Demographic history has shaped the strongly differentiated corkwing wrasse populations in Northern Europe

Morten Mattingsdal1 | Per Erik Jorde2 | Halvor Knutsen1,2 | Sissel Jentoft3 | Nils Christian Stenseth1,3 | Marte Sodeland1 | Joana I. Robalo4 | Michael M. Hansen5 | Carl André6 | Enrique Blanco Gonzalez1,7

Abstract
Understanding the biological processes involved in genetic differentiation and divergence between populations within species is a pivotal aim in evolutionary biology. One particular phenomenon that requires clarification is the maintenance of genetic barriers despite the high potential for gene flow in the marine environment. Such patterns have been attributed to limited dispersal or local adaptation, and to a lesser extent to the demographic history of the species. The corkwing wrasse (Symphodus melops) is an example of a marine fish species where regions of particular strong divergence are observed. One such genetic break occurred at a surprisingly small spatial scale ($F_{ST} \sim 0.1$), over a short coastline (<60 km) in the North Sea-Skagerrak transition area in southwestern Norway. Here, we investigate the observed divergence and purported reproductive isolation using genome resequencing. Our results suggest that historical events during the post-glacial recolonization route can explain the present population structure of the corkwing wrasse in the northeast Atlantic. While the divergence across the break is strong, we detected ongoing gene flow between populations over the break suggesting recent contact or negative selection against hybrids. Moreover, we found few outlier loci and no clear genomic regions potentially being under selection. We concluded that neutral processes and random genetic drift e.g., due to founder events during colonization have shaped the population structure in this species in Northern Europe. Our findings underline the need to take into account the demographic process in studies of divergence processes.

KEYWORDS
aquaculture, fish, landscape genetics, phylogeography, population genetics—empirical

INTRODUCTION

Many marine species present a pelagic stage during their life cycle (Hauser & Carvalho, 2008), with high potential for dispersal and gene flow. While such life cycles should generally result in panmixia and weak population divergence (Palsboll, Berube, & Allendorf, 2007), some species display genetic patterns of reproductive isolation indicative of barriers to random mating (Ravinet
et al., 2017; Storfer, Murphy, Spear, Holderegger, & Waits, 2010). The observed patterns of divergence may be characterized by: (a) isolation-by-distance, where spatially separated individuals are less likely to encounter and hence mate; (b) isolation-by-adaptation, where locally adapted populations produce maladaptive or unviable hybrids when faced with gene flow, including Dobzhansky-Muller models of hybrid incompatibility; and (c) isolation-by-colonization, where the path and colonization history across the seascapes and barriers may continue to restrict gene flow (Nadeau, Meirmans, Aitken, Ritland, & Isabel, 2016; Orsini, Vanoverbeke, Swillen, Mergeay, & De Meester, 2013; Spurgin, Illera, Jorgensen, Dawson, & Richardson, 2014).

After the last glacial maximum (~21 kya), serial colonization and founding events along the recolonization routes have shaped the biota on the northern hemisphere (Hewitt, 2000; Taberlet, Fumagalli, Wust-Saucy, & Cosson, 1998). Re-colonization has some common features among species, such as a general loss of genetic variation with increasing latitude, but the reconstructed histories tend to be quite complex and sometimes species-specific, involving glacial refugia, isolated pockets and secondary contact, as exemplified by terrestrial plants (Francois, Blum, Jakobsson, & Rosenberg, 2008; Kyrkjeide, Stenøien, Flatberg, & Hassel, 2014; Petit et al., 2002). Similarly, many marine species also carry clear genetic signals of post-glacial range-expansions (Jenkins, Castilho, & Stevens, 2018). During the last glacial maximum, cold-adapted fish species are believed to have persisted in Northern Europe, while temperate fish species, such as the wrasses, found refuge in the Mediterranean and the surrounding coast of the Iberian Peninsula (Kettle, Morales-Muñiz, Roselló-Izquierdo, Heinrich, & Øvlestad, 2011).

The genetic makeup of several temperate wrasse fish species follow this classical pattern of loss of genetic variation with increasing latitude, as seen for ballan wrasse, *Labrus bergylta*, (Almada et al., 2017) and corkwing wrasse, *Symphodus melops* (Robalo et al., 2012). The corkwing wrasse has emerged as a particularly interesting case due to two substantial genetic breaks, across the North Sea ($F_{ST} = 0.15$) and over a narrow coastal barrier with unsuitable sandy habitats (~60 km; $F_{ST} = 0.11$) in southwestern Norway (Blanco Gonzalez, Knutsen, & Jorde, 2016). In addition, as the corkwing wrasse is currently exploited as “cleaner fish” in aquaculture (Blanco Gonzalez & de Boer, 2017), this conspicuous genetic break demands clarification, in particular as individuals are translocated across the genetic break and members of the two populations can interbreed (Blanco Gonzalez et al., 2019; Faust, Halvorsen, Andersen, Knutsen, & Andre, 2018).

Genome-wide patterns of differentiation are particularly informative in elucidating if reproductive isolation is driven by directional selection (Feder & Nosil, 2010) or random genetic drift (Nielsen, 2005). Somewhat simplified, a classical strong selective sweep should display a local genomic signal, with “hitchhiking” neutral markers in proximity of the beneficial variant (Feder & Nosil, 2010). On the other hand, isolation-by-colonization should demonstrate a global and random pattern of genome-wide differentiation, a result of the stochastic fluctuations of variant frequencies imposed by for instance a founding event (Nielsen, 2005).

While population genetic methods are typically used to investigate patterns of population divergence, analyses using demographic inference to explicitly test different scenarios of divergence are rarely undertaken (Rougemont & Bernatchez, 2018). Here, we make use of whole genome resequencing methods to analyze the divergence between populations of corkwing wrasse in Northern Europe and to investigate demographic histories and putative patterns of reproductive isolation of this rocky shore marine fish.

## 2 | MATERIALS AND METHODS

### 2.1 | Samples and genotyping

Sixty-five corkwing wrasses were sampled from eight coastal locations from three regions: the British Isles, western and southern Scandinavia (Table 1). Samples from southern Norway were collected by beach seine, while those from the west coast of Norway, Sweden and the British Isles were collected by fish pots, as described in (Blanco Gonzalez et al., 2016). Muscle tissues were taken from fresh or frozen specimens and stored in 96% ethanol prior to DNA extraction. Total genomic DNA was extracted with the DNeasy kit (Qiagen) or the E.Z.N.A. Tissue DNA kit (Omega Bio-Tek) and resuspending the DNA in TE buffer. The extractions

### TABLE 1 Regional groupings of corkwing sampling locations by region, location, code, sampling year, latitude, longitude, sample size (n) and number of variable single nucleotide polymorphisms (SNPs) per site (minor allele count > 1)

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>Code</th>
<th>Year</th>
<th>Latitude</th>
<th>Longitude</th>
<th>n</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>British Isles</td>
<td>Ardtoe</td>
<td>ARD</td>
<td>2010</td>
<td>N 56.40</td>
<td>W 5.50</td>
<td>7</td>
<td>704,073</td>
</tr>
<tr>
<td>Western Scandinavia</td>
<td>Smola</td>
<td>SM</td>
<td>2015</td>
<td>N 63.32</td>
<td>E 8.11</td>
<td>8</td>
<td>592,767</td>
</tr>
<tr>
<td></td>
<td>Norheimsund</td>
<td>NH</td>
<td>2014</td>
<td>N 60.39</td>
<td>E 6.48</td>
<td>8</td>
<td>584,422</td>
</tr>
<tr>
<td></td>
<td>Stavanger</td>
<td>ST</td>
<td>2015</td>
<td>N 59.01</td>
<td>E 5.56</td>
<td>8</td>
<td>597,438</td>
</tr>
<tr>
<td>Southern Scandinavia</td>
<td>Egersund</td>
<td>EG</td>
<td>2008</td>
<td>N 58.45</td>
<td>E 5.53</td>
<td>8</td>
<td>480,354</td>
</tr>
<tr>
<td></td>
<td>Arendal</td>
<td>AR</td>
<td>2014</td>
<td>N 58.41</td>
<td>E 8.74</td>
<td>8</td>
<td>431,556</td>
</tr>
<tr>
<td></td>
<td>Tvedestrand</td>
<td>TV</td>
<td>2010</td>
<td>N 58.62</td>
<td>E 9.06</td>
<td>8</td>
<td>440,905</td>
</tr>
<tr>
<td></td>
<td>Gullmarsfjorden</td>
<td>GF</td>
<td>2008</td>
<td>N 58.18</td>
<td>E 11.32</td>
<td>10</td>
<td>463,654</td>
</tr>
</tbody>
</table>
were analyzed with Qubit (Thermo Fisher Scientific) for assessment of the DNA quality and concentration. After normalization to 1,200 ng with Qiagen EB buffer (10 mM Tris-cl; pH = 8.0) the samples were fragmented to ~350 bp using a Covaris S220 (Life Technologies). Library construction was performed using the Illumina TruSeq DNA PCR Free protocol and checked on Bioanalyzer High sensitivity chip and Tapestation (both Agilent) followed by Kapa Biosystems qPCR assay for Illumina libraries quantification.

Whole-genome resequencing was conducted on the Illumina HiSeq platform, generating 2 × 125 bp paired-end reads to an average depth of ~9.16× per sample (595× in total across the 65 sample libraries). The mean read insert size across samples was 347 (range: 246–404). Reads were mapped to the corkwing wrasse reference genome assembly (Mattingsdal et al., 2018) using bwa-mem (v0.7.5a; Li & Durbin, 2009) followed by duplicate removal by Picard (http://broadinstitute.github.io/picard/). Single nucleotide polymorphisms (SNPs) were called across all samples with freebayes (v1.0.2-33; Garrison & Marth, 2012), using the following quality control criteria: (a) quality >40; (b) minimum and maximum read depth of ×4 and ×30; (c) maximum 5% missing genotypes; (d) minimum minor allele count of 3 (MAF >2%). Two data sets were made: (a) all SNPs with ancestral states and (b) a thinned data set keeping random SNPs equally spaced by 10,000 bp and excluding rare variants (MAF >2%, thinned with “−bp-space 10,000”).

The ancestral allele states were inferred using whole-contig alignments between the corkwing and ballan wrasse (L. bergylta) genome assemblies (Lie et al., 2018; Mattingsdal et al., 2018) constructed by last (v923; Frith, Hamada, & Horton, 2010); both species are members of the Labridae family. First, the genomes were indexed specifying the “YASS” and “R11” options, optimizing for long and weak similarities and masking low-complexity regions. Then, a pairwise genome-wide alignment between corkwing- and ballan wrasses was made, setting minimum E-value to 0.05 and maximum matches per query position = 100. The “last-split” function was run twice to ensure 1-1 alignments. The multiple alignments were converted to bam format and SNP positions in the corkwing wrasse genome used to extract “genotypes” in the corkwing and ballan wrasse alignment using samttools and bcftools (Li et al., 2009). The inferred ancestral states were manually controlled and plink v1.90b3.40 (Purcell et al., 2007) was used to annotate the ancestral state as the reference allele. Missing data were imputed and phased using beagle default settings (Browning & Browning, 2013). To elucidate demographic relationships between the populations, we searched for identical-by-decent (IBD) haplotypes inferred by beagle (Browning & Browning, 2013), which accounts for haplotype phase uncertainty.

### 2.2 Population structure and admixture

Single nucleotide polymorphism-wise Fst values between populations were calculated using the Fst of Weir and Cockerham (1984) implemented in vcftools (v0.1.13; Danecek et al., 2011). Patterns of population structure were investigated by Multidimensional scaling (MDS) analysis and inbreeding coefficients using plink v1.90b3.40 (Purcell et al., 2007). Proportion of ancestry for each individual, Q, for each putative ancestral population, K, was estimated using admixture (v1.3.0; Alexander, Novembre, & Lange, 2009), making use of the integrated five-fold cross-validation scheme for 10 iterations each for K = 2–6, each using different random seed.

In an idealized diploid population, the identity-by-descent (IBD) haplotype lengths are exponentially distributed in an organism with a mean of 1/(2×generations) Morgans (Thompson, 2013). Therefore, IBD lengths and their distribution are of interest in inferring the ancestry of populations. Pairwise IBD segments between individuals were estimated by beagle (v. 08 Jun17; Browning & Browning, 2013) using a minimum segment length of 0.01 cM, LOD score >3, overlap = 100 and ibdlim = 40. To assess the extent and length of IBD segment sharing between populations, a subset of seven random individuals from the most distant sampling locations (ARD, SM and GF; cf. Table 1 for sample information) were selected.

Gene flow and diversity between locations relative to geographical distance were estimated using eems (Petkova, Novembre, & Stephens, 2016), which models effective rates of gene flow using the pairwise dissimilarity matrix calculated by the embedded bed2difs tool. The number of demes was set to 300 and several iterations were performed using default settings (100,000 burnin iterations, 200,000 MCMC iterations, 9,999 thinning interval) to ensure consistent and converging results. To detect admixture, the f3-statistic (Reich, Thangaraj, Patterson, Price, & Singh, 2009), implemented in treemix/threepop (v 0.1; Pickrell & Pritchard, 2012), was used as a formal test for admixture between all population triplets using a block size of 200.

### 2.3 Gene flow

Gene flow across the genetic break was estimated by calling discrete local ancestry using pcdmix (Brisbin et al., 2012). Individuals from the ST and EG sites defined the admixed sample (N = 16) and the remaining samples from Scandinavia used as "South" (N = 24) and "West" (N = 16) ancestors. Simplified, pcdmix uses a PCA based algorithm of phased SNPs in a sliding window projecting the admixed samples using PCA loadings from the ancestral populations and calls local ancestry. Here, we used our previously phased dense SNPs data set and specified a fixed window size of 50 Kbp (-wMb 0.05). To reduce bias introduced by artificial breakpoints by our genome assembly, we used SNPs located on contigs >N50 (461 Kbp). Fifty Kbp bins of southern origin were denoted by "0" and bins with a western origin “1”. We inserted a flag “9” to signify breaks introduced by a new contig. The r function "re" was used to count the length of consecutive “0” and “1” for each haplotype for each contig (R Core Team, 2017).
Demographic histories were estimated using two Markovian coalescent methods, PSIMC (v 0.6.5-67; Li & Durbin, 2011) and SMCP++ (v1.12.1; Terhorst, Kamm, & Song, 2017). In the PSIMC analysis, the minimum read depth was increased to \(\times 6\) (Alex Buerkle & Gompert, 2013) and a maximum missing rate increased to 20% (Nadachowska-Bryzyska, Burri, Smeds, & Ellegren, 2016). Population substructure can induce spurious signals of population bottlenecks and expansions (Mazet, Rodriguez, Grusea, Boitard, & Chikhi, 2016), so the analyses were performed separately for each of the three regions excluding possible admixed samples. In the PSIMC analysis, one random individual from each of the most geographically distant locations were selected (ARD15, SM111 and GF01) using a similar approach to individual from each of the most geographically distant locations (Mazet, Rodriguez, Grusea, Boitard, & Chikhi, 2016), so the analyses were performed separately for each of the three regions excluding possible admixed samples. In the PSIMC analysis, one random individual from each of the most geographically distant locations were selected (ARD15, SM111 and GF01) using a similar approach to the one described in (Barth, Damerau, Matschiner, Jentoft, & Hanel, 2011), setting minimum and maximum read depth at six and 30 and base quality \(>30\). Then the resulting fastq files were converted to PSIMC input format specifying quality threshold \(\times 20\). PSIMC was run using default parameters, followed by 100 bootstraps.

For the SMCP++ analysis four random individuals from each region were combined into a composite likelihood for each population (South: GF01, GF49, TV69, TV70. West: NH61, NH60, SM111, SM114 and the British Isles: ARD20, ARD21, ARD18 and ARD15). Only SNPs situated on contigs \(>50\) (461,652 bp) were included. SMCP++ was run using the options “thinning 50, unfold, knots 30”, specifying an unfolded frequency spectrum, reducing LD (approx. SNP density after thinning 1 SNP pr. 25,350 bp) and fixing the number of spline knots used in smoothing. For both PSIMC and SMCP++ the site mutation rate was set to \(1 \times 10^{-8}\) and generation time to 3 years (Halvorsen et al., 2016, 2017; Uglem, Rosenqvist, & Wasslavik, 2000).

A diffusion approximation method was implemented, DADI (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009), to determine the most likely population history scenario and its coalescent parameters. Four classical models were tested: (a) Strict Isolation (SI); (b) Isolation with Migration (IM); (c) Ancient Migration (AM); and (d) Secondary Contact (SC). The scenario obtaining the best Akaike information criterion (AIC) was deemed the most probable model. To reduce the effect of linkage disequilibrium (LD), we used the thinned data set. An optimization function (Optimize_Functions. Optimize_Routine) which sequentially refines the perturbation of parameters, was used (Portik et al., 2017). The optimization function included four rounds each with 10, 20, 30 and 40 replications, increasing maximum iterations (three, five, 10 and 15) and decreasing fold in parameter generation (3, 2, 2 and 1), resulting in 100 replications. We looped the aforementioned algorithm 10 times, yielding 1,000 local minima from the four models. The best model and its parameters were subjected to a goodness-of-fit test (Optimize_Functions_GOF) generating simulated parameters and these were used to assess the significance of the empirical parameters (Portik et al., 2017). Coalescent parameters were converted as follows: ancestral effective population size \(N_e\) was calculated by \(N_e = \theta/4\mu\), where \(\theta\) was the scaled population parameter, \(\mu\) was the mutation rate per site and per generations, and \(l\) the length of analyzed sequence. Thinning the data set to one SNP per 10 kbp effectively reduced the length of the analyzed sequence by a factor of \(18\), resulting in 35 Mbps. Migration was calculated as \(m = M/2N_e\) and time in years as \(t = 2TN_e\), using \(g = 3\) as generation time (Halvorsen et al., 2016, 2017; Uglem et al., 2000).

**2.5 Local patterns of differentiation and adaptation**

Selective sweeps should display localized, elevated and linked \(F_{ST}\) values between populations (Sabeti et al., 2006). SNP-wise Weir and Cockerham’s \(F_{ST}\) values were calculated by VCFTOOLS (v0.1.13; Danecek et al., 2011). In addition, the \(F_{ST}\) outlier test implemented in BAYESCAN was conducted (Foll & Gaggiotti, 2008) using default settings. Finally, a haplotype based test, HAPFLK (Fariello, Boitard, Naya, SanCristobal, & Servin, 2013), was also used. First, we calculated the Reynolds distance matrix using the thinned data set. No outgroups were defined, 20 local haplotype clusters (\(K = 20\)) were specified and the HAPFLK statistic computed using 20 EM iterations (\(nfit = 20\)). Statistical significance was determined though the script “scaling_chi2_hapflk.py”. To adjust for multiple testing, we set the false discovery rate (FDR) level to 5% using \(qvalue\) (Storey, Bass, Dabney, & Robinson, 2019). Samples from the western and southern locations were grouped into their respective groups (South, \(N = 34\) and West, \(N = 24\)) in all three tests.

**3 RESULTS**

**3.1 Genotyping**

The whole genome resequencing analysis generated a total of 3,048 million reads. Approximately 0.8% of these reads were duplicated and thus discarded. Of the remaining reads in the merged data set (3,024,360,818 reads), 97.19% mapped to the genome, and 93.27% were correctly paired. The mean depth of coverage per individual was \(>9.16\). In total, 13.2 million sequence variants were detected, of which, 5.55 million had a quality metric \(>40\). After applying min/max depth and maximum missing filters, 2.69 million variants were kept, of which 2.25 million SNPs were biallelic. We successfully inferred the ancestral state of 1,210,723 SNPs. Excluding rare SNPs, minor allele count (MAC) \(>3\), resulted in 836,510 SNPs. We denote this as the “all SNPs” data set. This highly dense data set was further reduced to keeping one SNP per 10 Kbp, using VCFTOOLS (“bp-thin 10,000”), yielding a reduced data set of 50,130 SNPs, denominated as the “thinned data set”. Due to a relatively low minimum read depth filter (\(\times 4\)) it is likely that the proportion of heterozygous SNPs is underestimated, which can introduce a systematic error especially in windowed analyses which rely on breakpoints like IBD haplotypes (Meynert, Bicknell, Hurles, Jackson, & Taylor, 2013).
3.2 Population structure and sequential loss of genetic variation

The number of SNPs within each sampling location suggests a pattern of sequential loss of diversity among regions, initially from the British Isles to western Scandinavia and followed by a further reduction to southern Scandinavia (Table 1). Of the 894 k SNPs (MAC >3 across all samples), ~704 k were found to be polymorphic (MAC >1) in the British Isles, ~590 k polymorphic in western Scandinavia (MAC >1) and ~450 k polymorphic in southern Scandinavia (MAC >1). We chose ARD (n = 7), SM (n = 8) and TV (n = 8) as representative samples to count the overlap and unique SNPs between populations. Of the 704 k SNPs detected in the British Isles, 69% (485 k) were found in the West (SM) and 51% (360 k) in the South (TV). The proportion of unique SNPs in the British Isles, western and southern regions were 18%, 6% and 3%, respectively. A total of 327 k SNPs (39%) were found to be polymorphic in all three populations. The dramatic loss of genetic variation in Scandinavia as compared to the British Isles, especially in southern Scandinavia, was also revealed by the pairwise FST estimates (Table S1).

The simulation of effective migration surfaces (Figure 1) and MDS plot (Figure 2) identified three distinct groups corresponding to the British Isles, southern and western Scandinavia, as previously reported (Blanco Gonzalez et al., 2016; Knutsen et al., 2013), with some evidence of contact between the western and southern populations at the ST-EG site of south-western Norway. The ADMIXTURE analysis suggested K = 3, as the most likely number of ancestral populations with lowest mean cross validation of 0.368. The mean cross validation error for each K-value were, K2 = 0.378, K3 = 0.368, K4 = 0.424, K5 = 0.461 and K6 = 0.471 (for K2 and K3, see Figure 3). The results from admixture added further evidence for some gene flow across the contact zone between southern and western Scandinavian sample localities. The f3-statistic test for admixture revealed that EG had the most negative f3-statistic and Z-score in any combination with western (SM, NH, ST) and southern samples (AR, TV, GF), suggesting the EG population as a candidate admixed population in Scandinavia (mean: −0.0024). The inbreeding coefficient (“plink –het”) also revealed that the EG site was somewhat less homozygous compared to the other southern Scandinavian sites (Figure S1).

3.3 Stochastic genome-wide differentiation

Searching for localized signals of differentiation and candidate regions of selection we explored the genome-wide pattern of variation between the two Scandinavian populations. The analysis revealed a strong and global genome-wide pattern of differentiation (Figure S2). Across the genome five regions showed FST values >0.9 and 32 regions FST >0.8. The haplotype based test, HAPFLK, returned a similar pattern but with more distinct candidate regions, albeit none passed the threshold for statistical significance (q-value <0.05). Testing for outliers between the western and southern populations, BAYESCAN results yielded two significant loci possibly under diversifying selection, SYMME_00001686_632632 and SYMME_00023564_399441 (Figure S3). The frequency of these two loci is 0.72 and 0.89 in the western population, and both loci were monomorphic in the southern population. The most differentiated SNPs can be informative in population discrimination and are listed in Table S2.

FIGURE 1  Map showing sampling locations (see also Table 1) together with estimation of effective migration surfaces inferred by EEMS, where brown colour indicates a reduction and cyan colour indicates an increase in gene flow on the log10 scale. Note that gene flow between differentiated populations would appear as a barrier (for example between ST and EG)

FIGURE 2  Multidimensional scaling (MDS) plot using PLINK and the thinned data set (SNPs = 50,130). Individuals sampled in the British Isles (yellow), western Scandinavia (red) and southern Scandinavia (blue). SNP, single nucleotide polymorphism
3.4 | Gene flow across the genetic break

We used the default parameters in `pcadmix` thereby removing SNPs in high LD ($r^2 > .8$) and monomorphic SNPs in the ancestral samples. Of the 501,177 SNPs located on large contigs, 123,831 SNPs passed `pcadmix` filters and were used for inference of local ancestry. They were located on 343 contigs, representing half the genome of the species (307 Mbp). Approximately 21.7 SNPs remained per bin of 50 Kbp ($N_{bins} = 5,695$), a SNP quantity per bin recommended in the `pcadmix` manual. A total of 27% of the genetic composition in the EG population was classified as "western" and 13% of the genetic composition in the ST sample was classified as "southern". The overall mean length of consecutive western haplotypes in EG was 9.28 bins or 464 Kbp ($SD = 7.2$, median = 9, mean bins = 246). The EG population has thus both longer and more regions of western origin than the ST population has of southern origin, clearly demonstrating introgression from the West into the South (Figure 4a). Some EG individuals appeared highly admixed (EG21 and EG24) with a 50.1% and 46.7% western ancestry, also suggested in the MDS plot (Figure 2) and admixture graph (Figure 3). Inspecting these individuals as potential F1 hybrids, revealed numerous heterozygous bins (~40%), but approximately ~60% of bins were homozygous from either southern or western ancestry, suggesting that these individuals were not F1 hybrids but instead admixed individuals. The ancestral calls for the ST and EG individuals can be obtained through http://doi.org/10.6084/m9.figshare.9741641.v1.

3.5 | Demographic history and founding events

The analysis of `psmc` and `smc++` is based on the "all SNP" data set, while `dadi` analysis was conducted using the "thinned SNPs" data set. Demographic history estimated by `psmc` suggests that all populations overall reduction of effective population size ($N_e$) in all populations during the last ice age approximately 50 kya (Figure 5a). The population in the British Isles experienced a more recent recovery (~5 kya), while the decline of $N_e$ continued in the Scandinavian wrasses. There is also a distinct phase shift between the Scandinavian and British Isles population. The `psmc` has limitations in inferring recent histories, as addressed by `smc++` (Terhorst et al., 2017). The inferred histories of `smc++` are remarkably similar (Figure 5b), suggesting that all populations have experienced a decline at different points in time, possibly reflecting sequential founding events. The most pronounced reduction of $N_e$ was in southern Scandinavian, approximately 10 kya (blue line, Figure 5b). `smc++` offers flexibility, thus we experienced variation in the results (data not shown) depending on the options used. However, some patterns remained constant regardless of settings and included: (a) decline in all populations started approximately 30 kya, first in the British Isles, then in western and southern Scandinavia; (b) the magnitude of the decline was smallest in the British Isles, followed by West and finally largest in southern Scandinavia. `smc++` seems to present limitations to detect the two independent declines presumably experienced by the southern corkwing population, due to the algorithmic smoothing of inferred history. Even though `smc++` allows a folded frequency spectrum, we experienced a one order of magnitude improvement of the log likelihood by inferring the ancestral states of SNPs and specifying an unfolded spectrum during the simulations.

The isolation with migration model was the most likely scenario for the three comparisons analyzed in `dadi`. Among all 2D (two populations) models, the secondary contact projection yielded the best log likelihood and AIC statistic (Table S3). Nevertheless, we observed increased residuals on the rare frequency range, suggesting difficulties in modelling the loss of variation (Figure S5). We converted the coalescent values for the best model, the West and South Scandinavia secondary contact, resulting in an ancestral population ($N_{ref}$) of 384. The size of the populations after the split were 3,980 and 1,275 for the West and South Scandinavia, respectively. The total time of divergence was $T = 2*N_{ref}*generation time*(T1 + T2) = 68,659$ years of which the first 63,102 were spent...
in isolation while during the most recent 5,527 years the populations have experienced gene flow. The estimated migration rate was quite low (i.e., $m_{12}/2N_{ref}$) as the proportion of new migrants are $4.1 \times 10^{-4}$ in the southern population and $6.7 \times 10^{-4}$ in the western population (Table S3).

4 | DISCUSSION

Several marine species display cryptic population structure in parts of their range, and uncovering the underlying mechanisms behind such genetic breaks are often nontrivial. Using whole genome sequencing and analyses of demographic history, we clarify the genetic underpinnings of reproductive isolation and differentiation of a marine fish, the corkwing wrasse. As a result of the cumulative evidence from our analyses, a clear picture of genetic drift has emerged as the dominant evolutionary force shaping contemporary patterns of population differentiation in corkwing wrasse.

The first line of evidence is the clear geographical pattern of global loss of genetic variation (number of polymorphic SNPs per sampling location, Table 1) and the increase in homozygosity from the British Isles, to western Scandinavia and finally to southern
Scandinavia (Figure S1). The loss of SNPs is dramatic, as ~700 k SNPs detected in the British Isles are reduced to ~590 k (~16% less) SNPs in western Scandinavia with a further reduction to ~450 k (~35% less) SNPs in southern Scandinavia (Table 1), suggesting the direction and sequence of possible founding events to follow the British Isles-western Scandinavia-southern Scandinavia route, as previously suggested (Robalo et al., 2012). The pattern of genome-wide divergence ($F_{ST}$ and $\text{hapflk}$; Figure S2) did not show any fixed variation or clearly localized genomic regions that may suggest hard selective sweeps. Instead it showed a stochastic pattern of differentiation, probably imposed by strong drift (Figure 5), indicative of historical events shaping contemporary populations. Distinguishing between the genomic effects of bottlenecks with that of selective sweeps remains unresolved and are even discouraged (Pavlidis & Alachiotis, 2017; Poh, Domingues, Hoekstra, & Jensen, 2014). That being said, a polygenic model of adaptation remains a possibility although notoriously hard to detect and intrinsically difficult to distinguish from drift and population structure (Hollinger, Pennings, & Hermisson, 2019). The patterns of sequential loss of variation and lack of any missing fixed SNPs are also demonstrated in the site-frequency-spectra (Figure S4).

A second line of evidence is associated with the reduction of the effective population sizes in line with founding events detected using the sequentially Markovian coalescent methods in $\text{psmc}$ and $\text{smc++}$ (Figure 5b). Our results suggest that corkwing wrasse colonized western Scandinavia about 11 kya, possibly from the British Islands. Based on the number of biallelic and heterozygous SNPs and inbreeding coefficients, Stavanger site (ST; Table 1, Figure 1) may be close to the point of entry into Scandinavia. Then, over a stretch of newly formed coastline, southern Scandinavia was subsequently (~10 Kya) colonized from the western Scandinavian population. The post-glacial colonization pattern in Scandinavia is similar to the colonization routes suggested for other marine species which depend on rocky habitats, such as seaweed, invertebrates and other fishes (Almada et al., 2017; Evankow et al., 2019; Hoarau, Coyer, Veldsink, Stam, & Olsen, 2007; Kettle et al., 2011; Maggs et al., 2008; Quintela et al., 2016). The colonization of Scandinavia ~10 kya, coincides with the deglaciation in western Norway (Stroeven et al., 2016). The demographic history date estimates inferred by the two Markovian approaches should be considered approximations, as the simulations rely on accurate generation time, mutation rates and sex ratio (Spence, Steinrücken, Terhorst, & Song, 2018). These values are intrinsically challenging for a species like corkwing wrasse, considering the variance in reproductive behaviour and generation time displayed by the species along the latitudinal gradient covered in this study (Halvorsen et al., 2017).
The fact that psmc and smc++ do not adjust for periods of gene flow between populations and assumes clean population splits demands some care when interpreting changes in effective population sizes, and validation of findings using other methods are encouraged (Beichman, Phung, & Lohmueller, 2017). The method using diffusion approximations of the joint frequency spectrum implemented in dadi is frequently used to model complex scenarios of gene flow between populations (Rougmont et al., 2017; Tine et al., 2014). Here, we only tested simple scenarios, as more complicated models (Rougex, Bernatchez, & Gagnaire, 2017) failed to converge and tended to produce artificial fits and parameters (data not shown). Using the site-frequency spectrum (SFS), the model which best fitted the empirical spectrum of all three comparisons was the secondary contact model (British Isles vs. western Scandinavia, British Isles vs. southern Scandinavia and western Scandinavia vs. southern Scandinavia, Table S3).

The third line of evidence is in the distribution of shared haplotypes (identical-by-decent) between the populations which corroborate the findings from the demographic history (Figure S6; Harris & Nielsen, 2013). Mean length of shared haplotypes was longer between the two Scandinavian populations, compared to the mean length between the British Isles and either Scandinavian population, suggesting a more recent split between the Scandinavian populations. The frequency of shared haplotypes also indicated the sequential loss of shared haplotypes and the direction of founding events.

Finally, we detected ongoing gene flow in both directions across the genetic break (Figure 4b). The contact was asymmetrical with increased gene flow from the West into the South. By using half the genome (>N50) and bin size of 50 Kbp, we detected 1,568 bins of a total of 5,694 bins (27.5%) were of western origin in the EG population, and 785 bins in the ST population were of southern origin (13.8%).

Gene flow across genetic breaks can be an indicator of secondary contact after divergence (Sexton, McIntyre, Angert, & Rice, 2009). This strongly suggests that the genetic break is a hybrid zone with ongoing secondary contact after divergence. Findings from the IBD analysis suggests that the southern population descends from the western population. Our limited geographical sampling scheme does not, however, exclude the scenario of a ring-like colonization pattern, possibly surrounding the Norwegian Trench, where the southern population could descend from an unstudied population from the coastline in the southern parts of the North Sea. The persistence of the break remains intriguing and suggests that the contact is recent or actively selected against (Abbott et al., 2013).

In conclusion, our findings shed new light on the dynamics underlying the presence of two genetic breaks of this species in Northern Europe (Blanco Gonzalez et al., 2016; Knutsen et al., 2013; Robalo et al., 2012). It also serves to remind us that more simple scenarios involving sequential recolonization and associated founder events combined with secondary contact could underlie instances of strong genetic breaks, without having to invoke more elaborate scenarios of selection and environmental adaptation (Hewitt, 1999; Ravinet et al., 2017; Schluter & Conte, 2009). Yet, while we could associate contemporary patterns of genetic differentiation to historical demographic events rather than adaptation, isolating mechanisms between western and southern Scandinavian populations still need further clarification, including a possible polygenic model of adaptation. In conclusion, corkwing wrasse could become an interesting future model for complementing and exploring the full span of possible dynamics that can lead to distinct contact zones, ranging from selectively neutral population history and structure to strong selection.

ACKNOWLEDGEMENTS
The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo supported by the Research Council of Norway and the Southeastern Regional Health Authorities. Computational work was performed on the Abel Supercomputing Cluster operated by the University of Oslo IT-department. This work was supported by funds from the Norwegian Research Council (grants 234328 and 280453), the Swedish Research Council FORMAS and by the Centre for Marine Evolutionary Biology at the University of Gothenburg (CeMEB).

AUTHOR CONTRIBUTIONS
The study was conceived and designed by P.E.J, H.K, S.J and E.B.G. Sample collection were performed by P.E.J, H.K, C.A and E.B.G and genomic analyses preformed by M.M. The manuscript was written by M.M with contributions from all authors. All authors read, revised and approved the manuscript.

ORCID
Morten Mattingsdal https://orcid.org/0000-0003-4440-0324
Per Erik Jorde https://orcid.org/0000-0001-5515-7257
Sissel Jentoft https://orcid.org/0000-0001-8707-531X
Michael M. Hansen https://orcid.org/0000-0001-5372-4828
Enrique Blanco Gonzalez https://orcid.org/0000-0002-2631-2331

DATA AVAILABILITY STATEMENT
Sequence reads are available through NCBI sequence read archive by accession number PRJNA354496. SNPs (Mattingsdal, 2019) can be obtained through: doi.org/10.6084/m9.figshare.7570907. v1.

REFERENCES


**SUPPORTING INFORMATION**

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