Contents lists available at ScienceDirect

Aquaculture Reports

journal homepage: www.elsevier.com/locate/aqrep

Genetic diversity insights from population genomics and machine learning tools for Nordic Arctic charr (*Salvelinus alpinus*) populations

Christos Palaiokostas[a,*], Khrystyna Kurta[a], Fotis Pappas[a], Henrik Jeuthe[a,b], Ørjan Hagen[c], José Beirão[c], Matti Janhunen[d], Antti Kause[e]

[a] *Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala 750 07, Sweden*
[b] *Aquaculture Center North, Åvägen 17, Kälarne 844 61, Sweden*
[c] *Faculty of Bioscience and Aquaculture, Nord University, Bodo 8049, Norway*
[d] *Natural Resources Institute Finland (Luke), Joensuu FI-80100, Finland*
[e] *Natural Resources Institute Finland (Luke), Breeding and Genomics, Jokioinen FI-31600, Finland*

ARTICLE INFO

Keywords:
Arctic charr
Genetic diversity
Population structure
DdRAD
SNPs
Haplotypes
Unsupervised machine learning

ABSTRACT

Arctic charr (*Salvelinus alpinus*) is a salmonid species of high ecological and commercial value in the Holarctic region. Nevertheless, more information is needed about its underlying genetic diversity and population structure in the Nordics, especially regarding farmed populations. High-throughput sequencing was applied in three Arctic charr populations of anadromous or landlocked origin from Finland, Norway and Sweden. More specifically, the animals from the Swedish and Norwegian populations originated from a major egg supplier and producer, respectively. Furthermore, in the case of the Finnish population, the sampled animals originated from the only active conservation program for Arctic charr in the country with a potential interest in farming. Using double-digest restriction site-associated DNA sequencing (ddRAD-seq) on more than 500 fish, over 2000 single nucleotide polymorphisms (SNPs), both in the form of individual SNPs and as read haplotypes, were used to study the genetic diversity and structure of those populations. Genetic diversity metrics were similar between the Norwegian and the Swedish populations. However, substantially lower (40–50 %) genetic diversity was found in the Finnish population. Moreover, considerable genetic differentiation was implied between the studied populations as the mean fixation index (F_{ST}) was above 0.1 in all pairwise comparisons. All populations were easily discernible through either principal component analysis (PCA) or discriminant analysis of principal components (DAPC). In addition, unsupervised machine learning models such as K-means, Gaussian and Bayesian Gaussian mixtures were assessed for their ability to detect genetic clusters. A preceding dimensionality reduction step by PCA resulted in all three models, suggesting that the most probable number of clusters was three. Overall, our study affirmed the utility of the developed ddRAD-seq genotyping method and unveiled the genetic structure of the studied populations, both of which could contribute to their more efficient management by captive breeding.

1. Introduction

Fecundity in fish is orders of magnitude higher than in livestock. Due to their very high fecundity, higher genetic gains are commonly observed in aquaculture breeding programs than in livestock (Gjedrem and Rye, 2016; Kause et al., 2022; Vandeputte et al., 2022). Nevertheless, the above makes it possible to reach short-term production volumes using a minimal number of broodfish. Owing to the inherent difficulty of keeping track of genetic relationships among individuals in aquaculture settings, inbreeding levels may rise rapidly and threaten the long-term sustainability of the farmed populations (Saura et al., 2021).

As such, deciphering the status of genetic diversity and unveiling population structure in farmed fish populations is essential for efficient management (Symonds et al., 2019). Moreover, aside from producing fish for human consumption, aquaculture practices often aim to assist in the recovery of endangered natural populations through restocking activities (Mizuta et al., 2023). In those cases, minimising the loss of genetic diversity is the main priority as it increases the chances for the released fish to adapt to the natural environment (Fisch et al., 2015).

High-throughput sequencing technologies have powered up

[*] Corresponding author.
E-mail address: christos.palaiokostas@slu.se (C. Palaiokostas).

https://doi.org/10.1016/j.aqrep.2024.102495
Received 4 January 2024; Received in revised form 28 October 2024; Accepted 5 November 2024
Available online 7 November 2024

population genetic studies in fish, allowing for genome-wide genetic diversity estimates. Genotyping by sequencing (GBS) approaches, such as double digest restriction-site associated DNA (ddRAD-seq) (Peterson et al., 2012), offer a cost-effective solution for the simultaneous *de novo* detection of genetic markers like single nucleotide polymorphisms (SNP) and genotyping using short-read sequencing platforms. Particularly in genetic diversity studies, the characteristics of GBS constitute a prominent advantage over SNP arrays as the former does not suffer from SNP ascertainment bias, at least to the same extent (Davey et al., 2013). Therefore, reliable genetic diversity metrics can be derived with insights regarding potentially hidden population structure. Due to its ease of library construction, ddRAD has been routinely used over the last decade in several genetic diversity studies in fish (Hosoya et al., 2018; Moses et al., 2019; Nyinondi et al., 2020; Takahashi et al., 2020; Saha et al., 2021; Palaiokostas et al., 2022; Naito et al., 2023).

Deriving inferences from the GBS data revolves almost entirely around individual SNPs identified in the obtained sequenced reads. In cases where several SNPs exist within the same read, it is customary to retain only one for further analysis due to their high linkage disequilibrium. Since paired-end sequence reads of 300–600 base pairs (bp) long are usually produced, additional information could be gained by considering the SNPs within the same sequenced read as phased haplotypes (Malinsky et al., 2018).

Furthermore, regardless of whether the analysis is centred around individual SNPs or haplotypes, most genetic diversity studies aim to derive information about population structure and identify underlying genetic clusters. Here, common analytic tools include the principal component analysis and clustering approaches, which can be either Bayesian-based (Pritchard et al., 2000) or in the form of discriminant analysis of principal components (Jombart et al., 2010). Detecting genetic clusters generally falls within the scope of a broad category of machine learning algorithms known as unsupervised learning. Even though machine learning algorithms have rarely been used in aquaculture genetics, promising results have recently been derived from aquaculture breeding studies (Bargelloni et al., 2021; Palaiokostas, 2021). Nevertheless, limited knowledge exists on whether unsupervised machine learning algorithms can help identify genetic clusters in SNP datasets from fish populations.

Arctic charr (*Salvelinus alpinus*) is a high-value salmonid inhabiting diverse Arctic ecosystems (Jacobs et al., 2020) with numerous genetically discrete populations across Europe (Klemetsen et al., 2003; Kottelat and Freyhof, 2007; Leskinen et al., 2013; Tiberti and Splendiani, 2019). Due to its inherent capacity for growth in cold waters even during winter, Arctic charr is particularly suited for commercial farming in the Nordics (Sæther et al., 2013). Moreover, conservation programs have been operating for several decades in Finland, aiming to support endangered wild charr populations through captive breeding and restocking (Primmer et al., 1999).

In the present study, we assessed the genetic diversity levels of three geographically distant Arctic charr populations from the Nordics and investigated their underlying population structure. The studied populations represent Arctic charr strains that either have a high impact on each country's farming industry (Norway and Sweden) or, in the case of Finland, could potentially form the base population for farming. More specifically, in the case of Sweden and Norway, the sampled populations originated from a major egg supplier and producer, respectively. On the other hand, the Finnish population originated from the country's only active conservation program, which holds potential for aquaculture purposes. Over 500 fish from those populations were genotyped using ddRAD-seq. Genetic diversity metrics and population-level coancestry coefficients were computed within and across populations using individual SNPs and phased read haplotypes. Moreover, we investigated the existence of genetic clusters using both commonly applied analytic tools in population genomics and unsupervised machine learning models.

2. Materials and methods

2.2. Background of studied populations

Arctic charr from the Swedish national breeding program, a commercial farm in Norway and the Finnish conservation program were used in our study. The Swedish breeding program for Arctic charr at Aquaculture Centre North (ACN) facilities in Jämtland, central Sweden (Fig. 1), has been active for approximately 40 years, supplying eggs to farms nationwide. The base population originates from the large Lake Hornavan in Northern Sweden, and a closed breeding nucleus has been kept since the beginning of the program (Eriksson et al., 2010). A representative sample of 170 brood fish (63 males, 107 females) from the 2013 year class was used. Additionally, samples from a Norwegian population of anadromous origin were obtained from a commercial farm in Sigerfjord, Norway (Fig. 1). The population founders originated primarily from the Hammerfest strain (~ 60–70 %) in northern Norway, previously shown to have a higher growth rate than landlocked populations in the country's south (Torrissen and Barnung, 1991). Secondarily, the Norwegian population originated from Svalbard (Norway) and Iceland (personal communication with Sigerfjord Fisk AC, January 2021). Mass spawning has been practised since 1995, with the breeding population consisting of 700 – 800 females and 150 males, on average, with seven generations recorded so far in captivity. In total, 164 breeding candidates (86 males and 78 females) from the 7th hatchery generation of the Norwegian population were used. The third sampled Arctic charr group represented a critically endangered landlocked population from Lake Kuolimo, Southeastern Finland. Due to the dramatic decline in population during the past century, natural reproduction has become scarce and is restricted only to some southern areas of the large Vuoksi watershed. Consequently, the Finnish population heavily depends on hatchery propagation and restocking activities performed by the Natural Resources Institute of Finland (Luke) (Primmer et al., 1999). Overall, 172 Finnish Arctic charr individuals were sampled from the wild (years 2013–2015) and hatchery broodstock (2019). The wild fish (27 males, 43 females and 14 immature fish with unknown sex) were captured from Lake Kuolimo and used to establish a new hatchery broodstock at Luke's Enonkoski aquaculture station. The rest of the fish from the Finnish population were sampled at the hatchery in 2019 ($n = 88$ fish) (Fig. 1).

2.3. DNA extraction and ddRAD library preparation

Genomic DNA was extracted from fin clips using a salt-based precipitation method (Palaiokostas et al., 2022). In summary, fin tissue was digested at 55 °C for 4 h using a lysis solution containing 200 μL SSTNE (50 mM Tris base, 300 mM NaCl, 0.2 mM each of EGTA and EDTA, 0.15 mM of spermine tetrahydrochloride, and 0.28 mM of spermidine trihydrochloride; pH 9; Sigma-Aldrich, Darmstadt, Germany), 10 % SDS (Bio-Rad, Hercules, USA), and 100 μg proteinase K. Thereafter, 5 μL RNaseA (Thermo Fisher, Vilnius, Lithuania) (2 mg/ml) was added followed by incubation at 37 °C for 60 min. Protein precipitation was performed by adding 0.7 vol of 5 M NaCl (Sigma-Aldrich, Darmstadt, Germany). With the addition of 0.7 vol of isopropanol and centrifugation (Pico 21, Thermo Fisher, Waltham, MA, USA) at 14 000 g for 5 min, a DNA pellet was formed. The DNA pellet was cleaned through overnight incubation with 75 % ethanol and dissolved in 30 μL of 5 mM Tris (pH 8.0; Sigma-Aldrich, Darmstadt, Germany). DNA content and quality metrics were obtained using a NanoDrop 8000 (Thermo Scientific, Waltham, MA, USA) spectrophotometer, agarose gel electrophoresis, and Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA). Finally, 5 mM Tris (pH 8.0) was used to dilute the DNA samples to 15 ng/μL.

The ddRAD library preparation was performed following a modified version of the original protocol described in detail by Palaiokostas et al. (2015). In total, four ddRAD libraries were prepared for 506 samples. 15 ng of each DNA sample were digested at 37 ℃ for 60 min with the

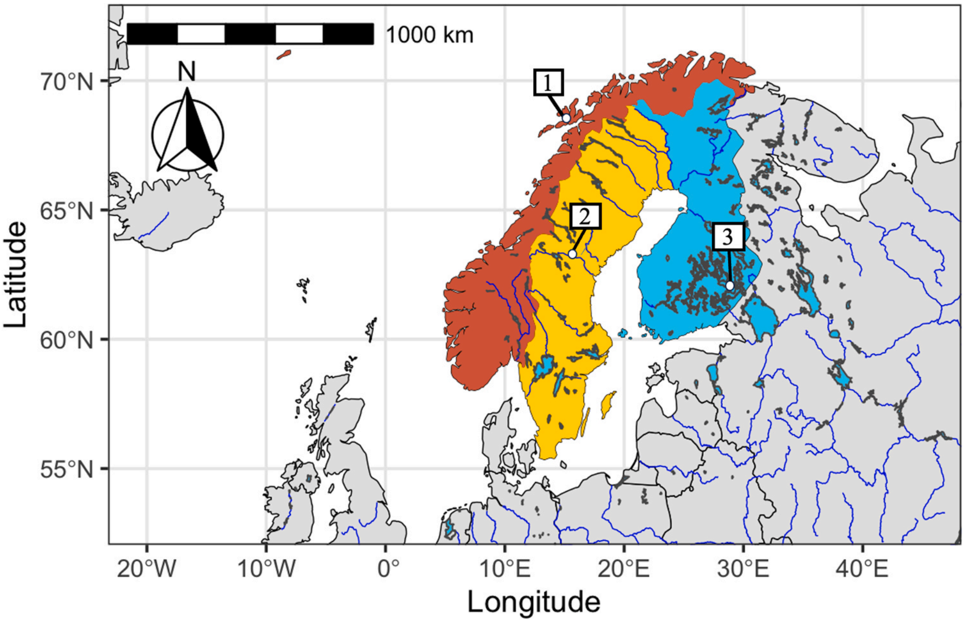

Fig. 1. Geographical overview of the used Arctic charr (*Salvelinus alpinus*) populations from 1 – Norway (n = 164), 2 – Sweden (n = 170), and 3 – Finland (n = 172). The Norwegian population was of admixed origin from Hammerfest-Svalbard in Norway and Iceland. The Swedish population originated from Lake Hornavan in Sweden, while the Finnish population was from Lake Kuolimo in Finland.

high-fidelity enzyme SbfI (CCTGCA|GG recognition motif) and the *Nla*III (CATG recognition motif) (New England Biolabs, UK). P1 and P2 adapters with a unique 5 or 7 bp barcode were ligated to the digested DNA and incubated at room temperature for 120 min. The addition of 2.5 vol of PB buffer (Qiagen, Hilden, Germany) terminated the ligation reaction, and the samples were combined in a multiplex pool and purified with a MinElute PCR Purification kit (Qiagen, Hilden, Germany).

Size selection (400 – 600 bp) was performed through electrophoresis on a 1.1 % TAE agarose gel. The gel was run at constant voltages of 45 V for 3 min, 60 V for 3 min, and 90 V for around 70 min. Following gel purification (QIAquick gel extraction kit; Qiagen, Hilden, Germany) library templates of 40 µl each were obtained, and PCR amplification was performed on a thermal cycler T100 (Bio-Rad, Redmond, WA, USA) using the following cycling conditions: 98 ◦C for 30 s, 13 – 14 PCR cycles of 98 ◦C for 10 s, 65 ◦C for 30 s, and 72 ◦C for 30 s, then a final step of 72 ◦C for 5 min. Each PCR amplified library was purified using an equal volume of AMPure XP beads (Beckman Coulter, USA) and eluted with 20 µL of EB buffer (MinElute Gel Purification Kit, Qiagen, Hilden, Germany). Finally, the libraries were sequenced in an Illumina NovaSeq 6000 using three lanes of two SP flow cells (150 base paired-end reads) at the National Genomics Infrastructure centre in Uppsala, Sweden. The sequenced reads were deposited in the National Centre for Biotechnology Information repository as fastq files under project ID PRJNA1044256.

2.4. Quality control of sequenced data - SNP identification and genotyping

Sequencing quality was assessed with FASTQC v0.11.8, while MultiQC (Ewels et al., 2016) v1.8 was used to produce a single quality report for all the samples. FASTP v0.22.0 (Chen et al., 2018) was used to trim adapter-oligomeric sequences and filter out reads with a Phred quality score below 20. Demultiplexing was performed using *process_radtags* from the Stacks software v2.5 (Rochette et al., 2019). The demultiplexed reads were aligned to the *Salvelinus sp.* reference genome assembly [Genbank accession number GCF_002910315.2] using the Bowtie2 program (Langmead and Salzberg, 2012).

Following SNP calling using the *gstacks* module of Stacks, genotypes were extracted using the *populations* module of the same software either in the form of a single SNP per ddRAD-tag or read haplotypes. The applied filtering parameters retained SNPs, having observed heterozygosity below 0.6, minor allele frequency (MAF) above 0.05 and calling rate above 80 % amongst all three populations. The above data processing was conducted using Snakemake v7.9 (Mölder et al., 2021) through a pipeline available at https://github.com/chpalaiokostas/Genetic-diversity-insights-Nordic-Arctic-charr-using-ddRAD under the *Variant_calling* folder.

2.5. Genetic diversity metrics and population structure based on individual SNPs

Prior to genetic analysis, the SNP data were further filtered using the R packages SNPfiltR v1.0.1 (DeRaad, 2022) and vcfR v1.14.0 (Knaus and Grünwald, 2017). More specifically, the provided pipeline of SNPfiltR was used with minor adjustments across all three populations. Following SNP filtering, generic diversity metrics like mean observed heterozygosity (H_o), expected heterozygosity (H_E), and Wright's F statistics, i.e. individual inbreeding coefficients (F_{IS}) and the fixation index (F_{ST}), were estimated using the *populations* module of Stacks.

A principal component analysis (PCA) was conducted using the R package adegenet v2.1.5 (Jombart, 2008) to decipher the studied populations' underlying genetic structure. During PCA, the optimal number of principal components (PCs) was limited to one less than the number of the sampled populations to capture maximum among-population variation and reduce model overfitting (Thia, 2023). Thereafter, a discriminant analysis of the PCs (DAPC) was performed to detect genetic clusters using the same software (Jombart et al., 2010).

DAPC was also used in a cross-validation scheme, where the utility of the SNP dataset for discriminating the charr samples among the three populations was tested. Specifically, the origin of 30 % of the fish from each population was masked and treated as a test set, while the rest of the dataset was used for model training purposes. Predictions regarding the animals' population origin in the test set were made and compared with the actual known data. To minimise stochastic error due to random allocations to test and training sets, the whole process was repeated 100 times (https://github.com/chpalaiokostas/Genetic-diversity-insights-Nordic-Arctic-charr-using-ddRAD).

2.6. Population structure using unsupervised machine learning models

Unsupervised machine learning (ML) models suitable for clustering tasks were fitted, aiming to assess their potential in identifying the underlying population structure. More specifically, K-means models were fitted, where the hyperparameter corresponding to *a priori* number of clusters had values ranging from 1 to 5. For the K-means models, silhouette coefficients were used to assess the most probable number of genetic clusters. The silhouette coefficients also provided confidence metrics regarding each sample's cluster allocation. Moreover, Gaussian and Bayesian Gaussian mixture models were fitted. In both cases, the hyperparameter corresponding to the number of components varied from 1 to 5. Ten replicates of each model were performed to minimise stochastics related to random initialisation, retaining the one with the highest log-likelihood. The above models were also assessed after a dimensionality reduction step with PCA was performed. More specifically, the first two PCs were retained, as with the DAPC analysis. The Python library scikit-learn v1.2.1 (Pedregosa et al., 2011) was used to fit the above ML models. The Python code used is available at https://github.com/chpalaiokostas/Genetic-diversity-insights-Nordic-Arctic-charr-using-ddRAD under the *ML_genetic_clusters* folder.

2.7. Haplotype-derived population structure

A haplotype-based analysis of the studied populations was performed with the RADpainter and fineRADstructure software (Malinsky et al., 2018). A coancestry matrix was estimated from all individuals using RADpainter followed by clustering using the Markov chain Monte Carlo (MCMC) algorithm of fineSTRUCTURE (Lawson et al., 2012). Finally, a heatmap depicting the clustered coancestry matrix was drawn with fineRADstructure.

3. Results

3.1. Sequencing output and SNP detection

Approximately 2.89 billion 150 bp paired-end reads were produced. Following the quality control, ~ 2.2 billion reads were retained. The executed pipeline identified 1.5 million loci with a mean sequence coverage of 44X (SD 25X). Out of those, 11785 loci found in at least 80 % of the genotyped fish were kept for downstream analysis, resulting in the detection of 5929 SNPs. Thereafter, a final round of filtering using the R packages SNPfiltR and vcfR resulted in 1902 SNPs for the analysis based on retaining a single SNP from each sequencing read. Regarding read haplotypes, haplotype-wise filtering was conducted using the Stacks populations module, resulting in 3638 SNPs. Moreover, 14 fish with more than 30 % missing genotypes were removed, with the final dataset consisting of 492 fish. The representation per population involved 166 fish from Finland, 161 from Norway and 165 from Sweden.

3.2. Generic genetic diversity metrics

The Finnish charr had the lowest observed and expected heterozygosity levels among the studied populations. This was observed in both the individual SNP and the haplotypic-based analysis (Table 1). On the other hand, the estimated heterozygosity metrics of the Swedish and Norwegian populations were practically equivalent. F_{IS} coefficients ranged between −0.001–0.06. Nevertheless, those estimates were unreliable as they were accompanied by high standard errors (0.07–0.24). Finally, according to the F_{ST} index, the Finnish population appeared to be the most genetically distant among all the pairwise comparisons. More specifically, the genetic distance between the Finnish-Swedish and Norwegian populations was 0.23 and 0.24 (the same F_{ST} values were obtained with either individual SNPs or haplotypes). On the other hand, the genetic distance between the Norwegian and the Swedish population was 0.14.

3.3. Population structure using PCA and DAPC fitted on the individual SNP dataset

The first two PCs from PCA were sufficient to separate the three populations, accounting for 31 % and 15 % of the explained variance, respectively (Fig. 2). A clear separation was shown between the Finnish and the Swedish–Norwegian populations across the first PC. The latter two populations were separated across the second PC.

DAPC was used to identify genetic clusters and provide insights regarding the underlying genetic structure of the three populations. Clustering was conducted by retaining the first two PCs in the discriminant analysis. Three clusters were generated, consisting of 166, 161 and 165 fish, respectively. A complete correspondence was found between the assigned cluster and the population of origin of each fish (Fig. 3).

Furthermore, the cross-validation scheme conducted using DAPC was 100 % accurate in allocating all samples from the test dataset to their origin population (Fig. 4).

3.4. Population structure using unsupervised machine learning models

The fitted ML models' ability to identify the most probable number of genetic clusters was assessed. Neither the Bayesian Gaussian nor the Gaussian mixture models identified the expected number of genetic clusters when the filtered SNP dataset was used *per se*. In the former case, no conclusive result was obtained, while in the latter, the most probable number of clusters was two (Supplementary Table S1). Conducting dimensionality reduction with PCA before fitting the above ML models resulted in the Bayesian Gaussian and the Gaussian mixture models aligning with the DAPC results and suggesting that the most probable number of genetic clusters was three (Supplementary Table S2).

In the case of the K-means models, a K of two and three were the most probable values when the models were fitted using the original SNP dataset (Fig. 5).

However, when a PCA preceded fitting the K-means models, the existence of three clusters was well supported (Fig. 6).

3.5. Population structure based on read haplotypes

The haplotype-derived coancestry values provided additional insights regarding the underlying genetic structure of the three populations. As expected, individuals within the Finnish population showed higher coancestry levels compared to the other two populations. The above agreed with the previous results suggesting lower genetic diversity in the Finnish population. Moreover, the produced heatmap showed that the coancestry levels between the Norwegian and the

Table 1
Genetic diversity metrics.

Populations	SNPs			Haplotypes		
	$H_o^{(SE)}$	$H_E^{(SE)}$	$F_{IS}^{(SE)}$	$H_o^{(SE)}$	$H_E^{(SE)}$	$F_{IS}^{(SE)}$
Finnish	$0.12^{(0.004)}$	$0.14^{(0.004)}$	$0.06^{(0.24)}$	$0.15^{(0.003)}$	$0.16^{(0.003)}$	$0.02^{(0.14)}$
Norwegian	$0.24^{(0.004)}$	$0.24^{(0.004)}$	$0.03^{(0.24)}$	$0.25^{(0.003)}$	$0.25^{(0.003)}$	$-0.001^{(0.14)}$
Swedish	$0.23^{(0.005)}$	$0.23^{(0.004)}$	$0.02^{(0.12)}$	$0.24^{(0.003)}$	$0.24^{(0.003)}$	$-0.01^{(0.07)}$

H_o refers to observed heterozygosity; H_E refers to the expected heterozygosity; F_{IS} refers to the inbreeding coefficient.

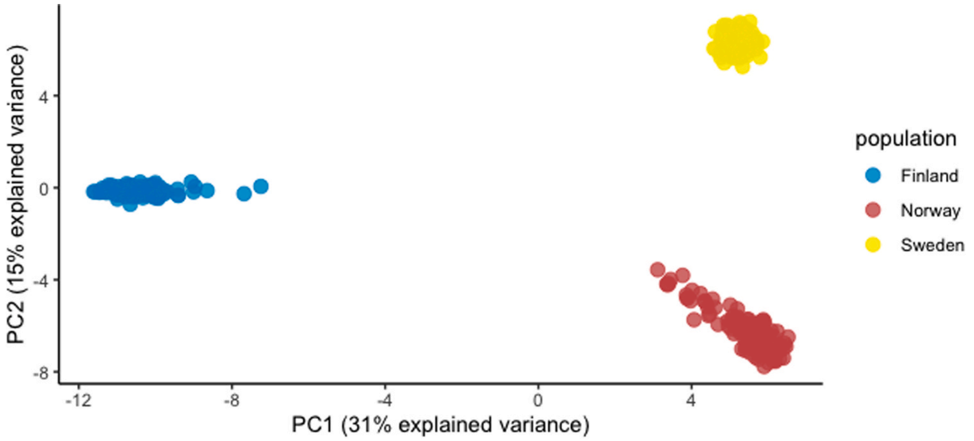

Fig. 2. Principal component analysis (PCA) of three Nordic Arctic charr populations. A unique colour is used to represent individuals belonging to the same population. The Norwegian population was of admixed origin from Hammerfest-Svalbard in Norway and Iceland. The Swedish population originated from Lake Hornavan in Sweden, while the Finnish population was from Lake Kuolimo in Finland.

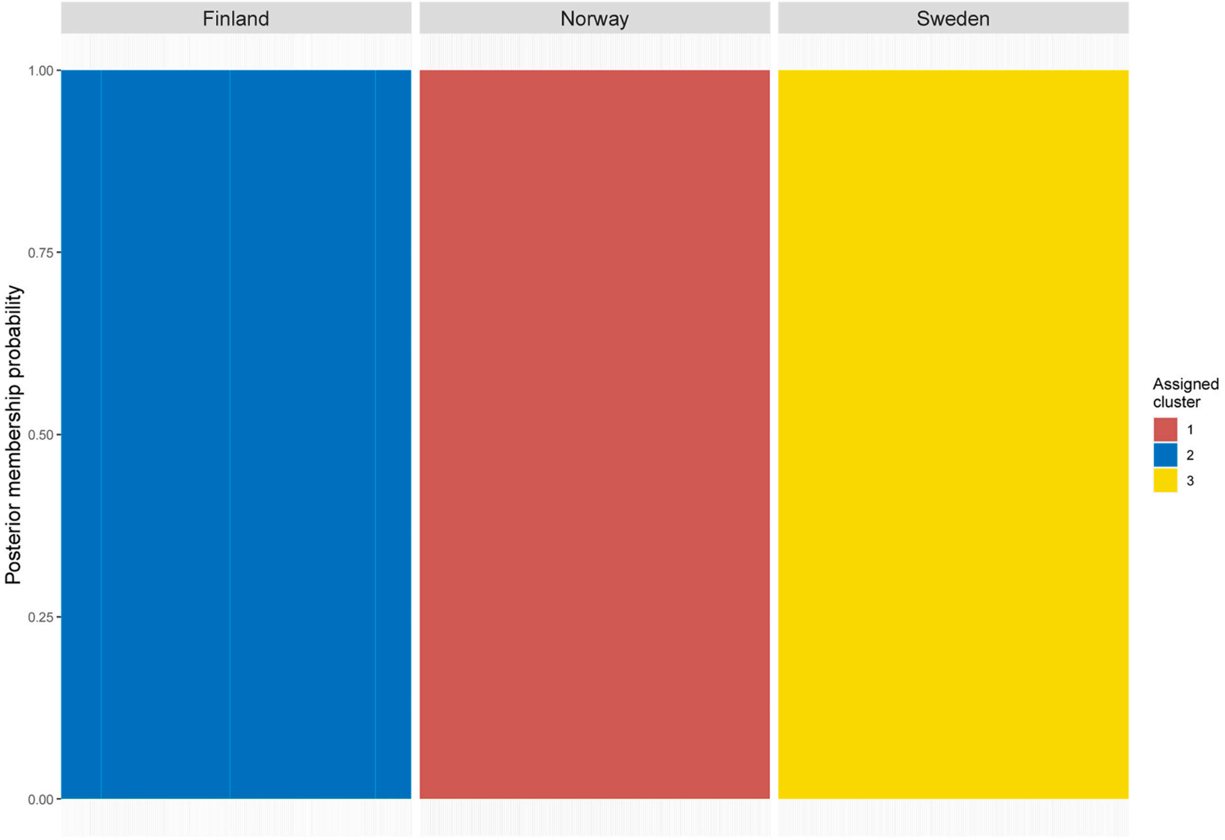

Fig. 3. Ancestry analysis assigning individual Arctic charr to genetic clusters. The top header of the figure represents the true population origin of each fish. Each single vertical bar represents an individual. The same colour indicates that the respective individuals belong to the same genetic cluster. The Norwegian population was of admixed origin from Hammerfest-Svalbard in Norway and Iceland. The Swedish population originated from Lake Hornavan in Sweden, while the Finnish population originated from Lake Kuolimo in Finland.

Swedish populations were higher than the Finnish population. Nonetheless, overall coancestry levels were estimated to be lower between the Finnish and the Norwegian populations than between the Finnish and the Swedish ones (Fig. 7).

4. Discussion

In contrast to farmed salmonids like Atlantic salmon (*Salmo salar*) and rainbow trout (*Oncorhynchus mykiss*), limited information exists about the genetic differentiation between Nordic populations of Arctic charr, especially regarding ones with potential value for aquaculture. The current study attempted to assess the genetic diversity status of Nordic Arctic charr populations from Norway, Sweden and Finland that either significantly impact the domestic industry or, in the case of Finland, represent the country's only active hatchery that supports a conservation program for this species. Whole-genome resequencing has previously suggested comparable levels of genetic diversity between the Norwegian and Swedish populations (Pappas et al., 2023). Nevertheless,

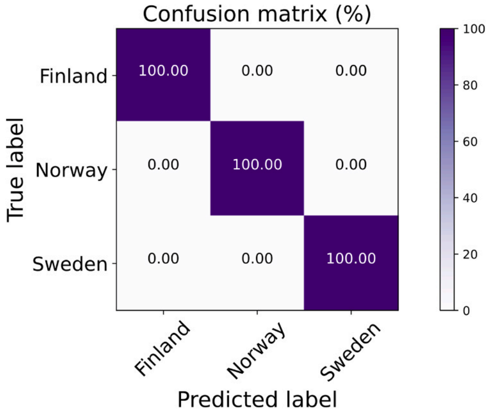

Fig. 4. Confusion matrix for prediction efficiency (% of successful classification) of the SNP dataset using cross-validation. A cross-validation scheme was followed to discriminate the three study populations. The origin of 30 % of randomly selected individuals from each population was masked and used as a test set. The above procedure was repeated 100 times to minimise potential bias due to stochastic sample allocation in the training/test dataset. The diagonal contains the mean percentage of correct population assignments for the cross-validation scheme. Off-diagonals contain the mean percentage of wrong population allocations for each particular case. Reported numbers have been standardised to range between 0 and 100. The Norwegian population was of admixed origin from Hammerfest-Svalbard in Norway and Iceland. The Swedish population originated from Lake Hornavan in Sweden, while the Finnish population was from Lake Kuolimo in Finland.

the above study used a relatively small number of animals from each population (n=24 fish) and might not have fully captured the entire spectrum of genetic diversity. Furthermore, no prior study attempted to use high-throughput sequencing on the Finnish Arctic charr population. As such, genotyping using ddRAD was performed in a relatively high number of animals from all three populations (~ 500 fish).

4.1. *Genetic diversity status of studied Nordic Arctic charr populations*

The Finnish population had significantly lower genetic diversity compared to the Norwegian and Swedish populations, as evidenced by both observed and expected heterozygosity metrics. In particular, the Finnish population's mean H_o and H_E metrics were nearly half those obtained from the Swedish and the Norwegian (0.12 – 0.16 versus 0.23 – 0.25). Overall, the obtained diversity values of the Finnish population were within the lower range reported in the literature (e.g. $H_E \sim 0.1 – 0.4$) in similar studies on farmed fish that used ddRAD (Jansson et al., 2016; Torati et al., 2019; Nyinondi et al., 2020; Palaiokostas et al., 2022).

Notably, in the case of the Swedish and Norwegian populations, the genetic heterozygosity metrics from either individual SNPs or read haplotypes were almost indistinguishable, with the later ones being only ~ 4 % higher. On the other hand, a prominent increase of 15–20 % was observed in the heterozygosity estimate of the Finnish population when read haplotypes were used. As by construction, haplotypes are expected to be more informative and discern relationships at a higher resolution than individual SNPs (Longo et al., 2024), obtaining higher diversity estimates from the latter is not so surprising. Most likely, the higher discrepancy observed between the two analyses in the Finnish population can be attributed to the combination of the sparse genotyping we followed and the low levels of genetic diversity among those fish. The above reasoning is further supported by the heterozygosity estimates in the Norwegian and Swedish populations closely aligned with the previously reported ones where considerably higher genotyping densities were used (Palaiokostas et al., 2022; Pappas et al., 2023). Nevertheless, it is essential to acknowledge that the reported metrics cannot be used to estimate levels of inbreeding, and they are also insufficient to conclude that the Finnish population is under inbreeding depression. A higher genotyping density would be required to estimate inbreeding through runs of homozygosity accurately (Yoshida et al., 2020).

4.2. *Population differentiation and underlying genetic structure using tools commonly applied in population genetics*

The ddRAD datasets sufficed to differentiate the studied charr populations. The F_{ST} index was exceptionally high among all pairwise comparisons, ranging from 0.14 to 0.24. As a rule of thumb, F_{ST} values in the range of 0.15 and above denote the existence of a high genetic differentiation (Wright, 1978). Considering our study populations' geographical and ecological differentiation, this pattern was anticipated. Natural allopatric Arctic charr populations are known to form highly distinctive genetic groups even at a relatively small spatial scale (e.g. Kapralova et al., 2011). For Alaskan charr populations with different historical biogeography (separate drainages), a mean weighted pairwise F_{ST} of 0.21 was estimated from a genomic analysis based on nearly 16k sequenced SNPs (Klobucar et al., 2021).

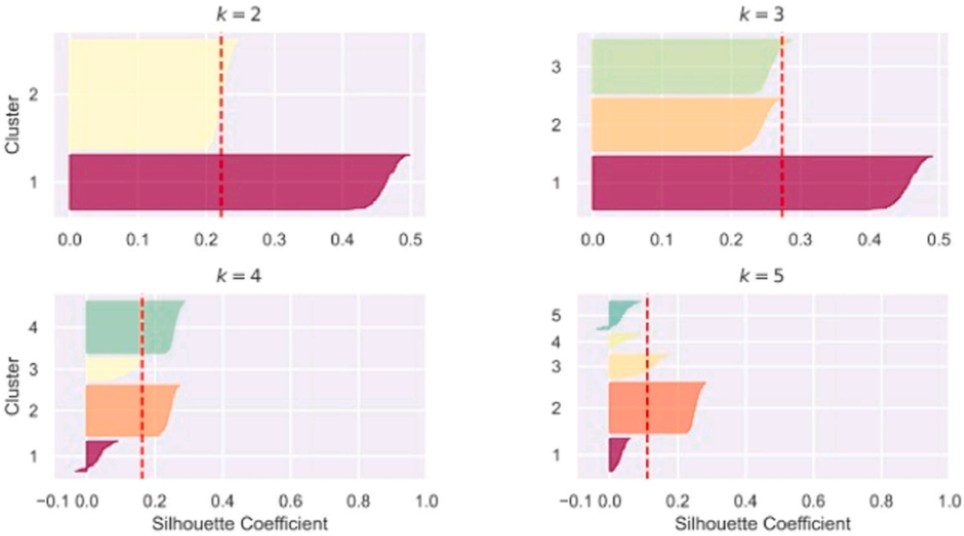

Fig. 5. Silhouette diagram for values of k (depicting the number of genetic clusters) between 2 and 5. Silhouette coefficients were estimated for different K-means models fitted to the SNP dataset. The dashed red lines represent the mean silhouette coefficients for each k value.

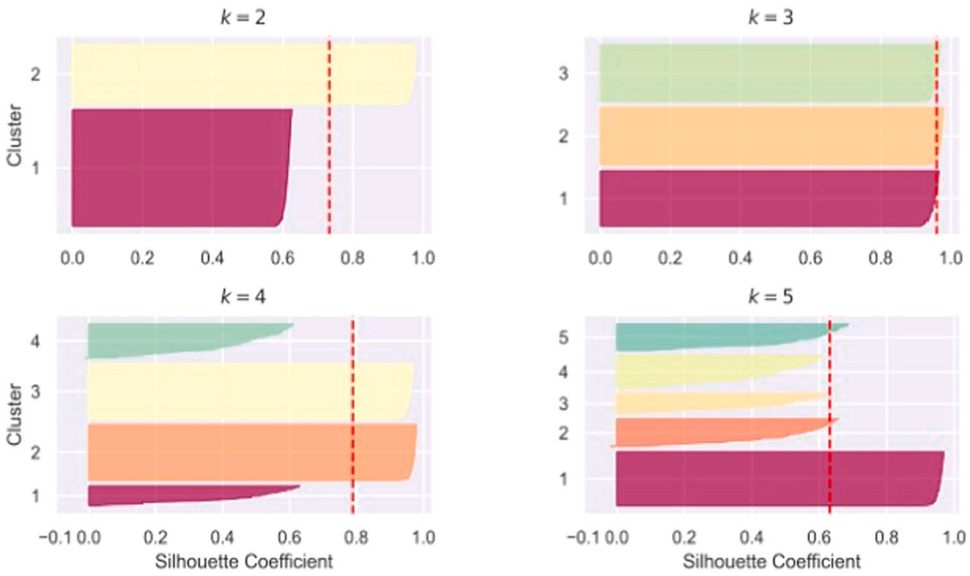

Fig. 6. Silhouette diagram for values of k between 2 and 5 (depicting the number of genetic clusters). Silhouette coefficients were estimated from K-means models following a dimensionality reduction to the original SNP dataset through PCA and by retaining the first two principal components. The dashed red lines represent the mean silhouette coefficients for each k value.

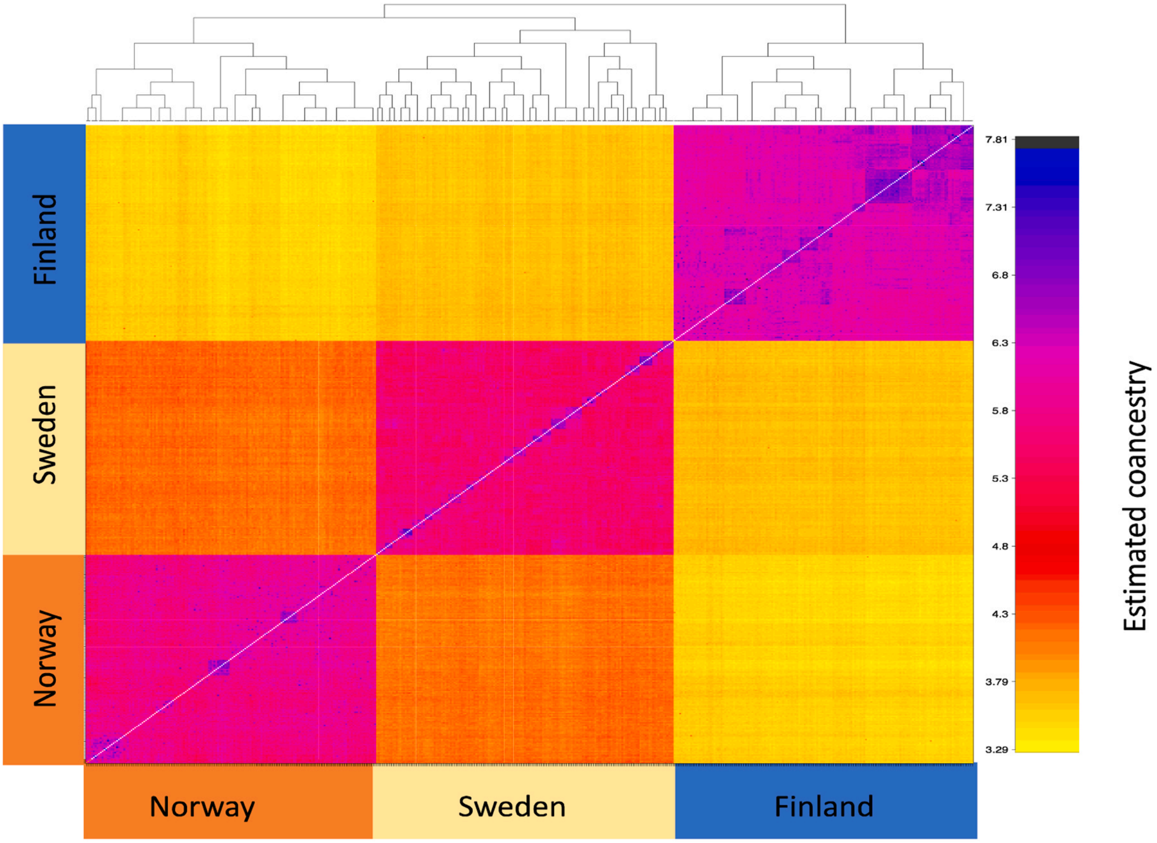

Fig. 7. Heatmap of the haplotype-derived coancestry matrix. The sidebar colour gradient shows the magnitude of shared coancestry between each pair of individuals. The Norwegian population was of admixed origin from Hammerfest-Svalbard in Norway and Iceland. The Swedish population originated from Lake Hornavan in Sweden, while the Finnish population originated from Lake Kuolimo in Finland.

Moreover, our results indicated that the Finnish population was more genetically distant than the other two, probably reflecting their different phylogeographic origins (i.e. Atlantic vs. Siberian lineages; see Brunner et al., 2001). The Finnish population is also critically endangered, with a suspected low effective population size (Primmer et al., 1999), and this is likely an additional reason. Therefore, a potential immediate farming endeavour would be advised only in a hypothetical scenario where the population shows clear evidence of beneficial phenotypic traits for farming, such as growth rate. Furthermore, a routine usage of genomic information would benefit all three populations as it would allow for a more efficient management of genetic diversity.

Complementing the typical single SNP-based genetic diversity analysis with haplotype-based information could offer additional insights, revealing previously undetected population structure. For instance, haplotype-based analysis using ddRAD recently suggested the independent and parallel evolution of different Arctic charr ecotypes (Jacobs et al., 2020). Our study's use of read haplotypes allowed for a finer resolution regarding the genetic relationships among the studied fish. In particular, the estimated coancestry coefficients from the haplotype-based analysis suggested that the Finnish population contained individuals more related to each other than in the other two populations. From a practical point of view, this information warrants careful decisions regarding the formation of mating pairs when new year classes are formed.

4.3. *Assessing the efficiency of unsupervised machine learning models for detecting genetic clusters*

Machine learning has recently been suggested to set a new paradigm in population genetics (Schrider and Kern, 2018). The availability of powerful algorithms suited for high-dimensional datasets fits ideally with the current norm of genomic datasets (López-Cortés et al., 2020). Even though the application of machine learning models in genomic studies is still in its infancy, it has already been part of the machinery of commonly used tools in population genetics. More specifically, the K-means algorithm constitutes the primary step of DAPC when a *de novo* clustering is performed (Miller et al., 2020). Nevertheless, an explicit assessment of machine algorithms in terms of their suitability for identifying genetic clusters is lacking, especially in the case of farmed fish. As such, we tested three machine learning algorithms of increasing complexity. Starting with the computationally lighter K-means and moving to more computationally demanding ones like the Gaussian and Bayesian Gaussian mixture models. Fitting the latter two directly to the SNP dataset failed to deliver reliable results. In comparison, K-means was more efficient even though it was not possible from its output to discern between the existence of two or three genetic clusters.

Detection of genetic clusters using the K-means algorithm requires the *a priori* definition of a suitable number of centroids. Naturally, the above unavoidably results in a certain level of subjectivity regarding the reported results. As a counterbalance, the custom data analysis approach relies on testing several *a priori* centroids and estimates, each using metrics such as inertia. Thereafter, the recorded metric is plotted as a function of the number of clusters K, with the resulting curve often displaying an inflexion point called the "elbow". This point is considered the most probable number of genetic clusters.

Nevertheless, the fact that the recorded metric of this workflow by construction decreases as the *a priori* number of centroids increases results in subjectivity in identifying the curve's inflexion point. Instead, estimating silhouette coefficients can be a more precise approach, though a more computationally demanding one (Géron, 2019). Simultaneously, estimating silhouette coefficients provides a metric of uncertainty regarding the assignment of each sample on each genetic cluster. As mentioned above, a dimensionality reduction step through PCA resulted in a silhouette diagram supporting the existence of three genetic clusters with a confident assignment of each individual. It would be fair to mention that in our study, a substantial genetic differentiation among the three populations appears to exist. Therefore, following the typical approach and attempting to identify the "elbow" of the curve depicting inertia against a predefined number of centroids would probably have been as efficient. Despite that, estimating silhouette coefficients is a valuable and surprisingly underused tool.

Regardless of which of the two analysis paths one chooses, the required *prior decision on* the number of centroids propagates a certain amount of subjectivity in the conducted analysis. Considering the above, algorithms like Bayesian mixture models are particularly appealing as, in this case, only an upper threshold in the number of genetic clusters can be defined (Lu, 2021). As such, only minimal prior knowledge is required. As with the Gaussian mixture model, a dimensionality reduction step through PCA also detected three genetic clusters by the Bayesian mixture model. Bayesian mixture models have been previously used in genetic studies to identify genetic variants involved in human diseases (Moser et al., 2015). However, except for the similar architecture of a Dirichlet mixture model that used genetic and spatial information for clustering (Reich and Bondell, 2011), Bayesian mixture models have not been used in similar studies to the best of our knowledge.

As suggested in our study, both the Bayesian and the Gaussian mixture models tend to suffer from overfitting. Therefore, direct application in genomic datasets will presumably not result in robust genetic cluster identification. Further, before concluding that either of the two models can efficiently detect genetic clusters, both should be tested in more challenging datasets where the studied populations demonstrate less striking genetic distances. Nevertheless, it should also be stressed that in comparison to K-means, where the algorithm is not efficient in detecting clusters of geometric shapes different from a circle (Géron, 2019), both the Bayesian and the Gaussian mixture models can be more flexible, implying that the latter ones could be applicable in more scenarios.

5. Conclusion

Our study identified distinct genetic differences amongst three Arctic charr populations that either have a substantial influence in the domestic industry (Norway and Sweden) or represent the only practical option for farming (Finland) using the genomic profile of each fish. The gene pool of the Finnish population appeared to be the most narrow one. Future genomic information for the management of all three populations should be applied. Unsupervised machine learning models could be worth considering for identifying genetic clusters. A dimensionality reduction step through PCA was beneficial towards increasing the robustness of the derived inference from each of the tested machine learning models.

Ethical statement

The current study was performed in accordance with the Swedish and European Union's legislation described in the Animal Welfare Act 2018:1192 (ethics permit: 5.2.18 – 09859/2019) and licence (No A08, 017) from the Norwegian Food Safety Authority (Mattilsynet) attributed to the Faculty of Bioscience and Aquaculture, Nord University, to perform experiments on animals.

CRediT authorship contribution statement

Matti Janhunen: Writing – review & editing. **Antti Kause:** Writing – review & editing, Funding acquisition, Conceptualization. **Ørjan Hagen:** Writing – review & editing, Funding acquisition, Conceptualization. **José Beirão:** Writing – review & editing, Funding acquisition, Conceptualization. **Fotis Pappas:** Writing – review & editing, Data curation. **Henrik Jeuthe:** Writing – review & editing, Conceptualization. **Christos Palaiokostas:** Writing – review & editing, Writing – original draft, Software, Formal analysis, Data curation, Conceptualization. **Khrystyna Kurta:** Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Christos Palaiokostas reports financial support was provided by Kolarctic. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge support from the Kolarctic CBC under the ARCTAQUA project (grant agreement 4/2018/095/KO4058). The Finnish samples were collected by local voluntary people and the staff of the Luke's Enonkoski aquaculture station. The Statutory Services of Luke are acknowledged for financial support of sample collection. Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.aqrep.2024.102495.

Data Availability

Sequence data have been uploaded to NCBI under project PRJNA1044256. The code of the conducted analysis can be found at https://github.com/chpalaiokostas/Genetic-diversity-insights-Nordic-Arctic-charr-using-ddRAD.

References

Bargelloni, L., Tassiello, O., Babbucci, M., Ferraresso, S., Franch, R., et al., 2021. Data imputation and machine learning improve association analysis and genomic prediction for resistance to fish photobacteriosis in the gilthead sea bream. Aquac. Rep. 20, 100661.

Brunner, P.C., Douglas, M.R., Osinov, A., Wilson, C.C., Bernatchez, L., 2001. Holarctic phylogeography of Arctic charr (*Salvelinus alpinus* L.) inferred from mitochondrial DNA sequences. Evolution 55 (3), 573–586.

Chen, S., Zhou, Y., Chen, Y., Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890.

Davey, J.W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., et al., 2013. Special features of RAD Sequencing data: implications for genotyping. Mol. Ecol. 22, 3151–3164.

DeRaad, D.A., 2022. snpfiltr: An R package for interactive and reproducible SNP filtering. Mol. Ecol. Resour. 22, 2443–2453.

Eriksson, L.-O., Alanärä, A., Nilsson, J., Brännäs, E., 2010. The Arctic charr story: development of subarctic freshwater fish farming in Sweden. Hydrobiologia 650, 265–274.

Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048.

Fisch, K.M., Kozfkay, C.C., Ivy, J.A., Ryder, O.A., Waples, R.S., 2015. Fish hatchery genetic management techniques: integrating theory with implementation. North Am. J. Aquac. 77, 343–357.

Géron, A., 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

Gjedrem, T., Rye, M., 2016. Selection response in fish and shellfish: a review. Rev. Aquac.

Hosoya, S., Kikuchi, K., Nagashima, H., Onodera, J., Sugimoto, K., et al., 2018. Assessment of genetic diversity in Coho salmon (*Oncorhynchus kisutch*) populations with no family records using ddRAD-seq. BMC Res Notes 11, 548.

Jacobs, A., Carruthers, M., Yurchenko, A., Gordeeva, N.V., Alekseyev, S.S., et al., 2020. Parallelism in eco-morphology and gene expression despite variable evolutionary and genomic backgrounds in a Holarctic fish. PLoS Genet. 16, e1008658.

Jansson, E., Taggart, J.B., Wehner, S., Dahle, G., Quintela, M., et al., 2016. Development of SNP and microsatellite markers for goldsinny wrasse (*Ctenolabrus rupestris*) from ddRAD sequencing data. Conserv. Genet Resour. 8, 201–206.

Jombart, T., 2008. Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24, 1403–1405.

Jombart, T., Devillard, S., Balloux, F., 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11, 94.

Kapralova, K., Morrissey, M., Kristjánsson, B., et al., 2011. Evolution of adaptive diversity and genetic connectivity in Arctic charr (*Salvelinus alpinus*) in Iceland. Heredity 106, 472–487.

Kause, A., Nousiainen, A., Koskinen, H., 2022. Improvement in feed efficiency and reduction in nutrient loading from rainbow trout farms: the role of selective breeding. J. Anim. Sci. 100, skac214.

Klemetsen, A., Amundsen, P.-A., Dempson, J.B., Jonsson, B., Jonsson, N., et al., 2003. Atlantic salmon Salmo salar L., brown trout Salmo trutta L. and Arctic charr *Salvelinus alpinus* (L.): a review of aspects of their life histories. Ecol. Freshw. Fish. 12, 1–59.

Klobucar, S.L., Rick, J.A., Mandeville, E.G., Wagner, C.E., Budy, P., 2021. Investigating the morphological and genetic divergence of arctic char (*Salvelinus alpinus*) populations in lakes of arctic Alaska. Ecol. Evol. 11, 3040–3057.

Knaus, B.J., Grünwald, N.J., 2017. vcfr: a package to manipulate and visualize variant call format data in R. Mol. Ecol. Resour. 17, 44–53.

Kottelat, M., Freyhof, J., 2007. Handbook of European freshwater fishes. Publications Kottelat.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Lawson, D.J., Hellenthal, G., Myers, S., Falush, D., 2012. Inference of Population Structure using Dense Haplotype Data. PLOS Genet. 8, e1002453.

Leskinen, P.K., Piironen, J., Primmer, C.R., 2013. Genetic characterization of a newly discovered Finnish Arctic charr (*Salvelinus alpinus*; Salmoniformes) population: stocked or natural? J. Ichthyol. 53, 183–190.

Longo, A., Kurta, K., Vanhala, T., Jeuthe, H., de Koning, D.-J., et al., 2024. Genetic diversity patterns in farmed rainbow trout (*Oncorhynchus mykiss*) populations using genome-wide SNP and haplotype data. Anim. Genet. 55, 87–98.

López-Cortés, X.A., Matamala, F., Maldonado, C., Mora-Poblete, F., Scapim, C.A., 2020. A Deep Learning Approach to Population Structure Inference in Inbred Lines of Maize. Front. Genet. 11.

Lu, J., 2021 A survey on Bayesian inference for Gaussian mixture model.

Malinsky, M., Trucchi, E., Lawson, D.J., Falush, D., 2018. RADpainter and fineRADstructure: population Inference from RADseq Data. Mol. Biol. Evol. 35, 1284–1290.

Miller, J.M., Cullingham, C.I., Peery, R.M., 2020. The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the DAPC method. Heredity 125, 269–280.

Mizuta, D.D., Froehlich, H.E., Wilson, J.R., 2023. The changing role and definitions of aquaculture for environmental purposes. Rev. Aquac. 15, 130–141.

Mölder, F., K.P. Jablonski, B. Letcher, M.B. Hall, C.H. Tomkins-Tinch *et al.*, 2021 Sustainable data analysis with Snakemake.

Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R., et al., 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLOS Genet. 11, e1004969.

Moses, M., Mtolera, M.S.P., Chauka, L.J., Lopes, F.A., de Koning, D.J., et al., 2019. Characterizing the genetic structure of introduced Nile tilapia (*Oreochromis niloticus*) strains in Tanzania using double digest RAD sequencing. Aquac. Int.

Naito, T., Nakayama, K., Takeshima, H., Hashiguchi, Y., Akita, T., et al., 2023. The detailed population genetic structure of the rare endangered latid fish akame Lates japonicus with extremely low genetic diversity revealed from single-nucleotide polymorphisms. Conserv Genet 24, 523–535.

Nyinondi, C.S., Mtolera, M.S.P., Mmochi, A.J., Lopes Pinto, F.A., Houston, R.D., et al., 2020. Assessing the genetic diversity of farmed and wild Rufiji tilapia (*Oreochromis urolepis urolepis*) populations using ddRAD sequencing. Ecol. Evol. 10, 10044–10056.

Palaiokostas, C., 2021. Predicting for disease resistance in aquaculture species using machine learning models. Aquac. Rep. 20, 100660.

Palaiokostas, C., Bekaert, M., Khan, M.G.Q., Taggart, J.B., Gharbi, K., et al., 2015. A novel sex-determining QTL in Nile tilapia (*Oreochromis niloticus*). BMC Genom. 16, 1–10.

Palaiokostas, C., Anjum, A., Jeuthe, H., Kurta, K., Lopes Pinto, F., et al., 2022. A genomic-based vision on the genetic diversity and key performance traits in selectively bred Arctic charr (*Salvelinus alpinus*). Evolut. Appl. 15, 565–577.

Pappas, F., Kurta, K., Vanhala, T., Jeuthe, H., Hagen, Ø., et al., 2023. Whole-genome re-sequencing provides key genomic insights in farmed Arctic charr (*Salvelinus alpinus*) populations of anadromous and landlocked origin from Scandinavia. Evolut. Appl. 16, 797–813.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., et al., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double Digest RADseq: an inexpensive method for De Novo SNP discovery and genotyping in model and non-model species (L. Orlando, Ed.). PLoS One 7, e37135.

Primmer, C.R., Aho, T., Piironen, J., Estoup, A., Cornuet, J.-M., et al., 1999. Microsatellite analysis of hatchery stocks and natural populations of arctic charr, *Salvelinus alpinus*, from the nordic region: implications for conservation. Hereditas 130, 277–289.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Reich, B.J., Bondell, H.D., 2011. A spatial dirichlet process mixture model for clustering population genetics data. Biometrics 67, 381–390.

Rochette, N.C., Rivera-Colón, A.G., Catchen, J.M., 2019. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol. Ecol. 28, 4737–4754.

Sæther, B.-S., Siikavuopio, S.I., Thorarensen, H., Brännäs, E., 2013. Status of arctic charr (*Salvelinus alpinus*) farming in Norway, Sweden and Iceland. J. Ichthyol. 53, 833–839.

Saha, A., Kent, M., Hauser, L., Drinan, D.P., Nielsen, E.E., et al., 2021. Hierarchical genetic structure in an evolving species complex: insights from genome wide ddRAD data in *Sebastes mentella*. PLOS ONE 16, e0251976.

Saura, M., Caballero, A., Santiago, E., Fernández, A., Morales-González, E., et al., 2021. Estimates of recent and historical effective population size in turbot, seabream, seabass and carp selective breeding programmes. Genet Sel. Evol. 53, 85.

Schrider, D.R., Kern, A.D., 2018. Supervised machine learning for population genetics: a new paradigm. Trends Genet. 34, 301–312.

Symonds, J.E., Clarke, S.M., King, N., Walker, S.P., Blanchard, B., et al., 2019. Developing successful breeding programs for new zealand aquaculture: a perspective on progress and future genomic opportunities. Front. Genet. 10.

Takahashi, T., Nagano, A.J., Kawaguchi, L., Onikura, N., Nakajima, J., et al., 2020. A ddRAD-based population genetics and phylogenetics of an endangered freshwater fish from Japan. Conserv Genet 21, 641–652.

Thia, J.A., 2023. Guidelines for standardizing the application of discriminant analysis of principal components to genotype data. Mol. Ecol. Resour. 23, 523–538.

Tiberti, R., Splendiani, A., 2019. Management of a highly unlikely native fish: The case of arctic charr Salvelinus alpinus from the Southern Alps. Aquat. Conserv.: Mar. Freshw. Ecosyst. 29, 312–320.

Torati, L.S., Taggart, J.B., Varela, E.S., Araripe, J., Wehner, S., et al., 2019. Genetic diversity and structure in Arapaima gigas populations from Amazon and Araguaia-Tocantins river basins. BMC Genet. 20.

Torrissen, K.R., Barnung, T.N., 1991. Genetic difference in trypsin-like isozyme pattern between two strains of Arctic charr (*Salvelinus alpinus*). Aquaculture 96, 227–231.

Vandeputte, M., Corraze, G., Doerflinger, J., Enez, F., Clota, F., et al., 2022. Realised genetic gains on growth, survival, feed conversion ratio and quality traits after ten generations of multi-trait selection in rainbow trout *Oncorhynchus mykiss*, fed a standard diet or a "future" fish-free and soy-free diet. Aquac. Rep. 27, 101363.

Wright, S., 1978 Variability within and among natural populations, in *Evolution and the genetics of populations*,.

Yoshida, G.M., Cáceres, P., Marín-Nahuelpi, R., Koop, B.F., Yáñez, J.M., 2020. Estimates of autozygosity through runs of homozygosity in farmed coho salmon. Genes 11, 490.