UiT The Arctic University of Norway

Faculty of Biosciences, Fisheries and Economics

Norwegian College of Fishery Science

Genomic characterization and insights of local adaptation in Norwegian juvenile lumpfish populations

Emilie Østby Granviken

Master's thesis in Marine Biotechnology (BIO-3901), May 2021

Preface

This thesis is submitted for the degree of Master of Science in Marine Biotechnology at UiT The Arctic University of Norway. The thesis constitutes 60 credits of the degree. The work was started in August 2020 and finished in May 2021, and was supervised by Kim Præbel, Enrique Blanco Gonzalez, Shripathi Bhat and Mathilde Horaud.

I would like to thank all my supervisors for invaluable guidance and feedback throughout the year. Particularly thanks to Enrique, who despite his relocation to UiA could continue to offer me his great supervision, which I could not have been without. And thanks to Kim, who was willing to be my main supervisor and kindly has offered me lots of great feedback. I owe Shripathi great thankfulness for all his help with the bioinformatics, which would have taken years if I were to figure out all on my own. I am also very grateful to Mathilde, who has followed me up and been a friendly support and collaborator all the way. In addition, I must thank Julie Bitz-Thorsen for teaching me everything in the laboratory. I feel very lucky to have had such a helpful and experienced team around me - without them I would have been completely lost.

Thank you to my super classmates for the nice time we have had together, though a certain pandemic put a limit to it this last year. It has been a pleasure to have someone to share all frustrations and joys (most frustrations) with. Also, many thanks to my boyfriend, who certainly has made me forget about most of the frustrations and brought me joy only. Equally many thanks to my family for all love and support, despite the fact that they have no clue of what I am doing (often I don't know myself either).

Lastly, thank you for having me, UiT. I admit there have been days when I have wondered why on earth I chose to study at this sometimes too dark, sometimes too light and almost always too cold place, but in the end, I am thankful for my time here and will never regret my choice.

Table of Contents

List of Abbreviations

BER	Bergen (sampling location)
FDR	False discovery rate
F_{IS}	Inbreeding coefficient
F_{ST}	The proportion of genetic variance within a (sub)population relative to the total genetic variance
H_{E}	Expected heterozygosity
H_{O}	Observed heterozygosity
HWE	Hardy Weinberg equilibrium
IMR	Institute of Marine Research
IBD	Isolation by distance
LYN	Lyngen (sampling location)
PC1/PC2	First/second principal component
PCA	Principal component analysis
SNP	Single nucleotide polymorphism
TRO	Tromsø (sampling location)
UiT	UiT The Arctic University of Norway

Abstract

The lumpfish (*Cyclopterus lumpus* Linnaeus, 1758) has for many years been an attractive target for the roe fisheries in Norway and has more recently become an important cleanerfish in the salmonid farming industry for the control of salmon lice (Copepoda: Caligidae). Despite this, there is a lack of knowledge about several life aspects of the species and studies on its genetic structure have been inconsistent. It is therefore uncertain how scenarios involving, for instance, overfishing or escaping of lumpfish from the fish farms might affect the natural stocks genetically and biologically. To initiate the investigations of how human activities might affect lumpfish in nature, this project aimed to clarify the genetic structure and reveal potential signatures of local adaptation in juvenile lumpfish in Norway. Whole genome sequencing was performed for 30 individuals from three different locations and a set of 607,663 SNPs were selected for downstream genetic structure analyses. Significant genetic differentiation was detected between the northern and southern parts of Norway, with further discrimination of two populations within one location in the north. Lumpfish separated by long geographical distances showed greater differentiation than those collected at locations more proximate to each other, but no significant correlation was detected. Loci putatively under selection were identified and revealed population structure at smaller geographical scales. These findings suggest that multiple driving forces may have contributed to population structuring in lumpfish. The study presents evidence for genetic structuring of Norwegian juvenile lumpfish and discriminate at least three populations, and further studies should therefore contribute to establishing sustainable management practices for the species.

Introduction

Cyclopterus lumpus (Linnaeus, 1758), commonly known as lumpfish or lumpsucker, is a fish species naturally found along the Norwegian coast that is much appreciated both by the fishery and fish farming industries (Eriksen et al., 2014; Holst, 1993). Lumpfish roe is consumed as a cheaper alternative to caviar (Johannesson, 2006), hence lumpfish fisheries have been going on since the 1940s (Kennedy et al., 2018). Since the Marine Stewardship Council (MSC) certification was obtained in 2017 (MSC, 2017), the demand and prices for Norwegian lumpfish roe have been rising (Norwegian Directorate of Fisheries, 2020a), and in 2020, a total of 219 tonnes of roe were landed by Norwegian fisheries (Durif, 2020). As the species has proven to graze on sea lice (Imsland et al., 2018), Caligidae, it has recently become an important resource as a cleanerfish in the production of farmed salmonids as well. Sea lice, used here as a collective term for the two species *Lepeophtheirus salmonis* and *Caligus elongatus*, are ectoparasitic copepods that make one of the greatest challenges to the salmonoid farming industry (Costello, 2009). The lice cause skin damage to the fish, making them susceptible to other secondary infections which in worst case may be lethal and result in high production losses (Costello, 1993). Moreover, sea lice transferred from the farmed fish are a serious threat to wild salmonids (Norwegian Scientific Advisory Committee for Atlantic Salmon, 2020; Thorstad et al., 2015), and this is one of the main factors impeding the continued growth of the industry (Norwegian Ministry of Trade, 2015). Different chemical treatments have traditionally been applied to fight the problem, but it remains a challenge that the sea lice develop resistance against them (Overton et al., 2019). Other treatments and methods for removal have consequently been adopted (Overton et al., 2019), among others the use of cleanerfish as a control agent against infestations. The main species applied for the purpose are different wrasses and lumpfish, of which the lumpfish is the most widely used (Norwegian Directorate of Fisheries, 2020b). Compared to wrasses, lumpfish have a shorter production cycle (Brooker et al., 2018) and a better performance at low temperatures (Geitung et al., 2020; Nytrø et al., 2014; Yuen et al., 2019), and is therefore a favourable choice particularly in Northern Norway and during the winter (L. Barrett et al., 2020). The use of lumpfish in the farming industry has thereof strongly increased in the last decade, and in 2019 nearly 43 million farmed individuals were sold to a total profit of more than 900 million NOK (Norwegian Directorate of Fisheries, 2020b).

The utilization of lumpfish as cleanerfish is however not without problems. The species have shown to be prone to a wide range of diseases (Erkinharju et al., 2021), and a large number die in the fish net pens (Stien et al., 2020). From the first large survey on cleanerfish mortality in the Norwegian fish farming industry it was reported that 46% of the lumpfish die during the production, mainly due to disease, the application of other lice-removal methods, and natural conditions (Stien et al., 2020). As the survival at the end of the production cycle often is not reported, it is suggested that the mortality might be even higher in reality. It is also suspected that some fish escape from the net pens (Herrmann et al., 2021), which has been reported to happen with wrasses (Faust et al., 2018) and is a familiar problem with farmed salmonids (Grefsrud et al., 2019). At present, the lumpfish used in aquaculture are juveniles farmed from wild broodstock (Powell et al., 2018), which may be collected from locations other than the farming areas. If the lumpfish escape, this possesses a risk for interference and interbreeding with local lumpfish populations that might be genetically different and which may be affected negatively by such interference. However, despite the active roe fisheries and high numbers of lumpfish that are being produced in aquaculture, the knowledge about their genetic composition is limited (Powell et al., 2018). To maintain sustainable fisheries and to be able to assess the consequences related to cleanerfish escapees, more studies on their genetics and biology in general are needed. Therefore, in this thesis, I will investigate the genetic composition of lumpfish in Norway, in order to get a better understanding of the existence of potential patterns in its population structure.

The lumpfish biology

Lumpfish belong to the family Cyclopteridae, of which members are characterized by their round body shape, a ventral suction disk and spiny tubercles covering the body (Mecklenburg & Sheiko, 2003). The lumpfish is the only species of the family that is being exploited commercially (Mecklenburg & Sheiko, 2003). The species is recognized by a tall spiny crest and three tubercle rows along each side of the body (Figure 1). The males grow to about 30 cm, whereas the females reach a length of around 50 cm (Albert et al., 2002; Davenport, 1985). The colour varies from green and green-yellow when they are in oceanic waters, to darker grey when entering coastal waters (Davenport & Thorsteinsson, 1989). The males do also for a shorter period attain a red colour, associated with their courtship (Goulet et al., 1986).

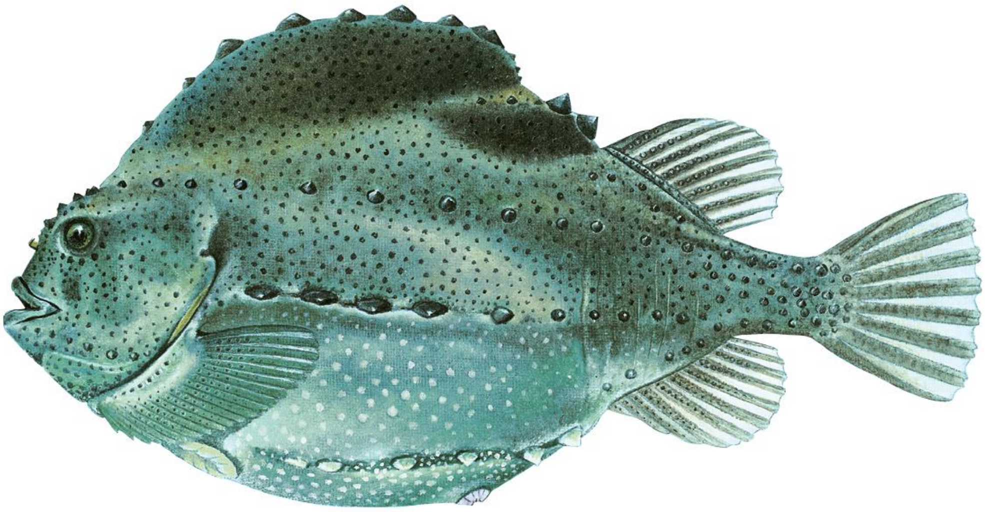


Figure 1: Illustration of adult lumpfish. By Jan Fekjan (https://artsdatabanken.no/Pages/F37435). CC BY-SA 4.0.

The lumpfish is distributed along the coasts of the North Atlantic, from Massachusetts (US) and Portugal in the south, to Svalbard in the north (Davenport, 1985) (Figure 2). Adult fish are pelagic, usually found far out from the shores in the upper 60 m of the water column (Blacker, 1983; Holst, 1993), where they feed on various large planktonic organisms (Davenport, 1985). In the spring, mature fish migrate long distances from their offshore feeding areas into shallower water to spawn (Kennedy et al., 2015; Schopka, 1974; Sigsgaard et al., 2017). The males mature presumably at the age of two to three, whereas the females mature one year older (Albert et al., 2002). The males arrive at the coasts first to establish their territories and build their nests, and the females follow subsequently to lay their eggs in the nests (Goulet et al., 1986). In experimental trials, it has been demonstrated that the females release two egg batches during one season (Kennedy, 2018), and it is suggested that these two batches are dispersed at separate spawning grounds in nature (Kennedy et al., 2015). Each male's territory may moreover contain eggs from different females, as they arrive at different times during the spawning period (Goulet et al., 1986). While the spent females disappear instantly afterwards, the males remain to take care of the eggs until hatching takes place after 1-2 months (Goulet et al., 1986). The juveniles remain close to the shores, among seaweed and other substrates they can attach to with their suction disc (Daborn & Gregory, 1983; Moring & Moring, 1991) until they adopt a pelagic lifestyle around the age of one (Bagge, 1964; Moring, 2001). There is lacking knowledge about the movements of the fish from when they migrate offshore to when they return to coastal waters again to spawn.

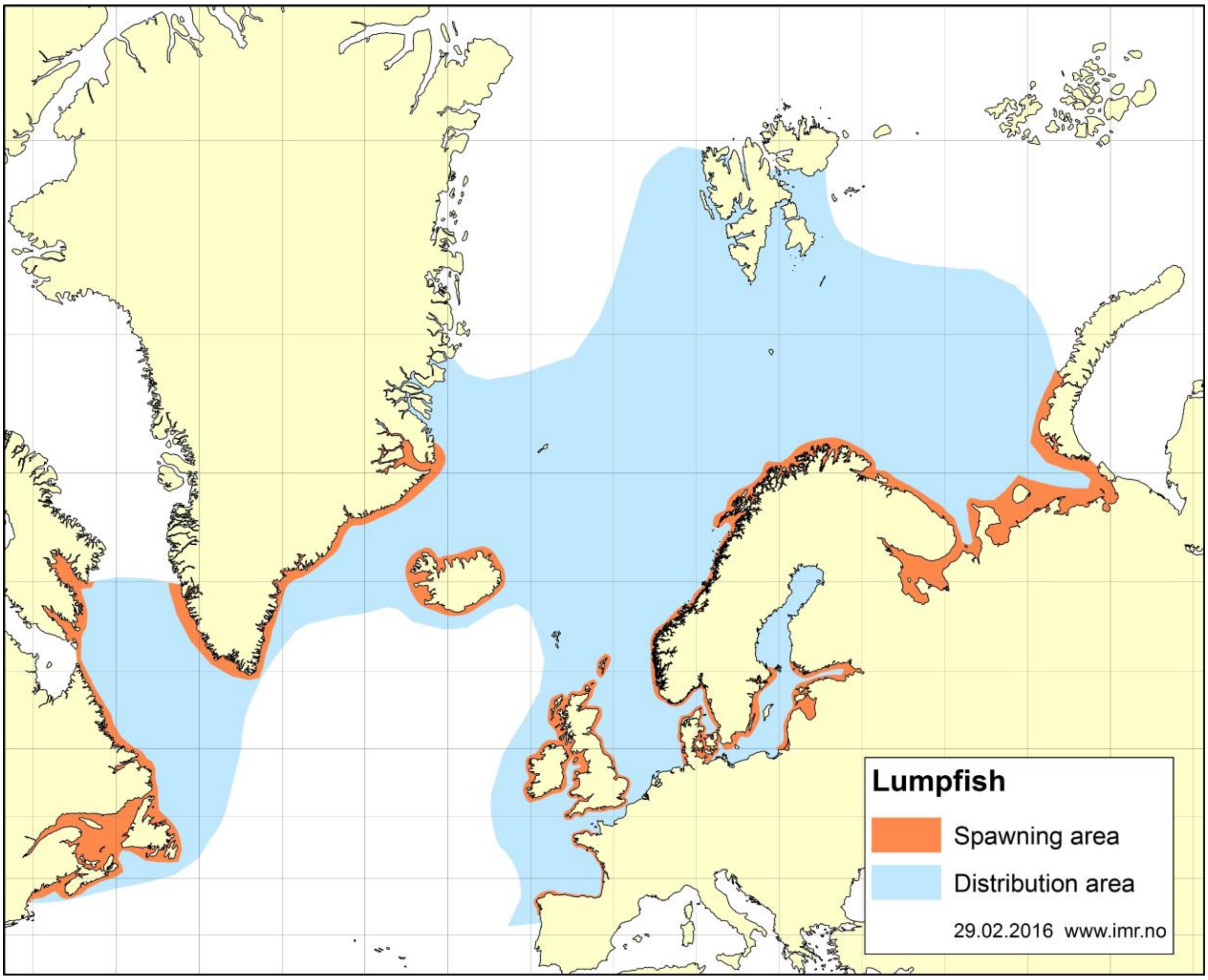


Figure 2: Map of the distribution of lumpfish. By K. E. Gjertsen, P. A. Horneland, E. M. Skulstad / The Institute of Marine Research, 2016. Used with permission.

The number of spawning seasons per individual seems to be one or two. Lumpfish spawning two subsequent seasons have been observed both in nature (Kennedy & Ólafsson, 2019) and in experimental trials in tanks (Imsland et al., 2019). Yet, the mortality appears to be considerably high after the first spawning period (Kasper et al., 2014), and the somatic production peaks at maturation in both sexes, suggesting that lumpfish are mostly semelparous (Hedeholm et al., 2014). It seems likely that the somatic production and extensive migrations prior to spawning is such a great exhaustion that most of the fish endure only one spawning in nature (Bagge, 1967; Hedeholm et al., 2014). For those fish that do spawn more than once however, it is not clear whether they spawn at the same location each time, and if this would be at the same location where they hatched. In several tagging studies, lumpfish were recaptured close to where they were tagged after one year at liberty, suggesting a homing behaviour (Bagge, 1967; Kennedy et al., 2015; Schopka, 1974). The fish showed furthermore extensive migration routes between feeding- and spawning areas (Kennedy et al., 2015; Schopka, 1974). Instead of moving into the nearest spawning grounds, the fish travelled long distances along the coast to grounds farther away. The proportion of tagged fish recaptured after one year was indeed only 0.6-1.4%, although it should be considered that some fish got recaptured in the fisheries already the same year as they were tagged. The low numbers of returns could yet support a theory that the fish mainly spawn once, and that those that manage to restore and mature again home into the same spawning grounds the year after.

Lumpfish is listed as a least-concern species in the Norwegian red list (Norwegian Biodiversity Information Centre, 2015), meaning that it is an abundant and non-threatened species (International Union for Conservation of Nature, 2001). The Norwegian Directorate of Fisheries is setting fishing quotas each year based on stock estimations from the Institute of Marine Research (IMR), and the stock in the Norwegian Sea was estimated to be historically high in 2020, although with some uncertainty (Durif, 2020). In the Barents Sea on the contrary, the stock has been decreasing since 2013 (Durif, 2020). This has not put constrains for an increase of the fishing quotas the last years, which increased from 3 tonnes unprocessed roe per fishing boat in 2013 to 5 tonnes in 2021 (Norwegian Directorate of Fisheries, 2020a). A presumption for the quotas recommended by IMR is that the lumpfish along the Norwegian coast belong to the same genetically distinct group. This could in worst case put a threat to the species' diversity, if smaller genetically distinct populations exist. Knowledge about the genetic structure of lumpfish along the Norwegian coast is therefore highly valuable by putting caution to stocks that could be vulnerable to over-exploitation.

The population genetics of lumpfish

In the first large genetic study on lumpfish that was performed, Pampoulie et al. (2014) suggested that there are three large, genetically distinct populations of lumpfish in the North Atlantic: (1) Maine–Canada–Greenland, (2) Iceland–Norway and (3) Baltic Sea. They reported one genetically homogeneous population along the Norwegian coast and Iceland accordingly, without further differentiation at a smaller geographic scale. However, a more recent study by Whittaker, Consuegra and de Leaniz (2018) showed that lumpfish from Averøy in Norway are genetically distinct from the fish from two other Norwegian locations, which were moreover different from the fish around Iceland. Another study published the same year found in contrast no differentiation between five Norwegian locations, including Averøy (Jónsdóttir et al., 2018). There is accordingly not full agreement between the studies done so far about the genetic structure of lumpfish in Norway, and more research is therefore demanded. As all the beforementioned studies have applied the same genetic marker type in their studies, namely microsatellites, it is further of interest to investigate what results other more powerful methods might reveal. The application of other genetic markers will make it possible to obtain different kinds of genetic data, which could possibly shed more light on population genetic studies of this species to a resolution that have been difficult to obtain in the previous studies.

Single nucleotide polymorphisms (SNPs) as a tool to study genetic structure

Microsatellites, the markers applied in the earlier genetic studies on lumpfish, are genomic regions with short tandemly repeated sequences made up of 1-6 base pairs (Chistiakov et al., 2006). Such sequence motifs have a high mutability and therefore a high degree of polymorphism (Chistiakov et al., 2006). This makes them useful as markers for studying genetic composition and divergence between individuals and populations, and the application has successfully detected genetic structure in numerous of species (Abdul-Muneer, 2014). More recently, the study of single nucleotide polymorphisms (SNPs) has become a common alternative approach in molecular studies. SNPs are variations of single base pairs and are widespread throughout the genome. When variation at a set of SNPs is compared among multiple individuals or populations, they can reveal significant patterns of genetic relationships or insight into patterns of adaptation to the local environmental conditions (Chen et al., 2018). While the study of microsatellites usually only covers a few and non-coding regions of the genome (Abdul-Muneer, 2014), the mapping of SNPs can reveal genome-wide patterns of variation. SNPs can thereby in many cases resolve genetic signatures that are not detected by microsatellites (Gärke et al., 2012; Lemopoulos et al., 2019; Sunde et al., 2020). One of the downsides with the studies of SNPs have been the high costs of sequencing, making it considerably more expensive than microsatellite studies, which do not require sequencing. The recent reduction on sequencing costs (Wetterstrand, 2020) and the development of more cost-effective methods and downstream bioinformatic analyses (Davey et al., 2011) have however made SNP studies a more accessible alternative to many researchers and has allowed to deepen into genomic studies on non-model species (Chen et al., 2018).

Population genomics and the driving forces of population structure

The studies of genetic structure involve in simplest terms the characterization and detection of differences in the genome between individuals or populations. While all individuals exhibit genetic variation that defines them as individuals, genetically distinct populations (or subpopulations) are characterized by genomic patterns that are shared between individuals within the population, but not with individuals from other populations (Hartl & Clark, 2007). In the studies of SNPs, this will imply that the individuals within the population have one or several specific variants of SNPs that are not expected to find in other groups of individuals. When allele frequencies become different between populations, we call it genetic differentiation (Hartl & Clark, 2007). To understand how different evolutionary processes have contributed to this variation, population genomics has emerged as a field (Luikart et al., 2003).

One fundamental law in population genetics in general is the Hardy-Weinberg principle. This says that allele frequencies remain constant from generation to generation in large populations when mating is random and there is no mutation, migration, or selection (Edwards, 2008). In such a situation, the genotype frequencies are said to be in Hardy-Weinberg equilibrium (HWE). However, the prerequisites for HWE do usually not hold in reality. The most basic mechanism behind genetic variation is mutation, which both microsatellites and SNPs arise from. Non-deleterious mutations might further by the effects of non-random mating, migration, and selection settle in groups of individuals and form genetically distinct populations. Driven by genetic drift, which is the genome-wide random distribution of alleles from generation to generation, the allele frequencies of neutral mutations may increase in isolated populations where the mating is limited, until alleles eventually become fixed (Hartl & Clark, 2007). In the same way may geographical distances between individuals limit the extent of random mating, and lead to isolation by distance (IBD; Wright, 1943). Mutations that confer an advantage to the individuals that possess them have a stronger and more rapid effect on population structure, by natural selection (Hartl & Clark, 2007). This occurs typically when individuals with a certain genotype have a better fitness than others in a given environment, and the individuals with alternative genotypes are outcompeted. Natural selection can either lead to reduced variation within a population by directional selection or maintain variation at levels higher than expected under neutrality by balancing selection (Nielsen, 2005). In practice, this means that either homozygotes or heterozygotes for the selected allele are favoured, respectively. The identification of allele frequencies that vary greatly from what can be expected to come from neutral variation alone can consequently contribute to discover genes that are important for local adaptation to specific environmental conditions (Luikart et al., 2003). We call SNPs that exhibit such properties outlier loci.

As non-random mating, mutation, migration, and selection are inevitable in nature, HWE rarely occurs. The reason for its esteem as a fundamental principle for population genetics is that it provides a baseline for where no other evolutionary forces than reproduction alone is acting (Hartl & Clark, 2007). Allele frequencies that deviate from HWE can thereby be used to detect the influence of other evolutionary forces, as the ones that have been mentioned. To measure population differentiation in a useful way, reliable bioinformatic tools that can handle large data sets containing thousands of markers have become invaluable. Yet, despite the rapid development within bioinformatics, the detection of which factors that are contributing to population structure is most often an intricate job (Oleksiak & Rajora, 2020). Population structure is normally a result of both physical elements and biological behaviour (Oleksiak, 2019), and knowledge about these factors are therefore essential to comprehend the whole. Indeed, at the same time as knowledge about the biology of a species helps to understand its genetic structure, do the elucidation of the genetic structure help to understand the biology (Oleksiak & Rajora, 2020). Population genomics has therefore become an important field serving both researchers, conservation governments and industries with much-needed knowledge.

Research objective

The main objective of this study was to resolve patterns of genetic structure and potential signatures of local adaptation in juvenile lumpfish in Norway. In particular, a genome-wide set of SNP markers was studied in order to clarify the existing incongruency among earlier genetic studies on lumpfish using microsatellite markers. Two hypotheses were put forward: 1) A genome-wide set of SNPs will reveal genetic structure among lumpfish along the Norwegian coast not found previously. 2) Patterns of genetic structure among lumpfish populations in Norway are related to geographical distance between locations, possibly due to IBD and/or local adaptation. The outcome of the study may provide useful knowledge for the general understanding of the species' biology and provide important insight both towards the sustainable exploitation by the fisheries and use of lumpfish as cleanerfish in the aquaculture industry.

Material and Methods

Sampling and DNA extraction

Lumpfish juveniles were collected among seaweeds between 5 m deep and the water surface, from three different locations in Norway (Figure 3): Bergen (BER), Tromsø (TRO) and Lyngen (LYN), in summer 2020. At each sampling location, ten juveniles were collected and preserved in 96% ethanol at -18 ˚C for later DNA extraction at UiT. For the extraction, a small tissue sample was cut from the tail or posterior part of the body of each individual. Total genomic DNA extraction was performed in spin columns following the DNeasy® Blood & Tissue Kit protocol from Quiagen (Hilden, Germany). The extracts were analysed by electrophoresis on a 1% agarose gel (UltraPure^{TM} Agarose, Invitrogen, Thermo Fisher Scientific, Carlsbad, CA, USA) to examine contamination and DNA degradation. Concentrations of double-stranded DNA were assessed with the Quant-iT™ PicoGreen™ dsDNA Assay Kit (Invitrogen, Thermo Fisher Scientific, Oregon, US). For the sequencing, a minimum of 0.2 µg DNA from each individual was required. All extracts were kept at -18˚C until sequencing.

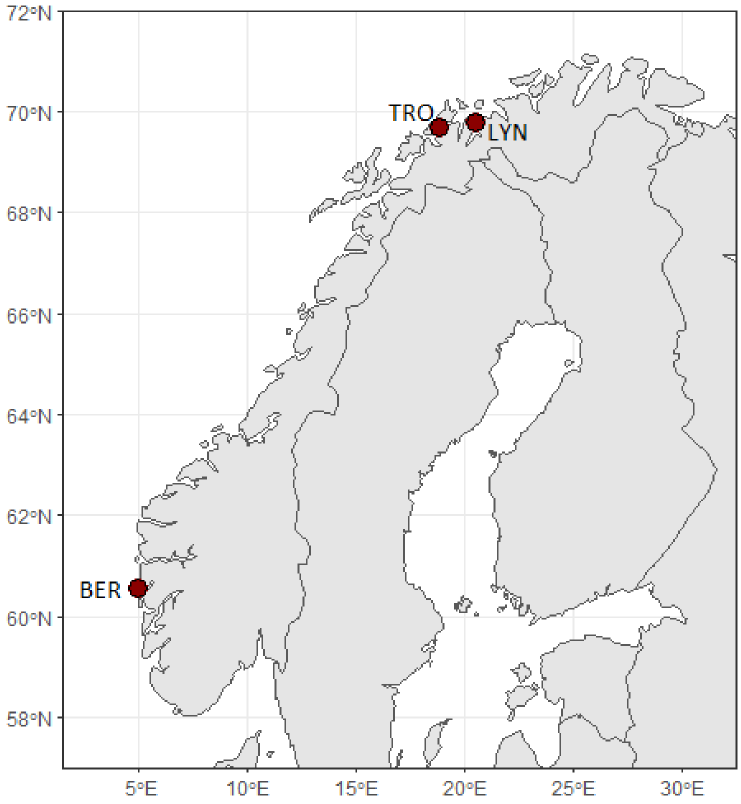


Figure 3: Sampling locations.

Whole genome sequencing (WGS)

DNA purification, next generation sequencing library preparation and sequencing were performed by Novogene Co., Ltd. (Hongkong, China). Briefly, the genomic DNA from each sample was fragmented to the size of approximately 350 base pairs and a DNA library was constructed according to Illumina paired-end protocols. Library quality was assessed by Agilent 2100 Bioanalyzer (Agilent Technologies, CA, US) for size distribution and qPCR was used to quantify DNA fragments with sequencing adapters. Using the Illumina NovoSeq 6000 sequencing platform (Illumina, Inc., CA, US), paired end reads of 150 base pairs were generated, with a minimum coverage of 10× and mean coverage of 30× of 99% of the genome.

SNP detection and filtering

Quality control and sequence alignment of the sequencing data was performed before SNP detection was carried out. The reads were mapped with the *BWA* software (v 0.7.8-r455; Li & Durbin, 2009) to the available lumpfish reference genome (*RefSeq* Accession GCA_009769545.1), using the maximal exact matches algorithm with a minimum seed length of 32 (parameters '*mem -k 32'*). *Picard* (v 1.111; http://broadinstitute.github.io/picard/) and *SAMtools* (v 1.3.1; Li et al., 2009) were used for subsequent processing and removal of PCR duplicates. SNPs were called with *SAMtools*, with indel candidates requiring a minimum of two gapped reads, and the minimum fraction of gapped reads set to 0.002 ('*mpileup -m 2 -F 0.002*'). The detected SNPs were filtered with *SAMtools* to obtain a minimum mapping quality of 20 and a minimum read depth of 4. The *ANNOVAR* software (v 2015Dec14; Wang et al., 2010) was applied to annotate the SNPs. The SNPs were then filtered by *VCFtools* (v 0.1.15; Danecek et al., 2011) for a minimum allele count of six ('*--mac 6*'), no missing genotypes ('*--max-missing 1*'), and no indels ('*--remove-indels*'). Only biallelic SNPs in the data set were kept ('*--min-alleles 2*' and '*--max-alleles 2*'). The SNPs were further filtered for significant deviation from HWE (cut-off *p*-value: 0.001) using a *Perl* script (https://github.com/jpuritz/dDocent/blob/master/scripts/filter_hwe_by_pop.pl). Lastly, the SNPs were filtered for linkage disequilibrium with *bcftools* (v 1.10.2; Li, 2011) with the minimum squared correlation coefficient set to 0.8 and a window size of 100,000 ('*+prune -l 0.8 -w 10000*'). This yielded the final data set for the downstream analyses, henceforth referred to as the 'full data set'.

Calculation of basic population statistics

Basic population genetic statistics were calculated in *R* version 4.0.4 (R Core Team, 2021). Observed heterozygosity (H_O) and expected heterozygosity (H_E) for each population were calculated with the *adegenet* package (v 2.1.3; Jombart, 2008). The inbreeding coefficient (F_{IS}) for each population was calculated with the *hierfstat* package (v 0.5-7; Goudet & Jombart, 2020). The Barlett's test integrated in *R* was further performed to assess whether differences between H_E and H_O were statistically significant. To test for deviations from HWE, the Monte Carlo procedure by Guo and Thompson (1992) implemented in the *pegas* package (v 0.14; Paradis, 2010) was applied with 1000 replicates. The significance level was set to 0.05.

Initial analyses of population structure

To initially assess the extent of population structure, pairwise global F_{ST} coefficients between the sampling locations were calculated. The F_{ST} coefficients measure the proportion of genetic difference that can be explained by variation between populations (Wright, 1965), providing measures of genetic distance between the sampling locations. The estimations were performed using the *StAMPP* package (v 1.6.1; Pembleton et al., 2013) in *R*, which applies the *F*-statistics by Weir and Cockerham (Weir & Cockerham, 1984). Estimations were performed first with the full data set, and subsequently with an outlier data set and neutral data set (please refer to the next section for explanation). As the sampling was performed in a large geographical region, it was further examined whether the observed genetic distances were associated with the geographical distances between the sampling locations. IBD analysis was performed using the function *mantel.randtest* in the *ade4 R* package (v 1.7-16; Dray & Dufour, 2007). The Mantel test assesses whether there is a correlation between two equally dimensioned distance matrices (Mantel, 1967), in this case containing the pairwise F_{ST} estimates for the full data set and geographical distances between sampling locations, respectively. The shortest distances by sea between the approximate coordinates for the sampling locations (Table 1) were measured using *Gule Sider* (https://kart.gulesider.no/). The test was performed with 100,000 repetitions.

Table 1: Approximate coordinates for the sampling locations (obtained from *Gule Sider*).

Location code	Location	Latitude	Longitude
BER	Bergen	N60.54401	E04.92102
TRO	Tromsø	N69.70951	E18.95528
LYN	Lyngen	N69.78061	E20.52057

Outlier loci detection

In order to find SNPs that potentially could be under selection and thereby related to local adaptation, different outlier loci detection methods were tested. A common property of most of the methods is that they involve locus-wise estimation of F_{ST} coefficients. This may reveal SNPs with allele frequencies differing significantly between populations and is therefore used to pinpoint SNPs putatively under selection. To prepare the input files for the various analyses, *PGDSpider* (v 2.1.1.5; Lischer & Excoffier, 2012) and the *vcf2bayescan Perl* script (https://github.com/santiagosnchez/vcf2bayescan) were used. Four different outlier detection tools were applied: 1) *BayeScan* (v 2.0; Foll & Gaggiotti, 2008), a Bayesian program using estimated posterior probabilities of one selection model and one neutral model to assess whether a SNP is under selection. The program was run with the prior odds set to 100,000, meaning that the neutral model was assumed to be 100,000 times more likely to explain the variation than the selection model. The results were interpreted in *R* and SNPs with a false discovery rate (FDR) < 0.05 were considered as outlier loci. 2) *Arlequin* (v 3.5.2.2; Excoffier & Lischer, 2010) with the '*Detect loci under selection*' function was applied using a non-hierarchical finite island model. *Arlequin* performs coalescent simulations to obtain a F_{ST} null distribution which is used to detect outlier F_{ST} values. The analysis was performed with 20,000 simulations and 5 demes. The FDR of the p-values was estimated using the *p.adjust* function in *R*, and loci were considered as outliers if the FDR was < 0.05. 3) The *OutFLANK R* package (v 0.2; Whitlock & Lotterhos, 2015), in which the approach is to identify SNPs under selection by distinguishing all other genetic variation from the selective variation, by inferring a F_{ST} null distribution. This null distribution is achieved by trimming the tails of a fitting chi-squared distribution. The left and right trim fractions were set to 0.05, and a minimum heterozygosity of 0.01 and q-threshold of 0.05 were applied. 4) The *pcadapt R* package (v 4.3.3; Privé et al., 2020). Outlier detection with *pcadapt* is based on principal component analysis (PCA) and does not make use of F_{ST} coefficients to detect outlier loci, but Mahalanobis distances. The PCA decomposes first the allele frequency differences into *K* principal components explaining the variation in the data set. The Mahalanobis distances calculated for each SNPs are subsequently used to identify SNPs that contribute excessively to variation and which are then considered as outlier loci. The first four principal components ($K = 4$) were chosen for the analysis, as these were shown to contribute the most to the variation (Appendix 1). As with *Arlequin*, the *p.adjust* function in *R* was used to estimate the FDR of the p-values, and SNPs with a FDR < 0.05 were considered as outliers.

As the outlier detection methods vary in terms of statistic approaches, they may have different outcomes and accuracy in detecting true outliers (Ahrens et al., 2018). As a measure to eliminate potential false positives, outlier loci detected with a minimum of two of the four approaches tested were here considered as “true outliers”. These loci were combined in an outlier data set, henceforth referred to as the ‘outlier data set’. The remaining loci were considered as neutral markers, constituting the ‘neutral data set’.

Clustering-based population structure analyses

To examine the patterns of population structure in more detail, PCA was performed as a distance-based approach, and the software *ADMIXTURE* (v 1.3.0; Alexander et al., 2009) was used as a model-based approach. PCA was performed with the *pcadapt* package (v 4.3.3; Privé et al., 2020) in *R*, following the same procedures as for the outlier loci detection. The analysis was first performed with the full data set, then further with the outlier data set and neutral data set to assess what effect the potentially selected SNPs had on the population structure. PCA plots obtained with the *adegenet* package (v 2.1.3; Jombart, 2008) and neighbour joining trees obtained with the *ape* package (v 5.4-1; Paradis & Schliep, 2019) in *R* were used to identify the individuals and label the *pcadapt* plots accordingly (Appendix 2).

ADMIXTURE, being a model-based approach, follows population genetic assumptions using a statistical model (Alexander et al., 2009). The software infers the number of genetically distinct clusters, K, of individuals independently of the sampling locations. Based on the multiloci genotypes of all the individuals, it estimates population allele frequencies and ancestry proportions using a maximum likelihood approach. For each individual, the fraction of which each K ancestral population contributes to its genome is estimated. By testing for different values of K, the pattern of structure can be explored at different levels and the value of K that best explains the population structure can be inferred (Pritchard et al., 2000). To convert the data to the *binary ped* format required for the analysis, *PLINK* (v 1.07; Purcell et al., 2007) was utilized. *ADMIXTURE* was then run for K ranging from 1 to 6, following the procedures of Evanno, Regnaut, and Goudet (2005), which tested for K up to three more than the true number of populations. Based on the number of sampling locations, it was assumed that there were three true populations. For each K, 10 runs were performed with different random seeds. *CLUMPAK* (Kopelman et al., 2015) was subsequently used to collate all the runs from *ADMIXTURE* and to infer the best K. For the latter, *CLUMPAK* calculates the delta K (ΔK) as proposed by Evanno et al. (2005). Based on all runs for each K, ΔK is given by the mean of the second order rate of change of the loglikelihood output by *ADMIXTURE*, divided by the standard deviation. With simulated data sets, Evanno et al. (2005) showed that the maximum ΔK is obtained for the true K, and it can thereby be used as an indicator of the best fitting K for the given data set. As input for the assessment, the loglikelihoods calculated for each K from all runs were used (except K = 1, for which ΔK cannot be calculated). All the procedures were first performed for the full data set, then for the outlier data set and neutral data set separately to investigate how the patterns in genetic structure may be affected by the presence of outlier loci, as with the PCA.

Results

Sequencing and genotyping

A total of 1,754,515,036 raw reads were obtained from the WGS, corresponding to a mean of 47,419,325 (SD ± 18,883,544) raw reads per individual. The proportion of reads mapped successfully to the genome of each individual ranged from 95.28% to 98.13%. The average mapping depth per individual was 11.59× (SD ± 0.77), and a minimum depth of 4× was achieved for 93.95 % of the genome on average (SD ± 1.07). SNP calling and filtering for minimum mapping quality of 20 and a minimum read depth of 4 yielded a SNP data set consisting of 2,508,556 loci. After all filtering steps, the final 'full data set' contained a total of 607,663 loci.

Basic population parameters

H_O was significantly greater than H_E for all locations ($p < 0.01$), ranging from 0.35 to 0.37 (Table 2). All locations had negative and almost equal F_{IS} values in the range -0.036 to -0.035. For each location, 1.41% to 1.57% of the SNPs deviated significantly from HWE (p <0.05).

Table 2: Average observed heterozygosity (H_O), expected heterozygosity (H_E), and inbreeding coefficient (F_{IS}) across loci for each sampling location. The percentage of loci deviating from Hardy-Weinberg Equilibrium (HWE) at a 0.05 significance level is given for each location.

	H_O	H_E	F_{IS}	**HWE deviation** ($p < 0.05$)
BER	0.348 *	0.320	- 0.035	1.57 %
TRO	0.366 *	0.336	- 0.036	1.45 %
LYN	0.365 *	0.335	- 0.036	1.41 %

* H_O significantly different from H_E ($p < 0.01$)

Initial indications of population structure

For the full data set, the estimated pairwise F_{ST} indicated significant differentiation between all pairwise locations ($p < 0.01$). The highest genetic divergence was observed between BER and the two northern locations, with $F_{ST} > 0.04$ (Table 3). A F_{ST} of 0.002 between TRO and LYN suggests minor genetic differences between the two northern locations. The F_{ST} estimates were considerably higher for the outlier data set. Between BER and the northern locations, both F_{ST} estimates were > 0.6 ($p < 0.05$). Between TRO and LYN, the F_{ST} was 0.06 ($p < 0.05$). The F_{ST} coefficients estimated for the neutral data set were similar to those for the full data set.

Table 3: Pairwise F_{ST} estimations. Values above the diagonal are for the full data set. The analysis of the neutral data set yielded similar results (not shown). Values below the diagonal are for the outlier data set. All values were significant with $p < 0.05$.

	BER	TRO	LYN
BER	-	0.042	0.044
TRO	0.632	-	0.002
LYN	0.640	0.063	-

When considering the F_{ST} estimates in relation to the geographical distances between the sampling locations, the F_{ST} was greater between the locations separated by longer geographical distances (BER vs. TRO, and BER vs. LYN). The Mantel test showed however that there was a non-significant correlation ($r = 0.999$, $p = 0.17$) between the pairwise F_{ST} estimates and the geographical distances, indicating that IBD is not contributing to population structuring.

Detected outlier loci

BayeScan detected 6 outlier loci with FDR < 0.05 when prior odds of 100,000 were chosen. With *Arlequin*, 1,543 outlier loci with FDR < 0.05 were identified, whereas *pcadapt* detected a total of 16,157 outliers with the same prerequisites. *OutFLANK* on the other side did not detect any outlier loci. Thus, 17,706 SNPs in total were suggested to be under selection by the different approaches. Of these, 40 SNPs were appointed as outlier loci by both *Arlequin* and *pcadapt*, and one SNP was detected by both *pcadapt* and *BayeScan*. No common outlier loci were found between *Arlequin* and *BayeScan*. The outlier data set consisting of outlier loci detected by a minimum of two approaches consisted thereby of 41 SNPs. All were putatively under directional selection, with F_{ST} values calculated in *Arlequin* ranging from 0.418 to 0.828. The remaining 607,622 SNPs were considered as neutral markers.

Population structure inferred from the cluster-based analyses

In the PCA performed with the full data set, the first principal component (PC1) separated BER from the two northern populations, TRO and LYN (Figure 4). PC1 explained 26.88% of the total variation in the data set. The second principal component (PC2) suggested further division within BER and LYN, with particularly two individuals from each location exhibiting high variation compared to the other individuals. From BER, these were BER01 and BER17. BER20 and BER25 were also partly separated from the main clustering of BER individuals by PC2. In the same way were LYN13 and LYN18 separated from the rest of the LYN and TRO individuals. A total of 20.88% of the variation was explained by PC2.

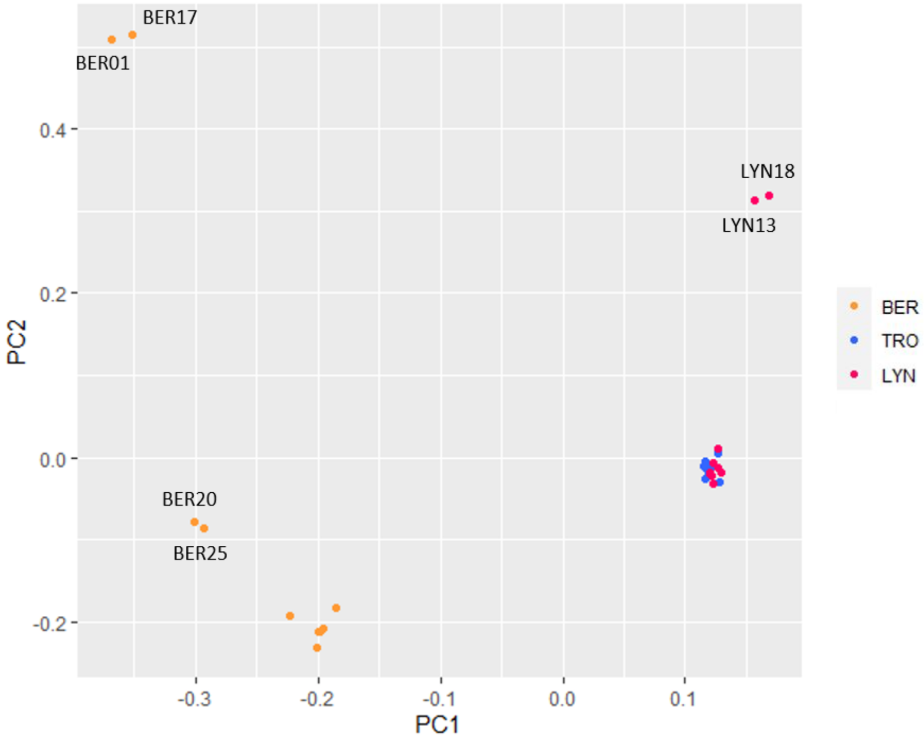


Figure 4: Principal component analysis (PCA) of the full data set performed with the *pcadapt* package in *R*. The plot displays the variation between samples explained by the first and second principal components (PC1 and PC2, respectively). Each sample are represented by a dot. The samples that differ the most from the rest are labelled.

The admixture analysis in *ADMIXTURE* revealed patterns that can be comparable to the PCA, with a clear division between BER and TRO/LYN for all values of K (Appendix 3). The optimal K by Evanno at al. (2005) was 3 (Figure 5 & Appendix 4). For K = 3, two LYN individuals made up their own cluster, which were the same two separated from the TRO/LYN cluster in the PCA (LYN13 and LYN18).

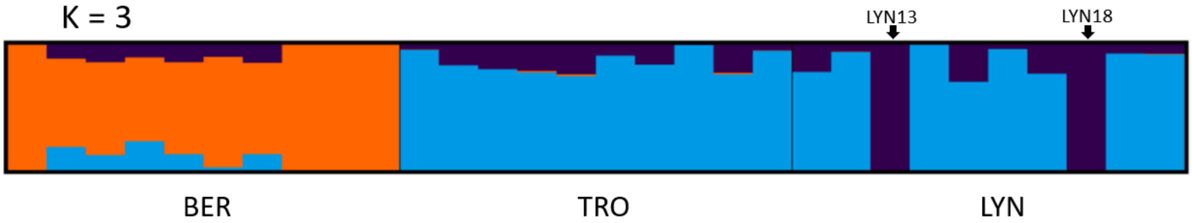


Figure 5: Bar plot obtained with *CLUMPAK* showing inferred population structures for K = 3 from *ADMIXTURE*, using the full data set. Each sample are represented by one bar (please refer to Appendix 3 for a list of sample names) and the colours represent different ancestries. Appointed samples are those that differ considerably from their respective location group.

When only the outlier loci were used as input for PCA, there was still a clear separation between BER and TRO/LYN by PC1 (Figure 6a). This component explained 98.77% of the total variation in the outlier data set. PC2 created a gradient along which the TRO and LYN individuals differentiated, with the TRO individuals mainly in the one end corresponding to the most variation, and the LYN individuals more towards the end with the least variation. The BER individuals were also spread out by PC2. 25.67% of the variance was explained by PC2. PCA performed with the neutral markers showed a similar pattern as the full data set (Figure 6b). There was a very slight decrease in the amount of variance explained by PC1 (26.86 %) when compared to the full data set, whereas PC2 explained a similar amount of difference (20.88 %).

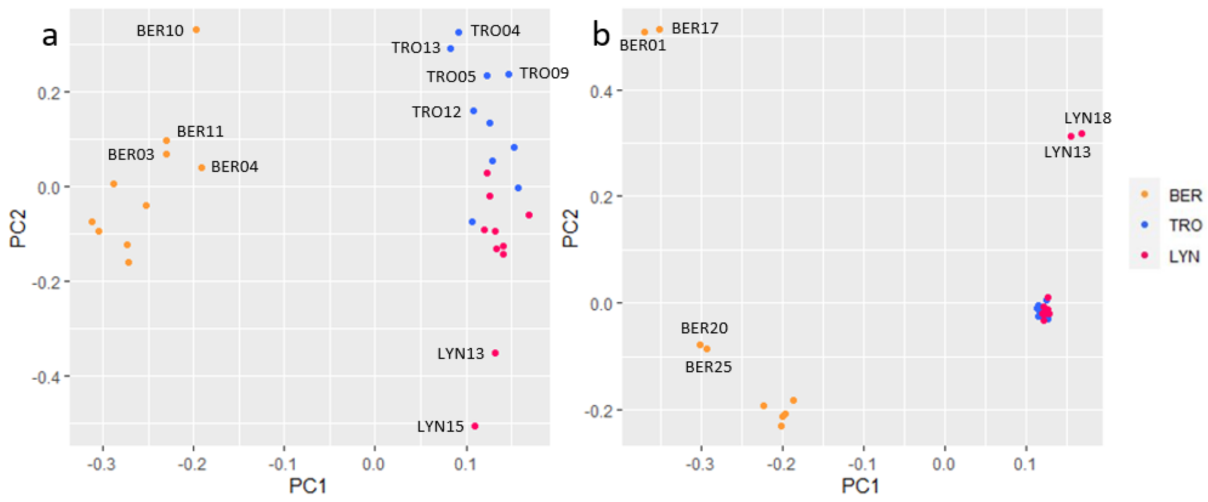


Figure 6: Principal component analysis (PCA) with a) the outlier data set and b) neutral data set, performed with the *pcadapt R* package. The variation between samples explained by the first and second principal components (PC1 and PC2, respectively) is plotted. The dots represent each sample and those that differ the most from the rest are labelled.

The outlier data set showed some differences in genetic structure compared to the full data set in the admixture analysis (Figure 7a). The optimal K by Evanno et al. (2005) was yet 3 (Appendix 4), for which there was still a clear separation between BER and the two northern locations. What was different from the full data set, was that four TRO individuals (TRO04, TRO05, TRO09, TRO12) made up their own cluster. These corresponded to the individuals observed in the upper end of the PC2 gradient in the PCA (Figure 6a). The remaining six TRO individuals and LYN made up the third cluster, but some individuals shared smaller proportions of ancestry with the four TRO individuals. The two LYN individuals that were discriminated from the TRO/LYN cluster by the full data set (Figure 4 & 5), did not stand out in the outlier loci analyses.

With the neutral data set, the structure appeared more like the full data set (Figure 7b). However, the optimal K by Evanno et al. (2005) was 5 for this data set (Appendix 4). When K = 5, the two BER individuals most separated from the BER cluster in the PCA (BER01 & BER17; Figure 6b) made up their own cluster, in addition to the two LYN individuals (LYN13 &

LYN18). The TRO and LYN individuals had further some proportions of ancestry that did not make up an own cluster.

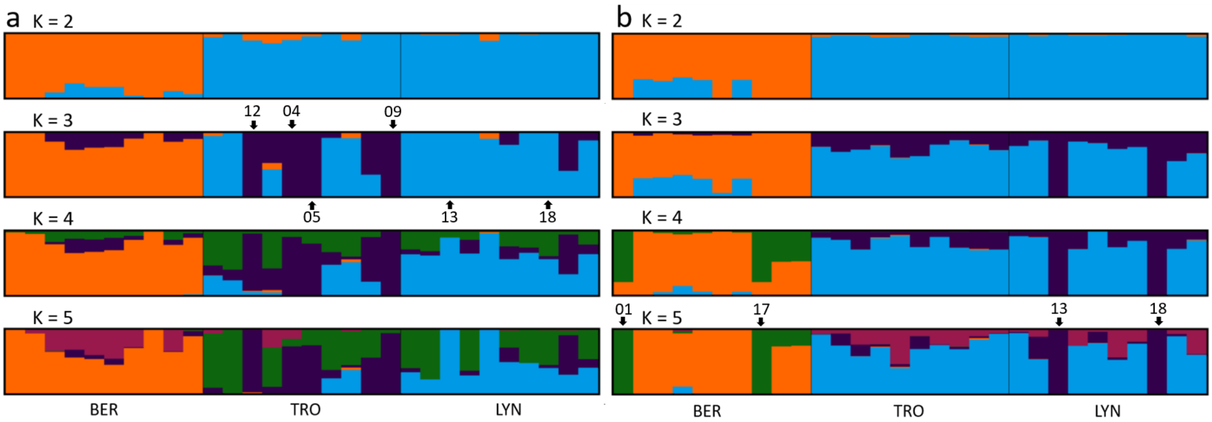


Figure 7: Population structures inferred with *ADMIXTURE* for a) the outlier data set and b) the neutral data set. The best K (as proposed by Evanno et al., 2005) was 3 for the outlier data set and 5 for the neutral data set. Each bar represents one sample (please refer to Appendix 3 for a list of sample names) and each colour represents different ancestries. The samples that differ the most from their respective location groups are pointed out with their numbers. Bar plots were obtained with *CLUMPAK*.

Overall, the different set of loci showed the same patterns, in that all results showed that there were large genetic differences between lumpfish from the south and the north of Norway. Some of the analyses revealed further differentiation within LYN, and the study of SNPs putatively under selection indicated in addition a differentiation between the two northern sampling locations.

Discussion

The aim of this study was to get a better understanding of the genetic structuring of lumpfish in Norway, and to investigate whether factors like IBD and/or local adaptation contributed to this structuring. WGS was applied to obtain a comprehensive set of SNPs, which was examined by various population structure analyses and subjected to outlier detection. The results did unambiguously show differentiation between lumpfish from one sampling location in southern Norway (BER), and two sampling locations in northern Norway (TRO and LYN). The results also showed genetic structuring at smaller geographical scales in all sample locations. It was not found a significant association between the genetic distances and geographical distances between sampling locations, and IBD could therefore not be confirmed as a driving factor of the population structuring. Studies of SNPs putatively under selection revealed a somewhat different population structure compared to the one detected by the full data set and the neutral data set. These findings open up for several theories about how the genetic structure of lumpfish in Norway might have been developed.

The large-scale population structure of Norwegian lumpfish

The genome-wide set of 607,633 SNPs implied that all locations had higher heterozygosities than expected and allele frequencies that deviated from HWE. The slightly negative F_{IS} values (-0.036 to -0.035) indicate that there is no relatedness between the individuals at any of the sampling locations. Taken together, this could indicate that factors like inbreeding or genetic drift are not the main drivers to the observed population structure, and that mating occur quite randomly within rather large groups of individuals. For a marine fish with migratory behaviour and few obvious barriers to gene flow this is not uncommon, and it has earlier been pointed out that the lumpfish has a great potential for active dispersal (Garcia-Mayoral et al., 2016). However, based on the observed population structure patterns, it seems possible that several populations are sampled at each location, and this would contradict these theories. The presence of several populations at the sampling locations would yet be able to explain the obtained population statistics. The observed subpopulation structures will be discussed more in detail in next section, after we first have looked at the population structure at the larger scale.

A clear differentiation between lumpfish from the south and the north of Norway was confirmed by several approaches. The estimated pairwise F_{ST} coefficients indicated a significant difference between BER and each of the two northern sampling locations, TRO (F_{ST} = 0.042,

$p < 0.05$) and LYN ($F_{ST} = 0.044$, $p < 0.05$). PCA and admixture analyses showed congruent patterns of differentiation (Figure 4-7). These findings are not in accordance with the study by Jónsdóttir et al. (2018), which did not detect any genetic structure in adult fish from five fishing grounds spread along the Norwegian coast from the north to the south. Our findings support however the previous findings by Whittaker et al. (2018), which were the first to suggest genetic differentiation in adult lumpfish within Norway. Whittaker et al. (2018) found fish from Averøy to be genetically different from fish from Rogaland and Namsen, which are south and north of Averøy, respectively. Potentially, it might be possible that the approach we have applied here could detect genetic differentiation between Rogaland and Namsen as well. Whittaker et al. (2018) applied 10 microsatellites which cover a maximum of 2,219 base pairs of the genome (Skirnisdottir et al., 2013). With the 607k SNPs applied here, more of the variation that the genome holds have been covered, and it seems reasonable to assume that this has allowed for detection of patterns of genetic structure that could not be detected by microsatellites. This has for instance also been demonstrated for the Atlantic herring, *Clupea harengus* (Lamichhaney et al., 2012). A more exhaustive sampling design should help to refine these geographical patterns of genetic population structure of lumpfish in Norway.

The pattern of genetic differentiation observed in the present study is expected in terms of the environmental heterogeneity along the Norwegian coastline and reported swimming distances for lumpfish (Kennedy et al., 2015). Lumpfish are poor swimmers (Hvas et al., 2018) and considering that females lay their eggs into nests guarded by males, dispersal during early life stages appears limited. Hence, unless ocean drifts greatly affect the individuals in the later life stages, high rates of gene flow between the southern and northern lumpfish populations seems unlikely. It was therefore hypothesised that genetic differentiation would increase with geographical distances between the sampling locations. The estimated pairwise F_{ST} coefficients support this theory, but there was not found a significant relationship between the geographical distances and the pairwise F_{ST} values (Mantel test, $p = 0.16$), rejecting the hypothesis of patterns of IBD. This does not implicitly mean that IBD does not contribute to the observed population structure in lumpfish. The lack of significance could likely be due to the low number of sampling locations (Jenkins et al., 2010) or an uneven geographical distribution of sampling sites. In a study of the genetic structure of lumpfish from six locations along the West Greenland coast, it was indicated significant IBD using microsatellite markers (Garcia-Mayoral et al., 2016). In the study by Whittaker et al. (2018), which sampled at 15 sites in the North Atlantic, it was also detected a weak, but significant IBD. There is accordingly reason to believe that a more exhaustive sampling design along the Norwegian coast may reveal a pattern of IBD.

Genetic structuring at smaller geographical scales

The applied methods allowed also for detection of genetic differentiation within some of the sampling locations. The individuals from TRO and LYN appeared to be quite genetically similar overall, but two individuals from LYN stood out from the main group. There is the possibility that these individuals have been driven by coastal currents from another place at the larval stage. Another reason could be that the sampling has taken place at the border between two populations, and that these individuals represent one of the populations. What seems even more plausible, is that the location has been under-sampled and that these two individuals represent a diversity that is present to a larger extent than what has been shown here. In general, this applies to all the three sampling locations. The negative F_{IS} values for all locations may indicate that the spawning stocks are relatively large in these areas, and ten individuals are then too few to represent the whole stocks. This, again, suggests that more studies with more samples should be performed in order to cover more variation and better understand the population structures seen at the smaller geographical scale.

From BER, four individuals in groups of two distinguished from the other six individuals. One factor that hypothetically could explain this variation is the different sea currents affecting the Norwegian coast. Diverging from the Atlantic current, the Norwegian current is a strong directional current moving towards the Polar areas, that potentially may affect pelagic fish in the north more than the fish in the south. As BER is somewhat in between where the Norwegian Current move towards the north and where some of it bends of towards the North Sea, it might be slightly random in terms of where the lumpfish end of spawning in this area. The northern part of Norway is more strongly affected by the Norwegian current, and it could therefore be less incidental where the fish move in to spawn. Considerable influence could potentially also be attributed the Norwegian Coastal Current, which can reach great velocities particularly outside the southern coast of Norway (Sætre & Ljøen, 1972). Spreading of eggs and larvae of fish species like Atlantic cod (*Gadus morhua*) and capelin (*Mallotus villosus*) are well-known to occur (André et al., 2016; Gjøsæter, 1998), but it is generally more uncertain how the currents may affect the migration patterns of adult fish (Lennox et al., 2019). To be able to assess this more comprehensively for the lumpfish, more samples and sampling locations are needed.

More tracking experiments could also undoubtedly help to better understand the migratory behaviour of the lumpfish.

Potential signs of local adaptation

Due to the long coastline, lumpfish in Norway are exposed to great variability in environmental conditions. Particularly the water temperature varies greatly between the north and the south (IMR, 2019), and this is a factor that potentially may affect the survival of lumpfish particularly at the early, crucial life stages. In one study, it was shown that juvenile lumpfish from different parts of Norway in the size range 154–426 g do not have different temperature preferences when reared in aquaculture facilities (Mortensen et al., 2020). We do however not know if the same applies to the fish in nature and how the temperature may affect the individuals before they reach this size. Latitudinal gradients in temperature should therefore not be ruled out as a potential driver of selection, contributing to locally adapted populations. The fish are furthermore exposed to different salinities, which increase with increasing latitude along the coastline (Sætre & Ljøen, 1972) and which is another environmental variable commonly associated with local adaptation in marine fish (Kijewska et al., 2016; Lamichhaney et al., 2012). Therefore, it was of interest to discover SNPs that could be subjected to selection in lumpfish. The four methods that were applied detected outlier loci in a range from zero to several thousand SNPs. With the great variation in statistical approaches constituting the different programs this was to be expected. The ability to detect true outlier loci is amongst others affected by the demographic history of the species, and particularly IBD might lead to high false positive rates (Lotterhos & Whitlock, 2014). *BayeScan* and *OutFLANK* has been reported to have the greatest power in detecting outlier loci under an IBD model (Whitlock & Lotterhos, 2015), and since no significant IBD was shown in the present study, this could explain why these programs detected the lowest numbers of outlier loci here (six and zero respectively). In the same study it was also stated that *OutFLANK* performs best when the numbers of samples and locations are not too low, which could be another possible explanation to why *OutFLANK* did not detect any outlier loci herein. For *BayeScan*, the low numbers of detected outliers might also be due to the presence of admixed individuals, which has shown to reduce its detection power (Luu et al., 2017). Another possibility could be that the prior odds were set too strict. The outlier detection procedure in *Arlequin* is based on a non-hierarchical finite island model, and more than one thousand loci potentially under selection were detected with this method. This method has, however, been shown to have very high false positive rates if the species possess a hierarchical population structure (Excoffier et al., 2009). As a hierarchical model seems more likely than an island model for lumpfish, there is thus a chance that some of the discovered outlier loci are false positives. The fourth approach, *pcadapt*, has the advantage of not being dependent on any prior group assignment to detect outlier loci and is supposed to have the greatest detection power for hierarchically structured populations (Luu et al., 2017). *pcadapt* showed undoubtedly the highest power in this study, with more than 16k outlier loci detected. Indeed, the relatively high significance level that was applied allows yet for a high number of false discoveries (5%), considering the large number of SNPs (607k) that were subjected to outlier detection.

Because it is impossible to validate which detected outlier loci that are true discoveries only using bioinformatic tools, only loci that were detected with a minimum of two of the methods were selected for the further analyses, to gain some confidence. A total of 41 SNPs were then considered as "true outlier loci". If there was only selection for these SNPs and no other factors creating the observed population structure, we would have expected the remaining putatively neutral markers to show a lack of population structure. This was not the case, instead the neutral data set showed the same structure as the full data set. However, the most likely number of clusters estimated by the method of Evanno et al. (2005) was five with the neutral data set, in difference to three for the full data set. This is possibly because the SNPs with the highest variation between locations are eliminated, so that SNPs that contribute less to variation have more effect on the genetic structure. For K = 5, some individuals from TRO and LYN had small proportions of ancestry that was not fully explained by the structure. This may again be due to under-sampling and that the source population therefore is not present in these samples. Or it could be that this is variation that among hundreds of individuals would be negligible but become prominent here due to the low sample number. It is important to keep in mind that the value of K regarded as the optimal one in the admixture analyses unlikely represent the true number of populations present, but may be considered as the number of different ancestries that best explains the most protruding variation in the given data (Lawson et al., 2018). Irrespective of this, what appears evident from the observed patterns of the outlier data set and neutral data set, is that the 41 chosen SNPs are not responsible for the population structuring alone. This does however not implicitly mean that lumpfish do not experience selection. From a total of 17,706 outlier loci detected by the different applied methods, the 41 SNPs were chosen as "true outliers" only on the basis of being detected by more than one method. It is fully possible that some of the excluded outlier loci, which then were regarded as neutral, are actually SNPs under selection. It is also possible that the neutral markers reflect patterns of local adaptation, despite not being under selection, as has been suggested for Atlantic salmon, *Salmo salar* (Moore et al., 2014). This may occur when selection lead to reproductive isolation and thereby limits the gene flow, which further may lead to ecological speciation (Thibert-Plante & Hendry, 2010). Such divergent selection seems indeed not very likely for lumpfish if we consider the 41 outlier loci as true subjects of selection. If that was the case, the neutral data set should have possessed a pattern more similar to the outlier data set. What was observed, was that the outlier data set exhibited more variation between individuals than the neutral data set (Figures 6 & 7). If the high differentiation of neutral loci were to be associated with the selection of the given outlier loci, the two data set should have showed more congruent patterns. More specifically, the outlier data set should not have exhibited more variation than the neutral data set, as the outlier loci would be the actual subjects of selection. What we at least can conclude from this, is that a larger number of SNPs than those that constitute the outlier data set are needed to provide the genetic resolution and power to detect this. If a rather large set of SNPs spread along the genome is needed to explain the large-scale genetic structure, it will further be reasonable to assume that the population structure of lumpfish has developed over a relatively long evolutionary time span and not due to selection of a limited number of loci.

Considering the outlier data set in isolation, there was still a strong differentiation between the southern sampling location and the two northern sampling locations ($F_{\mathrm{ST}} > 0.6$, $p < 0.05$). Beside this, the analyses indicated that there were significant differences between TRO and LYN ($F_{\mathrm{ST}} = 0.063$, $p < 0.05$), and some individuals from TRO showed be particularly differentiated from the rest (Figures 6 & 7). This suggests that selection might be contributing to population structure at local scales in lumpfish. To investigate this further, population structure analyses and outlier loci detection could have been performed between the two latter locations only. As the locations are geographically close, it is possible that the effects of classical neutral divergence scenarios (e.g. IBD) are relatively small and that selection therefore is easier to elucidate (Lotterhos & Whitlock, 2015). It should indeed be considered that other factors like for instance homing and reproductive timing also are alternative drivers of the observed divergence. A next step could yet be to investigate more specifically which SNPs contribute to the differentiation and if these somehow might relate to local adaptation.

Future perspectives

The presented results evidenced strong patterns of genetic population structure among Norwegian lumpfish populations, yet it emphasizes the need for further studies with better sampling coverage. As already mentioned, one of the limiting factors in this study has been the relatively small number of samples per sampling location and few sampling locations. The number of individuals sampled here (ten per location) are most likely not representative for all fish at the sampling locations, which for instance was demonstrated by the two genetically distinct individuals from LYN. Alternative methods, like for instance RADseq (restriction site-associated DNA sequencing; Etter et al., 2011), could have allowed for studying more individuals in this study. However, by using WGS, the full information about the genome is obtained, which in general provides more insight into how evolutionary processes may have contributed to genetic differentiation. Information that might become relevant in the future genetic studies on lumpfish is in that way already provided here. The WGS approach has indeed allowed for a non-random, consistent detection of population structure, despite the few numbers of individuals. More individuals will, however, better represent the variation that exists within the populations and thereby more precisely render the existing population structure (Fumagalli, 2013).

A next step could also, as mentioned, be to look at what genes are associated with the detected outlier loci and potential phenotypic effects of the variation. Approaches to detect outlier loci in relation to specific environmental variables, so-called environmental association analyses (Rellstab et al., 2015), are also available and should be considered. Such methods have shown to have more power than the common outlier methods, although at the cost of higher false positive rates (De Mita et al., 2013). This is however also just one way of detecting loci putatively under selection. In the end, no theory about local adaptation can be confirmed before the functional effects of the variation is studied in experimental trials (R. Barrett & Hoekstra, 2011).

One important thing to keep in mind when interpreting the presented results is that only juvenile lumpfish have been the object of this study. It is assumed that these are the result of spawning and hatching in the sampling areas and thereby provide an indication of the respective gene pools. Juveniles may also provide an indication of which individuals that survive the early life stages associated with high mortalities in the specific areas, and thereby potentially undergo selection. However, there is great uncertainty about how for instance coastal currents affect the juveniles, and it cannot be known for certain that they have not drifted by the currents from other spawning grounds. To understand the temporal stability of the population structure, it will be necessary to elucidate the genetic structure of lumpfish using adult individuals caught at the spawning grounds as well. The results presented here should therefore be interpreted with caution and not as a final answer.

The establishment of a baseline population genetic structure for Norwegian lumpfish would allow for evaluating the effects of exploitation by the fisheries and use of the species as a cleanerfish. The detection of genetic structure in the present study suggests that it might become necessary to regulate these activities differently in the future. For fisheries management, it appears that the lumpfish possibly should be treated as several management units demanding specific measures for sustainable exploitation. Regarding the use as a cleanerfish, it seems now possible that escaping might result in inference with local, genetically distinct populations if the broodstock is caught far from the farming locations. However, before these issues can be addressed more comprehensively, a solid genetic baseline is needed, which this study is not enough do provide alone. Hopefully, the presented findings could encourage both industries to evaluate anew the consequences of using lumpfish and to put more resources into the further elucidation of the population structure of lumpfish.

Conclusions

In this study, it was demonstrated that juvenile lumpfish from locations along the Norwegian coast exhibit genetic structure. The application of genome-wide SNPs allowed to reveal patterns of population structuring that has not been detected in the previous genetic studies on lumpfish, and it appears thus to be a better tool than microsatellite markers for this aim. A clear genetic differentiation was observed between the south and north of Norway, and it was further signs of differentiation at smaller geographical scales, which confirms the first hypothesis of the study. As also hypothesised, the genetic differentiation was greatest between sampling locations separated by longer distances, but IBD was not shown to be an explaining factor. The small-scale genetic differentiation could putatively be related to local adaptation. However, to better comprehend the extent of population structuring and the driving forces, there is a need for studies including more samples and sampling locations, and adult lumpfish collected at the spawning grounds.

References

Abdul-Muneer, P. M. (2014). Application of Microsatellite Markers in Conservation Genetics and Fisheries Management: Recent Advances in Population Structure Analysis and Conservation Strategies. *Genetics Research International, 2014*.

Ahrens, C. W., Rymer, P. D., Stow, A., Bragg, J., Dillon, S., Umbers, K. D. L., & Dudaniec, R. Y. (2018). The search for loci under selection: trends, biases and progress. *Molecular Ecology, 27*(6), 1342-1356.

Albert, O. T., Torstensen, E., Bertelsen, B., Jonsson, S. T., Pettersen, I. H., & Holst, J. C. (2002). Age-reading of lumpsucker (*Cyclopterus lumpus*) otoliths: dissection, interpretation and comparison with length frequencies. *Fisheries Research, 55*(1), 239-252.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research, 19*(9), 1655-1664.

André, C., Svedäng, H., Knutsen, H., Dahle, G., Jonsson, P., Ring, A.-K., . . . Jorde, P. E. (2016). Population structure in Atlantic cod in the eastern North Sea-Skagerrak-Kattegat: early life stage dispersal and adult migration. *BMC Research Notes, 9*(1), 63.

Bagge, O. (1964) Some observations on the biology of the lumpsucker (*Cyclopterus lumpus*). *ICES CM. 1964. Baltic Belt Seas Committee, No. 150:* 7 p. (mimeo)

Bagge, O. (1967). Some preliminary results from tagging of the lumpsucker (*Cyclopterus lumpus*) 1966. *ICES CM F. 1967/F:23. Demersal (N) Comittee*, 3 p. (mimeo)

Barrett, L. T., Overton, K., Stien, L. H., Oppedal, F., & Dempster, T. (2020). Effect of cleaner fish on sea lice in Norwegian salmon aquaculture: a national scale data analysis. *International Journal for Parasitology, 50*(10), 787-796.

Barrett, R. D. H., & Hoekstra, H. E. (2011). Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics, 12*(11), 767-780.

Blacker, R. W. (1983). Pelagic records of the lumpsucker, *Cyclopterus lumpus* L. *Journal of Fish Biology, 23*(4), 405-417.

Brooker, A. J., Papadopoulou, A., Gutierrez, C., Rey, S., Davie, A., & Migaud, H. (2018). Sustainable production and use of cleaner fish for the biological control of sea lice: recent advances and current challenges. *Veterinary Record, 183*(12), 383.

Chen, Z., Farrell, A. P., Matala, A., Hoffman, N., & Narum, S. R. (2018). Physiological and genomic signatures of evolutionary thermal adaptation in redband trout from extreme climates. *Evolutionary Applications, 11*(9), 1686-1699.

Chistiakov, D. A., Hellemans, B., & Volckaert, F. A. M. (2006). Microsatellites and their genomic distribution, evolution, function and applications: A review with special reference to fish genetics. *Aquaculture, 255*(1), 1-29.

Costello, M. J. (1993). Review of methods to control sea lice (Caligidae: Crustacea) infestations on salmon (*Salmo salar*) farms. *Pathogens of wild and farmed fish: sea lice, 219*, 252.

Costello, M. J. (2009). The global economic cost of sea lice to the salmonid farming industry. *Journal of Fish Diseases, 32*(1), 115-118.

Daborn, G., & Gregory, R. S. (1983). Occurrence, distribution, and feeding habits of juvenile lumpfish, *Cyclopterus lumpus* L. in the Bay of Fundy. *Canadian Journal of Zoology, 61*, 797-801.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158.

Davenport, J. (1985) *Synopsis of biological data on the lumpsucker,* Cyclopterus lumpus *(Linnaeus, 1758)* (FAO Fishery Synopsis No. 147). Retrieved from http://www.fao.org/3/ap950e/ap950e.pdf

Davenport, J., & Thorsteinsson, V. (1989). Observations on the colours of lumpsuckers, *Cyclopterus lumpus* L. *Journal of Fish Biology, 35*(6), 829-838.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics, 12*(7), 499-510.

De Mita, S., Thuillet, A.-C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., & Vigouroux, Y. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology, 22*(5), 1383-1399.

Dray, S., & Dufour, A.-B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software, 22*(4), 1-20.

Durif, C. M. F. (2020). *Regulering av fisket etter rognkjeks*. Institute of Marine Research. Retrieved from https://www.hi.no/resources/rad-rognkjeks.pdf

Edwards, A. W. F. (2008). G. H. Hardy (1908) and Hardy–Weinberg Equilibrium. *Genetics, 179*(3), 1143-1150.

Eriksen, E., Durif, C. M. F., & Prozorkevich, D. (2014). Lumpfish (*Cyclopterus lumpus*) in the Barents Sea: development of biomass and abundance indices, and spatial distribution. *ICES Journal of Marine Science, 71*(9), 2398-2402.

Erkinharju, T., Dalmo, R. A., Hansen, M., & Seternes, T. (2021). Cleaner fish in aquaculture: review on diseases and vaccination. *Reviews in Aquaculture, 13*(1), 189-237.

Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. A., & Cresko, W. A. (2012). SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In *Molecular methods for evolutionary genetics* (pp. 157-178). Humana Press.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology, 14*(8), 2611-2620.

Excoffier, L., Hofer, T., & Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity, 103*(4), 285-298.

Excoffier, L., & Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources, 10*(3), 564-567.

Faust, E., Halvorsen, K. T., Andersen, P., Knutsen, H., & André, C. (2018). Cleaner fish escape salmon farms and hybridize with local wrasse populations. *Royal Society Open Science, 5*(3), 171752.

Fekjan, J. (n.d) *Rognkjeks* Cyclopterus lumpus *Linnaeus, 1758*. Retrieved from https://artsdatabanken.no/Pages/F37435

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.

Fumagalli, M. (2013). Assessing the Effect of Sequencing Depth and Sample Size in Population Genetics Inferences. *PloS one, 8*(11), e79667.

Garcia-Mayoral, E., Olsen, M., Hedeholm, R., Post, S., Nielsen, E. E., & Bekkevold, D. (2016). Genetic structure of West Greenland populations of lumpfish *Cyclopterus lumpus*. *Journal of Fish Biology, 89*(6), 2625-2642.

Geitung, L., Wright, D. W., Oppedal, F., Stien, L. H., Vågseth, T., & Madaro, A. (2020). Cleaner fish growth, welfare and survival in Atlantic salmon sea cages during an autumn-winter production. *Aquaculture, 528*, 735623.

Gjertsen, K. E., Horneland, P. A., & Skulstad, E. M. (2016). *Lumpfish distribution map*. Received from the Institute of Marine Research.

Gjøsæter, H. (1998). The population biology and exploitation of capelin (*Mallotus villosus*) in the barents sea. *Sarsia, 83*(6), 453-496.

Goudet, J., & Jombart, T. (2020). hierfstat: Estimation and Tests of Hierarchical F-Statistics. R package version 0.5-7. Retrieved from https://CRAN.R-project.org/package=hierfstat

Goulet, D., Green, J. M., & Shears, T. H. (1986). Courtship, spawning, and parental care behavior of the lumpfish, *Cyclopterus lumpus* L., in Newfoundland. *Canadian Journal of Zoology, 64*(6), 1320-1325.

Grefsrud, E. S., Svåsand, T., Glover, K., Husa, V., Hansen, P. K., Samuelsen, O., . . . Stien, L. H. (2019). *Risikorapport norsk fiskeoppdrett 2019 - Miljøeffekter av lakseoppdrett* (Fisken og havet 2019-5). Retrieved from https://www.hi.no/hi/nettrapporter/fisken-og-havet-2019-5#sec-risikovurdering-av-velfer

Guo, S. W., & Thompson, E. A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics, 48*(2), 361-372.

Gärke, C., Ytournel, F., Bed'hom, B., Gut, I., Lathrop, M., Weigend, S., & Simianer, H. (2012). Comparison of SNPs and microsatellites for assessing the genetic structure of chicken populations. *Animal Genetics, 43*(4), 419-428.

Hartl, D. L., & Clark, A. G. (2007). *Principles of Population Genetics* (4th ed.). Sunderland, MA: Sinauer Associates, Inc.

Hedeholm, R., Blicher, M. E., & Grønkjær, P. (2014). First estimates of age and production of lumpsucker (*Cyclopterus lumpus*) in Greenland. *Fisheries Research, 149,* 1-4.

Herrmann, B., Sistiaga, M., & Jørgensen, T. (2021). Size-dependent escape risk of lumpfish (*Cyclopterus lumpus*) from salmonid farm nets. *Marine Pollution Bulletin, 162,* 111904.

Holst, J. C. (1993). Observations on the distribution of lumpsucker (*Cyclopterus lumpus*, L.) in the Norwegian Sea. *Fisheries Research, 17*(3), 369-372.

Hvas, M., Folkedal, O., Imsland, A., & Oppedal, F. (2018). Metabolic rates, swimming capabilities, thermal niche and stress response of the lumpfish, *Cyclopterus lumpus*. *Biology Open, 7*(9), bio036079.

Imsland, A. K. D., Hangstad, T. A., Jonassen, T. M., Stefansson, S. O., Nilsen, T. O., Hovgaard, P., . . . Reynolds, P. (2019). The use of photoperiods to provide year round spawning in lumpfish *Cyclopterus lumpus*. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology, 228*, 62-70.

Imsland, A. K. D., Hanssen, A., Nytrø, A. V., Reynolds, P., Jonassen, T. M., Hangstad, T. A., . . . Mikalsen, B. (2018). It works! Lumpfish can significantly lower sea lice infestation in large-scale salmon farming. *Biology Open, 7*(9), bio036301.

Institute of Marine Research [IMR] (2019). *Faste stasjoner*. Retrieved from http://www.imr.no/forskning/forskningsdata/stasjoner/view?station=

International Union for Conservation of Nature [IUCN] (2001). *IUCN Red List categories and criteria: version 3.1.* IUCN Species Survival Commission.

Jenkins, D. G., Carey, M., Czerniewska, J., Fletcher, J., Hether, T., Jones, A., . . . Tursi, R. (2010). A meta-analysis of isolation by distance: relic or reference standard for landscape genetics? *Ecography, 33*(2), 315-320.

Johannesson, J. (2006). *Lumpfish caviar – from vessel to consumer (FAO Fisheries Technical Paper. No. 485).* Food and Agriculture Organization of the United Nations.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics, 24*(11), 1403-1405.

Jónsdóttir, Ó. D. B., Schregel, J., Hagen, S. B., Tobiassen, C., Aarnes, S. G., & Imsland, A. K. D. (2018). Population genetic structure of lumpfish along the Norwegian coast: aquaculture implications. *Aquaculture International, 26*(1), 49-60.

Kasper, J., Bertelsen, B., Ólafsson, H. G., Holst, J. C., Sturlaugsson, J., & Jónsson, S. P. (2014). Observations of growth and postspawning survival of lumpfish *Cyclopterus lumpus* from mark-recapture studies. *Journal of Fish Biology, 84*(6), 1958-1963.

Kennedy, J. (2018). Oocyte size distribution reveals ovary development strategy, number and relative size of egg batches in lumpfish (*Cyclopterus lumpus*). *Polar Biology, 41*(6), 1091-1103.

Kennedy, J., Durif, C. M. F., Florin, A.-B., Fréchet, A., Gauthier, J., Hüssy, K., . . . Hedeholm, R. B. (2018). A brief history of lumpfishing, assessment, and management across the North Atlantic. *ICES Journal of Marine Science, 76*(1), 181-191.

Kennedy, J., Jónsson, S. Þ., Kasper, J. M., & Ólafsson, H. G. (2015). Movements of female lumpfish (*Cyclopterus lumpus*) around Iceland. *ICES Journal of Marine Science, 72*(3), 880-889.

Kennedy, J., & Ólafsson, H. G. (2019). Conservation of spawning time between years in lumpfish *Cyclopterus lumpus* and potential impacts from the temporal distribution of fishing effort. *Fisheries Management and Ecology, 26*(4), 389-396.

Kijewska, A., Kalamarz-Kubiak, H., Arciszewski, B., Guellard, T., Petereit, C., & Wenne, R. (2016). Adaptation to salinity in Atlantic cod from different regions of the Baltic Sea. *Journal of Experimental Marine Biology and Ecology, 478*, 62-67.

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources, 15*(5), 1179-1191.

Lamichhaney, S., Martinez Barrio, A., Rafati, N., Sundström, G., Rubin, C.-J., Gilbert, E. R., . . . Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences of the United States of America, 109*(47), 19345-19350.

Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications, 9*(1), 3258.

Lemopoulos, A., Prokkola, J. M., Uusi-Heikkilä, S., Vasemägi, A., Huusko, A., Hyvärinen, P., . . . Vainikka, A. (2019). Comparing RADseq and microsatellites for estimating genetic diversity and relatedness - Implications for brown trout conservation. *Ecology and evolution, 9*(4), 2106-2120.

Lennox, R. J., Paukert, C. P., Aarestrup, K., Auger-Méthé, M., Baumgartner, L., Birnie-Gauvin, K., . . . Cooke, S. J. (2019). One Hundred Pressing Questions on the Future of Global Fish Migration Science, Conservation, and Policy. *Frontiers in Ecology and Evolution, 7*(286).

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), 2987-2993.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics, 25*(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

Linnaeus, C. (1758). *Systema naturae* (Vol. 1, No. part 1, p. 532). Stockholm: Laurentii Salvii.

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics, 28*(2), 298-299.

Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Molecular Ecology, 23*(9), 2178-2192.

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology, 24*(5), 1031-1046.

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics, 4*(12), 981-994.

Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources, 17*(1), 67-77.

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research, 27*(2 Part 1), 209-220.

Marine Stewardship Council [MSC] (2017). *Nye norske MSC-sertifiserte arter.* Retrieved from https://www.msc.org/no/presse/pressemeldinger/nye-norske-msc-sertifiserte-arter

Mecklenburg, C. W., & Sheiko, B. A. (2003). Family Cyclopteridae Bonaparte 1831 - lumpsuckers. *California Academy of Sciences Annotated Checklists of Fishes, 6*, 1-17.

Moore, J. S., Bourret, V., Dionne, M., Bradbury, I., O'Reilly, P., Kent, M., . . . Bernatchez, L. (2014). Conservation genomics of anadromous Atlantic salmon across its North American range: outlier loci identify the same patterns of population structure as neutral loci. *Molecular Ecology, 23*(23), 5680-5697.

Moring, J. R. (2001). Intertidal growth of larval and juvenile lumpfish in Maine: a 20-year assessment. *Northeastern Naturalist, 8*(3), 347-354.

Moring, J. R., & Moring, S. W. (1991). Short-term movements of larval and juvenile lumpfish, *Cyclopterus lumpus* L., in tidepools. *Journal of Fish Biology, 38*(6), 845-850.

Mortensen, A., Johansen, R. B., Hansen, Ø. J., & Puvanendran, V. (2020). Temperature preference of juvenile lumpfish (*Cyclopterus lumpus*) originating from the southern and northern parts of Norway. *Journal of Thermal Biology, 89*, 102562.

Nielsen, R. (2005). Molecular Signatures of Natural Selection. *Annual Review of Genetics, 39*(1), 197-218.

Norwegian Biodiversity Information Centre (2015). Cyclopterus lumpus *Linnaeus, 1758*. Retrieved from https://artsdatabanken.no/Rodliste2015/rodliste2015/Norge/42937

Norwegian Directorate of Fisheries (2020a). *Regulering av fisket etter rognkjeks i 2021* (Høringer av reguleringer for 2021, Sak 13/2020). Retrieved from https://www.fiskeridir.no/Yrkesfiske/Dokumenter/Reguleringsmoetet2/Hoeringer-av-reguleringer-for-2021-reguleringsmoetet

Norwegian Directorate of Fisheries (2020b). *Use of cleanerfish 1998-2019.* Retrieved from https://www.fiskeridir.no/English/Aquaculture/Statistics/Cleanerfish-Lumpfish-and-Wrasse

The Norwegian Ministry of Trade, Industry and Fisheries (2015). *Forutsigbar og miljømessig bærekraftig vekst i norsk lakse- og ørretoppdrett* (Meld. St. 16 (2014–2015)). Retrieved from https://www.regjeringen.no/

Norwegian Scientific Advisory Committee for Atlantic Salmon (2020). *Status for norske laksebestander i 2020. Rapport fra Vitenskapelig råd for lakseforvaltning* (15). Retrieved from https://hdl.handle.net/11250/2657947

Nytrø, A. V., Vikingstad, E., Foss, A., Hangstad, T. A., Reynolds, P., Eliassen, G., . . . Imsland, A. K. (2014). The effect of temperature and fish size on growth of juvenile lumpfish (*Cyclopterus lumpus* L.). *Aquaculture, 434*, 296-302.

Oleksiak, M. F. (2019). Adaptation Without Boundaries: Population Genomics in Marine Systems. In O. P. Rajora (Ed.), *Population Genomics: Concepts, Approaches and Applications* (pp. 587-612). Cham: Springer International Publishing.

Oleksiak, M. F., & Rajora, O. P. (2020). Marine Population Genomics: Challenges and Opportunities. In M. F. Oleksiak & O. P. Rajora (Eds.), *Population Genomics: Marine Organisms* (pp. 3-35). Cham: Springer International Publishing.

Overton, K., Dempster, T., Oppedal, F., Kristiansen, T. S., Gismervik, K., & Stien, L. H. (2019). Salmon lice treatments and salmon mortality in Norwegian aquaculture: a review. *Reviews in Aquaculture, 11*(4), 1398-1417.

Pampoulie, C., Skirnisdottir, S., Olafsdottir, G., Helyar, S. J., Thorsteinsson, V., Jónsson, S. Þ., . . . Kasper, J. M. (2014). Genetic structure of the lumpfish *Cyclopterus lumpus* across the North Atlantic. *ICES Journal of Marine Science, 71*(9), 2390-2397.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics, 26*(3), 419-420.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics, 35*(3), 526-528.

Pembleton, L. W., Cogan, N. O., & Forster, J. W. (2013). StAMPP: An R package for calculation of genetic differentiation and structure of mixed‐ploidy level populations. *Molecular Ecology Resources, 13*(5), 946-952.

Powell, A., Treasurer, J. W., Pooley, C. L., Keay, A. J., Lloyd, R., Imsland, A. K., & Garcia de Leaniz, C. (2018). Use of lumpfish for sea-lice control in salmon farming: challenges and opportunities. *Reviews in Aquaculture, 10*(3), 683-702.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics, 155*(2), 945-959.

Privé, F., Luu, K., Vilhjálmsson, B. J., & Blum, M. G. (2020). Performing highly efficient genome scans for local adaptation with R package pcadapt version 4. *Molecular biology and evolution, 37*(7), 2153-2154.

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. *American Journal of Human Genetics, 81*(3), 559-575.

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology, 24*(17), 4348-4370.

Schopka, S. A. (1974) Preliminary results from tagging of lumpsucker (*Cyclopterus lumpus*), in Icelandic waters 1971–3. *ICES CM 1974/F: 18,* 6 p. (mimeo)

Sigsgaard, E. E., Nielsen, I. B., Carl, H., Krag, M. A., Knudsen, S. W., Xing, Y., . . . Thomsen, P. F. (2017). Seawater environmental DNA reflects seasonality of a coastal fish community. *Marine Biology, 164*(6), 128.

Skirnisdottir, S., Olafsdottir, G., Olafsson, K., Jendrossek, T., Lloyd, H. a. D., Helyar, S., . . . Kasper, J. M. (2013). Twenty-two novel microsatellite loci for lumpfish (*Cyclopterus lumpus*). *Conservation Genetics Resources, 5*(1), 177-179.

Stien, L. H., Størkersen, K. V., & Gåsnes, S. K. (2020). *Analyse av dødelighetsdata fra spørreundersøkelse om velferd hos rensefisk* (Rapport fra havforskningen 2020-6). Retrieved from https://www.hi.no/hi/nettrapporter/rapport-fra-havforskningen-2020-6

Sunde, J., Yıldırım, Y., Tibblin, P., & Forsman, A. (2020). Comparing the Performance of Microsatellites and RADseq in Population Genetic Studies: Analysis of Data for Pike (*Esox lucius*) and a Synthesis of Previous Studies. *Frontiers in Genetics, 11*(218).

Sætre, R., & Ljøen, R. (1972). The Norwegian Coastal Current. *Proceedings of the first international conference on port and ocean engineering under Arctic conditions, Trondheim, Norway Aug. 23-30,* 514-535.

Thibert-Plante, X., & Hendry, A. P. (2010). When can ecological speciation be detected with neutral loci? *Molecular Ecology, 19*(11), 2301-2314.

Thorstad, E. B., Todd, C. D., Uglem, I., Bjørn, P. A., Gargan, P. G., Vollset, K. W., . . . Finstad, B. (2015). Effects of salmon lice *Lepeophtheirus salmonis* on wild sea trout *Salmo trutta* - a literature review. *Aquaculture Environment Interactions, 7*(2), 91-113.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research, 38*(16), e164.

Weir, B. S., & Cockerham, C. C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution, 38*(6), 1358-1370.

Wetterstrand, K. A. (2020, 7th December). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Retrieved from www.genome.gov/sequencingcostsdata

Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of F_{ST}. *The American Naturalist, 186*(S1), S24-S36.

Whittaker, B. A., Consuegra, S., & de Leaniz, C. G. (2018). Genetic and phenotypic differentiation of lumpfish (*Cyclopterus lumpus*) across the North Atlantic: implications for conservation and aquaculture. *PeerJ, 6*, e5974.

Wright, S. (1943). Isolation by distance. *Genetics, 28*(2), 114.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, 395-420.

Yuen, J. W., Dempster, T., Oppedal, F., & Hvas, M. (2019). Physiological performance of ballan wrasse (*Labrus bergylta*) at different temperatures and its implication for cleaner fish usage in salmon aquaculture. *Biological Control, 135*, 117-123.

Appendices

Appendix 1

Scree plot showing that the four first principal components in PCA of the full data set explain most of the genetic variation.

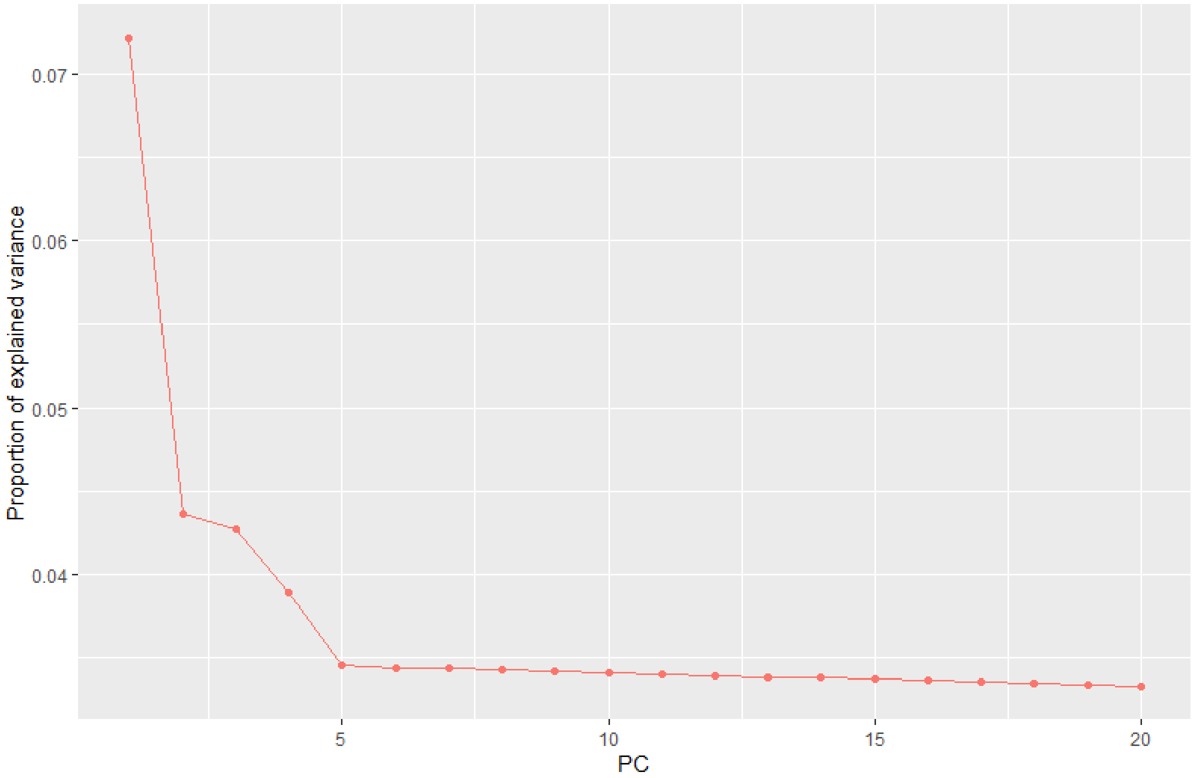


Figure A1.1: Scree plot obtained from *pcadapt* showing the proportions of variance explained by the 20 first principal components (PCs) in a principal component analysis (PCA) of the full data set. PC 1-4 explains most of the variation.

Appendix 2

Figures used to infer the identity of the samples in the *pcadapt* PCA plots.

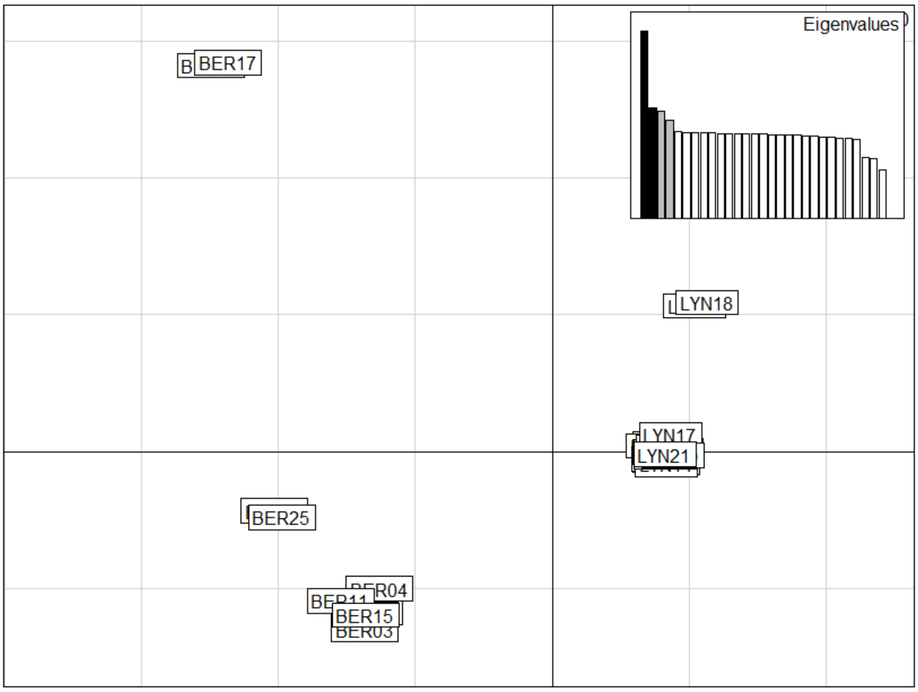


Figure A2.1: Principal component analysis (PCA) of full data set, performed with the *adegenet* package in *R.* PCA of the neutral data set showed similar results (not included).

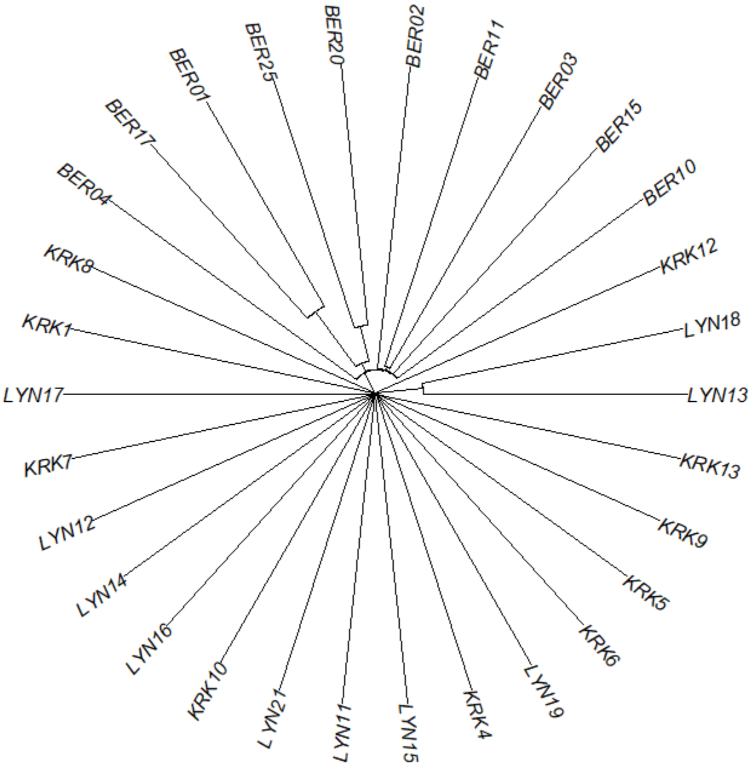


Figure A2.2: Neighbour joining tree of full data set, obtained with the *ape* package in *R.* The neutral data set showed similar patterns (not included).

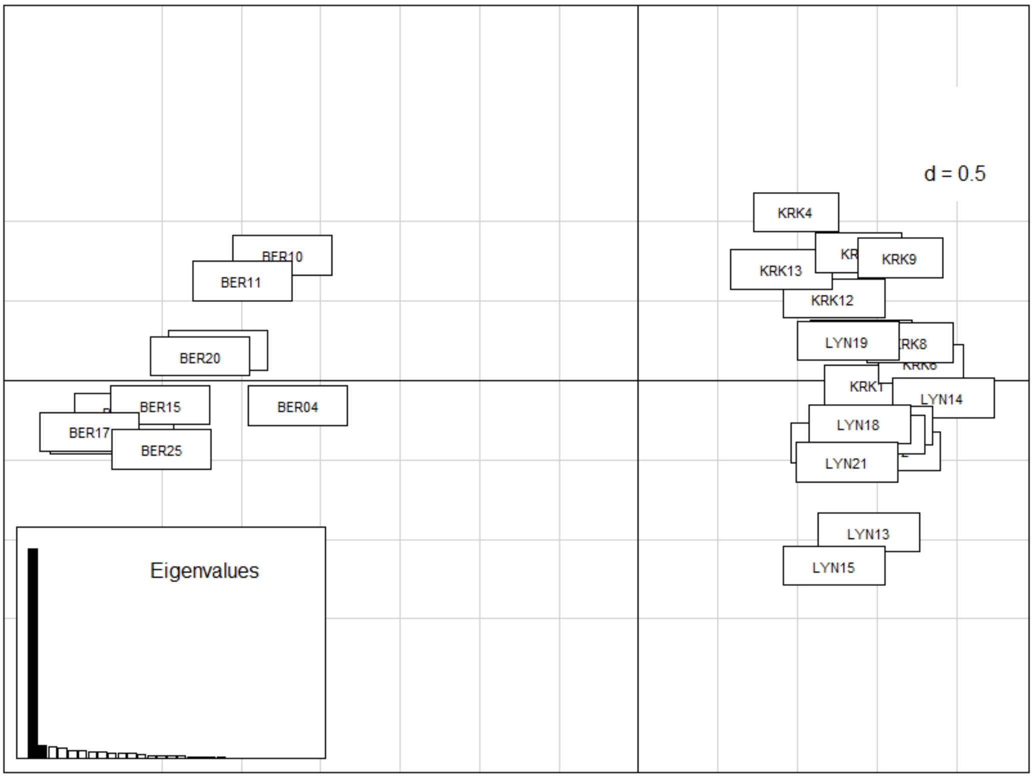


Figure A2.3: Principal component analysis (PCA) of outlier data set, performed with the *adegenet* package in *R.*

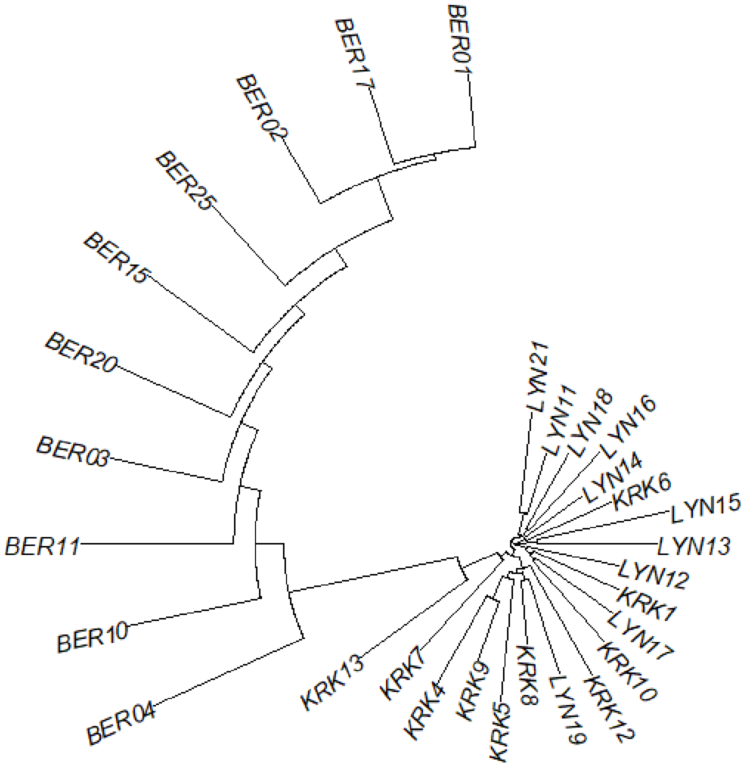


Figure A2.4: Neighbour joining tree of outlier data set, obtained with the *ape* package in *R.*

Appendix 3

Complete results from *ADMIXTURE/CLUMPAK*.

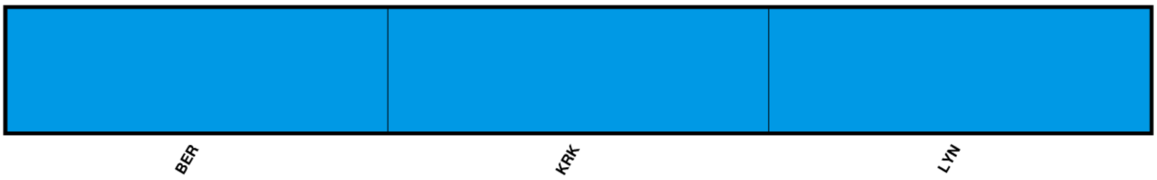
Table A3.1: Sample names rendered from left to right for *ADMIXTURE/CLUMPAK* bar plots. Each bar represents one sample. Please note that in the succeeding bar plots, KRK corresponds to TRO.

BER (Bergen)	TRO (Tromsø)	LYN (Lyngen)
BER01	TRO1	LYN11
BER02	TRO10	LYN12
BER03	TRO12	LYN13
BER04	TRO13	LYN14
BER10	TRO4	LYN15
BER11	TRO5	LYN16
BER15	TRO6	LYN17
BER17	TRO7	LYN18
BER20	TRO8	LYN19
BER25	TRO9	LYN21

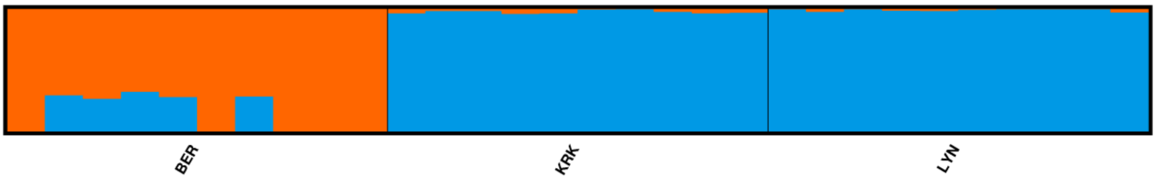
All bar plots obtained with *CLUMPAK* (main pipeline) for the full data set

Major modes for the uploaded data:

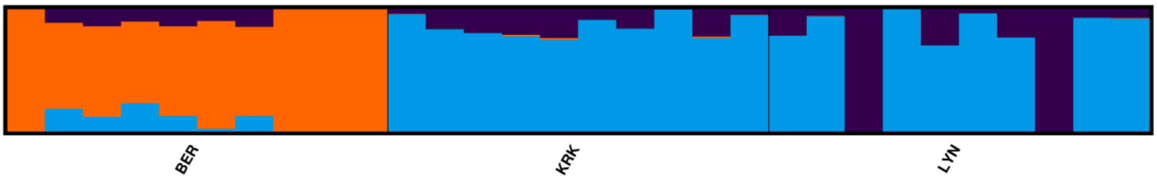
K=1



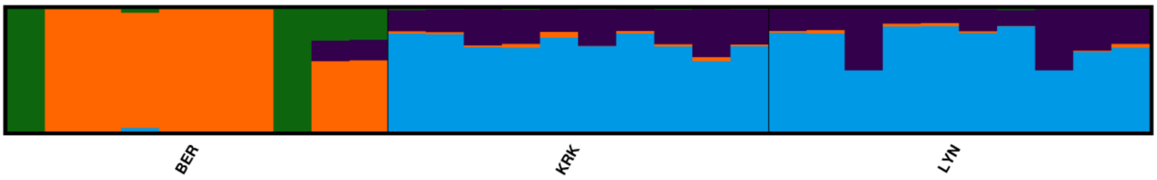
K=2



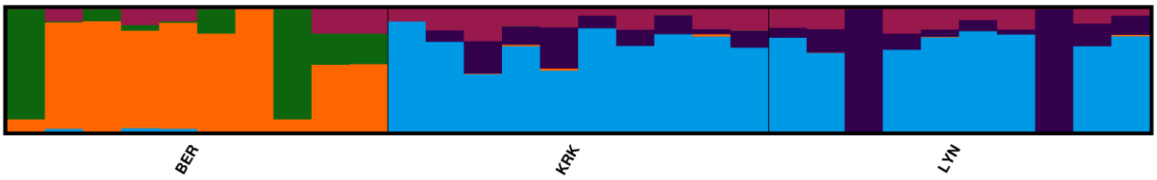
K=3



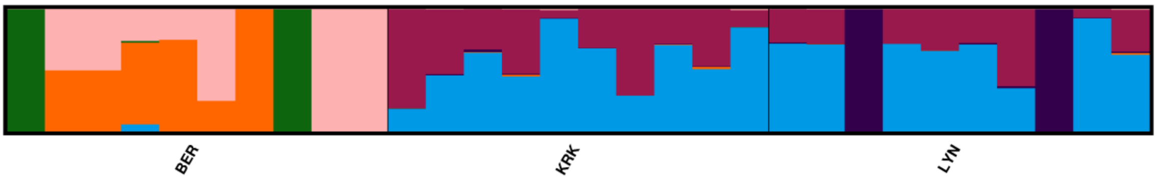
K=4



K=5

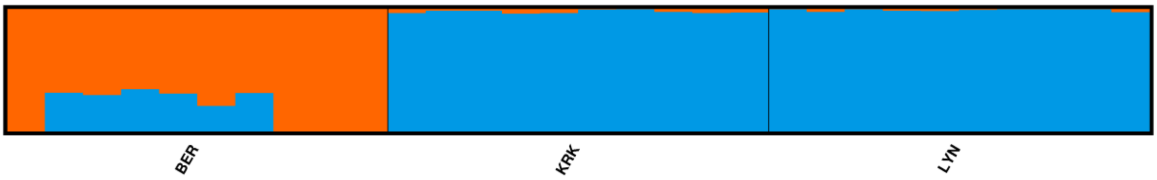


K=6

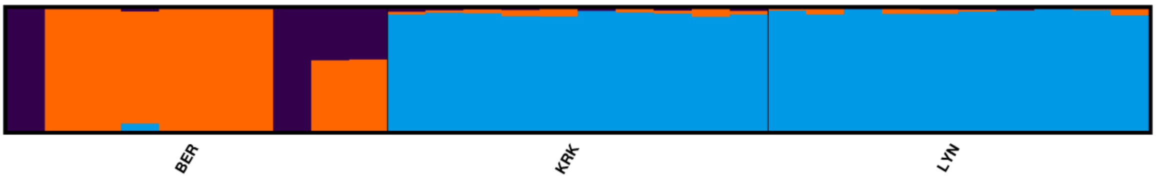


Minor modes for the uploaded data (for values of K with several minor modes, only the first one is included):

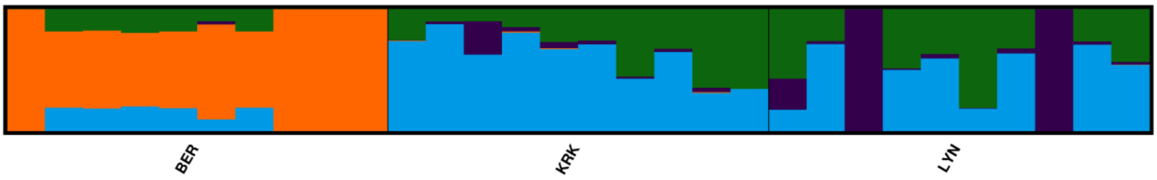
K=2 Minor Cluster 1



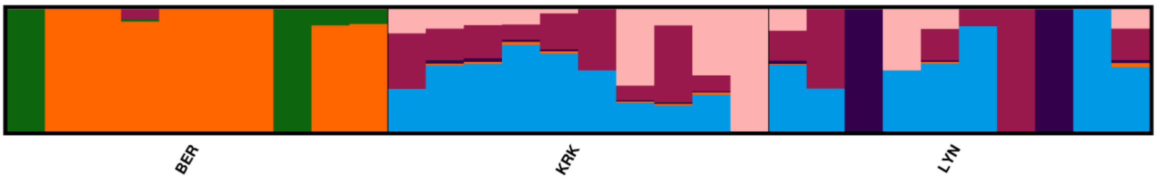
K=3 Minor Cluster 1



K=4 Minor Cluster 1



K=6 Minor Cluster 1



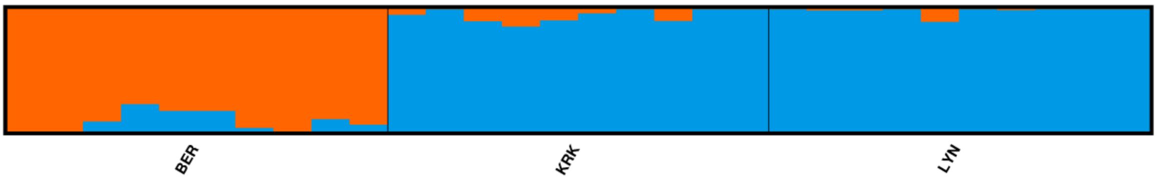
Division of runs by mode:
K=1 10/10
K=2 7/10, 3/10
K=3 7/10, 3/10
K=4 6/10, 4/10
K=5 10/10
K=6 4/10, 2/10, 1/10, 1/10, 1/10, 1/10

All bar plots obtained with *CLUMPAK* (main pipeline) for outlier data set

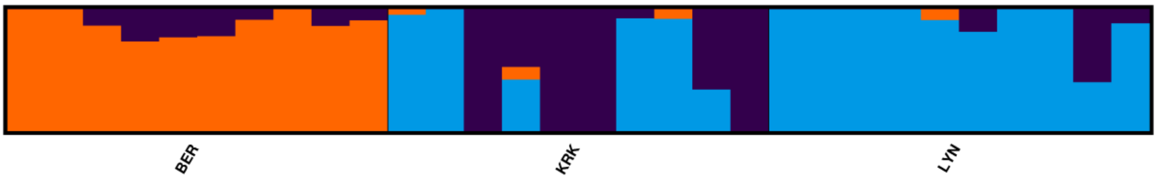
Major modes for the uploaded data:

K=1

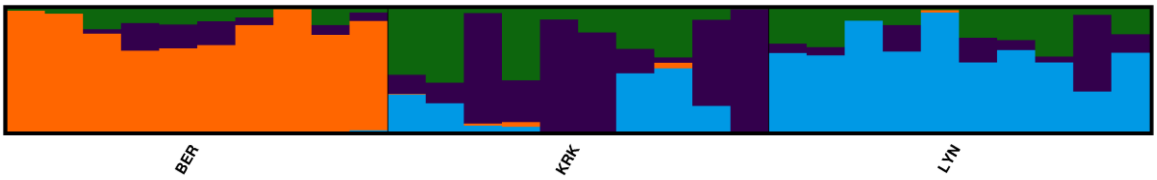
K=2



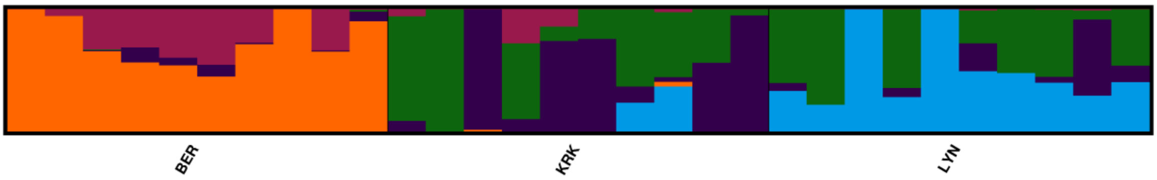
K=3



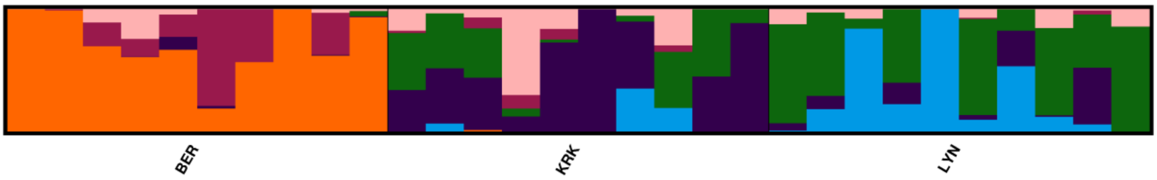
K=4



K=5

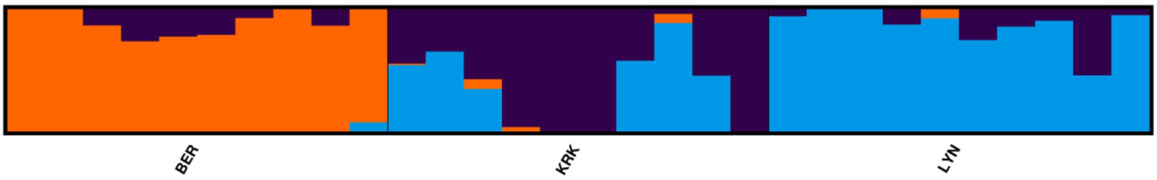


K=6

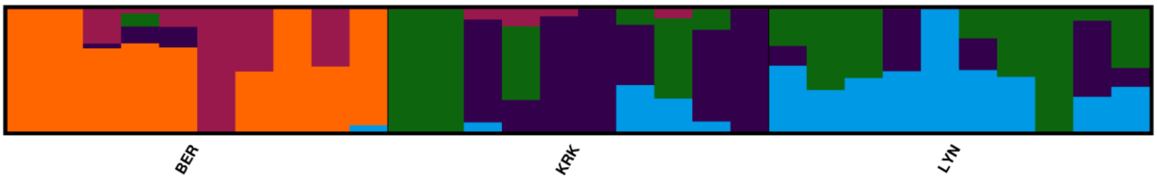


Minor modes for the uploaded data (for values of K with several minor modes, only the first one is included):

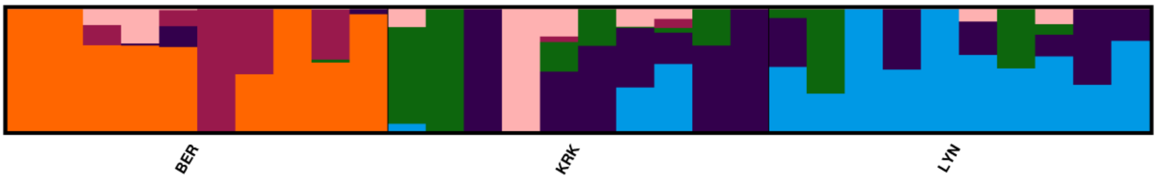
K=3 Minor Cluster 1



K=5 Minor Cluster 1



K=6 Minor Cluster 1



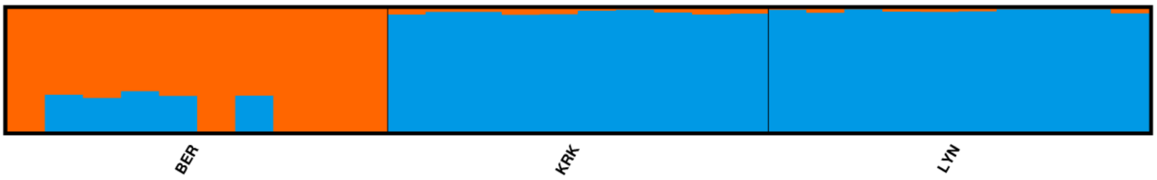
Division of runs by mode:
K=1 10/10
K=2 10/10
K=3 5/10, 3/10, 2/10
K=4 10/10
K=5 6/10, 2/10, 2/10
K=6 7/10, 2/10, 1/10

All bar plots obtained with *CLUMPAK* (main pipeline) for neutral data set

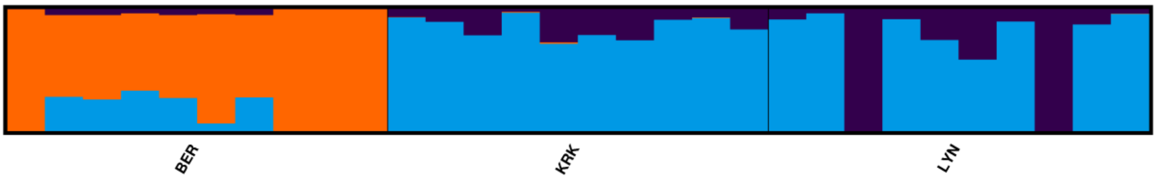
Major modes for the uploaded data:

K=1

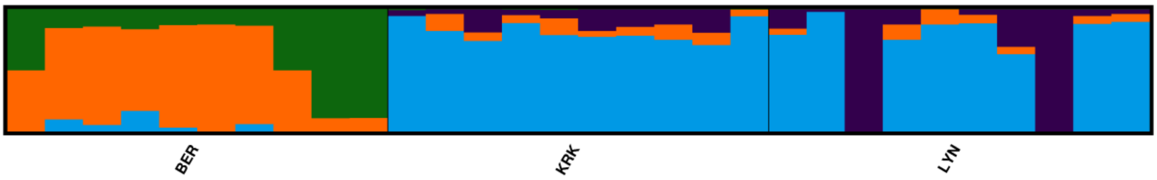
K=2



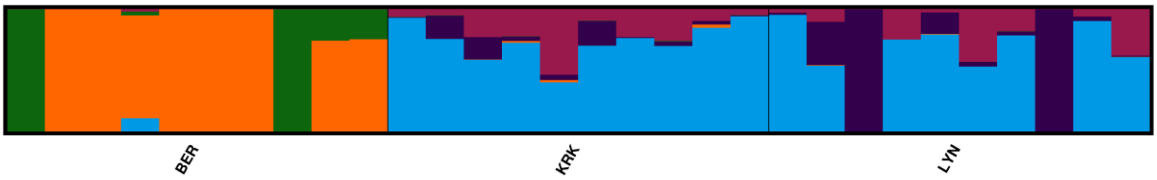
K=3



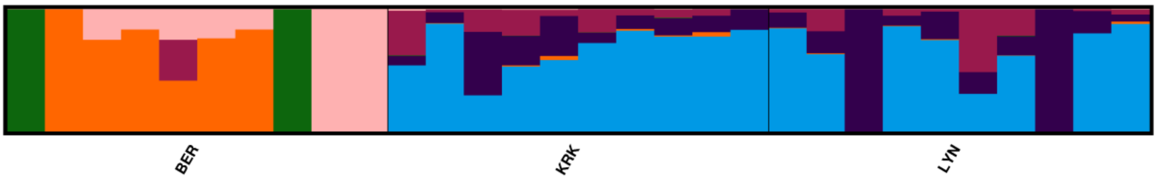
K=4



K=5

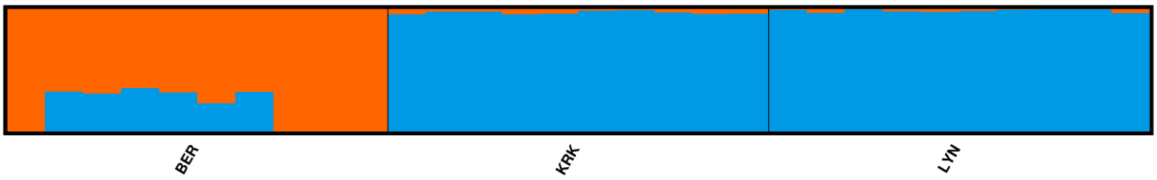


K=6

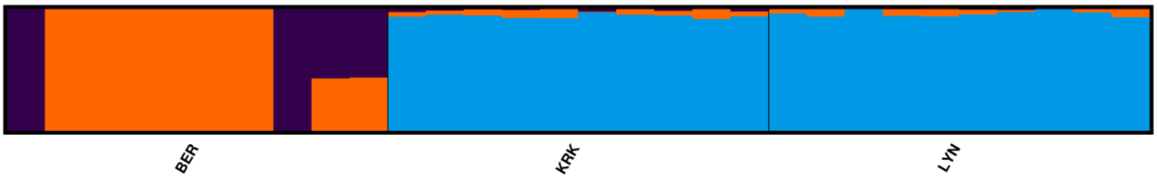


Minor modes for the uploaded data (for values of K with several minor modes, only the first one is included):

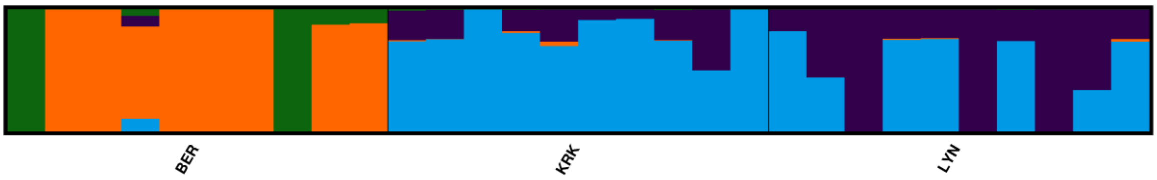
K=2 Minor Cluster 1



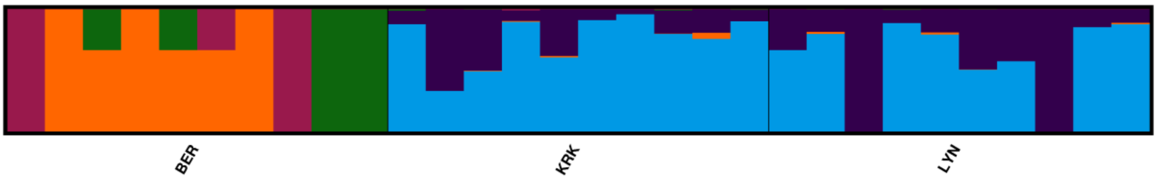
K=3 Minor Cluster 1



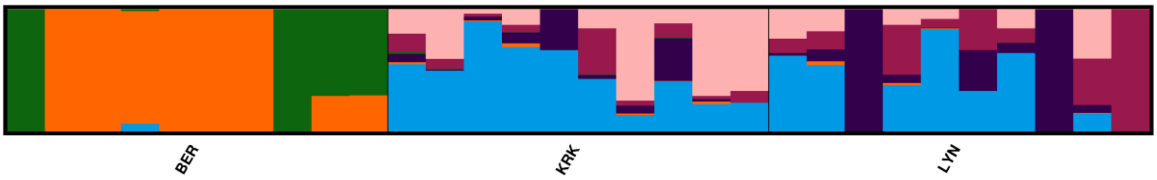
K=4 MinorCluster1



K=5 MinorCluster1



K=6 MinorCluster1



Division of runs by mode:
K=1 10/10
K=2 6/10, 4/10
K=3 6/10, 4/10
K=4 8/10, 2/10
K=5 6/10, 3/10, 1/10
K=6 6/10, 3/10, 1/10

Appendix 4

Graphs obtained from *CLUMPAK*, inferring the optimal K (maximum ΔK) from the *ADMIXTURE* results as proposed by Evanno et al. (2005).

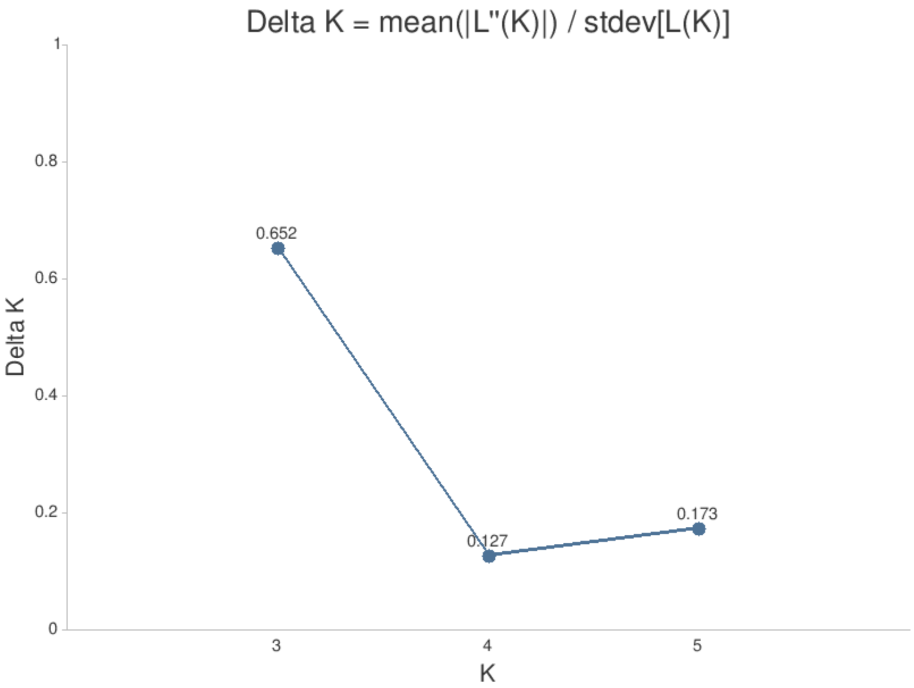


Figure A4.1: Graph showing the optimal K (maximum ΔK) as proposed by Evanno et al. (2005) for the full data set, inferred from *ADMIXTURE* results (calculations performed using *CLUMPAK*). Maximum ΔK = 3.

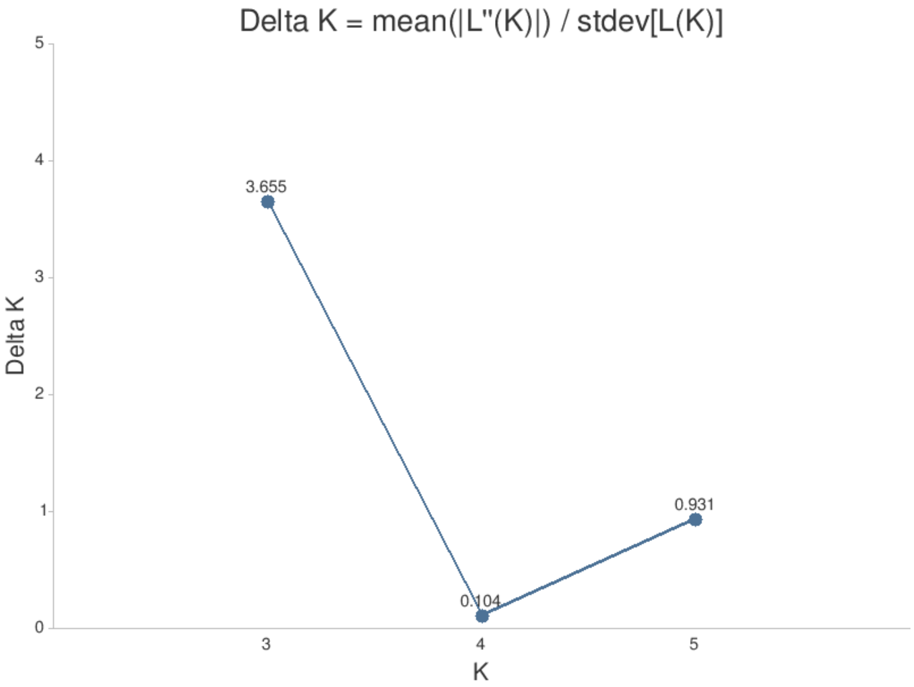


Figure A4.2: Graph showing the optimal K (maximum ΔK) as proposed by Evanno et al. (2005) for the outlier data set, inferred from *ADMIXTURE* results (calculations performed using *CLUMPAK*). Maximum ΔK = 3.

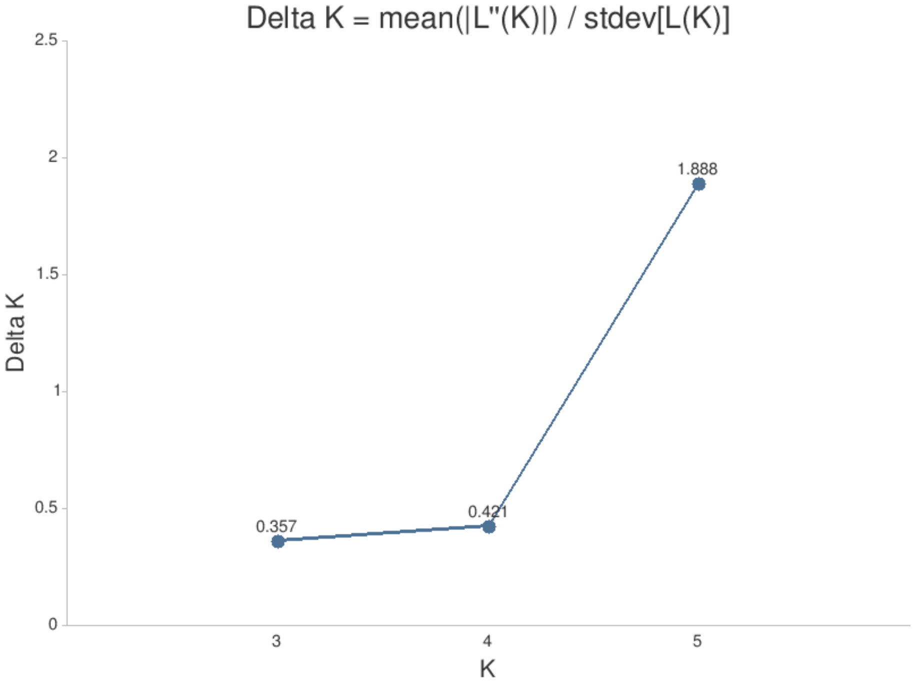


Figure A4.3: Graph showing the optimal K (maximum ∆K) as proposed by Evanno et al. (2005) for the neutral data set, inferred from *ADMIXTURE* results (calculations performed using *CLUMPAK*). Maximum ∆K = 5.