a genome-wide  $F_{ST}$  of 0.0047 but with an  $F_{ST}$  50 times higher (0.29) when we only considered the SNPs located in the Chr2–Chr9 rearrangement (Figure S9). This was higher than the differentiation observed between each haplogroup and the two other lineages for the same genomic region ( $F_{ST}$  haplogroup1 and GRE = 0.055;  $F_{ST}$  haplogroup1 and ARC = 0.012;  $F_{ST}$  haplogroup3 versus GRE = 0.072;  $F_{ST}$  haplogroup3 versus ARC = 0.148). The beta regression model also indicated that the heterozygosity of the SNPs from Chr2 and Chr9 differed among



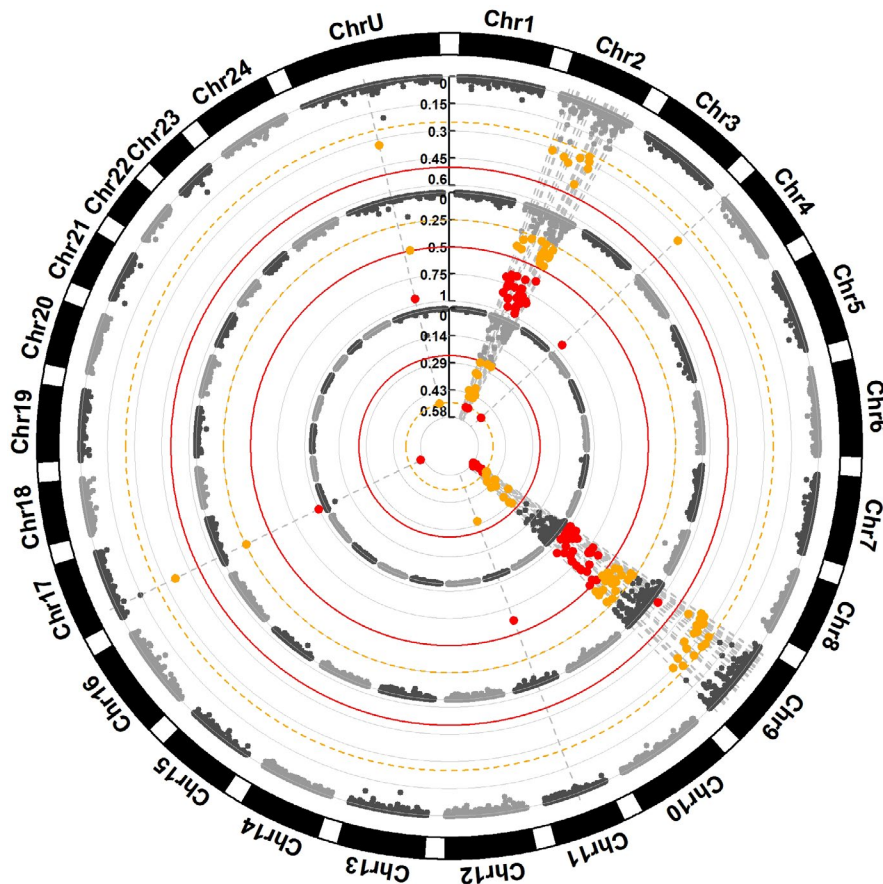


**FIGURE 3** Spatial variation of the putative chromosomal rearrangement (involving chromosomes 2 and 9) within the NWA lineage of capelin (*Mallotus villosus*). (a) Principal component analysis (axes 1 and 2) showing the three clusters of individuals corresponding to the three haplogroups. (b) Relative proportion of the three haplogroups among the 25 sampled spawning sites throughout the distribution area of the NWA lineage. (c) Ancestry proportion ordered by haplogroups from clustering analyses. (d) Frequency of the rearrangement in the demersal (deep-water sites + shallow-water sites) and beach spawning sites; we provide the results of an ANOVA testing the effect of the type of spawning site on the rearrangement frequency. (e) Heterozygosity analysis focused on the contigs contained in the potential chromosomal rearrangement in chromosomes Chr2 and Chr9. We provide the results of the beta regression model: intercept for haplogroup1, and the slope coefficients for haplogroup2 and haplogroup3, and their associated standard error and statistical test ( $z$  statistic and  $p$ -value) [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

haplogroups (Figure 3e; Table S2): it was highest in haplogroup2 and similar in haplogroup1 and haplogroup3. The clustering approach also supported the existence of these haplogroups. We detected two genetic groups ( $K = 2$ ) across all sampling sites. All individuals of haplogroup1 are contained in one genetic group whereas all the individuals of haplogroup2 and haplogroup3 are admixed with the other genetic group (Figure 3c). Moreover, when considering haplogroup1 and haplogroup3 as the homokaryotes and haplogroup2 as the heterokaryote for the putative rearrangement, there was no deviation from Hardy-Weinberg equilibrium at any sampling site (Table S12), suggesting Mendelian inheritance of markers found in this rearrangement. Altogether, these observations suggest that haplogroup2 may represent a group of heterozygotes bearing two,

perhaps partially nonrecombining, versions of haploblocks that exist in the homozygote state in haplogroups 1 and 3.

Haplogroups 1 and 2 were present in 100% of the sampling sites, while haplogroup3 occurred in only 78% of the sites. The frequency of haplogroups varied among sampling sites (Figure 3b): the proportion of individuals in haplogroup1 ranged from 35% to 73%, from 24% to 52% for haplogroup2, and from 0% to 21% for haplogroup3. Logistic models indicated that the probability that an individual was assigned to haplogroup1 or haplogroup2 (encompassing 95% of the individuals) was not influenced by its sex (haplogroup1:  $F = 0.33$ ,  $p = .33$ ; haplogroup2:  $F = 0.69$ ,  $p = .41$ ). The probability that an individual was assigned to haplogroup3 was sex-dependent ( $\beta_{\text{males}} = -0.97 \pm 0.34$ ,  $F = 8.73$ ,  $p = .003$ ), although



**FIGURE 4** Manhattan plots of the  $F_{ST}$  values between the three haplogroups within the NWA lineage of capelin (*Mallotus villosus*). The external circle shows the  $F_{ST}$  between haplogroups 1 and 2. The medium circle shows the  $F_{ST}$  between haplogroups 1 and 3. The central circle shows the  $F_{ST}$  between haplogroups 2 and 3. Capelin contigs were placed into 24 ancestral chromosomes based on synteny with four related species. Contigs that were not assigned to a chromosome were aggregated in the group ChrU. As the order of the contigs within the chromosomes was not conserved and was not always consistent among genome comparisons, contigs were ranked according to their size. SNPs having  $F_{ST} < 0.25$  are shown in grey, SNPs with  $F_{ST}$  ranging from 0.25 to 0.50 are shown in yellow, and SNPs with  $F_{ST}$  higher than 0.50 are shown in red [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

this was probably due to the small number of individuals assigned to haplogroup3. Furthermore, linear models indicated that the frequency of the putative rearrangement differed between spawning habitats (Figure 3d). The frequency of haplogroups 1 and 2 combined was higher in beach spawning than in demersal sites (both deep-water and shallow-water sites combined) ( $df = 24$ ,  $F = 6.74$ ,  $p = .02$ ). However, the rearrangement frequency was not influenced by temperature ( $df = 18$ ,  $F = 0.67$ ,  $p = .42$ ) or chlorophyll concentration ( $df = 18$ ,  $F = 0.08$ ,  $p = .83$ ). Additional PCA, performed for the three haplogroups separately, suggested the absence of underlying genetic structure in the NWA lineage (Figure S6).

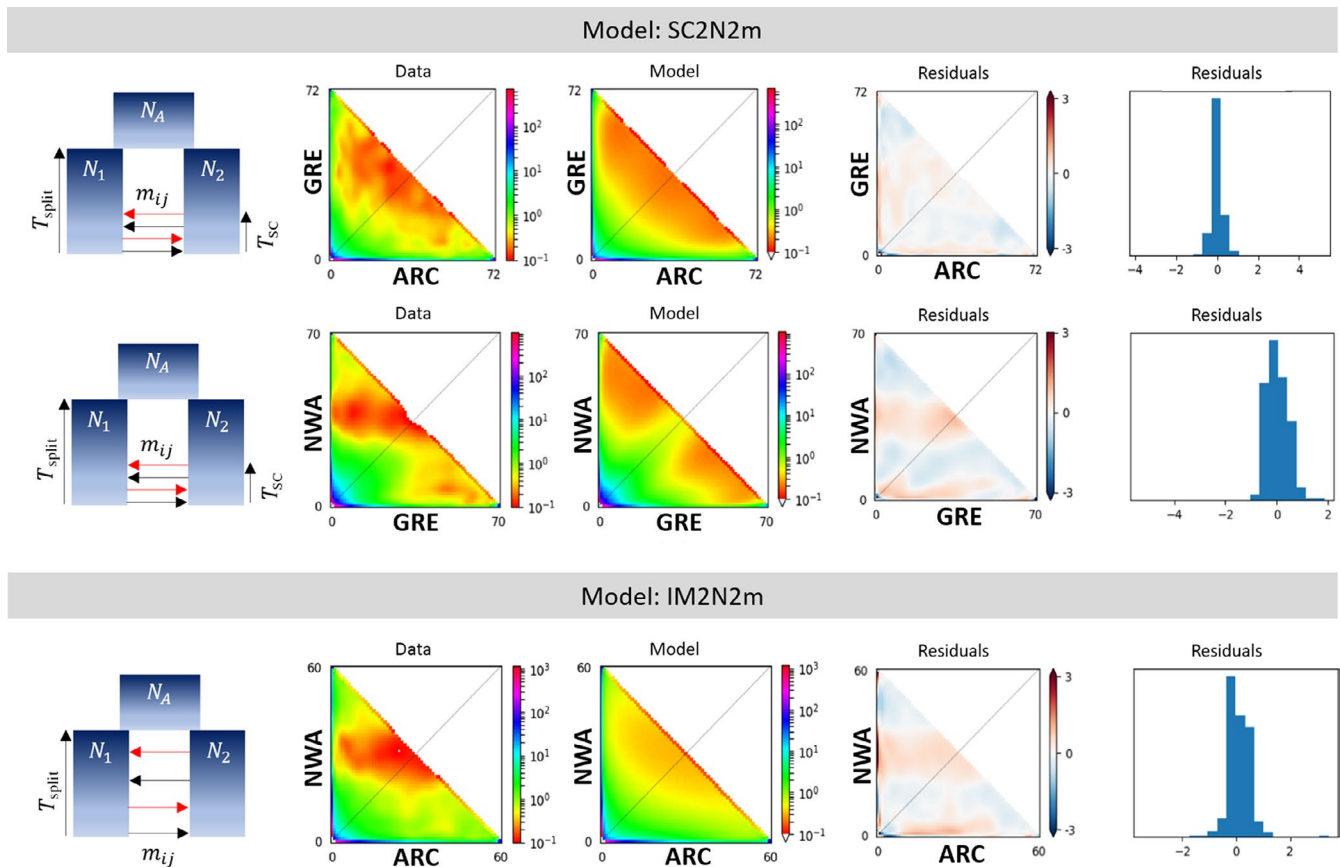
### 3.2 | Demographic history of divergence

#### 3.2.1 | History of divergence among glacial lineages

The divergence history inferred with  $\partial a d i$  revealed the occurrence of ongoing gene flow between all three lineages. In two interlineage comparisons, GRE versus NWA lineages and GRE versus ARC lineages, the models of secondary contact including both linked selection and variable introgression rate among loci ("SC2N2m") received the highest statistical support ( $wAIC = 1$ ; Table S3; Figure 5). In the comparison of the NWA versus ARC lineages, the model of isolation with migration including both linked selection and variable introgression

rate among loci ("IM2N2m") received the highest statistical support ( $wAIC = 1$ ; Table S3). These results were robust to the confounding effect of the putative chromosomal rearrangement defined above (Supporting Information methods section, Table S3 and Figure S8).

Parameter estimates revealed that the NWA and ARC lineages diverged  $\sim 3.8$  million years ago (Ma; 95% CI: 3.6–3.8 Ma). The GRE lineage was inferred to have diverged from the ARC lineage  $\sim 2.7$  Ma (95% CI: 2.1–3.2 Ma). The NWA versus GRE comparison indicated a divergence time of  $\sim 1.8$  Ma (95% CI: 0.8–2.8 Ma). These split times were older when we removed Chr2 and Chr9 and when we separated by haplogroup, with respective averaged values of 4.3 Ma for the NWA versus ARC split and 2.5 Ma for the NWA versus GRE split (Table S4). This suggests a more recent time to the most recent common ancestor (TMRCA) for Chr2 and Chr9. Gene flow was asymmetric and heterogeneous. Indeed, under SC, on average half of the genome displayed a reduced introgression rate, while our inference under IM2N2m indicated that 99% of the genome displayed reduced introgression. The rate of introgression under IM2N2m between the NWA and ARC lineages appears to be an order of magnitude lower than in other comparisons (Table S4). Finally, in all comparisons, our inference suggested that approximately half of the genome might be affected by selection at linked sites resulting in a reduced effective population size at such sites (Table 1). Our estimates of effective population sizes were high, varying from 220,000 ( $SD = 7,600$ ) in the ARC lineage under IM2N2m to more than 1,330,000



**FIGURE 5** Demographic divergence among lineages of capelin (*Mallotus villosus*): simplified representation of the best demographic model (left panel), observed and best fitted jSFS obtained from  $\partial a \partial i$  to unravel the demographic histories of the three lineages (middle panels) and residuals of the fitted models (right panels). The best model was the SC2N2m (model of secondary contact including both linked selection and variable introgression rate among loci) in the comparison between ARC versus GRE and GRE versus NWA, whereas the IM2N2m model (model of isolation with migration including both linked selection and variable introgression rate among loci) best fits the data in the ARC versus NWA comparison. Left panel:  $N_1$ ,  $N_2$  and  $N_A$  represent the respective effective population size of the two descending populations and of the ancestral lineage. The black and red arrows ( $m_{ij}$ ) represent the migration rate from population  $j$  into population  $i$ .  $T_{\text{split}}$  = splitting time.  $T_{\text{sc}}$  = time of secondary contact [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

( $SD = 82,000$ ) in the NWA lineage under the same model (Table 1; see also Table S4 for details of parameter estimates when separating by haplogroups).

### 3.2.2 | Candidate loci driving genetic divergence among lineages

Based on the parameters estimated with  $\partial a \partial i$ , we performed coalescent simulations to compute a neutral envelope of differentiation conditioned on expected heterozygosity. After excluding markers with reduced heterozygosity ( $H_E < 0.10$ ), a total of 2085, 622, and 330 markers departed from neutrality in the GRE versus ARC, NWA versus ARC and NWA versus GRE comparisons, respectively. These were all randomly distributed across the reconstructed chromosomes (Table S7). Of these, 276 were shared outliers between GRE versus ARC and NWA versus ARC, 130 were shared between NWA versus ARC and NWA versus GRE, and 114 were shared between

GRE versus ARC and NWA versus GRE. A total of seven outliers were shared in all three pairwise comparisons, such that only a few outliers should be affected by background selection (see Discussion). A total of 118 outliers were fixed (i.e.,  $F_{ST} = 1$ ) between the GRE and ARC lineages, no outliers were fixed in the other two pairwise comparisons, and  $F_{ST}$  values were less pronounced in comparisons involving the NWA clades. Significant BLAST searches (e-value  $< 1e^{-10}$  and length  $> 90$ ) were obtained for 688 loci, with a total of 43 putative nonsynonymous mutations. These significant hits were also randomly distributed in the genome (Table S7) and associated with various biological processes (Table S8). The SNPs associated with the putative chromosomal rearrangement on Chr2 and Chr9 were significantly enriched with outliers when compared to the rest of the genome for the comparison between GRE and NWA (Fisher's exact test,  $p < .0001$ ) as well as between ARC and NWA (Fisher's test,  $p = .0006$ ), but not in the comparison between ARC and GRE (Fisher's test,  $p = .7038$ ) as expected given that this rearrangement is absent in these lineages.

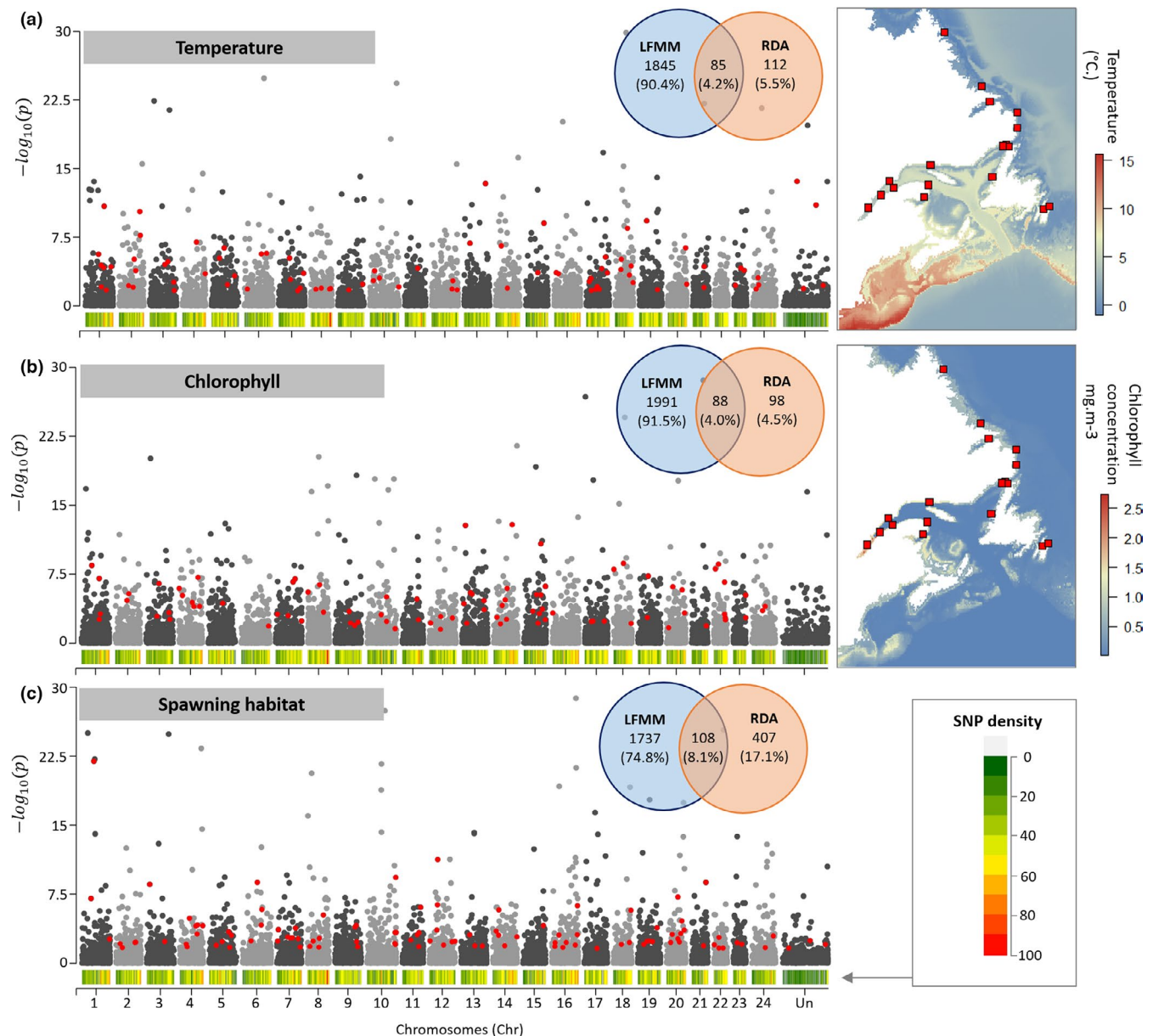


### 3.3 | Genotype–environment associations in the NWA clade

#### 3.3.1 | Outlier loci associated with spawning habitats

In the pRDA, genetic variance was significantly explained by spawning habitat (i.e., beach spawning, shallow-water demersal and deeper-water demersal sites; overall significance,  $df = 2$ ,  $F = 1.58$ ,  $p < .001$ ). We detected 407 and 1,737 outlier SNPs associated with

spawning habitat using pRDA (overall significance,  $df = 2$ ,  $F = 1.12$ ,  $p < .001$ ) and LFMM, respectively (Figure 6). A total of 108 SNPs were detected by the two methods; 26% of the SNPs detected with pRDA were also detected using LFMM. Respectively, 5% and 9% of the 108 loci were located in the genomic regions of Chr2 and Chr9 associated with the haplogroups. Outlier loci were in excess in these genomic regions compared to the rest of the genome (see Fisher's test results in Table 2). Furthermore, 85% and 55% of the outlier SNPs were polymorphic (allelic frequency [AF] > 0) in the GRE (mean AF = 0.28) and ARC lineages (mean AF = 0.26),



**FIGURE 6** Molecular adaptation to local environmental conditions in the capelin (*Mallotus villosus*). (a) SNP outliers associated with temperature: the Manhattan plot show  $p$ -values from LFMM analyses for SNPs detected as outliers by both pRDA and LFMM (shown in red) and the others (shown in grey). Venn diagrams show the outlier loci detected with either LFMMs or pRDA, or both—the percentage corresponds to the proportion of outliers detected by each method and the two methods combined. The map shows the spatial variation in temperature. (b) SNP outliers associated with chlorophyll concentration. (c) SNP outliers associated with spawning habitats. Note that contigs were placed on a chromosome but the order of contig themselves may not be accurate. The SNP positions here should not be interpreted as physical positions on the chromosomes [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

**TABLE 2** Enrichment of SNP outliers in the potential chromosomal rearrangement in chromosome Chr2 and Chr9 and in shared polymorphisms with the ARC and GRE lineages; also presented are the results of Fisher's tests (odds ratio and *p*-values)

Enrichment	Chlorophyll		Temperature		Type of spawning site	
	Ratio	<i>p</i>	Ratio	<i>p</i>	Ratio	<i>p</i>
Chromosomal rearrangement	0.35	.16	0.10	<.0001	0.09	<.0001
Shared polymorphism with ARC	0.31	<.0001	0.42	.0001	0.31	<.0001
Shared polymorphism with GRE	0.10	<.0001	0.19	<.0001	0.18	<.0001
Shared polymorphism with GRE and ARC	0.12	<.0001	0.36	<.0001	0.22	<.0001

respectively. Fisher's tests (Table 2) revealed an excess of shared polymorphism within both lineages at outlier loci associated with spawning habitat. We identified 112 transcripts in the 1-kbp window around the outlier loci enriched in various cellular and molecular processes (Table S9).

### 3.3.2 | Outlier loci associated with temperature and chlorophyll concentration within beach-spawning sites

In the pRDA, the genetic variance was significantly explained by the two environmental variables (overall significance,  $df = 2$ ,  $F = 1.09$ ,  $p < .001$ ). A total of 197 and 1,930 outlier SNPs were associated with temperature based on pRDA and LFMM, respectively (Figure 6). A total of 85 SNPs were identified by the two methods; 43% of the SNPs identified with pRDA were also detected using LFMM. Among those 85 SNPs, 7% and 2% respectively occurred within the genomic regions of Chr2 and Chr9 associated with the haplogroups. Fisher's tests indicated the outlier loci were in excess in these regions compared to the rest of the genome (Table 2). Furthermore, 76% and 47% of the SNPs were polymorphic in the GRE (mean AF = 0.220) and ARC lineages (mean AF = 0.22), respectively. Fisher's tests showed an excess of shared polymorphism for both lineages at these outlier loci (Table 2). Within the 1-kbp window around the outlier loci, we detected 77 transcripts that were enriched in cellular and molecular processes (Table S10).

We also identified 186 and 2,079 outlier SNPs associated with chlorophyll concentration with pRDA and LFMM, respectively (Figure 6). A total of 88 SNPs were identified using the two approaches; 47% of the SNPs detected with pRDA were also detected using LFMM. Among those 88 SNPs, 2% and 5% were present in the genomic regions Chr2 and Chr9 associated with the haplogroups. However, outlier loci were not in excess in these genomic regions (Table 2). Furthermore, 91% and 55% of the SNPs were polymorphic in the GRE (mean AF = 0.26) and ARC lineages (mean AF = 0.23) respectively. Fisher's tests showed an excess of shared polymorphism for both lineages in outlier loci (Table 2). We detected 93 transcripts within the 1-kbp window around the outlier loci that were enriched in cellular and molecular processes (Table S11).

## 4 | DISCUSSION

Our results revealed a complex demographic history of divergence among the three capelin glacial lineages. Models of secondary contact including both linked selection and variable introgression rates among loci received highest statistical support. They suggest that the lineages are old, having diverged from 3.8 to 1.8 Ma depending on the pair of clades considered. Using coalescent simulations, we also identified a set of candidate loci driving the genetic divergence among lineages and that might be involved in their reproductive isolation. These observations support the hypothesis that these lineages may in fact represent cryptic species of capelin or at the very least indicate that these lineages are part of an ongoing speciation process. Within lineages, our analyses revealed large  $N_e$  and weak genetic divergence among sampling sites, typical of other marine fishes. In the NWA lineage, a chromosomal rearrangement probably appeared during the interlineage divergence process, resulting in the existence of three haplogroups spread across most spawning sites of the NWA lineage, but showing differences in frequencies between capelin from different spawning habitats. Genotype–environment associations revealed molecular signatures of local adaptation to sea conditions prevailing at spawning sites. Importantly, our study suggested that shared polymorphism among glacial lineages and chromosomal rearrangements occurring in NWA might have broadly contributed to the genetic polymorphisms involved in local adaptation in the presence of high gene flow.

### 4.1 | An ongoing speciation process among lineages

Our modelling approach supports the hypotheses that (a) the three lineages have diverged in allopatry, (b) gene flow during the secondary contact eroded past differentiation and resulted in heterogeneous differentiation along the genome, and (c) such differentiation is also probably driven by linked selection in the form of either background selection (Charlesworth et al., 1993) or hitchhiking (Smith & Haigh, 1974). It was previously shown that neglecting these two effects can lead to bias in demographic inference in terms of model choice, false inference of ongoing gene flow or, by contrast, inference of the absence of gene flow while the genome might still be



semipermeable (Cruickshank & Hahn, 2014; Ewing & Jensens, 2016; Roux et al., 2016; Souza et al., 2013). The growing number of studies that have used this approach (e.g., Leroy et al., 2019; Rougemont & Bernatchez, 2018; Roux et al., 2014; Roux et al., 2016; Roux, Tsagkogeorga, Bierne, & Galtier, 2013) found increased support for models of secondary contact or have demonstrated difficulty in distinguishing models of secondary contact from models of sympatric divergence with gene flow (i.e., IM models). Moreover, the role of linked selection is increasingly recognized as being implicated in observed patterns of genome-wide divergence (Burri, 2017; Burri et al., 2015; Stankowski et al., 2019; Wang, Street, Scofield, & Ingvarsson, 2016). Here, our approach does not allow a fine quantitative estimation of linked selection and future studies could be undertaken to address this question.

Overall, our demographic reconstruction suggests a relatively long period of divergence, compatible with a level of divergence more typical of interspecific rather than intraspecific variation (Hedges, Marin, Suleski, Paymer, & Kumar, 2015; Roux et al., 2016). This divergence has resulted in pronounced contemporary genome-wide differentiation randomly distributed across the genome. Accordingly, coalescent simulations have suggested the existence of putative outliers (i.e., departing from the neutral envelop). These outliers were spread randomly among and within chromosomes, as expected under the genic view of speciation (Coyne & Orr, 2004; Wu, 2001). The divergence process has probably favoured the establishment of a barrier to gene flow resulting in restricted gene flow along the genome. To our knowledge, such levels of restricted gene flow between parapatric lineages have rarely been documented in a strictly marine species (Gagnaire et al., 2018; Le Moan et al., 2016), thus raising the hypothesis that these three lineages may actually be cryptic species of capelin occurring in the North Atlantic and Arctic oceans. Further refinement would be required to improve our demographic reconstructions. In particular, three-population models and the inclusion of ghost populations (Slatkin, 2005) may provide a richer picture of the divergence history of capelin lineages. Moreover, further experimental studies should formally test the hypothesis that limited gene flow and the absence of current admixture (based on our clustering analyses) among lineages result from pre- or post-zygotic reproductive isolation.

#### 4.2 | A potential chromosomal rearrangement in the NWA lineage

Our PCA revealed the existence of three haplogroups occurring at variable frequencies among sampling sites. This observation suggests the presence of a chromosomal rearrangement characterized by a lower recombination rate compared to the rest of the genome, although PCA is not the most effective tool to detect structural variants (Li & Ralph, 2019; but see Ma & Amos, 2012). Indeed, the signal left by inversions cannot easily be distinguished from long haplotypes under balancing selection or simply regions of reduced

recombination (Lotterhos, 2019). Here, however, our synteny analysis indicates that the loci included in the putative chromosomal rearrangement are located in Chr2 and Chr9 of the *Esox lucius* genome. The PCA executed for Chr2 and Chr9 separately also displayed the same haplogroup pattern. Assuming that Chr2 and Chr9 are completely fused in the capelin, this suggests the existence of a polymorphic chromosomal inversion in the NWA lineage. Alternatively, it could be two covarying chromosomal inversions or a polymorphic fusion.

Our analyses indicated that such genomic regions, regardless of their true nature, are absent in ARC and GRE lineages. Given the large  $N_e$  of these lineages, it is unlikely that polymorphism within this region was lost due to genetic drift during lineage divergence. Instead, we hypothesized that the putative rearrangement appeared in the NWA lineage ~2.8 Ma after the split from the ARC lineage but almost concomitantly with the GRE split from the NWA.

Future work based on whole genome resequencing and long-read data will allow us to clarify the nature of the putative rearrangement (chromosomal fusion or inversion) and quantify its exact size, and identify the genes that are located within this region. Furthermore, exciting questions regarding the age of the putative rearrangement, its ancestry, and the possible association with individual phenotype (e.g., morphological, physiological and life history traits) could be addressed with such additional sequencing data.

#### 4.3 | High gene flow among spawning sites within lineages

Our study has highlighted weak, yet significant genetic variation within the NWA lineage. These results are also congruent with many studies on marine organisms (excluding reef species) that reported low genetic variation across large geographical areas (e.g., fishes: Lamichhaney et al., 2012; crustaceans: Benestan et al., 2016; molluscs: Van Wyngaarden et al., 2017; for a review, see Kelley, Brown, Therkildsen, & Foote, 2016). The weak intralineage genetic differentiation within NWA partly results from large historical effective population sizes in the three lineages. Furthermore, the low IBD detected using neutral markers suggests high dispersal among spawning sites that led to pronounced gene flow.

#### 4.4 | Habitat matching choice and adaptation to local marine conditions

Beyond weak genetic structure at the scale of the NWA geographical range of distribution, we detected loci under putative divergent selection associated with spawning habitats. As adult capelin were captured at spawning sites during breeding season, this pattern could result from habitat matching choice (Edelaar et al., 2008). This mechanism implies that individuals disperse towards reproductive

sites that maximize their breeding performances, which allows local adaptation despite high gene flow (Edelaar et al., 2008; Jacob et al., 2017). As individual habitat choice does not always match optimal environment conditions (i.e., partial habitat matching) or because a population may contain both specialist and generalist individuals (Jacob et al., 2017), it may result in weak genetic structure due to large effective population size and/or gene flow between the two habitats.

Our results and previous studies suggest that capelin could be adapted to spawning site characteristics. The breeding sites of beach-spawning capelin located in the intertidal zone are exposed to weather variation (e.g., temperature, wind) that strongly affects embryonic and larval growth and survival, as well as swimming performance during larval dispersal (Frank & Leggett, 1981a, 1981b, 1982; Leggett et al., 1984). Beach spawning sites also display more variable and higher water temperatures than demersal sites (Nakashima & Wheeler, 2002), which may sometimes lead to dramatic mortality events before larval hatching (Leggett et al., 1984). Interestingly, a previous study has suggested that embryos from beach-spawning capelin have a faster growth rate than those of demersal capelins (Nakashima & Wheeler, 2002), which might indicate local adaptation allowing an accelerated development in the intertidal zone. In parallel, our results indicate that candidate loci potentially under divergent polygenic selection (i.e., covarying markers of small effects detected both by pRDA and LFMM) and associated with spawning habitat are involved in a broad range of biological processes (Table S9) driving cell resistance to environmental stress. Together, these results raise the hypothesis of habitat matching choice and adaptation to local environmental conditions prevailing at beach and demersal spawning sites. Yet, as the number of demersal sites was limited (six versus 18 beach sites) in our study due to sampling constraints, we encourage future work to identify additional genotype–environment associations to validate the putative candidate outliers detected in our study.

Our study also suggests a similar mechanism of adaptation to the conditions of temperature and trophic productivity experienced in beach spawning sites. We detected two markers under putative divergent selection associated with water temperature and localized in genomic regions (a narrow window of 1 kbp) that contain genes involved in growth regulation (thermodependent in fishes, Pauly, 1980) and response to UV radiation (Table S4). Overall, these results are in accordance with previous studies on capelin showing that water temperature and trophic productivity strongly affect embryonic and larval growth and survival, as well as the swimming capacity of larvae during the drift phase following hatching (Frank & Leggett, 1981a, 1981b, 1982; Leggett et al., 1984). However, although our findings are consistent with previous GBS studies suggesting an environmentally driven signature of molecular adaptation in marine organisms (e.g., Benestan et al., 2016; Bradbury et al., 2010; Stanley et al., 2018; Xuereb et al., 2018), they suffer from the limitations already reported in RAD-seq analyses, especially the nonexhaustive genome sampling that prevents a full assessment of candidate loci (Lowry et al., 2017) and are limited by the accuracy of genome scan

approaches (Pavlidis, Jensen, Stephan, Stamatakis, 2012). Moreover, the complex demographic history among the three lineages, have likely favored the accumulation of intrinsic reproductive barriers. These complicated further the distinction between intrinsic versus extrinsic barriers (i.e. ecologically driven) (Bierne, Welch, Loire, Bonhomme, & David, 2011). Indeed, the former might be reused across different geographic contexts and involved or confounded with the latter (Bierne et al. 2011, Riquet et al. 2019). Further whole genome resequencing combined with functional validations and experimental work may provide a more exhaustive picture of candidate loci and could help validate our findings.

#### 4.5 | Contribution of genomic background to local adaptation

Our analyses suggested that shared ancestral polymorphisms may play a central role in local adaptation to sea conditions, as predicted by theory (Barrett & Schluter, 2008; Hedrick, 2013). Indeed, 85% and 55% of the markers putatively under divergent selection and associated with the type of spawning site were polymorphic in the GRE and ARC lineages, respectively. In addition, 76% and 47% of the candidate outliers related to temperature, as well as 91% and 55% of the outliers associated with trophic productivity, were also polymorphic in the GRE and ARC lineages, respectively. Importantly, we consistently found an excess of shared polymorphisms with both lineages in this set of outlier loci. Therefore, these results, combined with the observed level of divergence (at least 2 million years despite all uncertainties) and low migration rate among lineages, support the hypothesis that these shared variants are probably standing variants inherited from the last common ancestors of these lineages (Lai et al., 2019). Alternatively, they may have originated in a single lineage and then spread among lineages via introgression following secondary contact (Welch & Jiggins, 2014). Lastly, as it is the case for any other studies based on reduced representation sequencing data, the SNP outliers are probably neutral hitchhiker loci linked to the direct target of selection (Bierne et al. 2011). To distinguish among the different hypotheses above, whole genome data, combined with methods to distinguish among soft versus hard selective sweep, would be needed (e.g. Kern & Schrider, 2018). Further use of coalescent methods would also allow us to directly infer the age of the variants relative to the most recent common ancestor of the lineage (Albers & McVean, 2020).

Our analyses also indicated that a non-negligible proportion of outlier markers associated with the type of spawning site (8%), temperature (7%), and trophic productivity (2%) is located in the putative chromosomal rearrangement. For both the type of spawning site and temperature, outlier loci were in excess in the genomic regions contained in the chromosomal rearrangement compared to the rest of the genome. These results thus appear to be consistent with previous work suggesting that chromosomal polymorphism may facilitate thermal adaptation despite high gene flow in marine fishes (Barth et al., 2017; Wellband et al., 2019). However, future work based on

whole genome resequencing data should be conducted to more accurately assess the contribution of the putative rearrangement to local adaptation, especially by identifying the type of rearrangement detected in our study and as well as its characteristics (i.e., size and content), and its effects on capelin phenotype and performance.

## ACKNOWLEDGMENTS

We thank biologists and technicians of the Department of Fisheries and Oceans Canada for their implication as well as all everyone who contributed to sampling throughout the study area. We are also grateful to Associate Editor Shawn Narum and four anonymous reviewers for their constructive comments on a previous version of the manuscript. This research was funded by a Strategic Project Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to L. Bernatchez, M. Clément and P. Sirois, a financial contribution of Ressources Aquatiques Québec and was also supported by in-kind contribution from many other organizations: Department of Fisheries and Oceans Canada, Nunatsiavut Government, NunatuKavut Community Council, Labrador Fishermen's Union Shrimp Company, Department of Fisheries and Aquaculture – Government of Newfoundland and Labrador, World Wildlife Fund Canada, St. Lawrence Global Observatory, Parc Marin du Saguenay–Saint-Laurent, and the Greenland Institute of Natural Resources. Whole genome sequencing and construction of the draft capelin genome assembly were funded by the Research Council of Norway (RCN) through the Nansen Legacy project (RCN no. 276730) and the ComparaCod project (RCN no. 222378). PacBio library creation and high-throughput sequencing were carried out at the Norwegian Sequencing Centre (NSC), University of Oslo, Norway. Genome assembly was performed on the Abel Supercomputing Cluster (Norwegian metacenter for High Performance Computing [NOTUR] and the University of Oslo) operated by the Research Computing Services group at USIT, the University of Oslo IT department (<http://www.hpc.uio.no/>).








## AUTHOR CONTRIBUTIONS

H.C. and Q.R. made the statistical analyses and wrote the paper. M.L., C.M., E.N. and Y.D. contributed to the bioinformatics and statistical analyses. L.B., P.S., M.C.A. and M.C.L. initiated the project, and L.B. conceptualized and coordinated the work. T.J. collected the DNA samples in the GRE lineage. The generation of the draft capelin genome assembly was initiated and coordinated by S.J., and samples were provided by K.P. The whole genome sequencing DNA extraction was performed by S.N.K.H. and construction of the draft genome assembly was done by O.K.T. and S.N.K.H.

## DATA AVAILABILITY STATEMENT

Genome assembly and associated reads are available on NCBI under accession no. PRJEB38139. Raw sequencing data for GBS libraries are available under accession no. PRJNA631144. Filtered *vcf* file and environmental data are available on Dryad (<https://doi.org/10.5061/dryad.hx3ffbgbp>).

## ORCID

Hugo Cayuela  <https://orcid.org/0000-0003-3250-6295>  
 Quentin Rougemont  <https://orcid.org/0000-0003-2987-3801>  
 Martin Laporte  <https://orcid.org/0000-0002-0622-123X>  
 Claire Mérot  <https://orcid.org/0000-0003-2607-7818>  
 Yann Dorant  <https://orcid.org/0000-0002-7295-9398>  
 Siv Nam Khang Hoff  <https://orcid.org/0000-0001-8113-338X>  
 Sissel Jentoft  <https://orcid.org/0000-0001-8707-531X>

## REFERENCES

- Albers, P. K., & McVean, G. (2020). Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biology*, 18, e3000586. <https://doi.org/10.1371/journal.pbio.3000586>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Barrett, R. D., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology and Evolution*, 23, 38–44. <https://doi.org/10.1016/j.tree.2007.09.008>
- Barth, J. M., Berg, P. R., Jonsson, P. R., Bonanomi, S., Corell, H., Hemmer-Hansen, J., ... André, C. (2017). Genome architecture enables local adaptation of Atlantic cod despite high connectivity. *Molecular Ecology*, 26, 4452–4466. <https://doi.org/10.1111/mec.14207>
- Barton, N. H. (1998). The effect of hitch-hiking on neutral genealogies. *Genetical Research*, 72, 123–133. <https://doi.org/10.1017/S0016672398003462>
- Barton, N., & Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, 57, 357. <https://doi.org/10.1038/hdy.1986.135>
- Beaumont, M. A., & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 263, 1619–1626.
- Benestan, L., Quinn, B. K., Maaroufi, H., Laporte, M., Clark, F. K., Greenwood, S. J., ... Bernatchez, L. (2016). Seascape genomics provides evidence for thermal adaptation and current-mediated population structure in American lobster (*Homarus americanus*). *Molecular Ecology*, 25, 5073–5092.
- Berg, P. R., Star, B., Pampoulie, C., Bradbury, I. R., Bentzen, P., Hutchings, J. A., ... Jakobsen, K. S. (2017). Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity*, 119, 418. <https://doi.org/10.1038/hdy.2017.54>
- Bierne, N., Welch, J., Loire, E., Bonhomme, F., & David, P. (2011). The coupling hypothesis: Why genome scans may fail to map local adaptation genes. *Molecular Ecology*, 20, 2044–2072. <https://doi.org/10.1111/j.1365-294X.2011.05080.x>
- Bitter, M. C., Kapsenberg, L., Gattuso, J. P., & Pfister, C. A. (2019). Standing genetic variation fuels rapid adaptation to ocean acidification. *Nature Communications*, 10, 1–10. <https://doi.org/10.1038/s41467-019-13767-1>
- Blanquart, F., & Gandon, S. (2011). Evolution of migration in a periodically changing environment. *The American Naturalist*, 177(2), 188–201. <https://doi.org/10.1086/657953>
- Blanquart, F., Kaltz, O., Nuismer, S. L., & Gandon, S. (2013). A practical guide to measuring local adaptation. *Ecology Letters*, 16(9), 1195–1205. <https://doi.org/10.1111/ele.12150>
- Bosch, S., Tyberghein, L., & De Clerck, O. (2017). *Sdm predictors: Species distribution modeling predictor datasets*. R package version 0.2. 6.
- Bradbury, I. R., Hubert, S., Higgins, B., Borza, T., Bowman, S., Paterson, I. G., ... Bentzen, P. (2010). Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society B: Biological Sciences*, 277, 3725–3734. <https://doi.org/10.1098/rspb.2010.0985>

- Bradbury, I. R., Laurel, B., Snelgrove, P. V., Bentzen, P., & Campana, S. E. (2008). Global patterns in marine dispersal estimates: The influence of geography, taxonomic category and life history. *Proceedings of the Royal Society B: Biological Sciences*, 275, 1803–1809.
- Buren, A. D., Koen-Alonso, M., Pepin, P., Mowbray, F., Nakashima, B., Stenson, G., ... Montevercchi, W. A. (2014). Bottom-up regulation of capelin, a keystone forage species. *PLoS One*, 9, e87589. <https://doi.org/10.1371/journal.pone.0087589>
- Burri, R. (2017). Linked selection, demography and the evolution of correlated genomic landscapes in birds and beyond. *Molecular Ecology*, 26, 3853–3856. <https://doi.org/10.1111/mec.14167>
- Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., ... Ellegren, H. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Research*, 25, 1656–1665.
- Camacho, C., & Hendry, A. P. (2020). Matching habitat choice: It's not for everyone. *Oikos*, 129(5), 689–699. <https://doi.org/10.1111/oik.06932>
- Carscadden, J. E., Frank, K. T., & Leggett, W. C. (2001). Ecosystem changes and the effects on capelin (*Mallotus villosus*), a major forage species. *Canadian Journal of Fisheries and Aquatic Sciences*, 58, 73–85.
- Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, 22, 3124–3140. <https://doi.org/10.1111/mec.12354>
- Charlesworth, B., Morgan, M. T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134, 1289–1303.
- Clobert, J., Le Galliard, J. F., Cote, J., Meylan, S., & Massot, M. (2009). Informed dispersal, heterogeneity in animal dispersal syndromes and the dynamics of spatially structured populations. *Ecology Letters*, 12, 197–209. <https://doi.org/10.1111/j.1461-0248.2008.01267.x>
- Colbeck, G. J., Turgeon, J., Sirois, P., & Dodson, J. J. (2011). Historical introgression and the role of selective vs. neutral processes in structuring nuclear genetic variation (AFLP) in a circumpolar marine fish, the capelin (*Mallotus villosus*). *Molecular Ecology*, 20, 1976–1987. <https://doi.org/10.1111/j.1365-294X.2011.05069.x>
- Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sunderland, MA: Sinauer.
- Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23, 3133–3157. <https://doi.org/10.1111/mec.12796>
- Csilléry, K., Rodríguez-Verdugo, A., Rellstab, C., & Guillaume, F. (2018). Detecting the genomic signal of polygenic adaptation and the role of epistasis in evolution. *Molecular Ecology*, 27, 606–612. <https://doi.org/10.1111/mec.14499>
- DeWoody, J. A., & Avise, J. C. (2000). Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. *Journal of Fish Biology*, 56, 461–473. <https://doi.org/10.1111/j.1095-8649.2000.tb00748.x>
- Dodson, J. J., Tremblay, S., Colombani, F., Carscadden, J. E., & Lecomte, F. (2007). Trans-Arctic dispersals and the evolution of a circumpolar marine fish species complex, the capelin (*Mallotus villosus*). *Molecular Ecology*, 16, 5030–5043. <https://doi.org/10.1111/j.1365-294X.2007.03559.x>
- Dorant, Y., Benestan, L., Rougemont, Q., Normandeau, E., Boyle, B., Rochette, R., & Bernatchez, L. (2019). Comparing Pool-seq, Rapture, and GBS genotyping for inferring weak population structure: The American lobster (*Homarus americanus*) as a case study. *Ecology and Evolution*, 9, 6606–6623.
- Dormann, C. F., Eliith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36, 27–46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Edelaar, P., Siepielski, A. M., & Clobert, J. (2008). Matching habitat choice causes directed gene flow: A neglected dimension in evolution and ecology. *Evolution*, 62, 2462–2472. <https://doi.org/10.1111/j.1558-5646.2008.00459.x>
- Ewing, G. B., & Jensen, J. D. (2016). The consequences of not accounting for background selection in demographic inference. *Molecular Ecology*, 25, 135–141. <https://doi.org/10.1111/mec.13390>
- Faria, R., Johannesson, K., Butlin, R. K., & Westram, A. M. (2019). Evolving inversions. *Trends in Ecology & Evolution*, 34, 239–248. <https://doi.org/10.1016/j.tree.2018.12.005>
- Faria, R., & Navarro, A. (2010). Chromosomal speciation revisited: Rearranging theory with pieces of evidence. *Trends in Ecology & Evolution*, 25, 660–669. <https://doi.org/10.1016/j.tree.2010.07.008>
- Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology*, 27, 2215–2233. <https://doi.org/10.1111/mec.14584>
- Frank, K. T., & Leggett, W. C. (1981a). Wind regulation of emergence times and early larval survival in capelin (*Mallotus villosus*). *Canadian Journal of Fisheries and Aquatic Sciences*, 38, 215–223.
- Frank, K. T., & Leggett, W. C. (1981b). Prediction of egg development and mortality rates in capelin (*Mallotus villosus*) from meteorological, hydrographic, and biological factors. *Canadian Journal of Fisheries and Aquatic Sciences*, 38, 1327–1338.
- Frank, K. T., & Leggett, W. C. (1982). Environmental regulation of growth rate, efficiency, and swimming performance in larval capelin (*Mallotus villosus*), and its application to the match/mismatch hypothesis. *Canadian Journal of Fisheries and Aquatic Sciences*, 39, 691–699.
- François, O., Martins, H., Caye, S. D. (2015). Schoville. Controlling false discoveries in genome scans for selection. *Molecular Ecology*, 25(2), 454–469.
- Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, 6, 925–929. <https://doi.org/10.1111/2041-210X.12382>
- Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, 30, 1687–1699. <https://doi.org/10.1093/molbev/mst063>
- Gagnaire, P.-A., Lamy, J.-B., Cornette, F., Heurtebise, S., Dégremont, L., Flahauw, E., ... Lapègue, S. (2018). Analysis of genome-wide differentiation between native and introduced populations of the cupped oysters *Crassostrea gigas* and *Crassostrea angulata*. *Genome Biology and Evolution*, 10, 2518–2534. <https://doi.org/10.1093/gbe/evy194>
- Graffelman, J. (2015). Exploring diallelic genetic markers: The hardy weinberg package. *Journal of Statistical Software*, 64, 1–23.
- Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLOS Genetics*, 5, e1000695. <https://doi.org/10.1371/journal.pgen.1000695>
- Haanel, Q., Roesti, M., Moser, D., MacColl, A. D., & Berner, D. (2019). Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in stickleback fish. *Evolution Letters*, 3, 28–42. <https://doi.org/10.1002/evl3.99>
- Hancock, A. M., Alkorta-Aranburu, G., Witonsky, D. B., & Di Rienzo, A. (2010). Adaptations to new environments in humans: The role of subtle allele frequency shifts. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1552), 2459–2468. <https://doi.org/10.1098/rstb.2010.0032>
- Hedges, S. B., Marin, J., Suleski, M., Paymer, M., & Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*, 32, 835–845. <https://doi.org/10.1093/molbev/msv037>



- Hedrick, P. W. (2013). Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, 22, 4606–4618. <https://doi.org/10.1111/mec.12415>
- Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18, 337–338. <https://doi.org/10.1093/bioinformatics/18.2.337>
- Jacob, S., Legrand, D., Chaîne, A. S., Bonte, D., Schtickzelle, N., Huet, M., & Clobert, J. (2017). Gene flow favours local adaptation under habitat choice in ciliate microcosms. *Nature Ecology & Evolution*, 1, 1407. <https://doi.org/10.1038/s41559-017-0269-5>
- Jombart, T. (2008). ADEGENET: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24, 1403–1405. <https://doi.org/10.1093/bioinformatics/btn129>
- Jones, F. C., Chan, Y. F., Schmutz, J., Grimwood, J., Brady, S. D., Southwick, A. M., ... Kingsley, D. M. (2012). A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Current Biology*, 22, 83–90. <https://doi.org/10.1016/j.cub.2011.11.045>
- Kaplan, N. L., Hudson, R. R., & Langley, C. H. (1989). The “hitchhiking effect” revisited. *Genetics*, 123, 887–899.
- Kawecki, T. J., & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters*, 7, 1225–1234. <https://doi.org/10.1111/j.1461-0248.2004.00684.x>
- Kelley, J. L., Brown, A. P., Therkildsen, N. O., & Foote, A. D. (2016). The life aquatic: Advances in marine vertebrate genomics. *Nature Reviews Genetics*, 17, 523. <https://doi.org/10.1038/nrg.2016.66>
- Kern, D., & Schrider, D. R. (2018). diploS/HIC: An updated approach to classifying selective sweeps. *G3: Genes, Genomes, Genetics*, 8, 1959–1970.
- Li, H., & Ralph, P. (2018). Local PCA Shows How the Effect of Population Structure Differs Along the Genome. *Genetics*, 211(1), 289–304. <https://doi.org/10.1534/genetics.118.301747>
- Lai, Y. T., Yeung, C. K., Omland, K. E., Pang, E. L., Hao, Y., Liao, B. Y., ... Hung, H. Y. (2019). Standing genetic variation as the predominant source for adaptation of a songbird. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 2152–2157. <https://doi.org/10.1073/pnas.1813597116>
- Lamichanay, S., Barrio, A. M., Rafati, N., Sundström, G., Rubin, C. J., Gilbert, E. R., ... Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 19345–19350. <https://doi.org/10.1073/pnas.1216128109>
- Laporte, M., Pavey, S. A., Rougeux, C., Pierron, F., Lauzent, M., Budzinski, H., ... Bernatchez, L. (2016). RAD sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic Eels. *Molecular Ecology*, 25, 219–237. <https://doi.org/10.1111/mec.13466>
- Lasky, J. R., Des Marais, D. L., McKay, J. K., Richards, J. H., Juenger, T. E., & Keitt, T. H. (2012). Characterizing genomic variation of *Arabidopsis thaliana*: The roles of geography and climate. *Molecular Ecology*, 21, 5512–5529.
- Le Moan, A., Bekkevold, D., & Hemmer-Hansen, J. (2019). Evolution at two-time frames shape structural variants and population structure of European plaice (*Pleuronectes platessa*). *BioRxiv*, 662577.
- Le Moan, A. L., Gagnaire, P.-A., & Bonhomme, F. (2016). Parallel genetic divergence among coastal-marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Molecular Ecology*, 25, 3187–3202. <https://doi.org/10.1111/mec.13627>
- Legendre, P., & Legendre, L. F. (2012). *Numerical ecology*. Amsterdam, the Netherlands: Elsevier.
- Leggett, W. C., Frank, K. T., & Carscadden, J. E. (1984). Meteorological and hydrographic regulation of year-class strength in capelin (*Mallotus villosus*). *Canadian Journal of Fisheries and Aquatic Sciences*, 41, 1193–1201.
- Lenormand, T. (2002). Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, 17, 183–189. [https://doi.org/10.1016/S0169-5347\(02\)02497-7](https://doi.org/10.1016/S0169-5347(02)02497-7)
- Leroy, T., Rougemont, Q., Dupouey, J.-L., Bodénès, C., Lalanne, C., Belser, C., ... Plomion, C. (2019). Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. *New Phytologist*, 214, 865–878.
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.
- Linnen, C. R., Kingsley, E. P., Jensen, J. D., & Hoekstra, H. E. (2009). On the origin and spread of an adaptive allele in deer mice. *Science*, 325, 1095–1109. <https://doi.org/10.1126/science.1175826>
- Lotterhos, K. E. (2019). The effect of neutral recombination variation on genome scans for selection. *G3: Genes, Genomes, Genetics*, 9, 1851–1867. <https://doi.org/10.1534/g3.119.400088>
- Lowe, W. H., & Addis, B. R. (2019). Matching habitat choice and plasticity contribute to phenotype–environment covariation in a stream salamander. *Ecology*, 100, e02661. <https://doi.org/10.1002/ecy.2661>
- Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Mol Ecol Resour*, 17, 142–152. <https://doi.org/10.1111/1755-0998.12635>
- Ma, J., & Amos, C. I. (2012). Investigation of inversion polymorphisms in the human genome using principal components analysis. *PLoS One*, 7, e40224. <https://doi.org/10.1371/journal.pone.0040224>
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 17, 10–12. <https://doi.org/10.14806/ej.17.1.200>
- Mérot, C., Berdan, E. L., Babin, C., Normandeau, E., Wellenreuther, M., & Bernatchez, L. (2018). Intercontinental karyotype–environment parallelism supports a role for a chromosomal inversion in local adaptation in a seaweed fly. *Proceedings of the Royal Society B: Biological Sciences*, 285, 20180519. <https://doi.org/10.1098/rspb.2018.0519>
- Monnahan, P. J., Colicchio, J., & Kelly, J. K. (2015). A genomic selection component analysis characterizes migration–selection balance. *Evolution*, 69, 1713–1727. <https://doi.org/10.1111/evo.12698>
- Moore, J. S., Harris, L. N., Le Luyer, J., Sutherland, B. J. G., Rougemont, Q., Tallman, R. F., & Bernatchez, L. (2017). Genomics and telemetry suggest a role for migration harshness in determining overwintering habitat choice, but not gene flow, in anadromous Arctic Char. *Molecular Ecology*, 26(24), 1713–6800. <https://doi.org/10.1111/mec.14393>
- Nakashima, B. S., & Wheeler, J. P. (2002). Capelin (*Mallotus villosus*) spawning behaviour in Newfoundland waters–The interaction between beach and demersal spawning. *ICES Journal of Marine Science*, 59, 909–916. <https://doi.org/10.1006/jmsc.2002.1261>
- Oziolor, E. M., Reid, N. M., Yair, S., Lee, K. M., Guberman VerPloeg, S., Bruns, P. C., ... Matson, C. W. (2019). Adaptive introgression enables evolutionary rescue from extreme environmental pollution. *Science*, 364, 455–457. <https://doi.org/10.1126/science.aav4155>
- Palumbi, S. R. (1992). Marine speciation on a small planet. *Trends in Ecology & Evolution*, 7, 114–118. [https://doi.org/10.1016/0169-5347\(92\)90144-Z](https://doi.org/10.1016/0169-5347(92)90144-Z)
- Pauly, D. (1980). On the interrelationships between natural mortality, growth parameters, and mean environmental temperature in 175 fish stocks. *ICES Journal of Marine Science*, 39, 175–192. <https://doi.org/10.1093/icesjms/39.2.175>
- Pavlidis, P., Jensen, J. D., Stephan, W., Stamatakis, A. (2012). A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, 29(10), 3237–3248. <https://doi.org/10.1093/molbev/mss136>
- Pembleton, L. W., Cogan, N. O., & Forster, J. W. (2013). STAMP: An R package for calculation of genetic differentiation and structure of



- mixed-ploidy level populations. *Molecular Ecology Resources*, 13, 946–952. <https://doi.org/10.1111/1755-0998.12129>
- Pettersson, M. E., Rochus, C. M., Han, F., Chen, J., Hill, J., Hill, J., ... Andersso, L. (2019). A chromosome-level assembly of the Atlantic herring genome-detection of a supergene and other signals of selection. *Genome Research*, 29, 1919–1928.
- Præbel, K., Christiansen, J. S., Kettunen-Præbel, A., & Fevolden, S. E. (2013). Thermohaline tolerance and embryonic development in capelin eggs (*Mallotus villosus*) from the Northeast Atlantic Ocean. *Environmental Biology of Fishes*, 96, 753–761. <https://doi.org/10.1007/s10641-012-0069-3>
- Præbel, K., Westgaard, J. I., Fevolden, S. E., & Christiansen, J. S. (2008). Circumpolar genetic population structure of capelin *Mallotus villosus*. *Marine Ecology Progress Series*, 360, 189–199. <https://doi.org/10.3354/meps07363>
- Prezeworski, M., Coop, G., & Wall, J. D. (2005). The signature of positive selection on standing genetic variation. *Evolution*, 59, 2312–2323. <https://doi.org/10.1111/j.0014-3820.2005.tb00941.x>
- Pritchard, J. K., & Di Rienzo, A. (2010). Adaptation—not by sweeps alone. *Nature Reviews Genetics*, 11, 665–667. <https://doi.org/10.1038/nrg2880>
- Purchase, C. F. (2018). Low tolerance of salt water in a marine fish: New and historical evidence for surprising local adaption in the well-studied commercially exploited capelin. *Canadian Journal of Fisheries and Aquatic Sciences*, 75, 673–681. <https://doi.org/10.1139/cjfas-2017-0058>
- Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16, 359–371. <https://doi.org/10.1038/nrg3936>
- Riquet, F., Liautard-Haag, C., Woodall, L., Bouza, C., Louisy, P., Hamer, B., ... Bierne, N. (2019). Parallel pattern of differentiation at a genomic island shared between clinal and mosaic hybrid zones in a complex of cryptic seahorse lineages. *Evolution*, 73, 817–835. <https://doi.org/10.1111/evo.13696>
- Rose, G. A. (2005). Capelin (*Mallotus villosus*) distribution and climate: A sea “canary” for marine ecosystem change. *ICES Journal of Marine Science*, 62, 1524–1530. <https://doi.org/10.1016/j.icesjms.2005.05.008>
- Rougemont, Q., & Bernatchez, L. (2018). The demographic history of Atlantic salmon (*Salmo salar*) across its distribution range reconstructed from approximate Bayesian computations. *Evolution*, 72, 1261–1277.
- Rougemont, Q., Carrier, A., Le Luyer, J., Ferchaud, A. L., Farrell, J. M., Hatin, D., ... Bernatchez, L. (2019). Combining population genomics and forward simulations to investigate stocking impacts: A case study of Muskellunge (*Esox masquinongy*) from the St. Lawrence River Basin. *Evolutionary Applications*, 12, 902–922.
- Rougemont, Q., Gagnaire, P.-A., Perrier, C., Genthon, C., Besnard, A.-L., Launey, S., & Evanno, G. (2017). Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. *Molecular Ecology*, 20, 142–162.
- Roux, C., Castric, V., Pauwels, M., Wright, S. I., Saumitou-Laprade, P., & Vekemans, X. (2011). Does speciation between *Arabidopsis halleri* and *Arabidopsis lyrata* coincide with major changes in a molecular target of adaptation? *PLoS One*, 6, e26872. <https://doi.org/10.1371/journal.pone.0026872>
- Roux, C., Fraïsse, C., Castric, V., Vekemans, X., Pogson, G. H., & Bierne, N. (2014). Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? A case study in a *Mytilus* hybrid zone. *Journal of Evolutionary Biology*, 27, 1662–1675. <https://doi.org/10.1111/jeb.12425>
- Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biology*, 14, e2000234. <https://doi.org/10.1371/journal.pbio.2000234>
- Roux, C., Tsagkogeorga, G., Bierne, N., & Galtier, N. (2013). Crossing the species barrier: Genomic hotspots of introgression between two highly divergent *Ciona intestinalis* species. *Molecular Biology and Evolution*, 30, 1574–1587. <https://doi.org/10.1093/molbev/mst066>
- Savolainen, O., Lascoux, M., & Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews Genetics*, 14, 807–820. <https://doi.org/10.1038/nrg3522>
- Selkoe, K. A., Aloia, C. C., Crandall, E. D., Iacchei, M., Liggins, L., Puritz, J. B., ... Toonen, R. J. (2016). A decade of seascape genetics: Contributions to basic and applied marine connectivity. *Marine Ecology Progress Series*, 554, 1–19. <https://doi.org/10.3354/meps11792>
- Skatkin, M. (2005). “Seeing ghosts: the effect of unsampled populations on migration rates estimated for sampled populations”. *Molecular Ecology*, 14, 67–73. <https://doi.org/10.1111/j.1365-294X.2004.02393.x>
- Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, 23, 23–35. <https://doi.org/10.1017/S0016672300014634>
- Sousa, V., Hey, J. (2013). Understanding the origin of species with genome-scale data: modelling gene flow. *Nat Rev Genet*, 14, 404–414.
- Stankowski, S., Chase, M. A., Fuiten, A. M., Rodrigues, M. F., Ralph, P. L., & Streisfeld, M. A. (2019). Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *BioRxiv*, 342352.
- Stanley, R. R. E., DiBacco, C., Lowen, B., Beiko, R. G., Jeffery, N. W., Van Wyngaarden, M., ... Bradbury, I. R. (2018). A climate-associated multi-species cryptic genetic cline in the northwest Atlantic. *Science Advances*, 4, eaq0929.
- Tigano, A., & Friesen, V. L. (2016). Genomics of local adaptation with gene flow. *Molecular Ecology*, 25, 2144–2164. <https://doi.org/10.1111/mec.13606>
- Tine, M., Kuhl, H., Gagnaire, P. A., Louro, B., Desmarais, E., Martins, R. S., ... Reinhardt, R. (2014). European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, 5, 5770. <https://doi.org/10.1038/ncomms6770>
- Van Wyngaarden, M., Snelgrove, P. V., DiBacco, C., Hamilton, L. C., Rodríguez-Ezpeleta, N., Jeffery, N. W., ... Bradbury, I. R. (2017). Identifying patterns of dispersal, connectivity and selection in the sea scallop, *Placopecten magellanicus*, using RAD seq-derived SNPs. *Evolutionary Applications*, 10, 102–117.
- Wang, J., Street, N. R., Scofield, D. G., & Ingvarsson, P. K. (2016). Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens. *Molecular Biology and Evolution*, 33, 1754–1767. <https://doi.org/10.1093/molbev/msw051>
- Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38, 1358–1370.
- Welch, J. J., & Jiggins, C. D. (2014). Standing and flowing: The complex origins of adaptive variation. *Molecular Ecology*, 23, 3935–3937. <https://doi.org/10.1111/mec.12859>
- Wellband, K., Mérot, C., Linnansaari, T., Elliott, J. A. K., Curry, R. A., & Bernatchez, L. (2019). Chromosomal fusion and life history-associated genomic variation contribute to within-river local adaptation of Atlantic salmon. *Molecular Ecology*, 28, 1439–1459. <https://doi.org/10.1111/mec.14965>
- Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, 33, 427–440. <https://doi.org/10.1016/j.tree.2018.04.002>
- Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNP s: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, 28, 1203–1209. <https://doi.org/10.1111/mec.15066>
- Westram, A. M., Rafajlovic, M., Chaube, P., Faria, R., Larsson, T., Panova, M., ... Butlin, R. (2018). Clines on the seashore: The genomic architecture underlying rapid divergence in the face of gene flow. *Evolution Letters*, 2, 297–309. <https://doi.org/10.1002/evl3.74>

- Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16, 97–159.
- Wu, C.-I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, 14, 851–865. <https://doi.org/10.1046/j.1420-9101.2001.00335.x>
- Xuereb, A., Kimber, C., Curtis, J., Bernatchez, L., & Fortin, M. J. (2018). Putatively adaptive genetic variation in the Giant California sea cucumber (*Parastichopus californicus*) as revealed by environmental association analysis of RADseq data. *Molecular Ecology*, 27, 5035–5048.
- Yeaman, S., & Otto, S. P. (2011). Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. *Evolution*, 65, 2123–2129. <https://doi.org/10.1111/j.1558-5646.2011.01277.x>

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Cayuela H, Rougemont Q, Laporte M, et al. Shared ancestral polymorphisms and chromosomal rearrangements as potential drivers of local adaptation in a marine fish. *Mol Ecol*. 2020;29:2379–2398. <https://doi.org/10.1111/mec.15499>