DR. MORTEN MATTINGSDAL (Orcid ID : 0000-0003-4440-0324)

DR. PER ERIK JORDE (Orcid ID : 0000-0001-5515-7257)

DR. SISSEL JENTOFT (Orcid ID : 0000-0001-8707-531X)

PROF. MICHAEL M. HANSEN (Orcid ID : 0000-0001-5372-4828)

DR. ENRIQUE BLANCO GONZALEZ (Orcid ID : 0000-0002-2631-2331)

Article type : Original Article

Corresponding author mail id : morten.mattingsdal@uia.no

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

Morten Mattingsdal[1], Per Erik Jorde[2], Halvor Knutsen[1,2], Sissel Jentoft[3], Nils Christian Stenseth[1,3], Marte Sodeland[1], Joana I. Robalo[4], Michael M. Hansen[5], Carl André[6], Enrique Blanco Gonzalez[1,7]

1 Centre for Coastal Research, Department of Natural Sciences, University of Agder, Kristiansand, Norway

2 Institute of Marine Research, Flødevigen, Norway

3 Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway

4 Marine and Environmental Sciences Centre, ISPA Instituto Universitário de Ciências Psicológicas, Sociais e da Vida, Rua Jardim do Tabaco,34, Lisboa, Portugal

5 Department of Bioscience, Aarhus University, Ny Munkegade 114, DK-8000 Aarhus C, Denmark

6 Department of Marine Sciences-Tjärnö, Göteborg University, 452 96 Strömstad, Sweden

7 Norwegian College of Fishery Science, UiT The Arctic University of Norway, Tromsø, Norway

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 occur at a surprisingly small spatial scale (F_{ST} ~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 re-colonization route, can explain the present population structure of the corkwing wrasse in the northeast Atlantic. While the divergence across the break is strong, we detect ongoing gene flow between populations over the break suggesting recent contact or negative selection against hybrids. Moreover, we find few outlier loci and no clear genomic regions potentially being under selection. We conclude 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 demographic process into account in studies of divergence processes.

1 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: 1) isolation-by-distance, where spatially separated individuals are less likely to encounter and hence mate; 2) 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 3) isolation-by-colonization, where the path and colonization history across the seascape 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 re-colonization routes have shaped the biota on the northern hemisphere (G. 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; Kyrkjeeide, 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, & Vøllestad, 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 (E. Blanco Gonzalez, Knutsen, & Jorde, 2016). In addition, as the corkwing wrasse is currently exploited as “cleaner fish” in aquaculture (Enrique 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 (E. 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 re-sequencing 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 & Methods

2.1 Samples and genotyping

Sixty-five corking 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 (E. 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, Hilden, Germany) or the E.Z.N.A. Tissue DNA kit (Omega Bio-Tek, Norcross, GA) and re-suspending the DNA in TE buffer. The extractions were analyzed with Qubit (Thermo Fisher Scientific) for assessment of the DNA quality and concentration. After normalization to 1200 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, USA). Library construction was performed using the Illumina TruSeq DNA PCR Free protocol and checked on Bioanalyser High sensitivity chip and Tapestation (both Agilent) followed by Kapa Biosystems qPCR assay for Illumina libraries quantification.

Whole-genome re-sequencing was conducted on the Illumina HiSeq platform, generating 2 × 125 bp paired-end reads to an average depth of ~9.16x per sample (595x 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: 1) quality > 40; 2) minimum and maximum read depth of x4 and x30; 3) maximum 5% missing genotypes; 4) minimum minor allele count of 3 (MAF > 2%). Two datasets were made: 1) all SNPs with ancestral states; and 2) a thinned dataset keeping random SNPs equally spaced by 10,000 bp and excluding rare variants (MAF >2%, thinned with “–bp-space 10000”).

The ancestral allele states were inferred using whole-contig alignments between the corkwing and ballan wrasse (*Labrus 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 SAMtools 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

SNP-wise F_{ST} values between populations were calculated using Weir & Cockerham's F_{ST} (Weir & 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 5 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 ibdtim=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 bed2diffs tool. The number of demes was set to 300 and several iterations were performed using default settings (100,000 burn-in iterations, 200,000 MCMC iterations, 9999 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 PCAdmix (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, PCAdmix 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 dataset 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). 50 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 “rle” was used to count the length of consecutive “0” and “1” for each haplotype for each contig (R Core Team, 2017).

2.4 Demographic history

Demographic histories were estimated using two Markovian coalescent methods, PSMC (v 0.6.5-r67) (Li & Durbin, 2011) and SMC++ (v1.12.1) (Terhorst, Kamm, & Song, 2017). In the PSMC analysis, the minimum read depth was increased to x6 (Alex Buerkle &

Gompert, 2013) and a maximum missing rate increased to 20% (Nadachowska-Brzyska, 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 3 regions excluding possible admixed samples. In the PSMC 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, 2017), setting minimum and maximum read depth at 6 and 30 and base quality > 30. Then the resulting fastq files were converted to PSMC input format specifying quality threshold >20. PSMC was run using default parameters, followed by 100 bootstraps.

For the SMC++ 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 > N50 (461,652 bp) were included. SMC++ 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 fixating the number of spline knots used in smoothing. For both PSMC and SMC++ the site mutation rate was set to $1x10^{-8}$ and generation time to 3 years (Halvorsen et al., 2017; Halvorsen et al., 2016; 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: 1) Strict Isolation (SI); 2) Isolation with Migration (IM); 3) Ancient Migration (AM); and 4) 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 dataset. An optimization function (Optimize_Functions.Optimize_Routine) which sequentially refines the perturbation of parameters, was used (Portik et al., 2017). The optimization function included 4 rounds each with 10, 20, 30 and 40 replications, increasing maximum iterations (3, 5, 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 1000 local minima from the four models. The best model and its parameters was subjected to a goodness-of-fit test (Optimize_Functions_GOF) generating simulated parameters and using these to assess the significance of the empirical parameters (Portik et al., 2017). Coalescent parameters were converted as follows: ancestral effective population size (Ne) was calculated by $Ne = \theta / 4\mu L$, where θ is the scaled population parameter, μ is mutation rate per site and per generations, and L is the length of analyzed sequence. Thinning the dataset to 1 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 $m = M / 2Ne$ and time in years as $t = 2TNe \times g$, using g = 3 as generation time (Halvorsen et al., 2017; Halvorsen et al., 2016; 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 dataset. 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/R (Storey JD, Bass AJ, Dabney A, & D, 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 re-sequencing analysis generated a total of 3048 million reads. Approximately 0.8% of these reads were duplicated and thus discarded. Of the remaining reads in the merged dataset (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 x9.16. 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 bi-allelic. We successfully inferred the ancestral state of 1,210,723 SNPs. Excluding rare SNPs, MAC (Minor Allele Count) >3, resulted in 836,510 SNPs. We denominate this as the “all SNPs” dataset. This highly dense dataset was further reduced to keeping one SNP per 10 Kbp, using VCFtools (“bp-thin 10000”), yielding a reduced dataset of 50,130 SNPs, denominated as the “thinned dataset”. Due to a relatively low minimum read depth filter (x4) 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 894k SNPs (MAC>3 across all samples), ~704k were found to be polymorphic (MAC>1) in the British Isles, ~590k polymorphic in western Scandinavia (MAC>1) and ~450k 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 704k SNPs detected in the British Isles, 69% (485k) were found in the West (SM) and 51% (360k) 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 327k 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 F_{ST} estimates (Supporting Information Table S1).

The simulation of effective migration surfaces (Fig. 1) and MDS plot (Fig. 2) identified three distinct groups corresponding to the British Isles, southern and western

Scandinavia, as previously reported (E. 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 Fig.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 (Supporting Information Fig. S1).

3.2 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 (Supporting Information Fig. S2). Across the genome five regions showed F_{ST} values > 0.9 and 32 regions F_{ST} > 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 (Supporting Information Fig. 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 Supporting Information Table S2.

3.3 Gene flow across the genetic break

We used the default parameters in PCAdmix thereby removing SNPs in high LD ($r^2 > 0.8$) 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 = 5695), 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=334) and southern haplotypes in ST was 6.34 bins or 317 Kbp (sd=5.2, median=5, 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 (Fig. 4). Some EG individuals appeared highly admixed (EG21 and EG24) with a 50.1% and 46.7% western ancestry, also suggested in the MDS plot (Fig. 2) and admixture graph (Fig. 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.

3.4 Demographic history and founding events

The analysis of PSMC and SMC++ is based on the “all SNP” dataset, while DADI analysis was conducted using the “thinned SNPs” dataset. 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 (Fig. 5 top). The population in the British Isles experienced a more recent recovery (~5kya), 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 (Fig. 5 bottom), 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, Fig. 5 below). 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: 1) decline in all populations started approximately 30 kya, first in the British Isles, then in western and southern Scandinavia; 2) 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 (Supporting information Table S3). Nevertheless, we observed increased residuals on the rare frequency range, suggesting difficulties in modeling the loss of variation (Supporting information Fig. S5). We converted the coalescent values for the best model, the West and South Scandinavia secondary contact, resulting in an ancestral population (Nref) of 384. The size of the populations after the split were 3980 and 1275 for the West and South Scandinavia, respectively. The total time of divergence was T = 2*Nref*generation time*(T1+T2) = 68,659 years of which the first 63,102 were spent in isolation while during the most recent 5527 years the populations have experienced gene flow. The estimated migration rate was quite low (i.e. m12/2*Nref) as the proportion of new migrants are 4.1×10^{-4} in the southern population and 6.7×10^{-4} in the western population (Supporting information 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 non-trivial. 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 (Supporting information Fig. S1). The loss of SNPs is dramatic, as ~700k SNPs detected in the British Isles are reduced to ~590k (~ 16% less) SNPs in western Scandinavia with a further reduction to ~450k (~ 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 hapFLK) (Fig. 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, likely imposed by strong drift (Fig. 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 fixated SNPs are also demonstrated in the site-frequency-spectra (Supporting information Fig. S4).

A second line of evidence is associated to the reduction of the effective population sizes in line with founding events detected using the sequentially Markovian coalescent methods in PSMC and SMC++ (Fig. 5). Our results suggest that corkwing wrasse colonized western Scandinavia about 11kya, possibly from the British Islands. Based on the number of bi-allelic and heterozygous SNPs and inbreeding coefficients, Stavanger site (ST; Table 1, Fig. 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 behavior 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 (Rougemont et al., 2017; Tine et al., 2014). Here, we only tested simple scenarios, as more complicated models (Rougeux, 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, Supporting Information 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 (Supporting Information Fig. 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 indicates the sequential loss of shared haplotypes and the direction of founding events.

Finally, we detect ongoing gene flow in both directions across the genetic break (Fig. 4). The contact is asymmetrical with increased gene flow from the West into the South. By using half the genome (>N50) and bin size of 50 Kbp, we detect 1568 bins of a total of 5694 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 suggest that the contact is recent or actively selected against (Abbott et al., 2013).

5 Conclusions

Our findings shed new light on the dynamics underlying the presence of two genetic breaks of this species in Northern Europe (E. 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 (G. M. 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).

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., . . . Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology, 26*(2), 229-246. doi:10.1111/j.1420-9101.2012.02599.x

Alex Buerkle, C., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low should we go? *Mol Ecol, 22*(11), 3028-3035. doi:10.1111/mec.12105

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res, 19*(9), 1655-1664. doi:10.1101/gr.094052.109

Almada, F., Francisco, S. M., Lima, C. S., FitzGerald, R., Mirimin, L., Villegas-Rios, D., . . . Robalo, J. I. (2017). Historical gene flow constraints in a northeastern Atlantic fish: phylogeography of the ballan wrasse Labrus bergylta across its distribution range. *R Soc Open Sci, 4*(2), 160773. doi:10.1098/rsos.160773

Barth, J. M. I., Damerau, M., Matschiner, M., Jentoft, S., & Hanel, R. (2017). Genomic differentiation and demographic histories of Atlantic and Indo-Pacific yellowfin tuna (Thunnus albacares) populations. *Genome Biology and Evolution*. doi:10.1093/gbe/evx067

Beichman, A. C., Phung, T. N., & Lohmueller, K. E. (2017). Comparison of Single Genome and Allele Frequency Data Reveals Discordant Demographic Histories. *G3 (Bethesda), 7*(11), 3605-3620. doi:10.1534/g3.117.300259

Blanco Gonzalez, E., & de Boer, F. (2017). The development of the Norwegian wrasse fishery and the use of wrasses as cleaner fish in the salmon aquaculture industry. *Fisheries Science, 83*(5), 661-670. doi:10.1007/s12562-017-1110-4

Blanco Gonzalez, E., Espeland, S. H., Jentoft, S., Hansen, M. M., Robalo, J. I., Stenseth, N. C., & Jorde, P. E. (2019). Interbreeding between local and translocated populations of a cleaner fish in an experimental mesocosm predicts risk of disrupted local adaptation. *Ecol Evol, 9*(11), 6665-6677. doi:10.1002/ece3.5246

Blanco Gonzalez, E., Knutsen, H., & Jorde, P. E. (2016). Habitat Discontinuities Separate Genetically Divergent Populations of a Rocky Shore Marine Fish. *PLoS One, 11*(10), e0163052. doi:10.1371/journal.pone.0163052

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., . . . Bustamante, C. D. (2012). PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum Biol, 84*(4), 343-364. doi:10.3378/027.084.0401

Browning, B. L., & Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics, 194*(2), 459-471. doi:10.1534/genetics.113.150029

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

Evankow, A., Christie, H., Hancke, K., Brysting, A. K., Junge, C., Fredriksen, S., & Thaulow, J. (2019). Genetic heterogeneity of two bioeconomically important kelp species along the Norwegian coast. *Conservation Genetics*. doi:10.1007/s10592-019-01162-8

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., & Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics, 193*(3), 929-941. doi:10.1534/genetics.112.147231

Faust, E., Halvorsen, K. T., Andersen, P., Knutsen, H., & Andre, C. (2018). Cleaner fish escape salmon farms and hybridize with local wrasse populations. *R Soc Open Sci, 5*(3), 171752. doi:10.1098/rsos.171752

Feder, J. L., & Nosil, P. (2010). The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution, 64*(6), 1729-1747. doi:10.1111/j.1558-5646.2010.00943.x

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics, 180*(2), 977-993. doi:10.1534/genetics.108.092221

Francois, O., Blum, M. G., Jakobsson, M., & Rosenberg, N. A. (2008). Demographic history of european populations of Arabidopsis thaliana. *PLoS Genet, 4*(5), e1000075. doi:10.1371/journal.pgen.1000075

Frith, M. C., Hamada, M., & Horton, P. (2010). Parameters for accurate genome alignment. *BMC Bioinformatics, 11*, 80. doi:10.1186/1471-2105-11-80

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint, arXiv:1207.3907*([q-bio.GN]).

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 Genet, 5*(10), e1000695. doi:10.1371/journal.pgen.1000695

Halvorsen, K. T., Larsen, T., Sørdalen, T. K., Vøllestad, L. A., Knutsen, H., & Olsen, E. M. (2017). Impact of harvesting cleaner fish for salmonid aquaculture assessed from replicated coastal marine protected areas. *Marine Biology Research, 13*(4), 359-369. doi:10.1080/17451000.2016.1262042

Halvorsen, K. T., Sørdalen, T. K., Durif, C., Knutsen, H., Olsen, E. M., Skiftesvik, A. B., . . . Vøllestad, L. A. (2016). Male-biased sexual size dimorphism in the nest building corkwing wrasse (Symphodus melops): implications for a size regulated fishery. *Ices Journal of Marine Science, 73*(10), 2586-2594. doi:10.1093/icesjms/fsw135

Harris, K., & Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet, 9*(6), e1003521. doi:10.1371/journal.pgen.1003521

Hauser, L., & Carvalho, G. R. (2008). Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish and Fisheries, 9*(4), 333-362. doi:doi:10.1111/j.1467-2979.2008.00299.x

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature, 405*(6789), 907-913. doi:10.1038/35016000

Hewitt, G. M. (1999). Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society, 68*(1), 87-112. doi:https://doi.org/10.1006/bijl.1999.0332

Hoarau, G., Coyer, J. A., Veldsink, J. H., Stam, W. T., & Olsen, J. L. (2007). Glacial refugia and recolonization pathways in the brown seaweed Fucus serratus. *Mol Ecol, 16*(17), 3606-3616. doi:10.1111/j.1365-294X.2007.03408.x

Hollinger, I., Pennings, P. S., & Hermisson, J. (2019). Polygenic adaptation: From sweeps to subtle frequency shifts. *PLoS Genet, 15*(3), e1008035. doi:10.1371/journal.pgen.1008035

Jenkins, T. L., Castilho, R., & Stevens, J. R. (2018). Meta-analysis of northeast Atlantic marine taxa shows contrasting phylogeographic patterns following post-LGM expansions. *PeerJ, 6*, e5684. doi:10.7717/peerj.5684

Kettle, A. J., Morales-Muñiz, A., Roselló-Izquierdo, E., Heinrich, D., & Vøllestad, L. A. (2011). Refugia of marine fish in the northeast Atlantic during the last glacial maximum: concordant assessment from archaeozoology and palaeotemperature reconstructions. *Clim. Past, 7*(1), 181-201. doi:10.5194/cp-7-181-2011

Knutsen, H., Jorde, P. E., Gonzalez, E. B., Robalo, J., Albretsen, J., & Almada, V. (2013). Climate Change and Genetic Structure of Leading Edge and Rear End Populations in a Northwards Shifting Marine Fish Species, the Corkwing Wrasse (Symphodus melops). *PLoS One, 8*(6), e67492. doi:10.1371/journal.pone.0067492

Kyrkjeeide, M. O., Stenøien, H. K., Flatberg, K. I., & Hassel, K. (2014). Glacial refugia and post-glacial colonization patterns in European bryophytes. *Lindbergia*, 47-59. doi:10.25227/linbg.01046

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature, 475*(7357), 493-496. doi:10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Lie, K. K., Torresen, O. K., Solbakken, M. H., Ronnestad, I., Tooming-Klunderud, A., Nederbragt, A. J., . . . Saele, O. (2018). Loss of stomach, loss of appetite? Sequencing of the ballan wrasse (Labrus bergylta) genome and intestinal transcriptomic profiling illuminate the evolution of loss of stomach function in fish. *Bmc Genomics, 19*(1), 186. doi:10.1186/s12864-018-4570-8

Maggs, C. A., Castilho, R., Foltz, D., Henzler, C., Jolly, M. T., Kelly, J., . . . Wares, J. (2008). Evaluating signatures of glacial refugia for North Atlantic benthic marine taxa. *Ecology, 89*(11 Suppl), S108-122.

Mattingsdal, M., Jentoft, S., Torresen, O. K., Knutsen, H., Hansen, M. M., Robalo, J. I., . . . Gonzalez, E. B. (2018). A continuous genome assembly of the corkwing wrasse (Symphodus melops). *Genomics, 110*(6), 399-403. doi:10.1016/j.ygeno.2018.04.009

Mattingsdal. M. (2019). *SNPs from Corkwing in Northern Europe*. Retrieved from: https://figshare.com/articles/SNPs_from_Corkwing_in_Northern_Europe/7570907/1

Mazet, O., Rodriguez, W., Grusea, S., Boitard, S., & Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution--lessons for ancestral population size inference? *Heredity (Edinb), 116*(4), 362-371. doi:10.1038/hdy.2015.104

Meynert, A. M., Bicknell, L. S., Hurles, M. E., Jackson, A. P., & Taylor, M. S. (2013). Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics, 14*, 195. doi:10.1186/1471-2105-14-195

Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Mol Ecol, 25*(5), 1058-1072. doi:10.1111/mec.13540

Nadeau, S., Meirmans, P. G., Aitken, S. N., Ritland, K., & Isabel, N. (2016). The challenge of separating signatures of local adaptation from those of isolation by distance and colonization history: The case of two white pines. *Ecol Evol, 6*(24), 8649-8664. doi:10.1002/ece3.2550

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu Rev Genet, 39*, 197-218. doi:10.1146/annurev.genet.39.073003.112420

Orsini, L., Vanoverbeke, J., Swillen, I., Mergeay, J., & De Meester, L. (2013). Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. *Mol Ecol, 22*(24), 5983-5999. doi:10.1111/mec.12561

Palsboll, P. J., Berube, M., & Allendorf, F. W. (2007). Identification of management units using population genetic data. *Trends Ecol Evol, 22*(1), 11-16. doi:10.1016/j.tree.2006.09.003

Pavlidis, P., & Alachiotis, N. (2017). A survey of methods and tools to detect recent and strong positive selection. *J Biol Res (Thessalon), 24*, 7. doi:10.1186/s40709-017-0064-0

Petit, R. J., Brewer, S., Bordács, S., Burg, K., Cheddadi, R., Coart, E., . . . Kremer, A. (2002). Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *Forest Ecology and Management, 156*(1), 49-74. doi:https://doi.org/10.1016/S0378-1127(01)00634-X

Petkova, D., Novembre, J., & Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet, 48*(1), 94-100. doi:10.1038/ng.3464

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

Poh, Y. P., Domingues, V. S., Hoekstra, H. E., & Jensen, J. D. (2014). On the prospect of identifying adaptive loci in recently bottlenecked populations. *PLoS One, 9*(11), e110579. doi:10.1371/journal.pone.0110579

Portik, D. M., Leache, A. D., Rivera, D., Barej, M. F., Burger, M., Hirschfeld, M., . . . Fujita, M. K. (2017). Evaluating mechanisms of diversification in a Guineo-Congolian tropical forest frog using demographic model selection. *Mol Ecol, 26*(19), 5245-5263. doi:10.1111/mec.14266

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi:10.1086/519795

Quintela, M., Danielsen, E. A., Lopez, L., Barreiro, R., Svasand, T., Knutsen, H., . . . Glover, K. A. (2016). Is the ballan wrasse (Labrus bergylta) two species? Genetic analysis reveals within-species divergence associated with plain and spotted morphotype frequencies. *Integr Zool, 11*(2), 162-172. doi:10.1111/1749-4877.12186

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlovic, M., . . . Westram, A. M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *J Evol Biol, 30*(8), 1450-1477. doi:10.1111/jeb.13047

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature, 461*(7263), 489-494. doi:10.1038/nature08365

Robalo, J. I., Castilho, R., Francisco, S. M., Almada, F., Knutsen, H., Jorde, P. E., . . . Almada, V. C. (2012). Northern refugia and recent expansion in the North Sea: the case of the wrasse Symphodus melops (Linnaeus, 1758). *Ecol Evol, 2*(1), 153-164. doi:10.1002/ece3.77

Rougemont, Q., & Bernatchez, L. (2018). The demographic history of Atlantic salmon (Salmo salar) across its distribution range reconstructed from approximate Bayesian computations. *Evolution, 72*(6), 1261-1277. doi:10.1111/evo.13486

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. *Mol Ecol, 26*(1), 142-162. doi:10.1111/mec.13664

Rougeux, C., Bernatchez, L., & Gagnaire, P. A. (2017). Modeling the Multiple Facets of Speciation-with-Gene-Flow toward Inferring the Divergence History of Lake Whitefish Species Pairs (Coregonus clupeaformis). *Genome Biology and Evolution, 9*(8), 2057-2074. doi:10.1093/gbe/evx150

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., . . . Lander, E. S. (2006). Positive natural selection in the human lineage. *Science, 312*(5780), 1614-1620. doi:10.1126/science.1124309

Schluter, D., & Conte, G. L. (2009). Genetics and ecological speciation. *Proc Natl Acad Sci U S A, 106 Suppl 1*, 9955-9962. doi:10.1073/pnas.0901264106

Sexton, J. P., McIntyre, P. J., Angert, A. L., & Rice, K. J. (2009). Evolution and Ecology of Species Range Limits. *Annual Review of Ecology, Evolution, and Systematics, 40*(1), 415-436. doi:10.1146/annurev.ecolsys.110308.120317

Spence, J. P., Steinrücken, M., Terhorst, J., & Song, Y. S. (2018). Inference of population history using coalescent HMMs: review and outlook. *Current Opinion in Genetics & Development, 53*, 70-76. doi:https://doi.org/10.1016/j.gde.2018.07.002

Spurgin, L. G., Illera, J. C., Jorgensen, T. H., Dawson, D. A., & Richardson, D. S. (2014). Genetic and phenotypic divergence in an island bird: isolation by distance, by colonization or by adaptation? *Mol Ecol, 23*(5), 1028-1039. doi:10.1111/mec.12672

Storey JD, Bass AJ, Dabney A, & D, R. (2019). qvalue: Q-value estimation for false discovery rate control. *R package version 2.16.0*.

Storfer, A., Murphy, M. A., Spear, S. F., Holderegger, R., & Waits, L. P. (2010). Landscape genetics: where are we now? *Mol Ecol, 19*(17), 3496-3514. doi:10.1111/j.1365-294X.2010.04691.x

Taberlet, P., Fumagalli, L., Wust-Saucy, A. G., & Cosson, J. F. (1998). Comparative phylogeography and postglacial colonization routes in Europe. *Mol Ecol, 7*(4), 453-464.

Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet, 49*(2), 303-309. doi:10.1038/ng.3748

Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics, 194*(2), 301-326. doi:10.1534/genetics.112.148825

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. *Nat Commun, 5*, 5770. doi:10.1038/ncomms6770

Uglem, I., Rosenqvist, G., & Wasslavik, H. S. (2000). Phenotypic variation between dimorphic males in corkwing wrasse. *Journal of Fish Biology, 57*(1), 1-14. doi:10.1111/j.1095-8649.2000.tb00771.x

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution, 38*(6), 1358-1370. doi:10.1111/j.1558-5646.1984.tb05657.x

Data Accessibility

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

Figure Legends

Figure 1: Map showing sampling locations (See also Table 1) together with estimation of effective migration surfaces inferred by EEMS, where brown color indicates a reduction and cyan color 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 dataset (SNPs = 50,130). Individuals sampled in the British Isles (yellow), western Scandinavia (red) and southern Scandinavia (blue).

Figure 3: ADMIXTURE results using the thinned SNP dataset (SNPs = 50,130). K = 3 represents the most likely number of putative ancestral populations. Here, we show results at K = 2 and K = 3.

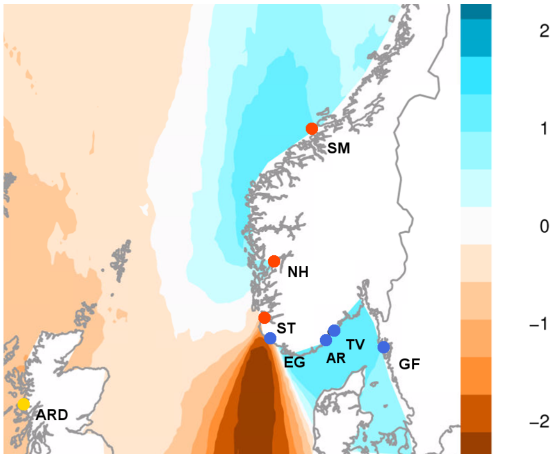
Figure 4: Introgression and local ancestry inferred by PCAdmix (Brisbin et al., 2012). The top figure displays the distribution of introgressed haplotype lengths in individuals from the sites across the genetic break (ST and EG). The distribution clearly shows larger haplotypes introgressed from the West into the EG site, including several large haplotypes > 1000 Kbp. The bottom figure displays the inferred local ancestry per haplotype using the largest contig as example region (SYMME_00023145, 5.4 Mbp).

Figure 5: Demographic history inferred using the sequentially Markovian coalescent implemented in PSMC and SMC++ using all SNPs, a generation time of 3 years and a mutation rate of $1x10^{-8}$. Top: results from PSMC with bootstraps using three individuals from the most distant sites. Below: Estimated histories in SMC++ using the composite likelihood of four individuals from each population. The yellow line represents the site

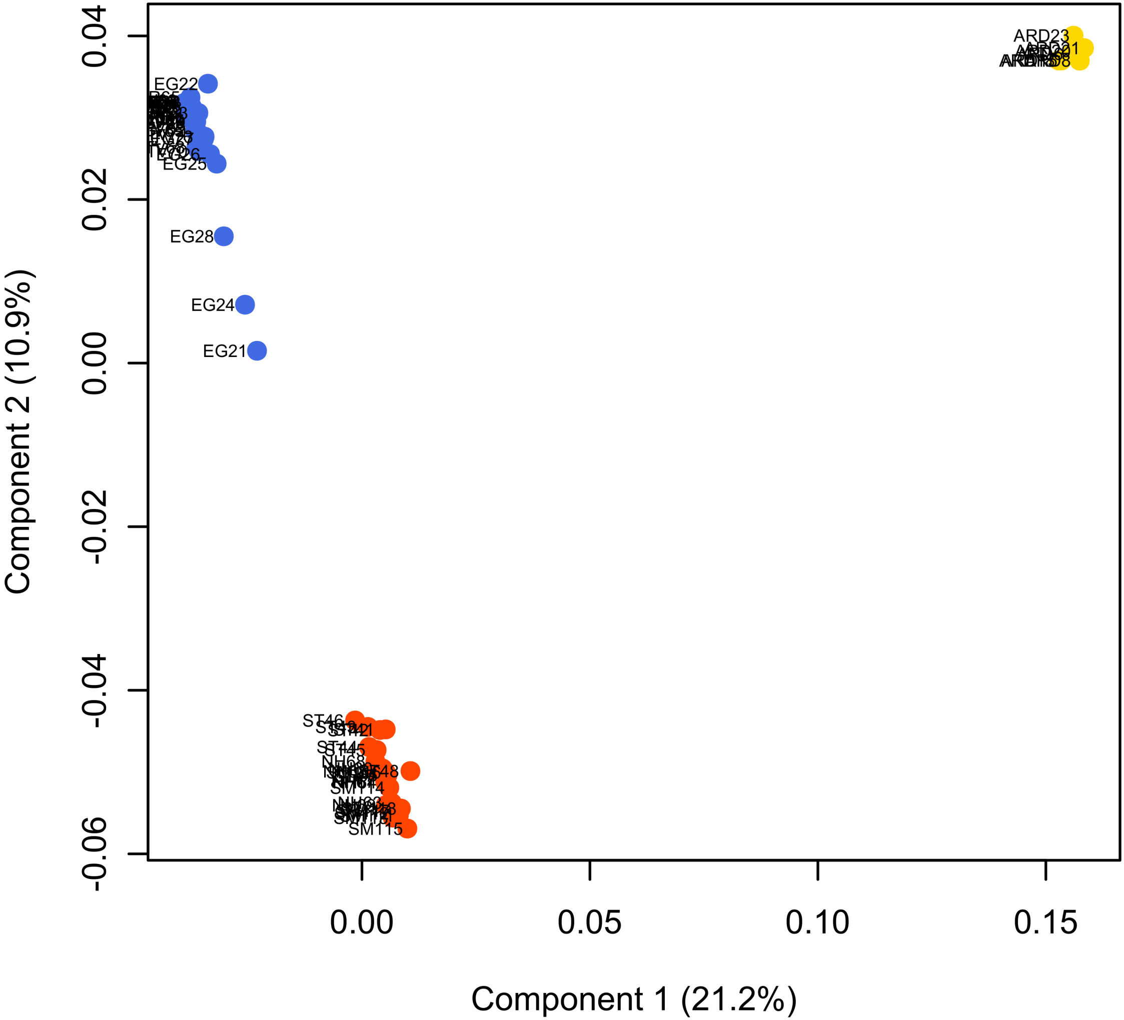
from the British Isles, the red line corresponds to western Scandinavian samples, and the blue line comprises southern Scandinavian samples. Beware the differences on the axes between the top and bottom figures, as these two methods capture variation in effective population sized through different time scales.

Table 1: Regional groupings of corkwing sampling locations by region, location, code, sampling year, latitude, longitude, sample size (n) and number of variable SNPs per site (Minor Allele Count > 1).

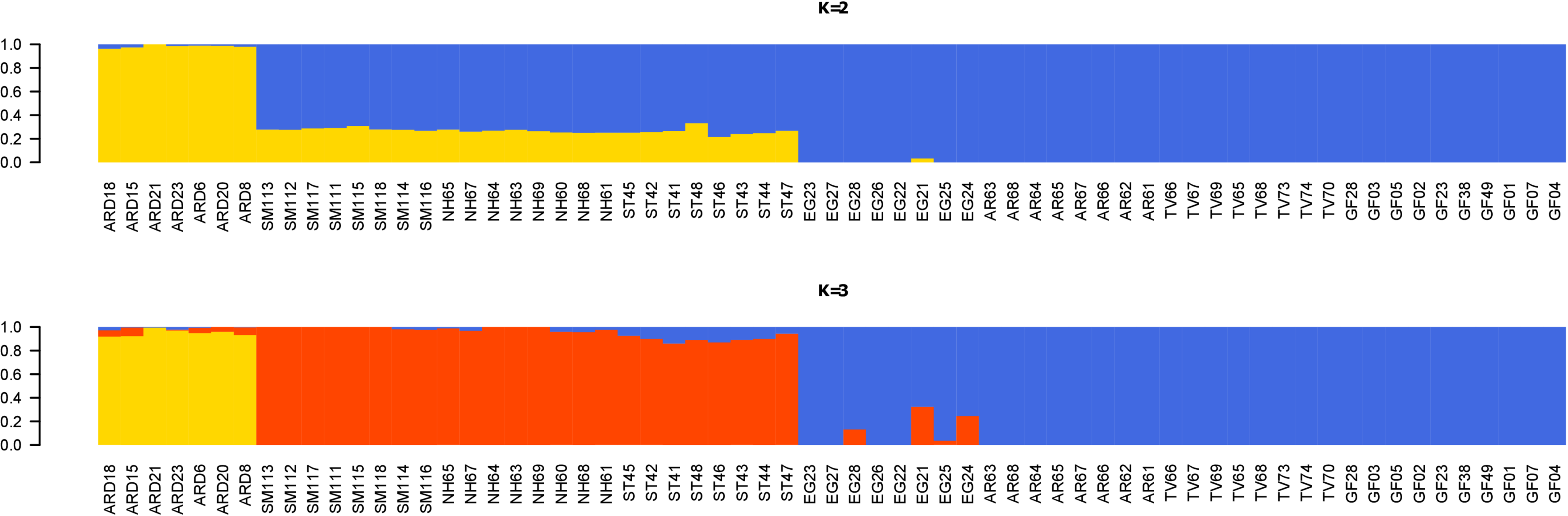
Region	**Location**	**Code**	**Year**	**Latitude**	**Longitude**	**n**	**SNPs**
British Isles	Ardtoe	ARD	2010	N 56.40	W 5.50	7	704,073
Western Scandinavia	Smøla	SM	2015	N 63.32	E 8.11	8	592,767
	Norheimsund	NH	2014	N 60.39	E 6.48	8	584,422
	Stavanger	ST	2015	N 59.01	E 5.56	8	597,438
Southern Scandinavia	Egersund	EG	2008	N 58.45	E 5.53	8	480,354
	Arendal	AR	2014	N 58.41	E 8.74	8	431,556
	Tvedestrand	TV	2010	N 58.62	E 9.06	8	440,905

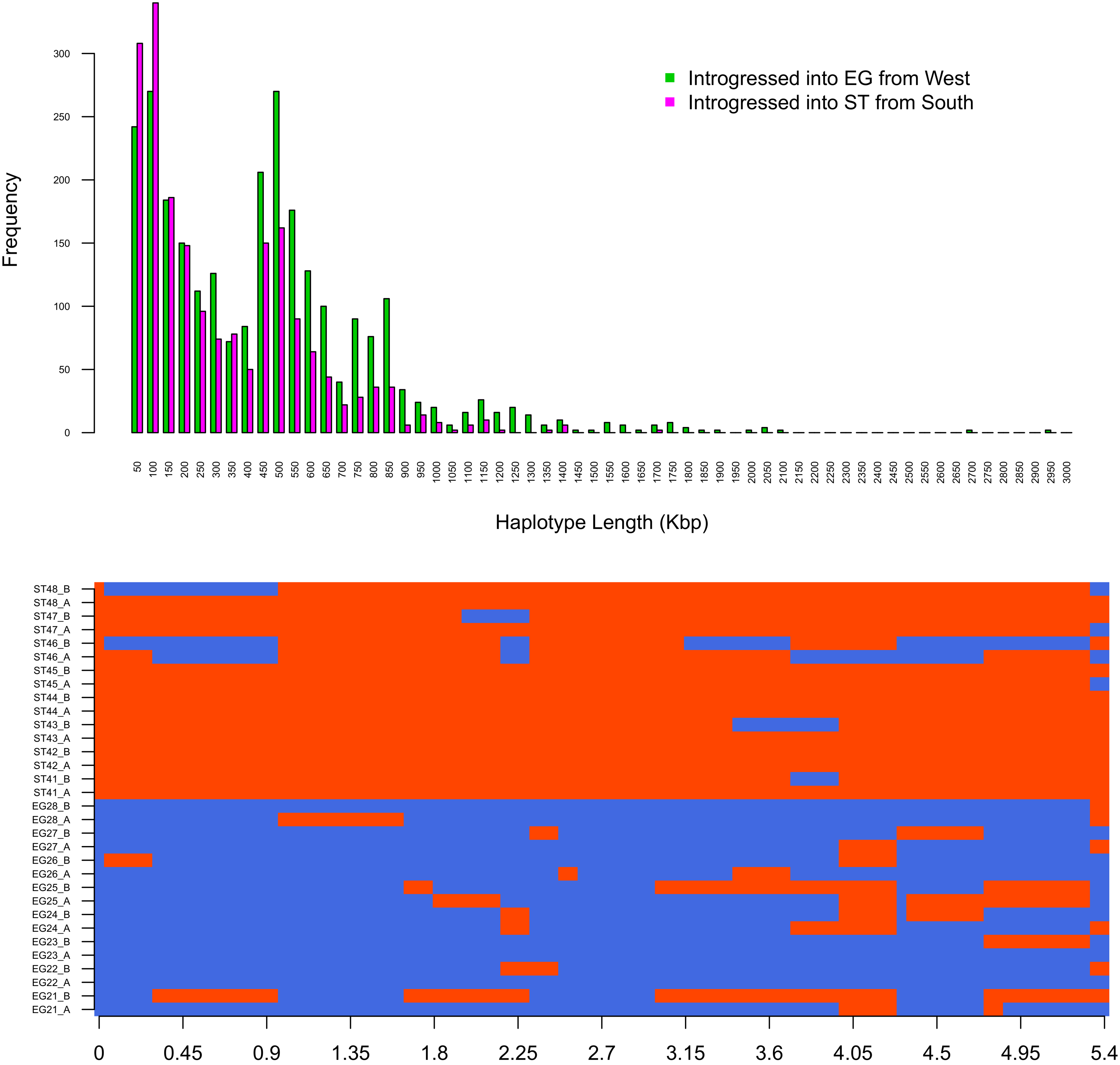


mec_15310_f1.png

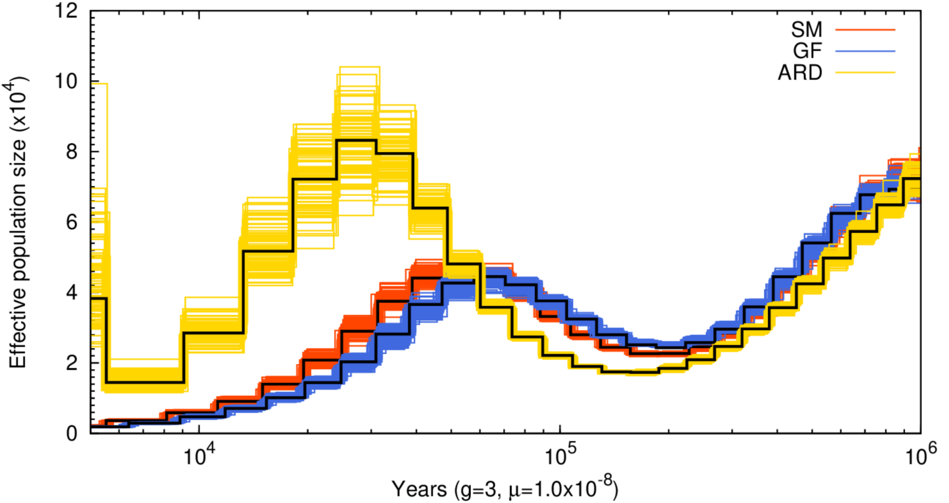


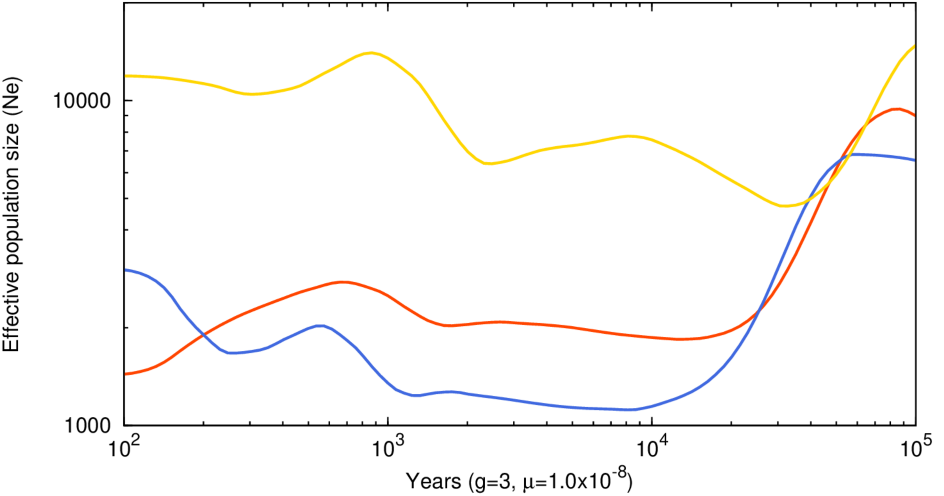
Component 2 (10.9%)
0.04
0.02
0.00
-0.02
-0.04
-0.06
Component 1 (21.2%)
0.00
0.05
0.10
0.15
EG22
EG26
EG25
EG28
EG24
EG21
ARD23
ARD21
ST46
ST45
NH63
SM114
SM115


K=2
1.0
0.8
0.6
0.4
0.2
0.0
ARD18 ARD15 ARD21 ARD23 ARD6 ARD20 ARD8 SM113 SM112 SM117 SM111 SM115 SM118 SM114 SM116 NH65 NH67 NH64 NH63 NH69 NH60 NH68 NH61 ST45 ST42 ST41 ST48 ST46 ST43 ST44 ST47 EG23 EG27 EG28 EG26 EG22 EG21 EG25 EG24 AR63 AR68 AR64 AR65 AR67 AR66 AR62 AR61 TV66 TV67 TV69 TV65 TV68 TV73 TV74 TV70 GF28 GF03 GF05 GF02 GF23 GF38 GF49 GF01 GF07 GF04
K=3
1.0
0.8
0.6
0.4
0.2
0.0
ARD18 ARD15 ARD21 ARD23 ARD6 ARD20 ARD8 SM113 SM112 SM117 SM111 SM115 SM118 SM114 SM116 NH65 NH67 NH64 NH63 NH69 NH60 NH68 NH61 ST45 ST42 ST41 ST48 ST46 ST43 ST44 ST47 EG23 EG27 EG28 EG26 EG22 EG21 EG25 EG24 AR63 AR68 AR64 AR65 AR67 AR66 AR62 AR61 TV66 TV67 TV69 TV65 TV68 TV73 TV74 TV70 GF28 GF03 GF05 GF02 GF23 GF38 GF49 GF01 GF07 GF04



Frequency
0
50
100
150
200
250
300
Introgressed into EG from West
Introgressed into ST from South
50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800 850 900 950 1000 1050 1100 1150 1200 1250 1300 1350 1400 1450 1500 1550 1600 1650 1700 1750 1800 1850 1900 1950 2000 2050 2100 2150 2200 2250 2300 2350 2400 2450 2500 2550 2600 2650 2700 2750 2800 2850 2900 2950 3000
Haplotype Length (Kbp)
ST48_B
ST48_A
ST47_B
ST47_A
ST46_B
ST46_A
ST45_B
ST45_A
ST44_B
ST44_A
ST43_B
ST43_A
ST42_B
ST42_A
ST41_B
ST41_A
EG28_B
EG28_A
EG27_B
EG27_A
EG26_B
EG26_A
EG25_B
EG25_A
EG24_B
EG24_A
EG23_B
EG23_A
EG22_B
EG22_A
EG21_B
EG21_A
0
0.45
0.9
1.35
1.8
2.25
2.7
3.15
3.6
4.05
4.5
4.95
5.4


SM
GF
ARD
Effective population size (x10^4)
0
2
4
6
8
10
12
10^4
10^5
10^6
Years (g=3, μ=1.0x10^{-8})


Effective population size (Ne)
1000
10000
10^2
10^3
10^4
10^5
Years (g=3, μ=1.0x10^{-8})