Estimating genetic diversity in a domesticated population of Norwegian Atlantic salmon (*Salmo salar*) using genomic data

Kjell Roymond Olsen

Master of Science in Developmental Biology and Physiology
Department of Biology

UNIVERSITY OF BERGEN

November 2017

Abstract

New developments of high throughput sequencing technologies have allowed for breeders to assess values related to loss of genetic diversity in domesticated animals on a genomic level for better accuracy than the pedigree used so far. This accuracy is important because loss of diversity has close ties to how much selection strength can be applied, and the sustainability of the population. In this thesis the genetic diversity, rate of inbreeding (ΔF), effective population size (N_e) and population structure in the nucleus of a domesticated population of Atlantic salmon (*Salmo salar*) was analysed. The results of this analysis were used to assess whether or not the same selection strength used today can be upheld for retention of genetic diversity in future generations. A total of 3596 animals were included, spread over 486 families and 5 year classes (YC) sequenced on two single nucleotide polymorphism (SNP) chips. To quality control the data, pruning for genotyping call rate (<90%), missing SNPs per individual (>5%) and Hardy-Weinberg equilibrium (p<1E-06) were performed, reducing the initial 30 318 SNPs to between 25k and 30k for the various data sets. Genetic diversity was measured as observed homozygosity ($\overline{F}_{HOM}$) and ranged from 62.3% (SE=±0.02%) to 64.2% (SE=±0.01%) in the YCs. ΔF per generation was measured as regression of a linearized F_{HOM} on YC to be 0.9% (p<0.05), and corresponded to an N_e of 58.4. The population structure was assessed as F_{ST} using the method developed by Weir and Cockerham (1984) both within- and among YC and ranged from 14.9% to 24.2% and 0.2% to 4.8%, respectively. Multidimensional scaling was furthermore used to assess the data basis and population structure. These results suggest that the population in question has retained genetic diversity, a sufficiently low ΔF and high N_e so that selection strength can be upheld at the same levels as today.

Acknowledgements

I would first and foremost like to thank my main supervisor Borghild Hillestad for help with this thesis whenever I needed it and Sergio Vela for much needed help with the data sets and software. I want to thank the friends who have supported me through my education at University of Bergen, both staff and students alike. A special thanks to my family back home and Natalie Johnsen for never-ending support.

I also want to direct a thank you to the following people for help with various challenges and issues: Geir Dahle, Hooman Khaleghi Moghadam and both SalmoBreed AS and Akvaforsk for sharing data.

Table of Contents

Introduction

Balancing the loss of genetic diversity and genetic gain (ΔG) in a domesticated population is important for both ecological and economical sustainability. To attain breeding progress through ΔG, a relatively small group of parents must be used to produce offspring, with the trade-off on possible higher rates of inbreeding and subsequent loss of genetic diversity (James and McBride, 1958, Woolliams et al., 2015). ΔG depends on the intensity and accuracy of the selection, genetic variation for the trait (σ_g^2) and generation interval (L) (Falconer and Mackay, 1996). Recent development of high throughput sequencing techniques like the single nucleotide polymorphism (SNP) chip has allowed for several of these factors to be monitored closely for each individual animal on a genomic level (Bernatchez et al., 2017). The strength of this technique is that the analysis does not need a pedigree, which has been shown to suffer from a threshold effect depending on depth and subsequently give artificially low estimates of inbreeding (Sonesson et al., 2012, Hillestad, 2015). This development provides an opportunity for novel insights into the genetics of a domesticated population of Atlantic salmon (*Salmo salar*) previously managed with pedigree, for further information to optimize the balance between the rate of inbreeding (ΔF) and ΔG through tweaking the intensity of selection, and the retention of genetic diversity.

The recent evolutionary history of Teleosts includes two whole genome duplication events, leading to a doubling of the genome twice, as opposed to humans only having one duplication (Jaillon et al., 2004, Meyer and Van de Peer, 2005). The consequences of these events is a larger repertory of raw genetic material for selection to work on and thus a high potential for adaptation and innovation towards breeding goals (Glasauer and Neuhauss, 2014). Recent evidence shows, that salmonids like the Atlantic salmon, has experienced a third genome duplication and although redundant genes are turned off, the Atlantic salmon provides a plethora of diversity for breeders to work with (Langham et al., 2004, Berthelot et al., 2014). Retaining sufficient diversity in a population while still exploiting the raw material resulting from several duplications has remained a focus area since we first started domesticating this species in the 1970s (Gjedrem et al., 1991). Severe inbreeding and thus loss of genetic diversity in a population can lead to inbreeding depression expressed by reduced growth, fecundity and

survival in salmonids, where insufficient genetic gain might not provide enough economic incentive to sustain the production (Kincaid, 1983, Su et al., 1996, Sonesson et al., 2003).

The foundation for selection response is the genetic diversity in a population and assessing this quality is the first step towards understanding the state of the population in question. The percent of observed homozygosity out of all SNPs in an individual can be known as the coefficient of inbreeding (F_{HOM}), and when its natural logarithm is regressed on year of birth it can be used to estimate change in diversity over time (Saura et al., 2013, Hillestad, 2015, Ellegren and Galtier, 2016). F_{HOM} does however have the drawback of not distinguishing between alleles identical by descent (IBD) and identical by state (IBS) that other methods used to calculate F like pedigree (F_{PED}) or runs of homozygosity (ROH) do. Inbreeding is however directly proportional to increase in F_{HOM} and has been shown to have high correlations to pedigree estimates (Wright, 1922, Bjelland et al., 2013, Silió et al., 2013). F_{HOM} thus remains a useful parameter for assessing the inbreeding and subsequent loss of diversity in the population

An increased F_{HOM} value represents a decrease in vigour and increase in genetic uniformity of the animal or population in question (Falconer and Mackay, 1996). F_{HOM} does however only represent a snapshot of the current gene dispersion and it is important to stress the aspect of time in populations with changing genotype frequencies and selection schemes. The change in F_{HOM} from one generation to the next, also known as ΔF_{HOM}, is more useful as it shows the dispersive process independently from initial gene frequencies and reflects the cumulating effect of genetic drift (Falconer and Mackay, 1996). The analysis of ΔF_{HOM} is based on the current state of the population due to changes from the previous generation and is not dependent on historic data which may not always be available. The prevalent factor affecting ΔF_{HOM} is genetic drift, whose rate over time can be explained by the effective population size (N_e). N_e was defined by Wright (1931) as 'the number of breeding individuals in an idealised population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration' (Charlesworth, 2009, Crow and Kimura, 1970). Quantifying N_e based on large amounts of genomic data has recently become possible with the development of high density SNP chips (Barbato et al., 2015). In the population in this study, the estimation of ΔF and

subsequent calculation of N_e was previously based on pedigree which can be shallow or with incomplete data creating an upwards skewed effective size compared to the actual effective size (Flury et al., 2010, Hillestad, 2015, Woolliams et al., 2015). N_e and ΔF is furthermore in direct proportion to the loss of diversity, where $\Delta\sigma_g^2 = \Delta F * \sigma_g^2$ (FAO, 2013). N_e thus remains important in optimizing the relationship between ΔG and loss of genetic diversity. The Food and Agriculture Organization of the United Nations (FAO) has recommended an N_e of 50, which corresponds to a ΔF of 1% to achieve a balance in this relationship for domesticated populations of animals (Woolliams et al., 1998).

Assessment of the genetic diversity and its change over time can be complimented with a study of the population structure for further insight into the implications of recent genetic management and data basis. Wright (1950) described the F value relative to various hierarchical structures of a population, calculating the diversity contained in a subpopulation relative to that of the total population (F_{ST}). F_{ST} describes the divergence of a subpopulation in question as a value between zero and one, where zero is no divergence, and one is complete divergence and no shared genetic diversity among the sub- and total population. When dealing with a population of animals an estimation of F_{ST}, coupled with a study of the underlying relationships between individuals can provide information about both divergence and how genetic diversity is contained in the population.

When dealing with substantial amounts of automatically sequenced data like with SNP chips, trust in results of the analysis is highly dependent on the quality control (QC) steps. Several signs like strong deviation from Hardy-Weinberg equilibrium (HWE) or low genotyping rate of both individuals and SNPs provides evidence for errors in either sequencing or in annotation. Depending on what kind of experiment is being performed, the QC steps include different thresholds for pruning SNPs and individuals of low quality. Pruning for SNPs in linkage disequilibrium (LD) is performed in several studies because a correlation between SNPs on different locus can interfere with bootstrap procedures of algorithms (Albrechtsen et al., 2010). For genome wide association studies QC has typically included removing SNPs under a certain threshold of minor allele frequency (MAF) (Laurie et al., 2010). For estimation of F_{HOM}, removing both LD and MAF has been shown to reduce the information available and thus been advised against (Hillestad, 2015). Other studies on genetic diversity and inbreeding like Visser

et al. (2016) has not followed this advice so variation within the field of research is apparent where QC parameters and thresholds has a high dependence on the kind of study and research group performing it.

Aims and hypothesis

Studying the genetic diversity in a population of domesticated Atlantic salmon with the tools described in the introduction will provide better grounds for further developing the breeding programmes they are in, towards either strengthening selection for higher ΔG, or lowering ΔF for retention of genetic diversity. Genetic diversity has previously been studied in this population using pedigree, which has been shown to be upward biased. This study will provide novel insights into the genetic diversity and possibilities for further genetic progress in the population at a genomic level using F_{HOM}, ΔF_{HOM} and N_e. By studying the population stratification using F_{ST} and multidimensional scaling (MDS), further insight into the data basis, population structure and dispersion of genetic diversity will be achieved as well. QC of the data will furthermore ensure that the results are accurate.

In this thesis, I want to use genomic data to test if the nucleus of Atlantic salmon in the SalmoBreed population has a sufficiently high genetic diversity and effective population size for continued breeding with the same selection strength as today.

Materials & methods

Origin of population

The SalmoBreed population originates from two strains of Atlantic salmon collected from rivers in 1975 and 1979, named Bolaks and Jakta. The Bolaks strain originate from Vosso river in Hordaland supplied with Mowi and Sunnlandsøra strains, while Jakta originate from Vosso and Årøyelva in Sogn og Fjordane (Gjedrem et al., 1991, SalmoBreed, 2017). In year 2001, 2002, 2003, and 2004 the strains were incorporated into the breeding program as four distinct subpopulations making up the total population of SalmoBreed (Table 1). The subpopulations are further divided into year classes (YC) and generations, where F0 makes up the base population when the pedigree and family selection was initiated. The population has been

bred with different goals and selection strengths for the various traits from generation to generation and was subjected to phenotype selection for approximately seven generations prior to inclusion in F0. This study includes a subset of individuals in YC 2009, 2010, and 2011 from generation F2, and YC 2013 and 2014 from generation F3.

Table 1. *An overview of the generations (F0-3) in the SalmoBreed population back to the base population (F0) when family selection and pedigree were initiated. The population is divided into four subpopulations, presented here as columns, and their YCs. This study included data from the 2009, 2010, 2011, 2013, and 2014 YC. YC 2009 and 2010 were merged to create 2013, the same with YC 2010 and 2011 to create 2014.*

Generation	Year			
F0	2001	2002	2003	2004
F1	2005	2006	2007	2008
F2	2009	2010	2011	
F3	2012	2013	2014	2015

Family selection has been ongoing for the last four generations where 300 families are created for each YC with 1000 individuals in each family. When the fry reaches a size of 2-3 cm, 3-500 individuals in each family are taken out and PIT tagged for tracing through the selection process (Figure 1). After PIT tagging, the families are divided to either producers or to testing for various traits. In the latter, the fish are tested for disease resistance and other traits for further calculation of estimated breeding values (EBV) based on their performance. At the producers, the fish are pre-selected based on pre-existing family values and phenotype, and tissue samples for genotyping are collected. This way each family can be scored and receive EBVs based on challenge tests and phenotype, as well as genomic EBV (GEBV) based on genotype data. When data is collected, and EBVs and GEBVs estimated, an index is created used for selecting the right individuals to produce the nucleus to achieve a high ΔG and a balanced ΔF. The best females, fertilized by males chosen based on the index, is then used to produce lines of in total 500 000 individuals whose eggs at an age of 3-4 years is sold as OVA for the market. Chosen individuals from each family are also used to create next year's families.

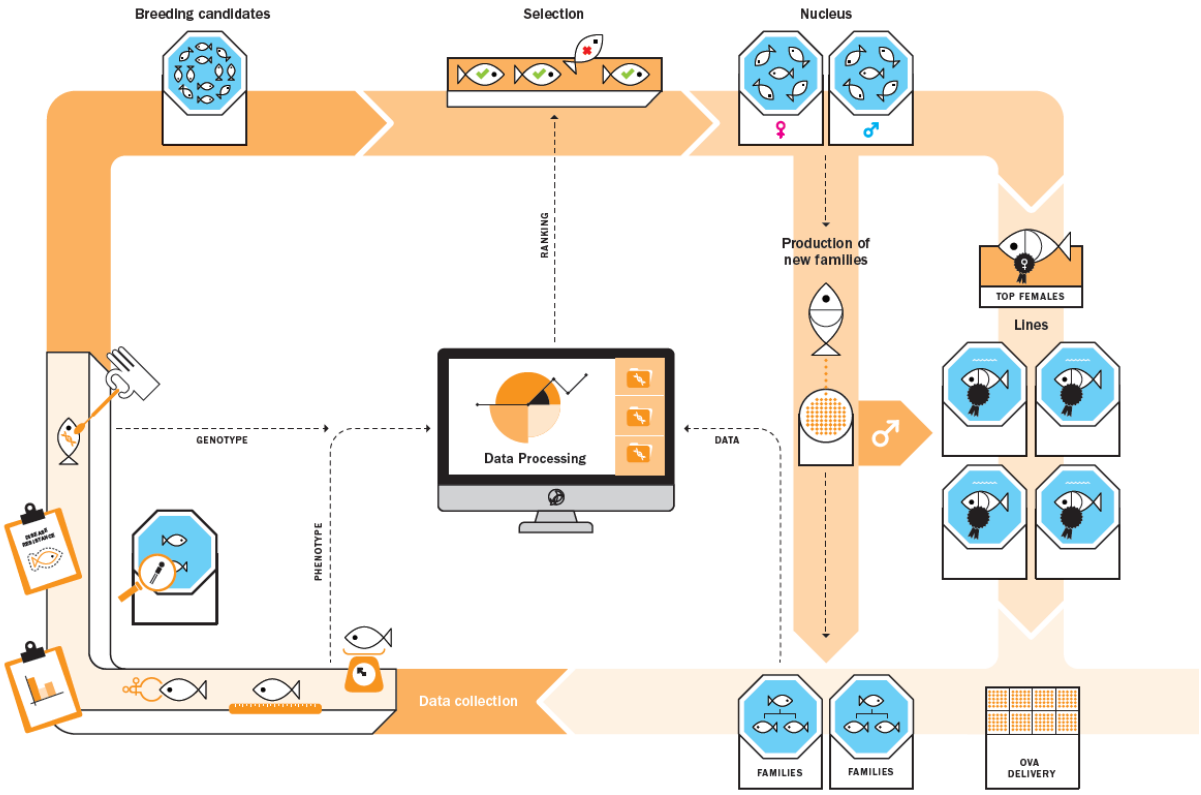

Figure 1. *The family selection scheme of SalmoBreed. 300 Families are created for each YC, where a subset in each family is PIT tagged and split between producers and tests. In the former the fish are pre-selected based on phenotype and pre-existing family values and are subsequently genotyped. In the latter, the families are subjected to tests for several traits. The information gathered from both these endeavours is used to calculate EBVs and GEBVs. Selection is then performed based on the information, and the nucleus is created. The nucleus is the individuals used for further production of lines for OVA to the costumer or new families. Figure provided by SalmoBreed.*

Genotyping quality controls

All samples were genotyped by SalmoBreed with the custom made Affymetrix SNP chips NOFSAL and NOFSAL02 developed by Nofima in collaboration with SalmoBreed, Marine Harvest, and Salmar (B Hillestad 2017, personal communication, 12 September). SNPs in these chips were developed from coding sequences in the transcriptome and has a good coverage of the genome where NOFSAL covered 35 894 SNPs (35K) and NOFSAL02 covered 57 053 SNPs (57K), respectively. The genotype data was subjected to QC to ensure that false-positives and false-negatives were avoided or reduced in number. When dealing with large-scale genomic data sets, appropriate software need to be used to do this accurately, for this purpose PLINKv1.09 was used in this study (Purcell et al., 2007, Chang et al., 2017).

Prior to QC, the data from both SNP chips were managed in R to create file sets for further analysis (R, 2016). A total of six data sets were created: one file including all individuals sorted

by which YC they belong to and five files containing each YC in a separate one sorted by the families it contains (Table 2). The two SNP chips were merged for all data sets, which resulted in 30318 non-overlapping SNPs.

Table 2. *Data sets constructed for analysis in this thesis, their SNP content, individuals, families, and genotyping rate prior to QC. The total population contains individuals divided into YCs, while the family data sets contains each YC with individuals divided into families. The genotyping rate explains the percentage coverage of SNPs for all individuals in the respective data set.*

Data set	SNPs	Individuals	Families	Genotyping rate
Total population	30318	3596	5*	0.9854
2009 Family	30318	45	26	0.9932
2010 Family	30318	195	48	0.9827
2011 Family	30318	147	52	0.9835
2013 Family	30318	1653	188	0.9886
2014 Family	30318	1556	172	0.9833

** YCs and not families*

In QC for all data sets, SNPs with call rate below 90% (missing more than 10% of expected alleles) were removed from further analysis to avoid these error points in the final data. In addition to removing SNPs with low call rate, individuals with too much missing genotype data (>5%) and SNPs not in HWE with a p-value<1E-06 were removed. The p-value in the latter indicate deviation from HWE and were calculated using an exact test (Wigginton et al., 2005).

Following QC, 3593 individuals out of 3596 remained in the total population data set. 3 individuals were removed due to low call rate (<90%), with a resulting high genotyping rate (Table 3). After QC on the family set, YC 2010 lost 3 individuals due to poor call rate (<90%) with a resulting 192 individuals left. The majority of lost SNPs for all QC performed were mainly based on markers so far outside HWE that they would be expected to be mistakes in sequencing. Call rate made a greater difference when QC was run on each separate YC in the family set.

Table 3. *Marker-based QC results and percentage out of initial SNP count in parenthesis in total population- and family data sets. The computations were performed with PLINKv1.09. Genotyping rate explains the percentage coverage of SNPs for all individuals in the respective data set.*

Data set	# SNP with call rate <90% (%)	# HWE p<1E-06 (%)	# SNPs remaining (%)	Genotyping rate post QC
Total pop.	1045 (3.45)	3753 (12.38)	25520 (84.17)	0.9913
2009 Family	204 (0.07)	9 (0.00)	30105 (99.30)	0.9942
2010 Family	1355 (4.47)	143 (0.05)	28820 (95.06)	0.9901
2011 Family	1505 (4.96)	87 (0.03)	28726 (94.75)	0.9908
2013 Family	811 (2.76)	2271 (7.49)	27236 (89.83)	0.9927
2014 Family	1418 (4.68)	2235 (7.37)	26665 (87.95)	0.9906

Genetic diversity

F_{HOM} was calculated using PLINKv1.09's calculation of observed number of homozygotes (O_{HOM}) divided by the amount of non-missing genotypes (N_{NM}) for each individual in the total population data set (Hillestad, 2015):

$$F_{HOM} = \frac{O_{HOM}}{N_{NM}} \quad [1]$$

The mean value for each YC and the total population was calculated from the results of all individuals analysed in the respective YCs to attain $\overline{F}_{HOM}$. In addition, $\overline{F}_{PED}$ values for each YC were provided by Akvaforsk from pedigree available for the population for comparative values.

Rate of inbreeding

ΔF_{HOM} was calculated regressing the natural logarithm of F_{HOM} ($\ln(1-F_{HOM})$) of all individuals on YC to find the linear slope of the regression line from 2009 to 2014, and subsequently multiplying by the average generation interval (3.5 years) following the formula (Hillestad, 2015):

$$\Delta F_{HOM} = \left(1 - e^{\beta}\right)\overline{L} \quad [2]$$

Where β is the slope of the regression line and $\overline{L}$ is the average generation interval. For comparative values based on pedigree, $\overline{F}_{PED}$ was regressed on YC and multiplied by $\overline{L}$ to obtain values of change for this parameter. ΔF provides information on how the genetic diversity of the population is changing from generation to generation. It can subsequently be used to calculate N_e.

Effective population size

N_e was calculated for both ΔF_{HOM} and ΔF_{PED} using the following formula (Falconer and Mackay, 1996):

$$N_e = \frac{1}{2 * \Delta F} \quad [3]$$

Population structure

F_{ST} was estimated within YCs in the family file set and between each YC in the total population data set with PLINKv1.09. The computation method used is developed by Weir and Cockerham (1984) and estimates both raw and weighted global means of F_{ST} for each autosomal diploid variant. Population structure of the SalmoBreed nucleus was further studied using MDS report performed on an inter-sample distance matrix in both the family- and total population data set using PLINKv1.09. MDS is an analysis that detects underlying dimensions in the data which when visualised can show genetic structure in the population. This function in PLINKv1.09 uses a singular value decomposition-based algorithm where two dimensions were chosen. This resulted in the creation of six MDS files, one for each YC based on family and one for the total population data file. These files were then subsequently loaded in to the software Haploview which plotted and visualised the MDS data (Barrett et al., 2005). All clusters were plotted with C1-values ranging from -0.15 to 0.12 and C2-values ranging from -0.145 to 0.16 to capture every individual in all data sets within the same threshold.

The analysis of the genetic diversity, its loss, and population structure will provide information about the state and history of the SalmoBreed population nucleus. Depending on the outcomes in this study, the selection pressure can either be strengthened or weakened for higher or lower ΔG, respectively.

Results

Genetic diversity

The genetic diversity within YCs and in the total population was measured as $\overline{F}_{HOM}$ and is supplemented with $\overline{F}_{PED}$ (Table 4).

Table 4. *The two F values for each YC and the total population used in this thesis. $\overline{F}_{HOM}$ represents average values from all individuals in each YC calculated with equation [1] with corresponding standard error (SE). $\overline{F}_{PED}$ is calculated by Akvaforsk from pedigree. A higher $\overline{F}_{HOM}$ value represents more similar animals, while $\overline{F}_{PED}$ represents the average pedigree inbreeding coefficient where higher values represent more inbreeding. $\overline{F}_{HOM}$ for the YCs are calculated from subsets within the total population data set.*

Data set	$\overline{F}_{HOM}$ (SE)	$\overline{F}_{PED}$
Total population	0.6359 (0.0003)	NA
2009 YC	0.6233 (0.0023)	0.0138
2010 YC	0.6416 (0.0012)	0.0075
2011 YC	0.6370 (0.0016)	0.0097
2013 YC	0.6388 (0.0005)	0.0231
2014 YC	0.6323 (0.0005)	0.0188

Rate of inbreeding

ΔF_{HOM} was calculated using a regression of $\ln(1-F_{HOM})$ on YC then multiplied with $\overline{L}$ following equation [2] (Figure 2). ΔF_{HOM} shows the development in homozygosity from generation to generation, where positive values translate to an increase, and negative values to a decrease. In addition, the $\overline{F}_{PED}$ values were plotted on YC and multiplied by $\overline{L}$ to assess their development over generations (Figure 3).

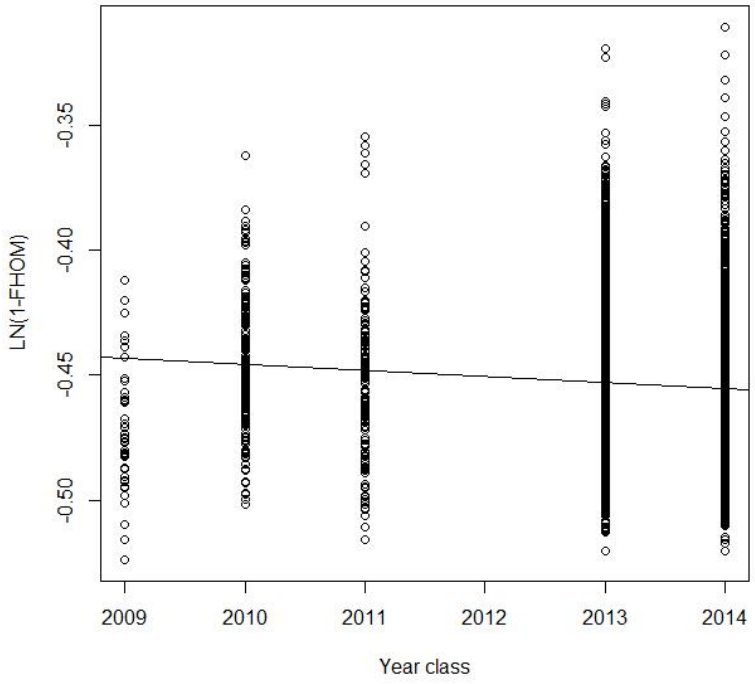

Figure 2. *Regression analysis of individual ln(1-F_{HOM}) values on YC. Following equation [2] the resulting slope (β) of this regression (-0.002449) was used to attain a ΔF_{HOM} of 0.00856 explaining the rate of genetic diversity loss each generation. The R-squared- and significance value of this slope was 0.007898 and 9.51E-08, respectively.*

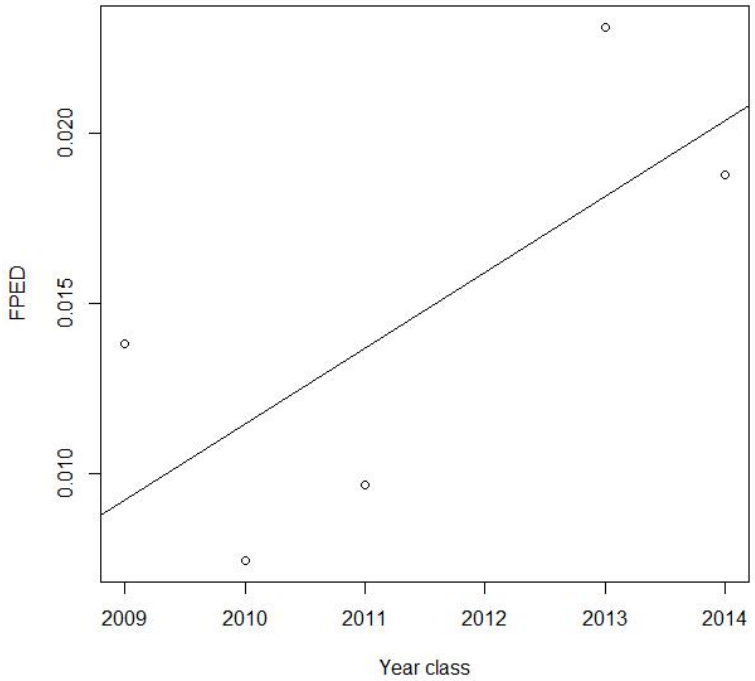

Figure 3. *Regression analysis of $\overline{F}_{PED}$ on YC. The slope (β) of this regression (0.002223) translates to the loss of diversity and increase in homozygosity from year to year and reflects ΔF_{PED} when multiplied by $\overline{L}$ (3.5). The R-squared- and significance value of this slope was 0.5161 and 0.1716, respectively.*

Table 5. *Summary of the two ΔF values calculated with regression analysis. ΔF_{HOM} was calculated with formula [2] and explains the rate of genetic diversity loss each generation. ΔF_{PED} was calculated regressing $\overline{F}_{PED}$ on YC then multiplied with $\overline{L}$ to attain per generation values of inbreeding based on pedigree in the whole population. The regression of ΔF_{HOM} was significant.*

Type of rate	Value
ΔF_{HOM}	0.0086*
$\Delta\overline{F}_{PED}$	0.0078

** Significant (p<0.05)*

Effective population size

N_e calculated with formula [3] was found to be above the recommended minimum values suggested by FAO (Table 6). The values used as ΔF varied based on how they were calculated (Table 5)

Table 6. *Current per-generation N_e calculated with formula [3] and the respective ΔF they were calculated with. Values from ΔF_{HOM} are based on observed homozygosity, and ΔF_{PED} from pedigree. These values represent the size an idealised population would need to have to show the same rate of genetic diversity loss as the real population in question.*

Type of rate	N_e
ΔF_{HOM}	58.4
ΔF_{PED}	64.3

Population structure

The divergence within each YC based on families was calculated as the weighted F_{ST} value and showed a high degree of divergence (Table 7). The weighted F_{ST} for the total population, calculated as genetic diversity in YCs compared to the total population, was low.

Table 7. *Summary of population structure values calculated on both total population- and family data files. F_{ST} is calculated based on the Weir and Cockerham (1984) method where the weighted value is reported in this table. Higher values represent more divergence between families and thus a distributed genetic diversity within the YC. For the total population, the values represent the total divergence among all YCs in the population. Between 2.3k and 5.3k SNPs were invalid for missing in one or more individuals and thus removed in PLINKv1.09's estimation. An ANOVA comparing values between each data set gave highly significant differences (p<1E-10) for all pairs.*

Data set	Weighted F_{ST} (# families)
Total population	0.0287 (5) *
2009 Family	0.2421 (26)
2010 Family	0.2019 (48)
2011 Family	0.2478 (52)
2013 Family	0.2228 (188)
2014 Family	0.1486 (172)

** Based on YC and not families*

F_{ST} was calculated between YCs in the total population data set with the same method as above to see the divergence between both subpopulations and YCs. The values coincide with the structure and management of the population shown in Table 1 and the relationship between the pairwise calculations are significant for all but two pairs (Table 8).

Table 8. *Pairwise F_{ST} values between each YC calculated based on the Weir and Cockerham (1984) method as the weighted mean value. The values in this table represent the degree of divergence, where a higher value translates to a higher percentage divergence between two YCs. The calculations were done on a subset of the total population data set including the respective YCs to be compared. In the analysis of F_{ST} PLINKv1.09 removed between 30 and 148 SNPs that were invalid for missing in one or more individuals. An ANOVA comparing the pairwise values with each other gave highly significant differences (p<0.001) between all but 2010-2014 (0.0397) and 2010-2011 (0.0408), and 2009-2014 (0.0440) and 2009-2010 (0.0481).*

	2009	**2010**	**2011**	**2013**
2010	0.0481			
2011	0.0441	0.0408		
2013	0.0109	0.0153	0.0310	
2014	0.0440	0.0397	0.0021	0.0310

The plotting of MDS calculated on the total population data set showed three major clusters, made up of YC 2009 (top left, (1)), 2010 (bottom left, (2)) and 2011 (far right, (3)) with their respective offspring 2013 (cluster 1 and 2) and 2014 (cluster 2 and 3) (Figure 4, see Appendix A for larger version). This clustering corresponds with the subpopulation admixture (Table 1).

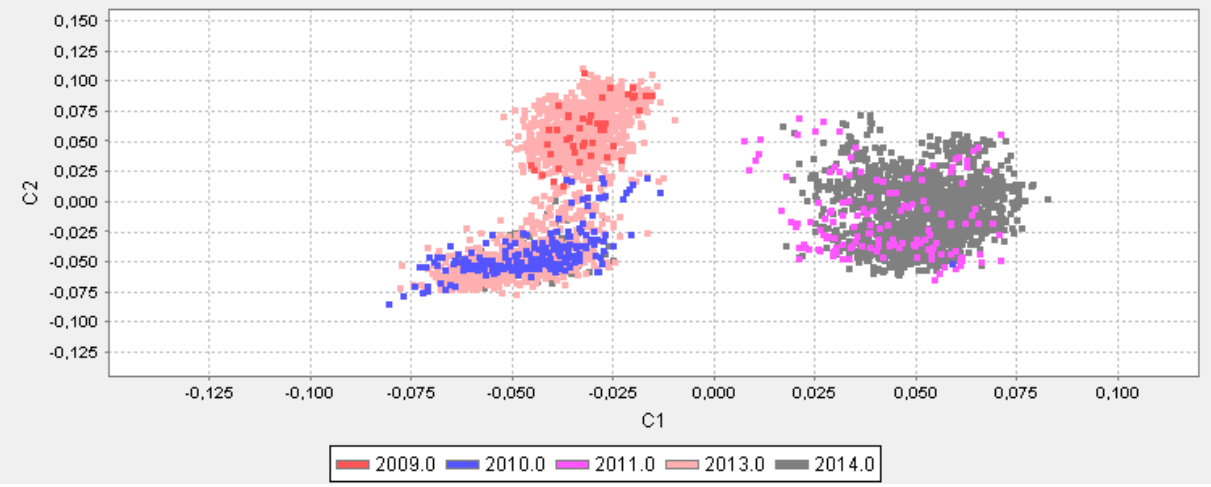

Figure 4. *Population structure in the total population data set through genetic similarity calculated using MDS visualized with HaploView. Each dot represents one individual for a total of 3593 individuals, and the assorted colours represent a different YC (5). The plot shows a distribution according to the subpopulation structure seen in Table 1, where YC 2009 and 2010 is the parents of 2013, and YC 2010 and 2011 of 2014. Three clusters can be seen: Top left is YC 2009 and its related offspring in 2013, bottom left is YC 2010 and its offspring in 2013 and 2014, and far right is 2011 and its offspring 2014. See appendix A for a larger figure.*

Plotting each YC based on family affiliation in the family data set using MDS provides insight into the spread of diversity within each YC. The 2009, 2010, and 2011 YC shows good spread and low clustering (Figure 5, 6, 7). It should be noted that these YCs are originally spread over up to 300 families, resulting in these figures only providing a glimpse of the actual distribution. YC 2013 and 2014 shows clustering corresponding to the structure of the total population, although the latter is prevalently made up of individuals in one cluster (Figure 8 and 9).

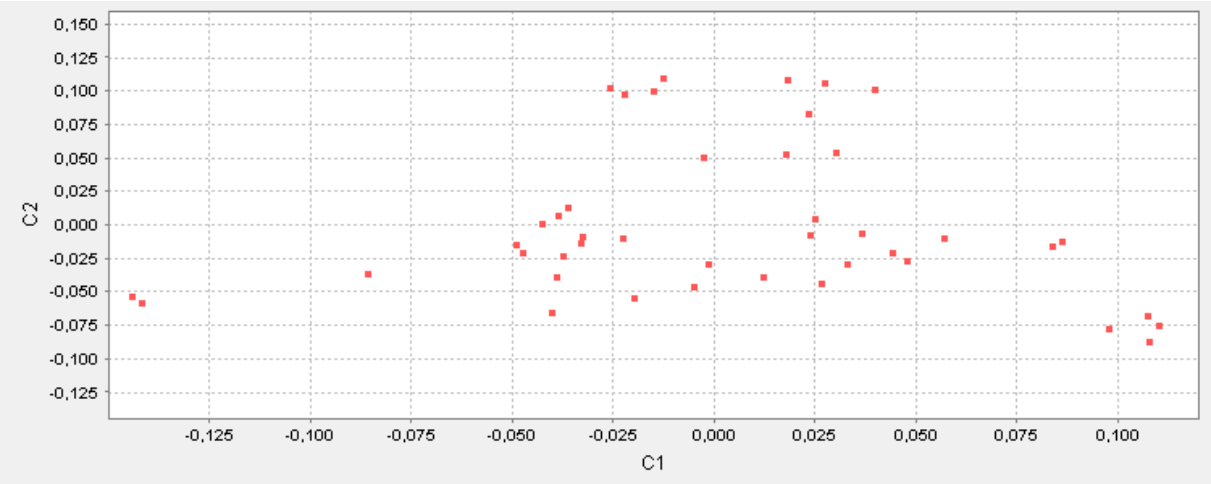

***Figure* 5.** *YC 2009 genetic similarity calculated from the family data set using MDS and visualization with HaploView. Each dot represents one individual for a total of 45 individuals in 26 families. The figure shows spread corresponding to variation between individuals. There is no apparent clustering in accordance with no known outcrossing of the subpopulation this YC belongs to (Table 1).*

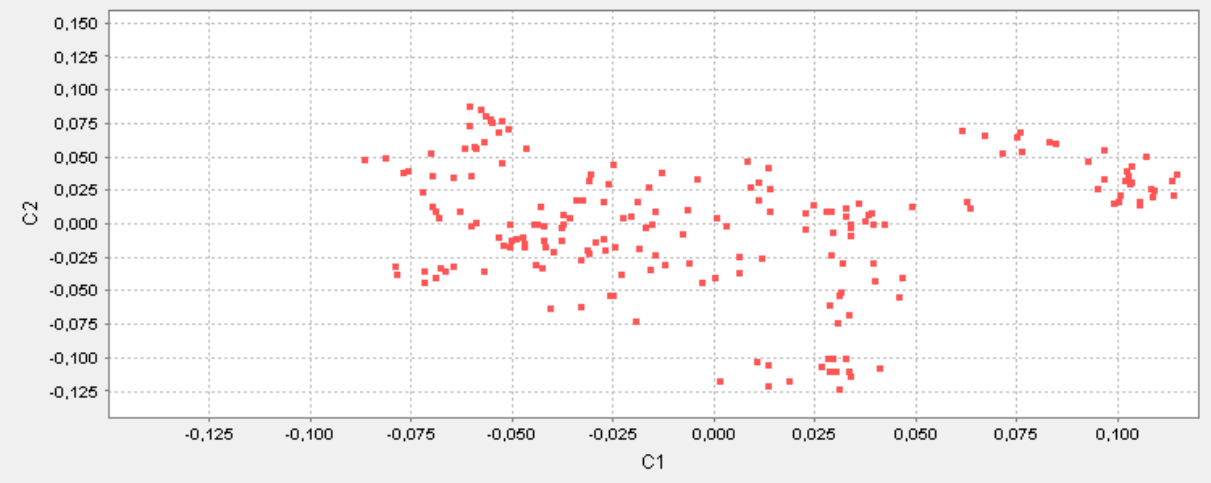

***Figure* 6.** *YC 2010 genetic similarity calculated from the family data set using MDS and visualization with HaploView. Each dot represents one individual for a total of 195 individuals in 48 families. There is no clear clustering in accordance with no known outcrossing of the subpopulation this YC belongs to (Table 1).*

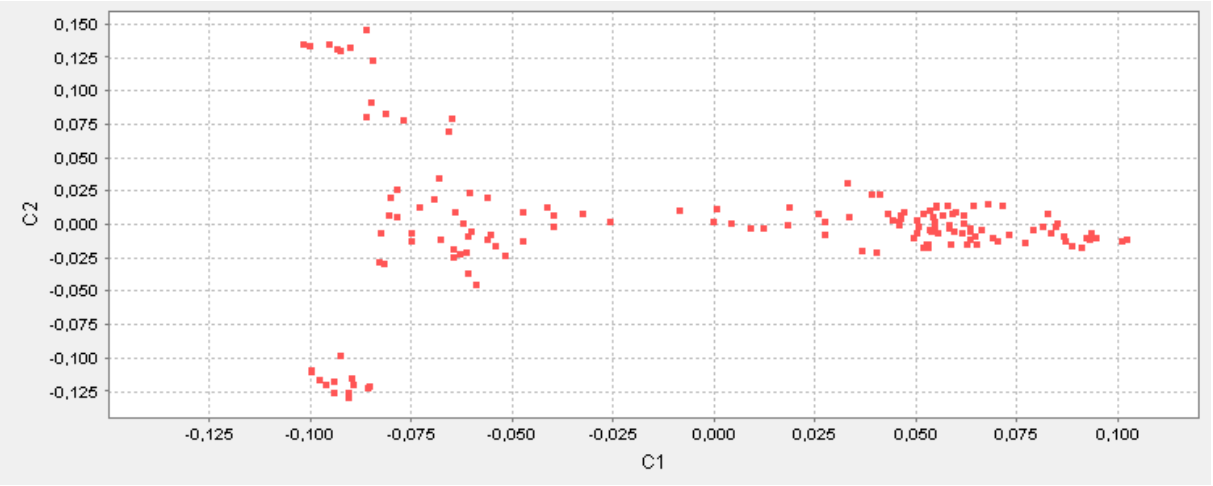

Figure 7. *YC 2011 genetic similarity calculated from the family data set using MDS and visualization with HaploView. Each dot represents one individual for a total of 147 individuals in 52 families. The figure shows clear clustering split at -0.025 on the C1-axis, showing the relationship between the individuals contained in the data set, where half comes from offspring of YC 2007 and the other half from 2008 (Table 1). The left clusters show indications of further clustering.*

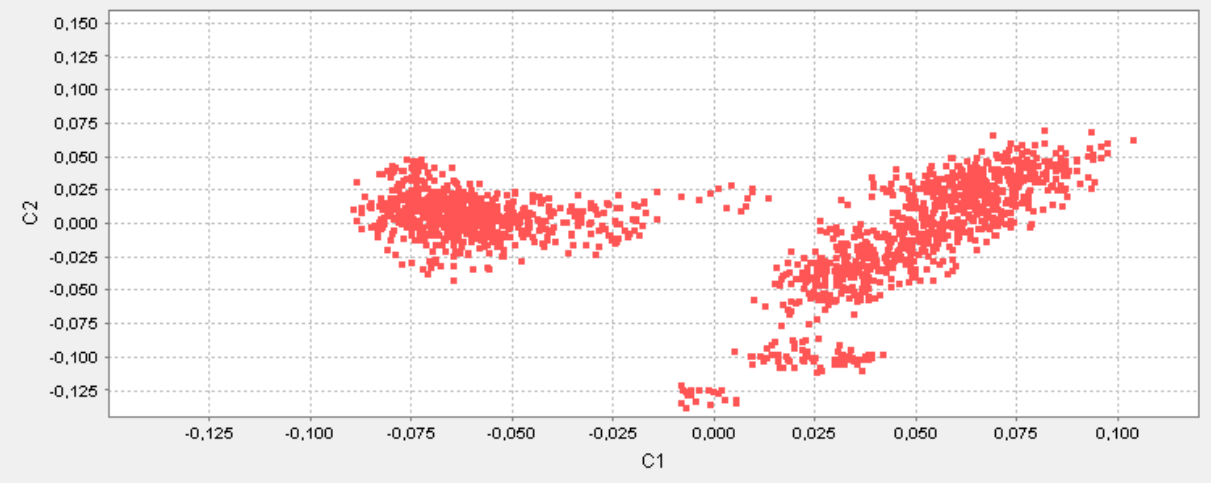

Figure 8. *YC 2013 genetic similarity calculated from the family data set using MDS and visualization with HaploView. Each dot represents one individual for a total of 1653 individuals in 188 families. This figure shows clear clustering between offspring stemming from YC 2009 and 2010.*

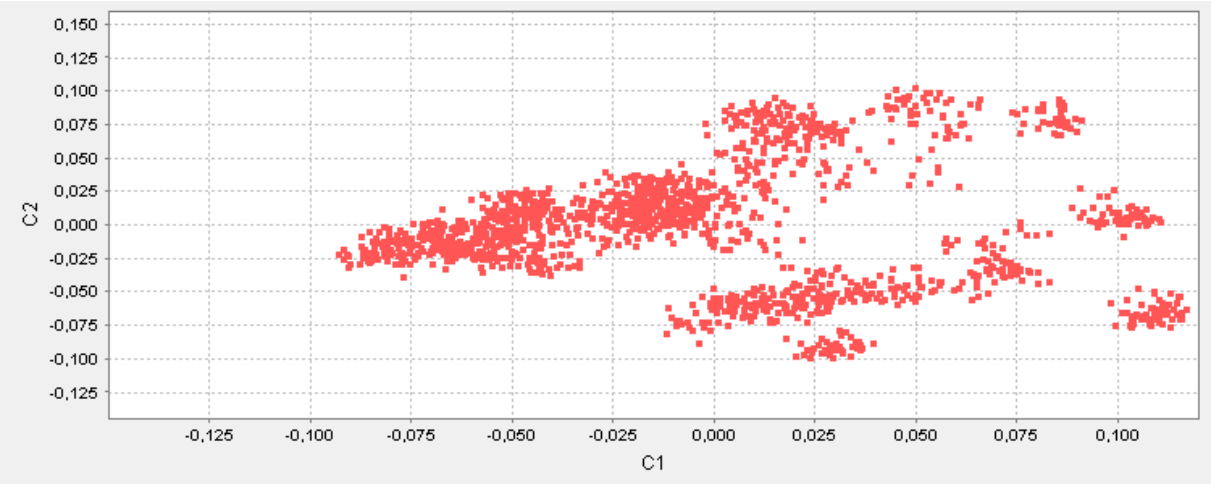

Figure 9. *YC 2014 genetic similarity calculated from the family data set using MDS and visualization with HaploView. Each dot represents one individual for a total of 1556 individuals in 172 families. There are indications of clustering with the top cluster containing substantially more individuals than the bottom one.*

Discussion

From the genomic analysis of the SalmoBreed population, the hypothesis that the nucleus has retained sufficiently high genetic diversity and N_e for sustaining the same selection strength as today, was supported. Genetic diversity was high where $\overline{F}_{HOM}$ ranged from 62.3% (SE=±0.02%) to 64.2% (SE=±0.01%) in the YCs. ΔF, and thus N_e, was above levels recommended by FAO (ΔF<1%, N_e>50). The divergence between families in each YC was high (F_{ST}>15%) showing containment of genetic diversity spread among the families (Table 7). The significant differences in F_{ST} between YCs in different subpopulations (from 0.2% to 4.8%) suggest that they are sorted into genetically distinct groups with variations based on recent migrations between them (Table 8). Although the N_e value was above recommended minimum levels it was close enough (N_e=58.4) so that the selection intensity should be maintained at current levels to maintain the available genetic diversity if the current selection scheme continue in the future (Table 6).

There are many ways to calculate F, and deciding which method to trust is important for achieving a result reflecting the real state of the population. The methods vary from estimations based on pedigree used in the original management of the population in this study, to calculations based on HWE values like Ritland's (1996) Method-of-moments used in PLINKv1.09's estimations of the inbreeding coefficient (F_{HWE}). Other methods are ROH and F_{HOM} used by Bjelland et al. (2013) and Silió et al. (2013). Hillestad (2015) showed that estimations based on pedigree suffered from a threshold effect depending on the amount of generations data was available, and thus gave too low estimates of $\overline{F}_{PED}$. They furthermore compared it to ROH and F_{HOM} to see which of the three were best. Their results suggest that due to too many assumptions, the ROH method is inferior to F_{HOM}. Therefore, F_{HOM} based on observed molecular data was given the most weight in this study.

The data basis for this study was limited to only two generations, limiting trust in methods estimating F with assumptions like HWE. Idealised assumptions like these are set up to create the simplest possible conditions where the dispersive processes like genetic diversity can be studied over time. In domesticated populations, these assumptions rely heavily on how many generations of data is available due to changing selection pressure and breeding goals

(Caballero, 1994). As this study contains only data from two generations, the HWE values typically used to calculate estimated heterozygosity (H_e) and F_{HWE}, are much more likely to be due to selection pressure rather than neutral genetic change happening over time. This effect becomes particularly clear when taking into consideration that the SNP chip contain coding sequences that are prone to selection pressure. Although markers outside HWE are pruned away, this only includes the most extreme cases (p<1E-06). The weighing for different traits like disease resistance, or growth varies from generation to generation, further creating a need for many generations before the data normalizes. Estimations of F_{HWE} and H_e based on HWE values are therefore less likely to reflect the actual dispersion in the population, further strengthening the choice for F_{HOM} based on observed homozygosity in this study.

Using ΔF_{HOM} to calculate N_e shows that the population size is lower than what has previously been estimated with pedigree. Sonesson et al. (2012) and Hillestad (2015) have shown a three- and fourfold increase in ΔF when going from pedigree to genomic data, but this is not the case in this study. There is however a difference in N_e of 5.9 between ΔF_{HOM} and $\Delta\overline{F}_{PED}$ that may be due to the assumption in the pedigree that animals in the base population are completely unrelated. This statement is very unlikely to be true given that some of the YCs are made up from the same rivers (both have ancestors from the Vosso river for instance) and that phenotype selection has been used for many generations prior to F0 (Table 6). Estimations based on observed genomic data like F_{HOM} captures the current state of the population independent of previous dispersion, leading to higher relatedness between the individuals and thus lower N_e compared to pedigree (Woolliams et al., 1998). Estimations in the study by Hillestad (2015) was furthermore conducted on cattle with pedigrees back to the 1800s, creating room for more mistakes. This is also true for Sonesson et al. (2012) who simulated 4000 generations. A further artificial increase in N_e based on $\Delta\overline{F}_{PED}$ compared to ΔF_{HOM} would thus be expected in this study over many generations, owing to assumptions in the estimation of pedigree based F-values (Woolliams et al., 2015).

The R-squared value in the regression analysis for calculating ΔF_{HOM} was low (0.8%) although highly significant (p=9.51E-08, Figure 2). R-squared indicates how much of the variability in the data the model explains. The differences seen between ΔF_{HOM} and $\Delta\overline{F}_{PED}$ (51.6%) are

mainly due to the amount of data included; $\overline{F}_{PED}$ were only average values from a relationship matrix and thus created a better regression fit, but was not significant (P=0.1716, Figure 3). When regressing the values of $\overline{F}_{HOM}$ on each YC the same way as the $\overline{F}_{PED}$ data was regressed, the R-squared increased by one tenfold. This value is still low, but the regression nonetheless shows a clear trend, it is significant, and can be used for strengthening the hypothesis. Furthermore, since data for YC 2012 is missing, results of the regression might be different for the actual population. Although the R-squared values in Figure 2 and 3 are low, non-significant for Figure 3, and data for YC 2012 is missing, the clear trend of the regression lines and highly significant values from ΔF_{HOM}, are deemed sufficient for concluding on the results.

Comparing $\overline{F}_{HOM}$ to other studies on the same species provides further insight into the relative genetic diversity in the SalmoBreed population. Although it is normal to report values in terms of heterozygosity in similar studies, for clarity and consistency all compared values from other studies below have been converted to homozygosity. Studies on the same species has given values of $\overline{F}_{HOM}$ between 79.8% and 84.3% (Vincent et al., 2013), although estimated with less SNPs (5k). Mäkinen et al. (2015) found observed values of $\overline{F}_{HOM}$ in both wild and captive North American populations to be between 61.7% and 65.7%, similar values like the ones observed in this study (Table 4). Other domesticated populations founded in Norway were found by Gutierrez et al. (2016) to have the similar $\overline{F}_{HOM}$ value of 65.4%, although they have not included any of the regular steps for QC of the genotype data. Expected homozygosity in wild populations in Norway has been found by Glover et al. (2013) to be between 62% and 67% with farmed individuals at 63%, further supporting the claim that the population in this study has a good retention of genetic diversity. This study got a low observed homozygosity amount (between 62.3% and 64.2%), an indication of the population having a good or similar amount of genetic diversity present in the population compared to other populations in Norway and the world, both wild and domesticated.

Merging several subpopulations as with YC 2009-2010 creating 2013, and subsequently YC 2010-2011 to create 2014 will have implications for the dispersion of genetic diversity in the population (Table 1). The main purpose of merging is lowering inbreeding, but Wright (1950) argued that substructured populations kept in partial isolation like the one in this study

provides the most favourable condition for transformation as a single species and maintenance of genetic diversity. For the population to maintain genetic diversity it is therefore important to not merge more than necessary as the subdivided population maintains more alleles at each locus than the total population. This can be seen in lower F_{ST} values between the two YCs combined by parents from two different sources (YC 2013 and 2014) which has a significantly lower degree of divergence (F_{ST}=3.1%) than that of completely unrelated YCs 2009 and 2010 (F_{ST}=4.8%) effectively leading to less divergent subpopulations (Table 8). On the other side, merging YCs to maintain σ_g^2 at the same level for two generations can be done to extrapolate data from one year's breeding value to the next. This saves both money and animal lives and is often done every other year in breeding (S Vela 2017, personal communication, 1 November). Wright's thesis provides the background to understand the values of F_{ST} in this study, where combined YCs shows significantly less divergence compared to the unrelated YCs and further warrants caution when merging too many subpopulations.

Comparing the F_{ST} values between YCs in this study with other similar populations shows the relative degree of divergence. Skaala et al. (2004) found F_{ST} between domesticated populations of Atlantic salmon ranging from 2% to 38.8%, indicating that the population in this study is less divergent (Table 8). The aforementioned study is however based on microsatellites with natural higher diversity than SNPs, so higher divergence is expected (Vignal et al., 2002, Morin et al., 2004). Mäkinen et al. (2015) compared the F_{ST} value between wild and captive strains of Atlantic salmon and found values ranging from 0.7% to 3.1%, corresponding to the differentiation between subpopulations and YCs in this study. The latter did however have half the time to deviate from its wild ancestors compared to this study, and a higher F_{ST} for them would be expected over time given the same population management. A comparison of YC divergence in a North American aquaculture strain was done by Liu et al. (2017) who found values comparable to this study. They found values ranging from 0.3% to 6.4% providing further evidence that the population in this study has the same amount of divergence as other domesticated populations. A comparison of F_{ST} values from this study to other similar studies on the same species further supports the hypothesis that the subpopulations and YCs are divergent and retain genetic diversity in subpopulations.

When assessing the F_{ST} values it is important to note that the YCs are calculated based on both the family- and the total population data set. The difference in the two data sets leads to results explaining two different traits of the population. The former shows that the genetic diversity is spread across families and not contained in few individuals with high values of up to 24.2% (Table 7), while the latter shows the divergence of the different subpopulations and YCs (Table 8). The F_{ST} results from the family data set indicates clearly that the differences among families are high, where this trend is still apparent in all but one YC independent of how many families are included in the data set (Table 2). 2014 is the only YC not following this trend, and is calculated to have a significantly lower F_{ST} value than the highest value of YC 2009 (14.9% vs. 24.2%). The two data sets (total population and family) were furthermore subjected to equal QC thresholds to ensure comparable results, but differences between individual YCs is still present. The total population file included all individuals at the time of QC while the family files only included individuals from the respective YC of which they were from. This led to slight deviations in the results of the QC owing to differences in genetic structures between the YCs (Table 3). Although the same thresholds were chosen in the QC for these two data sources, each YC would be different regarding how many missing SNPs there were per individual and per SNP. The results of this is that the family data sets showed a different distribution of QC results than the total data set, and furthermore represent two different traits of the population.

Genetic drift has the strongest effect on small populations, promoting caution when assessing and trusting the results about the YCs only containing a subset of the actual YC in terms of both individuals and families (Wright, 1931). This is particularly the expected case in YC 2009, 2010, and 2011 with 45, 195, and 147 individuals respectively (Table 2). In family selection, each YC consists of 300 families with on average 1000 individuals each, showing the small subset of individuals contained in these YCs in this thesis. Although the effect of drift would still be apparent in the relatively small subsets of 2013 and 2014, with 1653 and 1556 individuals respectively, this effect is expected to be much stronger in YC 2009, 2010, and 2011. The extent of this can be visually studied in the MDS report seen in Figure 4 (See appendix A for larger figure). MDS reflect underlying relationships in the data, where clustering and spread translates to the similarity of the animals represented as individual dots. Figure 4 shows the underlying relationship between the YC as seen in Table 1, where YC 2009

and 2010 are parents of 2013, and 2010 and 2011 are the parents of 2014. Apparent from this figure however, is that the data set in this study include few of the individuals in YC 2014 that are related to 2010, as we would expect to see more "grey" dots belonging to the cluster made up of 2010 and its other offspring 2013. On the other side, we see YC 2009 and 2010 overlapping with offspring 2013 in both parent YCs with good dispersion of offspring between the two parent YCs. The result of these sampling variations is that the diversity contained among families in YC 2014 would be expected to appear lower because many of the individuals that makes up half of the diversity resulting from a crossing of two subpopulations is not present in the data set. The latter can be seen in the significant difference in F_{ST} based on family data resulting from more similarity within the YC for 2014 (14.9%) compared to 2013 (22.3%, Table 7). The F_{ST} between YC 2010 and 2014 (3.97%) compared to between 2011 and 2014 (0.2%) further strengthens this hypothesis as YC 2010 appears to be almost unrelated to 2014 with F_{ST} values similar to other unrelated YCs like 2009 (F_{ST}=4.8%, Table 8). Because most of the individuals in YC 2014 related to 2010 is not present in the data set in this study, the F_{ST} value is expected to be lower for the family data set and the divergence between 2010 and 2014 higher than what is present in the real population. Genetic drift thus plays the largest role in YC 2014 due to data set composition and not number of individuals, where the other YCs show a distribution truer to what is expected in the real population.

While it has already been established why YC 2014 has a lower F_{ST} among families, the variations between the same values in other YCs can be explained by the spread and clustering of the MDS plots in Figure 5-9. YC 2009 shows no indications of clustering on a larger scale which is to be expected as there is no history of migration prior to this point for this subpopulation (Figure 5). This is further supported by the divergence from other YCs in the same generation (F2) being the highest among all YCs (4.4% and 4.8%, Table 8). The spread of individuals and clustering into smaller groups based on family affiliation does however explain why this YC shows a significantly higher divergence among families compared to other YC (F_{ST}=24.2%, Table 7). YC 2010 shows the same trends regarding family clustering as 2009, however it has lower spread of individuals resulting in the significantly lower F_{ST} value of 20.2% among families (Figure 6). Both YC 2011 and 2013 show clear indications of clustering on a larger scale, explained by them being made up of individuals stemming from two different origin subpopulations (Figure 7 and 8). The F_{ST} values among YC 2013 and their relatives in

2009 (3.1%) and 2010 (1.5%), is significantly lower than for unrelated YCs and although data from YC 2007 and 2008 is not available, they are expected to have the same trends in relation to YC 2011 (Table 8). These values can be used as evidence for distribution of offspring as well, where parents from 2010 makes up more of the genetic background in YC 2013 than 2009 explained by the significantly lower F_{ST}. The same trends where one of the parent YCs prevalently makes up the distribution of variation in the next generation is apparent in YC 2014 as explained in the last paragraph (Figure 9). The latter YC shows distribution into two major clusters, however the majority of individuals resides within one of them, further strengthening the conclusions based on F_{ST} values and Figure 4. The MDS plots in Figure 5-9 thus supports the explanation provided for the distribution of F_{ST} values and provides further insight into the data basis in this study.

The F_{ST} values were estimated with the method developed by Weir and Cockerham (1984) where a correlation between locus violates the assumption of the bootstrap procedure leading to possibly skewed or wrong results (Albrechtsen et al., 2010). This study did not prune for neither LD nor MAF, both QC steps advised by PLINKv1.09 to do for whole genome data (Purcell et al., 2007). Pruning away SNPs in LD makes for lower computational load as well as complying to bootstrap procedures as mentioned before. When a quality control check for LD in PLINKv1.09 was run, 16 854 of the SNPs were found to be in LD with each other owing to the dense SNP chips. Out of the original 25 520 SNPs in the cleaned data set only 8057 was in approximate linkage equilibrium. It has been shown by Hillestad (2015) that increased density of SNP chips result in a slightly better fit for natural logarithms of F_{HOM}, so pruning for both LD and MAF and thus removing this density was decided against in this study. In addition, since this is a selected population and the SNP chips contain coding sequences, SNPs are expected to be in LD. In some similar studies, pruning for LD has been performed to ensure an approximate non-random association between loci for the calculation of heterozygosity and individual inbreeding (Visser et al., 2016). However, in other studies, measuring the same parameters this pruning step has been left out (Bjelland et al., 2013, Saura et al., 2013, Hillestad, 2015). In this study, I chose not to prune for neither LD nor MAF, in order not to lose the data basis for calculations further down the pipeline.

The SNP chip is designed to be biallelic, because in the great majority of cases SNPs occur in two alleles, and thus represents an either-or case in regards to diversity (Nowak et al., 2009). The call rate, or genotyping efficiency, is a measure of the fraction of missing calls per SNP per sample over the total number of SNPs in the dataset, and provides information about how many sites that do not show either of the two alleles (Laurie et al., 2010). The reason for a site not showing either of the two can be due to multiple factors, including base-calling and alignment errors (Nielsen et al., 2011). Errors like this is important to remove as they can lead to an upward bias in homozygosity and inbreeding estimates (Wang et al., 2012). The QC threshold for call rate in this study was thus chosen to reflect a balance in loss of samples dropped due to poor genotyping efficiency and accuracy in the results (Turner, 2011).

The HWE threshold of p<1E-06 was chosen because markers that deviate strongly from HWE are suspected to do so due to technical problems, and not evolutionary forces (Wiggans et al., 2009, Edriss et al., 2013). This threshold is low enough so that only the most outlier SNPs are excluded, and that most SNPs not in HWE due to selection, mutation or migration are still included. The QC for HWE in the total population data set did however remove 12.4% of the SNPs leading to only 84.17% of the SNPs remaining (Table 3). The reason for this amount is that a bred population like the one under study would be assumed to be outside of HWE on many alleles in LD with markers and traits selected for. Although many SNPs were pruned away in the HWE, concluded to be due to both selection and genotyping errors, the data set still consisted of sufficiently dense markers for this study as estimations of F_{HOM} has been shown to not be as sensitive to the amount of markers as other similar estimations (Hillestad, 2015).

Initially the YCs in this study was sequenced on two different SNP chips, one of 35k and one of 57k markers. When a new SNP chip is made, the SNPs from the old ones are updated from current literature, and incorporated. The SNP chips in this study were merged, and since markers in the old chip might not always be as accurate as in the new one, some SNPs that are regarded as two SNPs because they are far away from each other in the genome might be the same SNP leading to a duplicate. To correct for this possibility both chips were merged and overlapping SNPs were removed. Because overlapping SNPs between the 35k and 57k

chips could lead to duplicate locus and thus false positive or negative results they were merged with a resulting the 30k SNPs available.

Today there exists a plethora of software for genetic analysis, where choosing the right one depends on type of data, the amount of data and what type of analysis is to be done. Making the right choice in software is crucial as it will not only affect your results, but also how much time it takes to get them. Estimations of F_{HOM} are straightforward, as they only use observed values. When it comes to F_{ST} however, several options and software are available. The method of moments used to calculate F_{ST} in this study developed by Weir and Cockerham (1984) and implemented in PLINKv1.09 has been widely used, and has a high robustness (Holsinger and Weir, 2009). The other widely used method is Bayesian estimates, which is more computationally demanding and requires that sample sizes are equal (Samanta et al., 2009). Although the two have not been extensively compared, the experiences of Holsinger and Weir (2009) suggests that the differences are small depending on number of individuals and populations in the data set. The decision to use the former in this study was made mainly due to the functionalities of PLINKv1.09 to cover the other analysis to be done in this study, and speed of the F_{ST} analysis. There is furthermore a lack of software that converts large SNP data sets to other useful formats, where the commonly used PGDSpider did not work due to too many SNPs and individuals. The latter challenge of a large data set and conversion was apparent in several other software packages like Structure and NeEstimator further supporting PLINKv1.09 as the software of choice in this thesis. The data format used in PLINKv1.09 is furthermore used in other useful software designed to analyse SNP chips like SNeP, of which the historic N_e of many generations (>13) based on LD was calculated, but not included in the results (Barbato et al., 2015). The choice of PLINKv1.09 in this study was thus on a basis of speed, available computation methods, file type used in other software, and that it was designed for working with large SNP data sets.

Conclusion and further work

In this thesis, I supported the hypothesis that the nucleus of Atlantic salmon in the SalmoBreed population has a sufficiently high genetic diversity and N_e for continued breeding with the same selection strength as today. The divergence among families was high and the F_{ST} values

between YCs supported by MDS plots showed clear subpopulation stratification and containment of genetic diversity. The levels of homozygosity are furthermore good compared to other similar populations, both wild and domesticated. These results illustrate that the population has been bred sustainably for retention of genetic diversity and has a balanced level of inbreeding. The same amount of selection pressure can be upheld, however an elevation is not recommended due to inbreeding rates close to the threshold set by FAO.

Further recommended work is to use genomic values sorted by chromosome to assess ΔF for each chromosome to study if selection for certain traits contributes to more genomic values of inbreeding than others. This information can be used to tweak selection strength for traits related to different locations on the genome. SalmoBreed is furthermore this year implementing optimum contribution selection, a strategy that lessens the impact of increased selection strength on inbreeding (Henryon et al., 2015, Woolliams et al., 2015). The latter implementation will allow for increased selection strength even though the N_e calculated in this study was close to the recommended threshold.

References

ALBRECHTSEN, A., NIELSEN, F. C. & NIELSEN, R. 2010. Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. *Mol Biol Evol,* 27**,** 2534-47.

BARBATO, M., OROZCO-TERWENGEL, P., TAPIO, M. & BRUFORD, M. W. 2015. SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Front Genet,* 6.

BARRETT, J. C., FRY, B., MALLER, J. & DALY, M. J. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics,* 21**,** 263-265.

BERNATCHEZ, L., WELLENREUTHER, M., ARANEDA, C., ASHTON, D. T., BARTH, J. M. I., BEACHAM, T. D., MAES, G. E., MARTINSOHN, J. T., MILLER, K. M., NAISH, K. A., OVENDEN, J. R., PRIMMER, C. R., YOUNG SUK, H., THERKILDSEN, N. O. & WITHLER, R. E. 2017. Harnessing the Power of Genomics to Secure the Future of Seafood. *Trends in Ecology & Evolution,* 32**,** 665-680.

BERTHELOT, C., BRUNET, F., CHALOPIN, D., JUANCHICH, A., BERNARD, M., NOËL, B., BENTO, P., DA SILVA, C., LABADIE, K., ALBERTI, A., AURY, J.-M., LOUIS, A., DEHAIS, P., BARDOU, P., MONTFORT, J., KLOPP, C., CABAU, C., GASPIN, C., THORGAARD, G. H., BOUSSAHA, M., QUILLET, E., GUYOMARD, R., GALIANA, D., BOBE, J., VOLFF, J.-N., GENÊT, C., WINCKER, P., JAILLON, O., CROLLIUS, H. R. & GUIGUEN, Y. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. 5**,** 3657.

BJELLAND, D. W., WEIGEL, K. A., VUKASINOVIC, N. & NKRUMAH, J. D. 2013. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. *J Dairy Sci,* 96**,** 4697-706.

CABALLERO, A. 1994. Developments in the prediction of effective population size. *Heredity,* 73**,** 657-679.

CHANG, C., CHOW, C., VATTIKUTI, S., TELLIER, L. & LEE, J. 2017. *PLINK 1.90 beta* [Online]. Available: https://www.cog-genomics.org/plink/1.9/ [Accessed 13.11 2017].

CHARLESWORTH, B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet,* 10**,** 195-205.

CROW, J. F. & KIMURA, M. 1970. An introduction to population genetics theory. *An introduction to population genetics theory.*

EDRISS, V., GULDBRANDTSEN, B., LUND, M. S. & SU, G. 2013. Effect of marker-data editing on the accuracy of genomic prediction. *Journal of Animal Breeding and Genetics,* 130**,** 128-135.

ELLEGREN, H. & GALTIER, N. 2016. Determinants of genetic diversity. *Nature Reviews Genetics,* 17**,** 422-433.

FALCONER, D. & MACKAY, T. 1996. *Introduction to quantitative genetics,* London, Longman Group Ltd.

FAO 2013. In vivo conservation of animal genetic resources. *FAO Animal Production and Health Guidelines.* 14 ed. Rome.

FLURY, C., TAPIO, M., SONSTEGARD, T., DROGEMULLER, C., LEEB, T., SIMIANER, H., HANOTTE, O. & RIEDER, S. 2010. Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium. *J Anim Breed Genet,* 127**,** 339-47.

GJEDREM, T., GJØEN, H. M. & GJERDE, B. 1991. Genetic origin of Norwegian farmed Atlantic salmon. *Aquaculture,* 98**,** 41-50.

GLASAUER, S. M. K. & NEUHAUSS, S. C. F. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular Genetics and Genomics,* 289**,** 1045-1060.

GLOVER, K. A., PERTOLDI, C., BESNIER, F., WENNEVIK, V., KENT, M. & SKAALA, Ø. 2013. Atlantic salmon populations invaded by farmed escapees: quantifying genetic introgression with a Bayesian approach and SNPs. *BMC Genetics,* 14**,** 74.

GUTIERREZ, A. P., YÁÑEZ, J. M. & DAVIDSON, W. S. 2016. Evidence of recent signatures of selection during domestication in an Atlantic salmon population. *Marine Genomics,* 26**,** 41-50.

HENRYON, M., OSTERSEN, T., ASK, B., SØRENSEN, A. C. & BERG, P. 2015. Most of the long-term genetic gain from optimum-contribution selection can be realised with restrictions imposed during optimisation. *Genetics Selection Evolution,* 47**,** 21.

HILLESTAD, B. 2015. *Inbreeding determined by the amount of homozygous regions in the genome.* Phd, Norwegian University of Life Sciences.

HOLSINGER, K. E. & WEIR, B. S. 2009. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics,* 10**,** 639.

JAILLON, O., AURY, J.-M., BRUNET, F. & PETIT, J.-L. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature,* 431**,** 946.

JAMES, J. & MCBRIDE, G. 1958. The spread of genes by natural and artificial selection in closed poultry flock. *Journal of Genetics,* 56**,** 55-62.

KINCAID, H. L. 1983. Inbreeding in fish populations used for aquaculture. *Aquaculture,* 33**,** 215-227.

LANGHAM, R. J., WALSH, J., DUNN, M., KO, C., GOFF, S. A. & FREELING, M. 2004. Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics,* 166**,** 935-45.

LAURIE, C. C., DOHENY, K. F., MIREL, D. B., PUGH, E. W., BIERUT, L. J., BHANGALE, T., BOEHM, F., CAPORASO, N. E., CORNELIS, M. C., EDENBERG, H. J., GABRIEL, S. B., HARRIS, E. L., HU, F. B., JACOBS, K., KRAFT, P., LANDI, M. T., LUMLEY, T., MANOLIO, T. A., MCHUGH, C., PAINTER, I., PASCHALL, J., RICE, J. P., RICE, K. M., ZHENG, X. & WEIR, B. S. 2010. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol,* 34**,** 591-602.

LIU, L., ANG, K. P., ELLIOTT, J. A. K., KENT, M. P., LIEN, S., MACDONALD, D. & BOULDING, E. G. 2017. A genome scan for selection signatures comparing farmed Atlantic salmon with two wild populations: Testing colocalization among outlier markers, candidate genes, and quantitative trait loci for production traits. *Evolutionary Applications,* 10**,** 276-296.

MÄKINEN, H., VASEMÄGI, A., MCGINNITY, P., CROSS, T. F. & PRIMMER, C. R. 2015. Population genomic analyses of early-phase Atlantic Salmon (Salmo salar) domestication/captive breeding. *Evolutionary Applications,* 8**,** 93-107.

MEYER, A. & VAN DE PEER, Y. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays,* 27**,** 937-945.

MORIN, P. A., LUIKART, G., WAYNE, R. K. & THE, S. N. P. W. G. 2004. SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution,* 19**,** 208-216.

NIELSEN, R., PAUL, J. S., ALBRECHTSEN, A. & SONG, Y. S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet,* 12**,** 443-51.

NOWAK, D., HOFMANN, W. K. & KOEFFLER, H. P. 2009. Genome-wide Mapping of Copy Number Variations Using SNP Arrays. *Transfus Med Hemother,* 36**,** 246-51.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M A R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P I W., DALY, M J. & SHAM, P C. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet,* 81**,** 559-75.

R, C. T. 2016. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*.

RITLAND, K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research,* 67**,** 175-185.

SALMOBREED. 2017. *Our Fish* [Online]. Available: http://salmobreed.no/our-fish/ [Accessed 22.08 2017].

SAMANTA, S., LI, Y. J. & WEIR, B. S. 2009. Drawing Inferences about the Coancestry Coefficient. *Theor Popul Biol,* 75**,** 312-9.

SAURA, M., FERNÁNDEZ, A., RODRÍGUEZ, M. C., TORO, M. A., BARRAGÁN, C., FERNÁNDEZ, A. I. & VILLANUEVA, B. 2013. Genome-wide estimates of coancestry and inbreeding in a closed herd of ancient Iberian pigs. *PLoS One,* 8**,** e78314.

SILIÓ, L., RODRÍGUEZ, M. C., FERNÁNDEZ, A., BARRAGÁN, C., BENÍTEZ, R., ÓVILO, C. & FERNÁNDEZ, A. I. 2013. Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. *Journal of Animal Breeding and Genetics,* 130**,** 349-360.

SKAALA, Ø., HØYHEIM, B., GLOVER, K. & DAHLE, G. 2004. Microsatellite analysis in domesticated and wild Atlantic salmon (Salmo salar L.): allelic diversity and identification of individuals. *Aquaculture,* 240**,** 131-143.

SONESSON, A. K., JANSS, L. L. G. & MEUWISSEN, T. H. E. 2003. Selection against genetic defects in conservation schemes while controlling inbreeding. *Genetics Selection Evolution,* 35**,** 353.

SONESSON, A. K., WOOLLIAMS, J. A. & MEUWISSEN, T. H. 2012. Genomic selection requires genomic control of inbreeding. *Genet Sel Evol,* 44**,** 27.

SU, G.-S., LILJEDAHL, L.-E. & GALL, G. A. 1996. Effects of inbreeding on growth and reproductive traits in rainbow trout (Oncorhynchus mykiss). *Aquaculture,* 142**,** 139-148.

TURNER, S. 2011. Quality Control Procedures for Genome Wide Association Studies. Chapter**,** Unit1.19.

VIGNAL, A., MILAN, D., SANCRISTOBAL, M. & EGGEN, A. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution,* 34**,** 275.

VINCENT, B., DIONNE, M., KENT, M. P., LIEN, S. & BERNATCHEZ, L. 2013. Landscape genomics in Atlantic salmon (Salmo salar): searching for gene–environment interactions driving local adaptation. *Evolution,* 67**,** 3469-3487.

VISSER, C., LASHMAR, S. F., VAN MARLE-KÖSTER, E., POLI, M. A. & ALLAIN, D. 2016. Genetic Diversity and Population Structure in South African, French and Argentinian Angora Goats from Genome-Wide SNP Data. *PLoS One,* 11.

WANG, C., SCHROEDER, K. B. & ROSENBERG, N. A. 2012. A Maximum-Likelihood Method to Correct for Allelic Dropout in Microsatellite Data with No Replicate Genotypes. *Genetics,* 192**,** 651-69.

WEIR, B. S. & COCKERHAM, C. C. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution,* 38**,** 1358-1370.

WIGGANS, G. R., SONSTEGARD, T. S., VANRADEN, P. M., MATUKUMALLI, L. K., SCHNABEL, R. D., TAYLOR, J. F., SCHENKEL, F. S. & VAN TASSELL, C. P. 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. *J Dairy Sci,* 92**,** 3431-6.

WIGGINTON, J. E., CUTLER, D. J. & ABECASIS, G. R. 2005. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet,* 76**,** 887-93.

WOOLLIAMS, J., BERG, P., DAGNACHEW, B. & MEUWISSEN, T. 2015. Genetic contributions and their optimization. *Journal of Animal Breeding and Genetics,* 132**,** 89-99.

WOOLLIAMS, J. A., GWAZE, D. P., MEUWISSEN, T. H. E., PLANCHENAULT, D., RENARD, J.-P., THIBIER, W. & WAGNER, H. 1998. Secondary Guidelinses for Development of National Farm Animal Genetic Resources Management Plans. *Initiative for Domestic Animal Diversity.* Food and Agriculture Organization of the United Nations.

WRIGHT, S. 1922. Coefficients of inbreeding and relationship. *The American Naturalist,* 56**,** 330-338.

WRIGHT, S. 1931. Evolution in Mendelian Populations. *Genetics,* 16**,** 97-159.

WRIGHT, S. 1950. Genetical structure of populations. *Nature,* 166**,** 247-9.

Appendix

Appendix A: Larger view of Figure 4