

Genome wide analysis reveals genetic divergence between Goldsinny wrasse populations

Eeva Jansson (✉ eeva.jansson@hi.no)

Havforskningsinstituttet <https://orcid.org/0000-0002-7840-7201>

Francois Besnier

Havforskningsinstituttet

Ketil Malde

Havforskningsinstituttet

Carl André

Goteborgs Universitet

Geir Dahle

Havforskningsinstituttet

Kevin A. Glover

Havforskningsinstituttet

Research article

Keywords: assignment, *Ctenolabrus rupestris*, marker validation, population genomics, resequencing, SNP

Posted Date: September 21st, 2020

DOI: <https://doi.org/10.21203/rs.2.22229/v2>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published on October 9th, 2020. See the published version at <https://doi.org/10.1186/s12863-020-00921-8>.

Abstract

Background: Marine fish populations are often characterized by high levels of gene flow and correspondingly low genetic divergence. This presents a challenge to define management units. Goldsinny wrasse (*Ctenolabrus rupestris*) is a heavily exploited species due to its importance as a cleaner-fish in commercial salmonid aquaculture. However, at the present, the population genetic structure of this species is still largely unresolved. Here, full-genome sequencing was used to produce the first genomic reference for this species, to study population-genomic divergence among four geographically distinct populations, and, to identify informative SNP markers for future studies.

Results: After construction of a *de novo* assembly, the genome was estimated to be highly polymorphic and of ~600Mbp in size. 33 235 SNPs were thereafter selected to assess genomic diversity and differentiation among four populations collected from Scandinavia, Scotland, and Spain. Global F_{ST} among these populations was 0.015–0.092. Approximately 4% of the investigated loci were identified as putative global outliers, and ~1% within Scandinavia. SNPs showing large divergence ($F_{ST}>0.15$) were picked as candidate diagnostic markers for population assignment. 173 of the most diagnostic SNPs between the two Scandinavian populations were validated by genotyping 47 individuals from each end of the species' Scandinavian distribution range. 69 of these SNPs were significantly ($p<0.05$) differentiated (mean $F_{ST_173_loci}=0.065$, $F_{ST_69_loci}=0.140$). Using these validated SNPs, individuals were assigned with high probability ($\geq 94\%$) to their populations of origin.

Conclusions: Goldsinny wrasse displays a highly polymorphic genome, and substantial population genomic structure. Diversifying selection likely affects population structuring globally and within Scandinavia. The diagnostic loci identified now provide a promising and cost-efficient tool to investigate goldsinny wrasse populations further.

1. Background

Thanks to the rapid development of whole-genome sequencing methods during the last decade (van Dijk *et al.* 2014; Goodwin *et al.* 2016), genome wide data is now relatively cost-effective to produce and is becoming increasingly commonplace to study evolutionary questions even for non-model organisms (Fonseca *et al.* 2016). For the conservation and management of wild populations, studies employing high-throughput sequencing technologies may provide better estimates than traditional population genetics tools for key parameters such as effective population size, genetic structure, and connectivity (for reviews, see Allendorf *et al.* 2010; Shafer *et al.* 2015). Greater resolution from using large numbers of genetic markers and/or by pre-selection of highly divergent markers (i.e., outliers) is especially useful in population genetic studies of marine organisms (Russello *et al.* 2012; Gagnaire *et al.* 2015) which are often characterized by very large census sizes and high levels of connectivity leading to generally low levels of population divergence (Hauser & Carvalho 2008). However, because the efficiency of natural selection is dependent on the (effective) population size (Charlesworth 2009; Allendorf *et al.* 2014), adaptive genetic differences are possible – or even likely – in large marine populations inhabiting

heterogeneous environments (Berg *et al.* 2015; Guo *et al.* 2015 Barrio *et al.* 2016). Therefore, finding biologically meaningful differences among populations, and the possible genetic factors underlying these differences, will help define appropriate management units, and thus sustainably exploit marine species. Better detection and understanding of human-mediated introgression (Crispo *et al.* 2011), a common problem in many wild fish populations (e.g. Glover *et al.* 2013; Blanco Gonzalez *et al.* 2015; Lamer *et al.* 2015; Ozerov *et al.* 2016), is also of high conservation priority, and now more feasible with the modern sequencing methods at hand.

Goldsinny wrasse, *Ctenolabrus rupestris*, is a small (<18 cm) inshore marine fish belonging to the *Labridae* family that includes over 500 described species worldwide. It is a common species in the Eastern Atlantic coastal waters from Morocco to Norway, and is also found in the Mediterranean and Black Sea. Traditionally, goldsinny wrasse had no commercial value and thus avoided significant exploitation (Darwall *et al.* 1992; Halvorsen *et al.* 2017). However, in response to its increasing demand as a cleaner-fish in the aquaculture industry for delousing cage-reared Atlantic salmon (*Salmo salar*) and rainbow trout (*Oncorhynchus mykiss*), the species is now extensively harvested together with other wrasse species (Skiftesvik *et al.* 2015). The demand for cleaner fish in the aquaculture industry has grown almost exponentially after 2007 (Skiftesvik *et al.* 2014; 2015) due to emerged resistance to delousing agents of the salmon louse (*Lepeophtheirus salmonis*) (Besnier *et al.* 2014; Kaur *et al.* 2016). In Norway alone, ~20 million wrasses were caught in 2014-2019 and used in commercial aquaculture. Between 8 to 12 million of these were goldsinny wrasses, which together with corkwing wrasse makes them the numerically most significant of the wild-captured cleaner fish used. In addition to wrasses caught in local waters, Norwegian salmon farms received wrasses caught from the Swedish west coast (about a million fish per year; Mortensen *et al.* 2017). The high fishing pressure combined with the high breeding-site philopatry of goldsinny wrasse (Hilldén 1984), and their slow growth rate (4–5 years for minimum commercial size of 11 cm; Skiftesvik *et al.* 2014) indicate that this species is likely to be sensitive to overexploitation. A study by Halvorsen *et al.* (2017) showed that intensive wrasse fisheries could have considerable impact on the target populations: the abundance of goldsinny wrasse was significantly lower on harvested than on control sites (marine protected areas, MPAs) in the Skagerrak region. The authors suggest that such negative fishery effects might be more severe in western Norway, where there are no MPAs and the wrasse fishery is much more intense, and that this reduction in population densities might even lead to cascade effects in the coastal ecosystems. Overexploitation also increases the risk of loss of genetic variation and potentially locally adaptive variants (e.g. Allendorf *et al.* 2014). Genetic integrity and adaptability of local populations can also be compromised via manmade gene flow from genetically diverged populations. This concern is of particular importance in fisheries management where large-scale population augmentations through deliberate or inadvertent releases of translocated, captively raised or domesticated individuals occurs (Laikre *et al.* 2010; Waples *et al.* 2016; Glover *et al.* 2017). Wild populations of wrasses are not intentionally augmented but receive human-mediated gene flow via large-scale translocations and release or escape of wrasse from fish farms (Faust *et al.* 2018; Blanco Gonzalez *et al.* 2019).

Given the current practise of using cleaner wrasses in great numbers, and the associated heavy fishing pressure placed on wild populations (Halvorsen *et al.* 2017), more knowledge of the population genetic structure of wrasses is required. Thus far, the best-studied of these species is the corkwing wrasse (*Symphodus melops*) which displays clear population subdivisions on different geographic scales (Robalo *et al.* 2012; Knutsen *et al.* 2013; Blanco Gonzalez *et al.* 2016; Mattingdsal *et al.* 2020). Genetic studies of ballan wrasse (*Labrus bergylta*) have concentrated on investigation of population structure on large geographic scales (D'Arcy *et al.* 2013; Almada *et al.* 2017), and on genetic divergence of different morphotypes (Quintela *et al.* 2016; Almada *et al.* 2016). A newly published study by Seljestad *et al.* (2020) revealed subdivision on a more local scale as well, dividing the Scandinavian ballan wrasse population into two distinct genetic clusters. An early genetic study of goldsinny wrasse with allozyme markers showed significant differences between the southern and mid-Norway (Sundt & Jørstad 1998). The only recent study, using a combination of 14 microsatellite and 36 SNP markers (Jansson *et al.* 2017), revealed clear population divergence ($F_{ST} \sim 0.02-0.05$) across the North Sea but only modest differences ($F_{ST} \leq 0.02$) that increased with geographic distance (i.e. isolation-by-distance, IBD) within Scandinavian populations. These patterns are concordant with restricted migration and gene flow as the main factor creating genetic patterns for the species but does not rule out other possible factors such as selection and/or demographic history shaping the population structures seen today (see e.g. Mattingdsal *et al.* 2020). In addition, aquaculture-mediated translocation of wrasses might affect both the donor and the recipient population. The primary area to move wrasses is mid-Norway where there are plenty of fish farms but insufficient local supply of cleaner fish. For corkwing wrasse, which exhibits a clear genetic population subdivision between southern and western Norway ($F_{ST} = 0.107$; Blanco *et al.* 2016), it has been shown that translocated fish from southern Scandinavia have escaped and hybridized with local populations in mid-Norway (Faust *et al.* 2018). For goldsinny wrasse, unequivocal determination of escape and/or hybridization with local populations has not yet been possible to determine due to weak level of genetic difference observed between the export and import areas (Jansson *et al.* 2017).

The present study had the following main aims: I) to develop *a de novo assembly* for goldsinny wrasse, II) to use a subset of high-quality SNP markers to conduct a first genomic population study for the species using population samples collected along the species entire distribution range (North and South Scandinavia, Scotland and Spain; Fig. 1), III) to compare these results with the results obtained with limited number of markers (in Jansson *et al.* 2017), as well as IV) to identify and validate putatively diagnostic markers between the two Scandinavian populations to be used in future studies.

2. Results

2.1. *De novo* assembled reference

The longest continuous scaffolds were obtained using the maximum K-mer size of 127. The best assembly consisted of approximately 3 million contigs (N=2974923), with 9.7% reaching 1kb or longer, and a *N50* of 874 bp. Despite this high level of fragmentation, when reads from all the 60 fish were aligned against the reference, on average, 98.4% (SD ± 0.38) were mapped. BUSCO (Simão *et al.* 2015)

search for the ray-finned fish core genes found 774 complete and 911 fragmented BUSCOs from the reference, comprising 36.8% of the searched genes. Based on the K-mer distribution, estimates for the size of the goldsinny wrasse genome ranged from 580 to 600 Mb. The obtained K-mer profile shows a clear two-peak pattern characteristic to a highly heterozygous genome (Fig. 2). Based on the distribution, the number of heterozygous loci in the goldsinny wrasse reference genome was estimated to be ~10 million.

Due to computational constraints that the high fragmentation of the reference causes, SNPs were only called on contigs >20kb. There were 1222 such contigs, covering in total ~45.4Mbp (~7.6% of the genome). Additional information of SNPs (frequency and divergence) in the selected contigs is given in supplementary table S4.

2.2. Population genomics using 33k SNPs data set

Genotype correlation was measured between all pairs of loci from the same contig. The average correlation was $r^2=0.19\pm0.09$. In comparison, the 100 000 pairs of loci randomly sampled across contigs had an average correlation of $r^2=0.12\pm0.15$, indicating a generally higher degree of linkage disequilibrium for the SNPs linked on the same contig compared to SNPs located on different contigs. When excluding the correlation between the nearest n SNPs (with $n=1$ to $n=15$), we observed a gradual reduction of LD from $r^2=0.18\pm0.14$ ($n=5$) to $r^2=0.16\pm0.13$ ($n=15$). The selected loci had an average coverage of 327 ± 100 per site among samples from Bodø, 313 ± 99 from Spain, 329 ± 102 from Scotland and 168 ± 46 from Varberg. Expected heterozygosity across all loci and populations was 0.321, observed heterozygosity 0.319 (Fig. S3), and F_{IS} 0.006. This overall deficiency of heterozygotes was statistically highly significant ($t = 43.322$, $df = 33234$, $p < 2.2e-16$), and likely due to population subdivision (see results below). Expected heterozygosity was rather uniform across populations, but realized distribution of variation within populations differed from each other (Table 1). Heterozygote deficiency was observed in the samples from Bodø, Scotland, and Spain, however, the sample from Varberg displayed a clear heterozygote excess.

Populations were clearly differentiated, and the mean F_{ST} over all four populations based on the 33k dataset was 0.062, while pair-wise values ranged between 0.015-0.092 (Table 2). Confidence intervals did not include zero in any pair-wise comparison indicating statistically significant differences between all population pairs.

The pairwise genetic distances between samples showed clear, highly supported (100% of bootstraps) divergence into East Atlantic (Spain and Scotland) and Scandinavian clades (Fig.3). The majority of individuals in each population clustered together with few exceptions: two individuals (40, 45) from Spain had the closest resemblance to basal Scottish samples, and one sample (1) from northern Norway, Bodø clustered together with Varberg samples from southern Scandinavia.

Inspecting clustering patterns within the main clades, some interesting patterns appeared. First, East Atlantic populations were clearly divided into two populations (Spain and Scotland), but many individuals within these populations were not very different from each other on a genomic scale (i.e. separating branch nodes between them were short and/or not supported by bootstrapping). Very different patterns emerged for the Scandinavian samples, however: instead of clear subclades, genetic distances between Bodø and Varberg increased more gradually. Also, differences between individuals within populations were in many cases smaller than in the East Atlantic clade, and especially so in Varberg.

Population clustering analysis based on discriminant analysis of principal components (Fig. 4) showed a very similar pattern to the analyses described above; largest separation between north and south, clear distinction between Scotland and Spain, as well as Scandinavian populations rather close each other but still as distinguishable clusters, and Bodø closer to East Atlantic populations than Varberg. All individuals were re-assigned back to their populations of origin with no signs of admixture (Figs. S3a and b).

Tests for evidence of selection indicated the presence of many outliers in both datasets with both approaches (see Figs. S4a-b, 5a-d and 6a-d). For the whole dataset including all four populations, 1379 SNPs (4.2%) were suggested as outliers with PCadapt, and 1209 (3.7%) BayeScan. 413 of the selected loci displayed concordance between methods. When considering Scandinavian populations only, 372 SNPs (1.2%) deviated from expectations under neutrality according to PCadapt analysis and 203 (0.7%) with BayeScan. 83 of these loci displayed concordance between the methods. Despite the applied distance filter of at least 1000bp between SNPs, some contigs (0.9%) contained more than one or two outlier SNPs (Supp. Table 5); possibly as a sign of stronger selection affecting many SNPs along the same sequence. Based on these results, it is possible that selection plays a role in population differentiation on large geographic scale as well as within Scandinavia. However, with no annotated reference genome available for goldsinny wrasse (or any closely-related species), the possible biological significance related to the detected outlier SNPs remains to be investigated in the future.

2.3. Validation of selected 173 SNPs to separate Scandinavian populations

A total of 231 SNPs was pre-selected as possibly diagnostic based on the estimated high divergence from sequence data between the two Scandinavian populations (Table S1a). Of these, 173 (74.9%) produced reliable genotypes with the used genotyping platform. These loci provided independent genetic information, i.e. no significant linkage between them after FDR correction was found (Fig. S7). Mean expected heterozygosity for the loci was high (0.387), and rather similar for both populations (Table 1; for locus-wise information, see Table S3), but observed heterozygosities were much lower showing significant mean overall heterozygote deficit ($F_{IS} = 0.228$; 95% CI 0.183-0.272), as well as in both populations separately (Varberg: $F_{IS} = 0.196$; 95% CI 0.131-0.232 and Bodø: $F_{IS} = 0.263$; 95% CI 0.222-0.279). Many of the used loci showed significant deviations from HWE after FDR correction: 37 loci (21.4%) in Varberg, and 46 (26.6%) in Bodø. Closer inspection by eye of the deviating loci revealed that for many (but not all) non-HWE loci alternate alleles were predominant in south and north (data not shown).

Using the reduced set of 173 putatively diagnostic SNP markers, the mean F_{ST} between Varberg and Bodø was 0.065 (Table 2), almost three times higher than the average estimated with the 33k dataset. However, compared with the used pre-selection criterion of the SNPs ($F_{ST} \geq \sim 0.15$ based on sequence data from 30 individuals), the observed divergence from 94 genotypes for these same loci was surprisingly low (Table S3), and the correlation between these two methods was weak even when F_{ST} was calculated from an identical set of individuals and markers (Fig. S8a; $R^2=0.028$, $p=0.029$). Locus-wise F_{ST} measurements between the populations showed that only 69 of the 173 loci (39.9%) were significantly differentiated ($p < 0.05$; not corrected for multiple comparisons; Table S3) and thus contributed to the overall divergence (Fig. S9). If considering only these 69 loci with significant divergence between the populations, the mean F_{ST} was 0.140. The discrepancy between the methods reduced when a second step of individual-based quality filtering was applied: The observed frequencies of matching, missing and non-matching genotypes were respectively 64%, 9% and 27% for the initial set of 173 SNPs, whereas it was 76%, 6% and 18% for the set of 74 SNPs. Similarly, the correlation between obtained F_{ST} from the sequence data and genotype data passed from $R^2=0.028$ with the 173 initial SNPs to $R^2=0.344$ ($p < 0.001$) with the set of 74 SNPs (Figs. S8b-c). The better reproducibility of 74 loci did not hold, however, when the larger dataset of 47x2 fish was compared with the sequence data ($R^2=0.047$, $p=0.064$; Fig. S8d). Locus-wise F_{ST} s derived from small (15x2) and large (47x2) datasets derived from genotyping were strongly and significantly correlated (for 173 loci, Fig. S8e: $R^2=0.391$, $p < 0.001$; for 74 loci, Fig. S8f; $R^2=0.500$, $p < 0.001$) suggesting that the used genotyping method gives more consistent estimates, and that the observed discrepancy likely stems from sequences. Possible explanations and implications of discrepancies between methods are discussed below.

The 173 SNPs provided a high level of accuracy in assigning individuals back to their populations of origin. DAPC-based analysis gave a mean assignment probability of 97.9% (Fig. 5); 100% for samples originating from Varberg and 95.7% for samples from Bodø. Two fish (4.3%) sampled in Bodø had high membership probability (≥ 0.9) into the southern population, but otherwise individuals showed very little admixture. Evaluation of the assignment across 360 tests from Monte-Carlo cross-validation showed also high accuracy. Mean assignment for Bodø was 0.95 (± 0.07 S.D.) and for Varberg 0.94 (± 0.08). Assignment accuracy was in general high ($\sim 90\%$ or higher), but varied somewhat depending on the number of loci used and samples analyzed (Fig. 6).

3. Discussion

This is the first population-genomic study of the goldsinny wrasse, a marine fish that has recently reached high economic value and harvest exploitation due to its importance as a cleaner-fish in commercial salmonid aquaculture throughout the North Atlantic. Based on the production of a *de novo* assembly, whole genome re-sequencing and identification of SNPs, we demonstrated that goldsinny wrasse displays a highly polymorphic genome of ~ 600 Mbp, and substantial population genomic structure throughout its native range. We were also able to identify and validate sub-sets of loci that were collectively diagnostic between samples of goldsinny wrasse from northern Norway and Sweden. These

SNPs now provide an efficient tool for investigating inadvertent translocations and possible non-native introgression of wrasse from the harvest and export regions in southern Norway and western Sweden to the import aquaculture region in mid-Norway.

Teleost fish genomes are known to vary a lot in size (see e.g. Malmstrøm *et al.* 2017), and thus generalizations of genomes are hard to make. However, the two other species in the *Labridae* family that are also used as cleaner fish in commercial salmonid aquaculture, have recently been assembled and their genomes were 805 Mbp for the ballan wrasse (Lie *et al.* 2018) and 614 Mbp for the corkwing wrasse (Mattingsdal *et al.* 2018). Thus, the goldsinny wrasse genome is rather similar in size with its closest relatives that have been studied. From an evolutionary perspective, however, these cleaner wrasses are distantly related. It has been estimated that the basal split between the genus *Ctenolabrus* (including goldsinny wrasse) and the other two genera (*Labrus* with ballan wrasse and *Symphodus* with corkwing wrasse) occurred at least 14 MYA (Hanel *et al.* 2002), so their cross-species usability in e.g. reference-assisted genome assembly is likely limited. Genome-level compatibility between the species was tested in connection with this study (data not shown) by aligning the used 1222 >20k contigs against the available reference genomes in the GenBank® for both species (ballan wrasse; https://www.ncbi.nlm.nih.gov/assembly/GCF_900080235.1, corkwing wrasse; https://www.ncbi.nlm.nih.gov/assembly/GCA_002819105.1). Sequence uniformity was on average 71.68% between goldsinny and corkwing wrasse, and 77.87% between goldsinny and ballan wrasse. There was very wide variability in the similarity between pairs of sequences, however, ranging from ~10% to >90% match between sequences suggesting largely differing genomic compositions between the species.

The high genomic variability revealed here (Fig.2 showing bimodal kmer distribution and estimated ~10 million variable sites in the reference) is consistent with an abundant marine fish covering large distribution area and with partly pelagic eggs (Hilldén estimated in 1984 that ~10% of the eggs float and may be transported by currents). This is because large (effective) population size and/or high connectivity between populations are known to be positively correlated with genetic diversity (e.g. Frankham 1996; Allendorf & Ryman 2014).

High genetic variability was also observed in all the four studied populations, with an average of 30% heterozygosity or more (Table 1). This is well in line with the detected levels of variability in our previous study (0.349-0.367; see Table 2 in Jansson *et al.* 2017) employing a set of 36 SNPs. Due to the protocol implemented here to discover polymorphic SNPs however, it is likely that rare allele variants were not included, thus inflating the observed mean level of variability, and that some part of the observed differences in variability between populations are due to sampling biases (see e.g. Lachance & Tishkoff 2013; Maroso *et al.* 2018). Our *de novo* reference was assembled from a fish caught in Varberg, a population which also displayed the highest heterozygosity levels (Table 1), and contrary to other populations (which showed slight but significant heterozygote deficiencies), significant general heterozygote excess on the genomic scale. To rule out if biological processes, such as non-random mating and gene flow (Allendorf *et al.* 2010), account for these observed general genetic patterns and the

differences observed between the populations, population studies including more samples, would be needed (Flesch *et al.* 2018). Interestingly, in our previous study (Jansson *et al.* 2017) with 36 SNPs and 14 microsatellites, the same set of samples from Varberg ($N=94$) showed significant heterozygosity deficiency with both marker types. These 50 markers were developed from ddRAD sequences without any reference genome (Jansson *et al.* 2016) using four other populations than Varberg, implying that the deviant heterozygosity pattern observed here for Varberg, could be a technical artefact.

Results from analyses using the genome wide panel of 33k SNPs revealed that goldsinny wrasse populations from different parts of the species distribution area (Fig. S1) are clearly and significantly differentiated from each other. In fact, the observed level of pairwise divergence ($F_{ST}= 0.015-0.092$; Table 2) is somewhat higher than in our previous study with 36 SNPs ($F_{ST}=0.013-0.049$; see Table 3 in Jansson *et al.* 2017), and in general quite high compared with many other marine organisms with pelagic life stages (e.g. Benestan *et al.* 2015; Carreras *et al.* 2017; Westgaard *et al.* 2017; Dahle *et al.* 2018a). Divergence between goldsinny wrasse populations is not necessarily due to geographic distance and restricted gene flow alone. A notable proportion, $\sim 1\%$ within Scandinavia and $\sim 4\%$ globally, of the diverged SNPs were identified as outliers likely under selection (Figs. S4-6). However, GenBank® nucleotide searches for the top-outlier SNPs did not retrieve any hits. Thus, their possible biological role (see e.g. Nielsen *et al.* 2009; Diopere *et al.* 2018) remains to be studied in the future when an annotated genomic reference, and/or genetic linkage map is available.

The constructed individual-based genomic phylogeny tree (Fig. 3) suggests that Scottish and Spanish populations are closer to North than South Scandinavian populations, and thus, that goldsinny wrasse in South-Scandinavia might stem from populations higher north along the coastline. This result is somewhat surprising because even though shortest oceanographic distances between the Scottish and both Scandinavian populations are roughly similar (Fig. 1), direct gene flow across the North Sea is highly unlikely (Jansson *et al.* 2017), and thus South Scandinavia is more reachable along the coastline. Moreover, within Scandinavia, passive drift – and thus possibly also gene flow is predominantly unidirectional from south to north along the Norwegian Coastal Current (Jansson *et al.* 2017). Considering a longer-time evolutionary context could elucidate the result, however. After the last ice age, the Scandinavian Ice Sheet retreated gradually around 20 000-10 000 years ago, starting from the Danish and Norwegian west coasts (Hughes *et al.* 2016). Depending on colonization routes and modes of dispersal, it is possible that many fish species first recolonized the southwestern corner of the present-day Norway, and spread from there north and/or south when more of the coastline re-emerged. Such population history of step-wise colonization would explain the observed phylogeny for goldsinny wrasse (see also e.g. Gysels *et al.* 2015; Almada *et al.* 2017). In a recent study, Mattingsdal and colleagues (2020) showed that the current-day population structure of corkwing wrasse in Scandinavia characterized by a substantial division between western and southern Scandinavia can mainly be explained with past demographic events followed by reproductive isolation and genetic drift. Similar colonization history was newly suggested for ballan wrasse (Seljestad *et al.* 2020). Unlike corkwing and ballan wrasse for which Scandinavian western and southern populations are quite isolated (but see Mattingsdal *et al.* 2020

proving some gene flow across the genetic break), the Scandinavian goldsinny population is characterized by extensive gene flow following oceanic currents, weak general population structure, increasing genetic divergence with oceanographic distance (i.e. isolation-by-distance, IBD), and with no clear breaks (Jansson *et al.* 2017). We conclude that the present-day population structure of goldsinny wrasse is therefore likely to represent a combination of past demographic processes shaping structures on larger scale, and dispersal mainly between nearby areas leading to limited gene flow and IBD. Even though demographic history and genetic drift would be the dominant evolutionary processes shaping the contemporary population patterns, other factors like isolation-by-adaptation, polygenic selection, and human interference may also play an important role and require further study.

Genome-wide datasets represent powerful tools with which to detect subtle population genetic differentiations (Benestan *et al.* 2015; Carreras *et al.* 2017; Mullins *et al.* 2018), patterns of selection (Vitti *et al.* 2013; Besnier *et al.* 2014; Price *et al.* 2018), as well as other selective responses such as genome rearrangements (Kirubakaran *et al.* 2016; Sodeland *et al.* 2016). However, at the present, genomic analysis does not provide a cost-effective approach for many fisheries management purposes (Martinsohn *et al.* 2019) where screening many individuals with fewer targeted or diagnostic loci is often more feasible approach (Glover *et al.* 2013; Malde *et al.* 2017; Dahle *et al.* 2018b; Johansen *et al.* 2018). The panel of 173 SNPs developed in the present study to separate Scandinavian goldsinny wrasse populations from north and south proved to have high assignment accuracy. Compared with the random genomic panel with an average F_{ST} of 0.024 between the populations (and 0.017 measured with 36 SNPs in Jansson *et al.* 2017), the putatively diagnostic loci developed here showed almost threefold average differences in divergence ($F_{ST} = 0.065$). The divergence based on sequence data was much higher, however (Table S1a), and there were large differences between loci between the expected and observed level of F_{ST} (Table S3, Fig. S8a). These discrepancies are probably due to different reasons: First, the sequence data consisted only of 15 fish per population compared with 47 per population genotyped for validation. Small sample sizes are adequate for estimating general genetic differentiation between populations on a genomic level (Willing *et al.* 2012), but for any specific locus likely much more sensitive to biases. Also, the SNP selection itself can introduce systematic upward bias when loci are screened for maximum divergence for population assignment without appropriate cross-validation procedures (Anderson 2010). Furthermore, NGS methods are prone to diverse errors (e.g. Pfeiffer *et al.* 2018; Maroso *et al.* 2018) also likely partly accounting for the lower than expected observed divergence, and lack of correlation between the two methods. Post hoc analysis of the selected SNPs revealed that individual filtering of SNPs based on coverage and quality could improve reproducibility between methods (Fig. S8a-c). This is in line with the earlier observation that variance in read coverage between individuals and between loci in the same individual introduce biases (Maroso *et al.* 2018). However, based on results from this study, larger sample sizes are likely also needed in search of reliable outliers between populations (Fig. S8c cf. S8d; see also Anderson 2010). Using another sequencing approach, such as some reduced representation method, could have given us larger samples per population and thus higher precision of observed allele frequencies. On the other hand, this in turn would have come on the expense of genomic coverage, and

thus possibility to detect (any) adaptive outliers (Hoban *et al.* 2016), and could also introduce other type of errors and biases (see e.g. Maroso *et al.* 2018; Wright *et al.* 2019; Graham *et al.* 2020).

From the individual assignment aspect, it does not matter if allelic differences between populations have arisen from neutral or adaptive evolutionary processes. However, SNPs associated with adaptive processes, and potentially local adaptations are more likely to remain divergent even in the face of moderate or even high levels of gene flow (e.g. Dennenmoser *et al.* 2017). Such SNPs are thus biologically and from a management perspective important. In this study, two individuals caught in Bodø (4.2%; Fig. 5) were strongly assigned to the southern population (Varberg). Natural, direct gene flow between so distant locations is very unlikely (Jansson *et al.* 2017) raising a question of possible translocations of wrasse from aquaculture. Faust and colleagues (2018) showed transported wrasses can survive and hybridize with local fish. In their study, of the sampled 40 wild corkwing wrasses in Flatanger area in mid-Norway, 2 were translocated fish, 1 first-generation hybrid and 12 possible second-generation hybrids. Considering that corkwing and goldsinny wrasse are by far the most abundant catches (between 8 to 12 million each annually during the last five years; Directorate of Fisheries; www.fiskeridir.no), and transported in great numbers, human-mediated gene flow and hybridization is a plausible sequence of events. But to decide whether these two fish are indeed translocated fish from south and/or hybrids between translocated and local fish, and how frequent such introgression would be in recipient areas in mid-Norway, the population structure and composition between these two sampling localities must be investigated in detail using the diagnostic SNPs developed here.

4. Conclusions

In this study, full-genome sequencing was used to produce a *de novo* assembly and thereafter investigate the population genomic structure of goldsinny wrasse in four geographically distinct locations. This is an economically important marine fish that is subjected to high harvest rates in some regions, and translocation to fish farms in other regions where it is used to delouse farmed salmonids. We demonstrated that the goldsinny wrasse genome, ~600Mbp in size, possesses high level of genetic variation, and that populations on different sides of the species distribution area are genetically significantly differentiated. Based on the conducted screening of outlier loci, some of the genetic differences observed among these populations are likely to be associated with functional divergence. For the two Scandinavian populations, we tested and evaluated a panel of putatively diagnostic SNP loci that proved to provide high resolution. Such markers now provide a tool with which to study human-mediated translocation between geographic areas through aquaculture practice.

5. Methods

In this study, new genetic markers for goldsinny wrasse were developed and validated by: I) sequencing one goldsinny wrasse sample with high coverage to build a genomic reference *de novo*, II) selecting the longest contigs from the assembly, and III) mapping individual reads from 60 fish from four populations (15 from each) against the reference contigs to call SNPs. A random subset of the high-quality SNPs

obtained (1 SNP/~kbp, in total 33 235), were then used to, IV) run population genetic analyses to explore basic population parameters, and to characterize genetic differentiation among the populations as well as to look for signs of selection. Finally, V) a set of 173 candidate SNP loci showing large divergence between northern and southern Scandinavian populations was selected and validated using Agena MassARRAY® iPLEX platform with an additional set of samples ($N=94$).

5.1. Samples and sequencing

60 goldsinny wrasse from four locations were used to search for and develop SNP markers against the reference contigs from one individual. 16 individuals (including a single fish used to build a reference genome) were collected from Varberg (VAR) in south-western Sweden, 15 from Bodø (BOD) in Northern Norway, 15 from Isle of Mull in Scotland, UK (SCO), and 15 from Galicia region (GAL) in north-western Spain (Fig. 1). These samples were collected in 2014-2016, and represent a subset of the samples used in a previous study of population genetic structure (Jansson *et al.* 2017). Samples were collected in compliance with EU Directive 2010/63/EU, and the national legislations in each country. Fish were killed upon catch and samples were taken immediately or killed and whole fish stored frozen until sampling in laboratory facilities. Details on data collection and DNA extraction are provided in Jansson *et al.* 2017. Samples were selected I) to cover the species' north-eastern Atlantic distribution area, II) to represent genetically most diverged populations from this area (based on results in Jansson *et al.* 2017), and III) being of good DNA quality and quantity. DNA quality was assessed by running samples on a 1% agarose gel, measuring their absorbance by a Nanodrop spectrophotometer, and by estimating DNA quantity with Qubit Quant-iT kit from Invitrogen. Superior samples (size $>10\ 000$ bp, $Abs_{260/280} = 1.8-2.0$ and $Abs_{260/230} = 1.8-2.4$) from each population were selected for sequencing (see below). 4 μ g of DNA from the selected reference individual (*VAR-ref-77*; sex unknown) and 2.5 μ g from each selected individual from the four populations were sent to the Norwegian Sequencing Centre (NSC) for sequencing. At the NSC, Illumina TruSeq adapter ligation was used to construct libraries, DNA fragmented to 300 bp target size, and barcoded to enable individual identification. Sequencing was done using an Illumina HiSeq X instrument producing 2x150 bp paired-end (PE) reads. The reference individual was run in a single lane, whereas the rest of the samples (60) were pooled in four lanes according to their origin (15 individuals/lane, all individual barcoded). Each lane produced between 302-316 GB of data as FASTQ files; for the reference individual, we obtained 451.1 million read pairs, and for pooled population samples 455.8–474.5 million read pairs per lane (on average $30.7 \pm 12.9 \times 10^6$ read pairs per individual, range 16.0-102.4 $\times 10^6$).

5.2. Building and evaluating the *de novo* reference

Read quality was checked with FastQC (Andrews 2010), and followed by trimming of sequences with Trimmomatic (v. 0.32; Bolger *et al.* 2014) where adapter contaminants and low-quality reads were removed using the following setting: phred score ≥ 33 , leading and trailing low-quality bases removed, scanning in 4 bp windows and cutting when the average quality per base drops below 15, minimum length for a sequence to be retained = 36 bp. Only correctly paired, high-quality reads were retained for the next phase, and consisted of 92.5% (~417 million) of the initial raw reads. Genome size and average

sequencing coverage was estimated using K-mer statistics with a K-mer size of 32 and software *kmx* (K-Mer indeXing; available at: <https://github.com/ketil-malde/kmx>). Expectation-maximization was used to fit a negative binomial distribution to the error K-mers, and Poisson distributions for haploid, diploid, and repeat K-mers.

To build a *de novo* reference, we used SOAPdenovo2 (v.2.0.4; Luo *et al.* 2012) with default parameter values and increasing K-mer sizes (bp) from 43 to 123 (every tenth), and with the assembler's possible maximum value of 127. Of these 10 assemblies, the assembly with K-mer size 127 produced the longest continuous scaffolds (based on scaffold length and *N50*), and was selected as our genomic reference. The reference quality was assessed based on contig lengths and their total amount (i.e. fragmentation), content of core genes (i.e. how many percent of ray-finned fish core genes were found using BUSCO (Benchmarking Universal Single-Copy Orthologs, v. 2.0.1; Simão *et al.* 2015)), and by checking the mapping success of each individual against the reference (as % of reads mapped).

5.3. SNP calling, selection and validation

Individual reads from all 60 fish were aligned against the *de novo* reference genome using *mem* method in BWA (v. 0.7.5a; Li & Durbin, 2009a), using default parameters. Sorting, indexing and variant calling was done with SAMtools (v.0.1.19; Li & Durbin, 2009b), and followed by filtering with BCFtools (v. 0.1.17; Danecek *et al.* 2011). Supplementary alignments were removed, and due to computational constraints, only reference scaffolds >20kbp were included. The *vcfR* package (Knaus & Grünwald 2017) in *R* (version 3.2.2; R core team 2015) was used to scrutinize quality and quantity of the aligned sequences, and the following filtering criteria (per population) was chosen for any SNP to be accepted: min_QUAL=600, min_DP=20, max_DP=999, min_MQ=0, and max_MQ=90. Due to observed discrepancies between the expected genotypes from mapping data and the observed genotypes from the genotyping platform (see 2.3 in Results), an additional *a posteriori* step of SNP selection was implemented. The output of the mapping was re-examined, and the average of the individual phred-scaled quality score was calculated for each of the 173 SNPs within Bodo and Varberg samples separately. The average individual phred-scaled mapping quality ranged from 0 to 100 in the Bodo sample and from 0 to 80 in the Varberg sample. SNPs with both average quality larger than 50 in Bodo sample and larger than 40 in Varberg samples, were selected as a subset of the 74 most robust markers.

Two separate data sets were created to: 1) Study population genomic patterns and divergence among the populations. For this, a random subset of the high-quality SNPs was selected along the longest scaffolds so that the physical distance between them was ≥ 1000 bp. This resulted in a data set of 33 866 SNPs. 2) Select putatively diagnostic SNPs for pair-wise separation of the four populations. To do this, the SNPs showing highest pairwise *F_{st}* between each population pair were selected from each scaffold. These SNPs were thereafter ranked from the highest to lowest pairwise *F_{ST}* (for each pair separately) down to a value of ~ 0.15 , leaving a few hundred candidate loci per pair (~ 450 -850; Tables S1a-f). To validate the SNPs that separated the Northern and Southern Scandinavian populations, 231 of the top-SNPs were selected and organized into eight multiplex groups (28-30 SNPs in each; Supplementary Table 2) using

the MassARRAY® Typer 4.0 Assay Design software (Agena Bioscience). Putatively diagnostic SNPs for other pair-wise comparisons were developed (Supplementary Tables 1b-f), but not validated in this study. Genotyping to validate the diagnostic SNPs developed between the northern and southern Scandinavian samples was performed on a MassARRAY® Typer 4.0 Analyser (Agena Bioscience). Only loci that were polymorphic and produced clear clustering patterns, were selected leaving 173 SNPs. 47 samples from Bodø in Northern Norway and Varberg on the Swedish West coast representing roughly the current edges of the species' distribution area in Scandinavia, were genotyped and analysed.

5.4. Population analyses of genome wide data

First, 609 of the 33 866 SNPs showing no or very little variation ($MAF < 0.01$) were removed. Remaining loci were checked for possible deviations in Hardy-Weinberg equilibrium (HWE) using *pegas* package v. 0.11 (Paradis 2010) in *R*. 22 SNPs not in HWE after False Discovery Rate correction (FDR; Benjamini & Hochberg 1995) in at least two of the studied populations, were removed from the final dataset ($N_{loci} = 33\ 235$). Linkage disequilibrium was tested by computing the correlation coefficient (r^2) between all pairs of loci that were physically linked on the same scaffold. In addition, a random set of 100 000 pairs of loci was sampled across scaffold, to produce a reference distribution of r^2 values without physical linkage. The *R* package *hierfstat* (v. 0.04-28; Goudet 2005) was used to calculate gene diversity (H_S), and observed heterozygosity (H_o) and overall F_{IS} within populations. Pairwise F_{ST} values (Weir & Cockerham 1984) were estimated for each population pair and over all samples using the same package, and their statistical significance was determined comparing the observed values with 95% confidence limits determined by 1000 bootstrap repeats. For examination of genetic relatedness between samples, *R* packages *poppr* (v.2.8.1; Kamvar *et al.* 2014) and *ape* (v. 5.2; Paradis *et al.* 2004) were utilized to reconstruct a genetic distance tree based on information from all 33 235 SNPs using UPGMA algorithm and 100 bootstrap replicates to assess branch support.

To investigate population structure further, and test assignment probability of individuals back to populations based on their genotypes, we performed a Discriminant Analysis of Principal Components (*DAPC*; Jombart *et al.* 2010) using the *R* package *adegenet* (v. 2.1.1; Jombart 2008; Jombart and Ahmed 2011). This is a multivariate method useful for exploring separation of groups in large genomic datasets, even when the general level of divergence between populations is low (Jombart *et al.* 2010). The method identifies structuring alleles, and maximizes among-group variation by first transforming the genotype data into principal components (PCs), followed by discriminant analysis (DA) to define the groups. The number of clusters (K) was determined by the *find.clusters* function with a maximum K set to 10. The lowest BIC (Bayesian Information Criterion) value was obtained with $K=4$. To avoid model overfitting, the optimal number of PCs was determined via α -optimization to be five (Fig. S2), which was then used in the following *DAPC* analysis and to define group membership (i.e. individual assignment probabilities to predefined populations).

To estimate if and to what extent natural selection could explain divergence between populations, the *pcadapt* (v. 3.0.4; Luu *et al.* 2017) package in *R* was employed to perform a genome scan to detect

markers potentially influenced by selection. This was done for the whole dataset, and separately for the Scandinavian samples only. The method first performs a principal component analysis (PCA) to ascertain the underlying population structure, and thus allows uncertainty of origin and admixed individuals. Based on obtained 'scree plots' from the PCA analysis, the optimal value of PCs was four for the whole dataset (Fig. S5a), and two for the Scandinavian dataset (Fig. S6a), and $K=4$ and 2, respectively, were used for the subsequent *pcadapt* analyses. SNPs with minor allele frequency below 0.05 were removed from the analysis leaving 32 712 SNPs in the whole and 31 941 SNPs in the Scandinavian datasets. A locus was considered as an outlier if its q-value threshold was below 0.1. Q-value estimation includes FDR and adjust p-values accordingly. Q-values were determined by *qvalue* package (v. 2.12.0; Storey *et al.* 2015) in *R*. Another selection test, BayeScan (v.2.1; Foll and Gaggiotti, 2008) was run for comparison for both datasets. Default parameter setting was used (prior odds 10, samples size 5000, thinning interval 10000, pilot runs 20, pilot run length 5000 and additional burn-in 50000). The decision whether a locus was under selection was based on q values (<0.05 suggests selection). Possible clustering of multiple (>2) selected SNPs along same contigs was checked.

5.5. Validation and population analysis using putatively diagnostic loci

All 94 genotyped samples produced reliable genotypes with more than 70% of the used 173 loci and were included in the analyses. Because these SNPs were picked as candidates for separating Scandinavian populations best, neutrality and thus HWE cannot be assumed. It was nevertheless investigated, together with other basic population parameters using same methods and approaches as given for the genome wide data analysis. Moreover, each locus pair was tested for linkage disequilibrium using the *snpStats* (v.; 1.32.0; Clayton 2018) package in *R* and visualized with *LDheatmap* (v. 0.99-5; Shin *et al.* 2006).

As for the genomic dataset, population subdivision and individual assignment was inspected with the *DAPC* package. To evaluate the discriminatory power for the baseline data, an additional Monte-Carlo cross-validation through resampling procedure was run with *R* package *assignPOP* (v. 1.1.4; Chen *et al.* 2018). The resampling scheme contained 50%, 70% and 90% of the individuals from both populations, and top (based on F_{ST}) 10%, 25%, 50% or all loci. Each resampling event was repeated 30 times.

List Of Abbreviations

bp	base pair
df	degrees of freedom
DNA	deoxyribonucleic acid
FDR	false discovery rate
HWE	Hardy-Weinberg equilibrium
IBD	isolation-by-distance

MPA marine protected area

MYA million years ago

N number (of)

NGS next generation sequencing

SD standard deviation

SNP single nucleotide polymorphism

Declarations

Ethics approval and consent to participate

Samples used in this study were collected by local fishermen and/or research personnel trained to take samples. All samples used were collected in compliance with EU Directive 2010/63/EU (<https://eur-lex.europa.eu/eli/dir/2010/63/oj>), and the national legislations in each country. Fish were killed upon catch and samples were taken immediately, or killed and whole fish stored frozen until sampling in laboratory facilities.

Consent for publication

Not applicable.

Availability of data and materials

Sequence reads from this study are available through NCBI archive from the following web page: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA508986>. The final genotype datasets used in this study are available from the Dryad repository: <https://doi.org/10.5061/dryad.rjdfn2z7s>. Lists of putative diagnostic SNPs identified in this study are given in Tables S1a-f.

Funding

This study was funded by the Norwegian Ministry for Trade, Industry and Fisheries, and the Swedish Research Council FORMAS and the Linnaeus Centre for Marine Evolutionary Biology (CeMEB). The Swedish Cultural Foundation in Finland (Svenska Kulturfonden) is acknowledged for personal grant to EJ during the early phases of this project. The funding sources were not involved in the study design or performance in any way.

Acknowledgements

We would like to thank Bjørghild Breistein for her help in the laboratory, and Stein Mortensen for acquiring the samples from Scandinavia and the British Isles. We acknowledge the support of the Galician Council

for Marine Affairs, as well as R. Barreiro, B. Carro and C. Caramelo from the University of A Coruña, for their help to obtain the samples from Galicia, Spain. Sequencing library creation and high-throughput sequencing was conducted at the Norwegian Sequencing Centre (NSC), Oslo, Norway.

Competing Interests

The authors declare there are no competing interests.

Authors' contributions

The study was designed by KG, EJ, and CA. EJ was responsible for the lab work with help from GD. FB, KM, and EJ were responsible for handling and analysis of the data. EJ wrote the manuscript with input from other authors. All authors read and approved the final manuscript.

References

1. Allendorf FW, Berry O, Ryman N. So long to genetic diversity, and thanks for all the fish. *Molecular Ecology*. 2014;23:23–25.
2. Allendorf FW, Hohenlohe PA, Luikart G. Genomics and the future of conservation genetics. *Nature Reviews Genetics*. 2010;11:697–710.
3. Almada F, Francisco SM, Lima CS, FitzGerald R, Mirimin L, Villegas-Rios D, et al. Historical gene flow constraints in a northeastern Atlantic fish: phylogeography of the ballan wrasse *Labrus bergylta* across its distribution range. *Royal Society Open Science*. 2017;4(2):160773.
4. Almada F, Casas L, Fransisco SM, Villegas-Rios D, Saborido-Rey F, Irigoien X, Robalo JI. On the absence of genetic differentiation between morphotypes of the ballan wrasse *Labrus berylta* (Labriadae). *Marine Biology*. 2016;163:86.
5. Anderson EC, Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources*, 2010;10:701–710.
6. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
7. Benestan L, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L. RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology*. 2015;24:3299–3315.
8. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1995;57(1): 289–300.

9. Berg PR, Jentoft S, Star B, Ring KH, Knutsen H, Lien S, et al. Adaptation to low salinity promotes genomic divergence in Atlantic cod (*Gadus morhua* L.). *Genome Biology and Evolution*. 2015;7:1644–1663.
10. Besnier F, Kent M, Skern-Mauritzen R, Lien S, Malde K, Edvardsen RB, et al. Human-induced evolution caught in action: SNP-array reveals rapid amphi-atlantic spread of pesticide resistance in the salmon ectoparasite *Lepeophtheirus salmonis*. *BMC Genomics*. 2014;15:937.
11. Blanco Gonzalez, E, Espeland, SH, Jentoft, S, et al. Interbreeding between local and translocated populations of a cleaner fish in an experimental mesocosm predicts risk of disrupted local adaptation. *Ecology and Evolution*. 2019; 9: 6665–6677.
12. Blanco Gonzalez E, Knutsen H, Jorde PE. Habitat discontinuities separate genetically divergent populations of a rocky shore marine fish. *PLoS ONE*. 2016;11(10).
13. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114-2120.
14. Carreras C, Ordóñez V, Zane L, Kruschel C, Nasto I, Macpherson E, Pascual M. Population genomics of an endemic Mediterranean fish: differentiation by fine scale dispersal and adaptation. *Scientific Reports*. 2017;7:43417.
15. Charlesworth B. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*. 2009;10:195–205.
16. Chen K-Y, Marschall EA, Sovic MG, Fries AC, Gibbs HL, Ludsin SA. *assignPOP*: Population Assignment using Genetic, Non-Genetic or Integrated Data in a Machine Learning Framework. *Methods in Ecology and Evolution*. 2018;9:439–446.
17. Clayton D. snpStats: SnpMatrix and XSnMatrix classes and methods. 2018. R package version 1.32.0.
18. Costello MJ. Review of the biology of wrasse (Labridae: Pisces) in northern Europe. *Progress in Underwater Science*. 1991;16:29–51.
19. Crispo E, Moore J-S, Lee-Yaw JA, Gray SM, Haller BC. Broken barriers: Human-induced changes to gene flow and introgression in animals. *Bioessays*. 2011;33:508-518.
20. da Fonseca RR, Albrechtsen A, Themudo GE, Ramos-Madrugal J, Sibbesen JA, Maretty L, et al. Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine Genomics*. 2016;30:1–11.
21. Dahle G, Quintela M, Johansen T, Westgaard JI, Besnier F, Aglen A, et al. Analysis of coastal cod (*Gadus morhua* L.) sampled on spawning sites reveals a genetic gradient throughout Norway's coastline. *BMC Genetics*. 2018a;19:17.
22. Dahle G, Johansen T, Westgaard JI, Aglen A, Glover KA. Genetic management of mixed-stock fisheries "real-time": The case of the largest remaining cod fishery operating in the Atlantic in 2007-2017. *Fisheries Research*. 2018b;205:77–85.
23. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools, *Bioinformatics*. 2011;27(15):2156–2158.

24. D'Arcy J, Mirimin L, FitzGerald R. Phylogeographic structure of a protogynous hermaphrodite species, the ballan wrasse *Labrus bergylta*, in Ireland, Scotland, and Norway, using mitochondrial DNA sequence data. *ICES Journal of Marine Science*. 2013;70(3):685–693.
25. Darwall WRT, Costello MJ, Donnelly R, Lysaght S. Implication of life-history strategies for a new wrasse fishery. *Journal of Fish Biology*. 1992;41(Supp. B):111–123.
26. Dennenmoser S, Vamosi S M, Nolte AW, Rogers S. M. Adaptive genomic divergence under high gene flow between freshwater and brackish-water ecotypes of prickly sculpin (*Cottus asper*) revealed by Pool-Seq. *Molecular Ecology*. 2017;26:25–42.
27. Diopere E, Vandamme SG, Hablützel PI, Cariani A, Van Houdt J, Rijnsdorp A, et al. Seascape genetics of a flatfish reveals local selection under high levels of gene flow. *ICES Journal of Marine Science*. 2018;75(2):675–689.
28. Faust E, Halvorsen KT, Andersen P, Knutsen H, André C. Cleaner fish escape salmon farms and hybridize with local wrasse populations. *Royal Society Open Science*. 2018;5(3):171752.
29. Flesch EP, Rotella JJ, Thomson JM, Graves TA, Garrott RA. Evaluating sample size to estimate genetic management metrics in the genomics era. *Molecular Ecology Resources*. 2018;18:1077–1091.
30. Foll M, Gaggiotti OE. A genome scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*. 2008;180:977–993.
31. Frankham R. Relationship of genetic variation to population size in wildlife. *Conservation Biology*. 1996;10:1500–1508.
32. Gagnaire P-A, Broquet T, Aurelle D, Viard F, Souissi A, Bonhomme F, et al. Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evolutionary Applications*. 2015;8:769–786.
33. Glover KA, Pertoldi C, Besnier F, Wennevik V, Kent M, Skaala Ø. Atlantic salmon populations invaded by farmed escapees: quantifying genetic introgression with a Bayesian approach and SNPs. *BMC Genetics*. 2013;14:4.
34. Glover KA, Solberg MF, McGinnity P, Hindar K, Verspoor E, Coulson MW, et al. Half a century of genetic interaction between farmed and wild Atlantic salmon: status of knowledge and unanswered questions. *Fish and Fisheries*. 2017;18:890–927.
35. Goodwin S, McPherson JD, McCombie R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*. 2016;17:333–351.
36. Goudet J. Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*. 2005;5:184–186.
37. Graham CF, Boreham DR, Manzon RG, Stott W, Wilson JY, Somers CM. How “simple” methodological decisions affect interpretation of population structure based on reduced representation library DNA sequencing: A case study using the lake whitefish. *PLoS ONE*. 2020;15(1):e0226608
38. Guo B, DeFaveri J, Sotelo G, Nair A, Merilä J. Population genomic evidence for adaptive differentiation in Baltic Sea three-spined sticklebacks. *BMC Biology*. 2015;13:19.

39. Gysels ES, Helllemans B, Pampoulie C, Volckaert FAM. Phylogeography of the common goby, *Pomatoschistus microps*, with particular emphasis on the colonization of the Mediterranean and the North Sea. *Molecular Ecology*. 2004;13:403–417.
40. Halvorsen KAT, Larsen T, Sørvalen TK, Vøllestad LA, Knutsen H, Olsen EM. Impact of harvesting cleaner fish for salmonid aquaculture assessed from replicated coastal marine protected areas. *Marine Biology Research*. 2017;13:359–369.
41. Hanel R, Westneat MW, Sturmbauer C. Phylogenetic relationships, evolution of broodcare behavior, and geographic speciation in the wrasse tribe Labrini. *Journal of Molecular Evolution*. 2002;55:776–789.
42. Hauser L, Carvalho GR. Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish and Fisheries*. 2008;9:333–362.
43. Hilldén N-O. Behavioural ecology of the Labrid fishes (Teleostei: Labridae) at Tjärnö on the Swedish west coast. 1984. Doctoral dissertation, University of Stockholm, Sweden.
44. Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, et al. Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *The American Naturalist*. 2016;188(4):379–397.
45. Hughes ALC, Gyllencreutz R, Lohne ØS, Mangerud J, Svendsen JI. The last Eurasian ice sheets – a chronological database and time-slice reconstruction, DATED-1. *Boreas*. 2016;45:1–45.
46. Jansson E, Quintela M, Dahle G, Albretsen J, Knutsen H, André C, et al. Genetic analysis of goldsinny wrasse reveals evolutionary insights into population connectivity and potential evidence of inadvertent translocations via aquaculture. *ICES Journal of Marine Science*. 2017;74:2135–2147.
47. Jansson E, Taggart JB, Wehner, S, Dahle G, Quintela M, Mortensen S, Kvamme BO, Glover KA. Development of SNP and microsatellite markers for goldsinny wrasse (*Ctenolabrus rupestris*) from ddRAD sequencing data. *Conservation Genetics Resources*. 2016;8:201–206.
48. Johansen T, Westgaard JI, Seliussen BB, Nedreaas K, Dahle G, Glover KA, et al. "Real-time" genetic monitoring of a commercial fishery on the doorstep of an MPA reveals unique insights into the interaction between coastal and migratory forms of the Atlantic cod. *Ices Journal of Marine Science*. 2018;75:1093–1104.
49. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*. 2008;24:1403–1405.
50. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*. 2010;11:94.
51. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*. 2011;27:3070–3071.
52. Kamvar ZN, Tabima JF, Grünwald NJ. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*. 2014;2:e281.
53. Kaur K, Jansen PA, Aspehaug VT, Horsberg TE. Phe362Tyr in AChE: A Major Factor Responsible for Azamethiphos Resistance in *Lepeophtheirus salmonis* in Norway. *PLOS ONE*. 2016;11(2):e0149264.

54. Kirubakaran TG, Grove H, Kent MP, Sandve SR, Baranski M, Nome T, et al. Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*. 2016;25:2130–2143.
55. Knaus BJ, Grünwald NJ. VCFR: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*. 2017;17:44–53.
56. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays*. 2013;35(9):780-786.
57. Laikre L, Schwartz MK, Waples RS, Ryman N, The GeM Working Group. Compromising genetic diversity in the wild: unmonitored large-scale release of plants and animals. *Trends in Ecology and Evolution*. 2010;25(9):520–529.
58. Lie KK, Tørresen OK, Solbakken MH, Rønnestad I, Tooming-Klunderud A, Nederbragt AJ, et al. 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*. 2018;19(1):186.
59. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 2012;1,18.
60. Luu K, Bazin E, Blum MGB. *pcadapt*: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*. 2017;17:67–77.
61. Malde K, Seliussen BB, Quintela M, Dahle G, Besnier F, Skaug HJ, et al. Whole genome resequencing reveals diagnostic markers for investigating global migration and hybridization between minke whale species. *BMC Genomics*. 2017;18:11.
62. Malmstrøm M, Matschiner M, Tørresen OK, Jakobsen KS, Jentoft S. Whole genome sequencing data and *de novo* draft assemblies for 66 teleost species. *Scientific Data*. 2017;4:160132.
63. Maroso F, Hillen JJ, Pardo BG, Gkagkavouzis K, Coscia I, Hermida M, Franch M, Hellemans B, Van Houdt J, Simionati B, Taggart JB, Nielsen EE, Maes G, Ciavaglia SA, Webster LMI, Volckaert FAM, Martinez P, Bargelloni L, Ogden R. Performance and precision of double digestion RAD (ddRAD) genotyping in large multiplexed datasets of marine fish species, *Marine Genomics*. 2018;39:64–72.
64. Martinsohn JTh, Raymond P, Knott T, Glover K A, Nielsen E E, Eriksen LB, Ogden R, Casey J, Guillen J. DNA analysis to monitor fisheries and aquaculture: too costly? *Fish and Fisheries* 2019;20:391–401.
65. Mattingsdal M, Jorde PE, Knutsen H, Jentoft S, Stenseth NC, Sodeland M, Robalo JI, Hansen MM, André C, Blanco Gonzalez E. Demographic history has shaped the strongly differentiated corkwing wrasse populations in Northern Europe. *Molecular Ecology*. 2020;160–171.
66. Mattingsdal M, Jentoft S, Tørresen OK, Knutsen H, Hansen MM, Robalo JI, et al. A continuous genome assembly of the corkwing wrasse (*Symphodus melops*). *Genomics*. 2018;110(6):399–403),
67. Mortensen S, Skiftesvik AB, Bjelland R, Karlsbakk E, Durif C, Sadlund N. Bruk av rensefisk i laksoppdrett. In Svåsand T, Grefsrud ES, Karlsen Ø, Kvamme BO, Glover KA, Husa V, Kristiansen TS. (Eds.), 2017;Risikoreport norsk fiskeoppdrett 2017 (pp. 162–173). *Fisken og havet, særnummer 2–*

2017. (in Norwegian, available at: http://www.imr.no/filarkiv/2016/04/risikovurdering_2016.pdf/nb-no)
68. Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D. Population genomics of marine fishes: identifying adaptive variation in space and time. *Molecular Ecology*. 2009;18:3128–3150.
 69. Paradis E. *pegas*: an R package for population genetics with an integrated-modular approach. *Bioinformatics*. 2010;26:419–420.
 70. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics Applications Note*. 2004;20:289–290.
 71. Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, Mayer G. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*. 2018;8:10950.
 72. Price N, Moyers BT, Lopez L, Lasky JR, Monroe JG, Mullen JL, et al. Combining population genomics and fitness QTLs to identify the genetics of local adaptation in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*. 2018;115(19):5028-5033.
 73. Quintela M, Danielsen EA, Lopez L, Barreiro R, Svåsand T, Knutsen H, Skiftesvik AB, Glover KA. Is the ballan wrasse (*Labrus bergylta*) two species? Genetic analysis reveals within-species divergence associated with plain and spotted morphotype frequencies. *Integrative Zoology*. 2016;11(2):162–172.
 74. R Core Team. R: A language and environment for statistical computing. 2015. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
 75. Russello MA, Kirk SL, Frazer KK, Askey PJ. Detection of outlier loci and their utility for fisheries management. *Evolutionary Applications*. 2012;5:39–52.
 76. Seljestad GW, Quintela M, Faust E, Halvorsen KT, Besnier F, Jansson E, Dahle G, Knutsen H, André C, Folkvord A, Glover KA. “A cleaner break”: Genetic divergence between geographic groups and sympatric phenotypes revealed in ballan wrasse (*Labrus bergylta*). *Ecology and Evolution*. 2020;00:1–16. <https://doi.org/10.1002/ece3.6404>
 77. Shafer ABA, Wolf JBW, Alves PC, Bergström L, Bruford MA et al. Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*. 2015;30:78–87.
 78. Shin J-H, Blay S, McNeney B, Graham J. LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms. *J Stat Soft*. 2006;16 Code Snippet 3.
 79. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31(19):3210–3212.
 80. Skiftesvik AB, Blom G, Agnalt A-L, Durif CMF, Browman HI, Bjelland RM, et al. Wrasse (*Labridae*) as cleaner fish in salmonid aquaculture – The Hardangerfjord as a case study. *Marine Biology Research*. 2014;10(3):289–300.

81. Skiftesvik AB, Durif CMF, Bjelland RM, Browman HI. Distribution and habitat preferences of five species of wrasse (Family Labridae) in a Norwegian fjord. *ICES Journal of Marine Science*. 2015;73(3):890–899.
82. Sodeland M, Jorde PE, Lien S, Jentoft S, Berg PR, Grove H, et al. "Islands of Divergence" in the Atlantic Cod Genome Represent Polymorphic Chromosomal Rearrangements. *Genome Biology and Evolution*. 2016;8:1012–1022.
83. Storey JD. (with contributions from Bassm AJ, Dabney A, Robinson D.) qvalue: Q-value estimation for false discovery rate control. 2015. Available from: <http://github.com/jdstorey/qvalue>
84. Sundt RC, Jørstad KE. Genetic population structure of goldsinny wrasse, *Ctenolabrus rupestris* (L.), in Norway: implications for future management of parasite cleaners in the salmon farming industry. *Fisheries Management and Ecology*. 1998;5:291–302.
85. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends in Genetics*. 2014;30:418–426.
86. Vitti JJ, Grossman SR, Sabeti PC. Detecting Natural Selection in Genomic Data. *Annual Review of Genetics* 2013;47(1):97–120.
87. Waples RS, Hindar K, Karlsson S, Hard JJ. Evaluating the Ryman-Laikre effect for marine stock enhancement and aquaculture. *Current Zoology*. 2016;62(6):617–627.
88. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*. 1984;38:1358–1370.
89. Westgaard J-I, Saha A, Kent M, Hellerud Hansen H, Knutsen H, Hauser L, et al. Genetic population structure in Greenland halibut (*Reinhardtius hippoglossoides*) and its relevance to fishery management. *Canadian Journal of Fisheries and Aquatic Sciences*. 2017;74:475-485.
90. Willing EM, Dreyer C, van Oosterhout C. Estimates of Genetic Differentiation Measured by F_{ST} Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers. *PLOS ONE*. 2012;7(8):e42649.
91. Wright BR, Grueber CE, Lott MJ, Belov K, Johnson RM, Hogg CJ. Impact of reduced-representation sequencing protocols on detecting population structure in a threatened marsupial. *Molecular Biology Reports*. 2019;46:5575–5580.

Tables

Table 1 Genetic diversity in four goldsinny wrasse populations based on 33k SNPs and 173 selected loci

	33235 SNP loci			173 SNP loci		
Population	<i>N</i>	<i>H_s</i>	<i>H_o</i>	<i>N</i>	<i>H_s</i>	<i>H_o</i>
Varberg	15	0.321	0.348	47	0.389	0.313
Bodø	15	0.319	0.307	47	0.385	0.284
Scotland	15	0.325	0.315			
Spain	15	0.319	0.307			
Mean		0.321	0.319		0.387	0.299

Table 2 Pairwise F_{ST} estimates (below diagonal) and corresponding 95% confidence intervals (above diagonal) based on 1000 bootstraps. Pairwise F_{st} value and its CI for the Scandinavian Varberg–Bodø pair is also given (in parenthesis) for the putatively diagnostic 173 SNPs genotyped on an extended set of samples.

	Varberg	Bodø	Scotland	Spain
Varberg	*	0.023-0.026 (0.050-0.085)	0.078-0.082	0.091-0.095
Bodø	0.024 (0.065)	*	0.071-0.075	0.082-0.085
Scotland	0.080	0.073	*	0.014-0.017
Spain	0.092	0.084	0.015	*

Figures



Figure 1

Map of sampling locations. For detailed information, see 5.1.

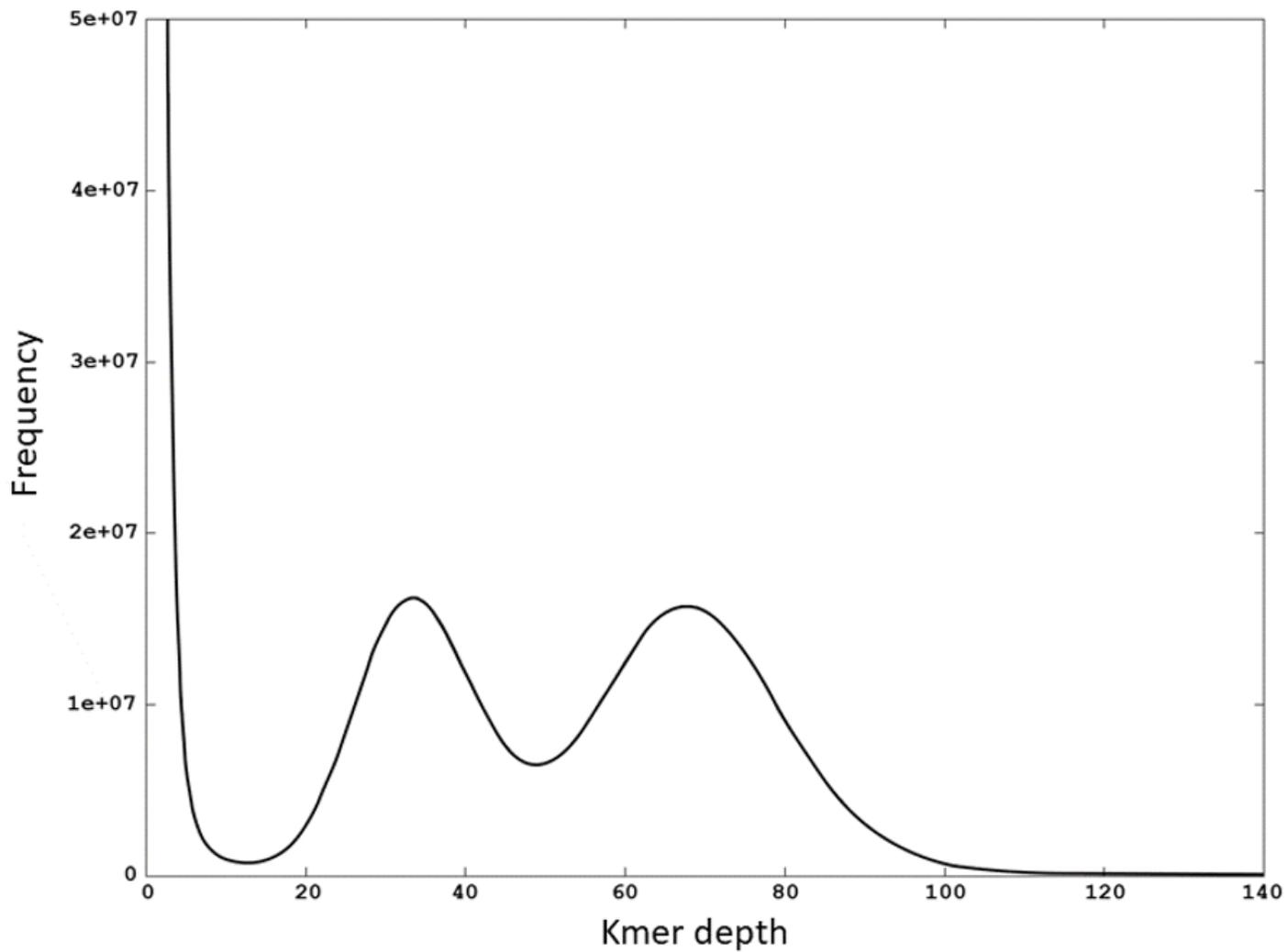


Figure 2

K-mer frequency distribution

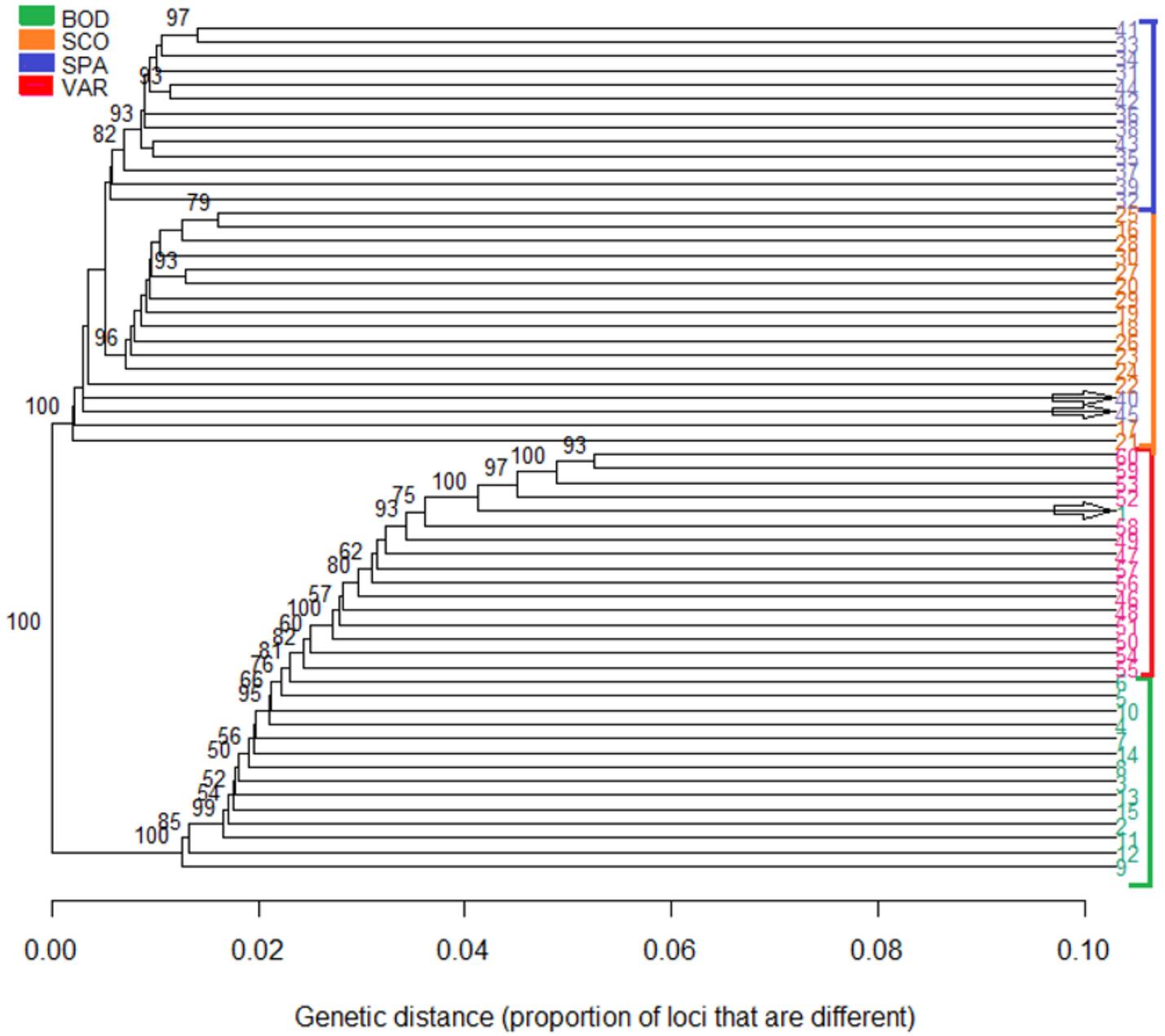


Figure 3

Genetic distance tree between samples based on 33 235 SNPs. Branch nodes supported by $\geq 50\%$ of bootstrap replicates are shown. Samples clustering with other populations than own are marked with arrows.

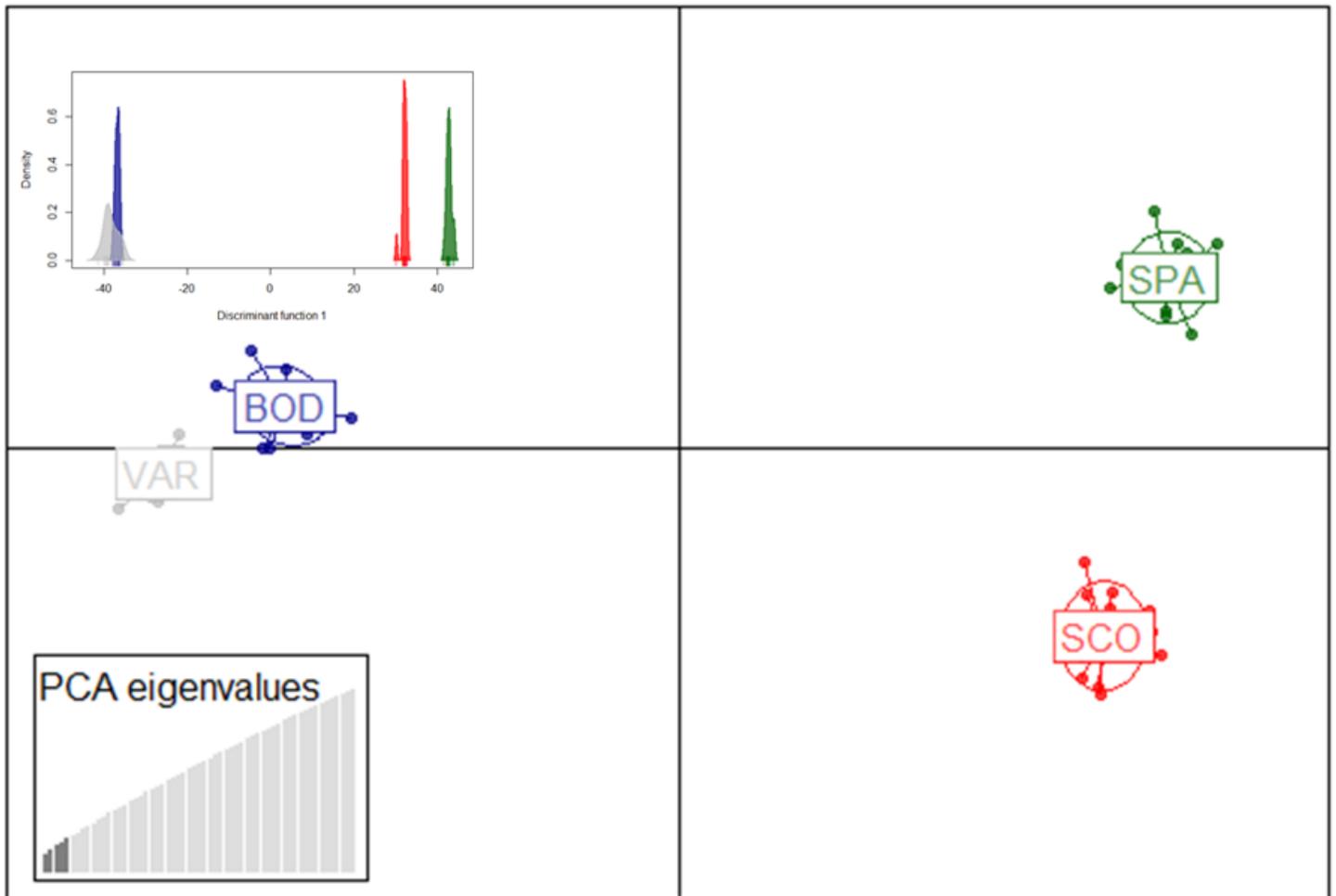


Figure 4

DAPC plot with 33k data. Optimized number of PCs (5; see Fig. S1), was used together with 3 (main figure) or 1 discriminant function(s) (small figure on upper left corner). Colour coding for populations: Varberg = grey, Bodø = blue, Scotland = red, and Spain = green.

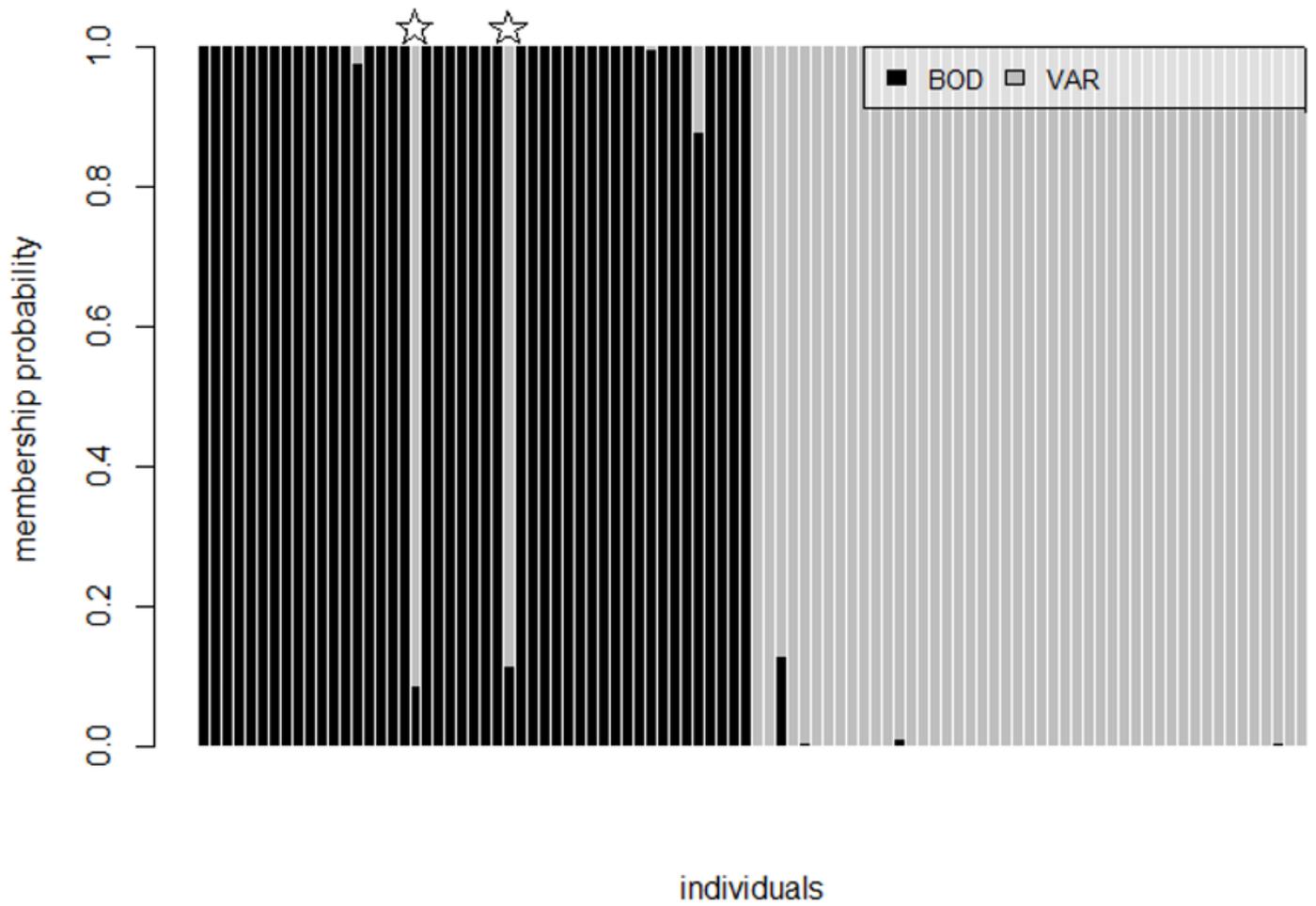


Figure 5

Compo plot of membership probability of the genotyped 94 individuals based on 173 loci. Individuals sampled in Bodø but assigned strongly to Varberg are marked with stars.

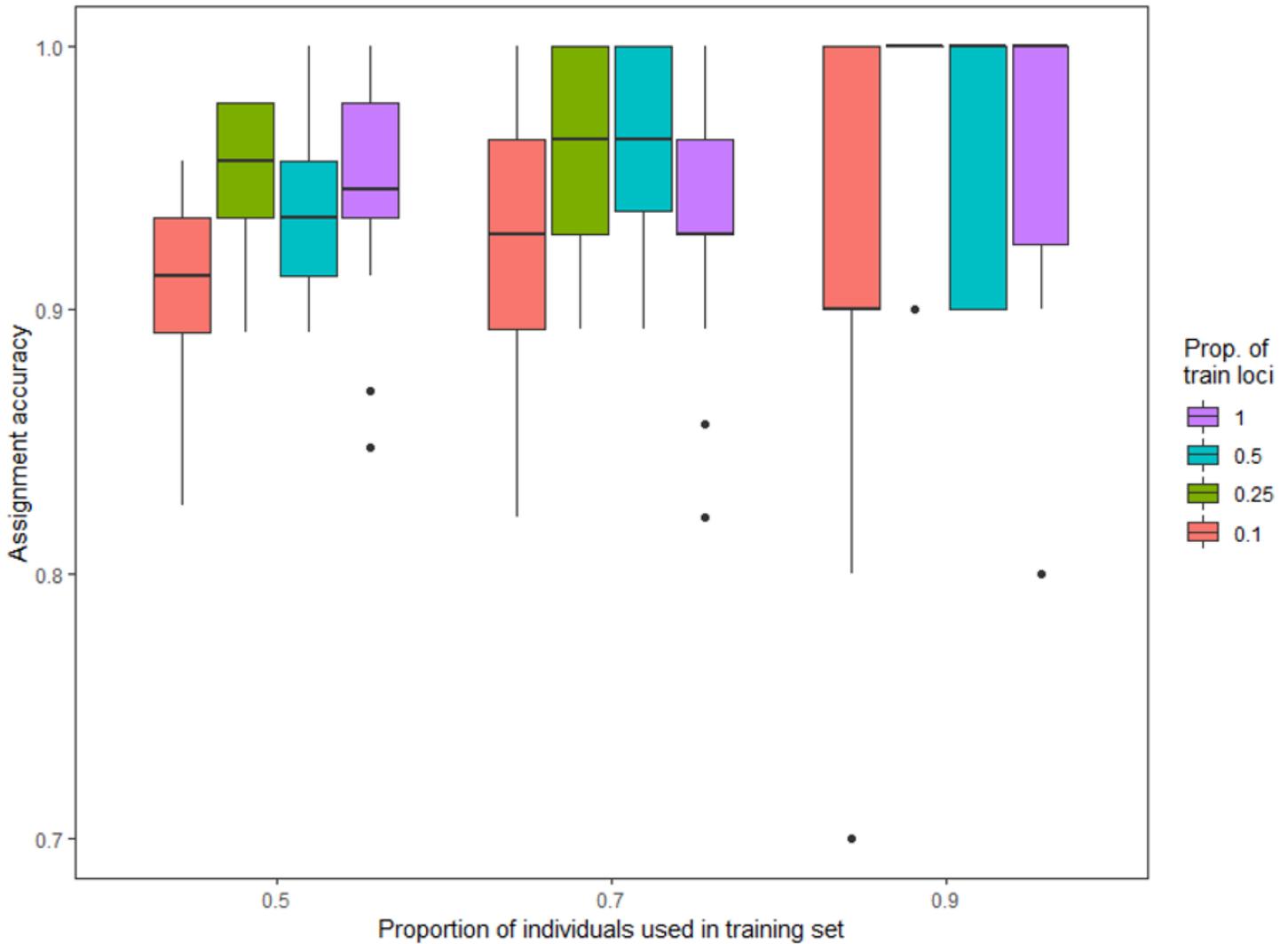


Figure 6

Overall assignment accuracies estimated via Monte-Carlo cross-validation. Three levels of training individuals (50%, 70% and 90% of individuals from both populations, on x-axis) were crossed by four levels of training loci (top 10%, 25% and 50% highest FST loci and all loci in color-coded boxes) by 30 resampling events. Results divided by populations are given in Fig.S10.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SuppTable5.xlsx](#)
- [SuppTable2.xlsx](#)
- [SuppTable3.xlsx](#)
- [SuppTable4.xls](#)

- [SuppTable1af.xlsx](#)
- [supplementaryfiguresS1S10.docx](#)