Norwegian University
of Life Sciences

Master's Thesis 2023 30 ECTS
Faculty of Biosciences (BIOVIT)

IMPUTATION TO AND USE OF WHOLE GENOME SEQUENCING FOR FINE MAPPING AND GENOMIC PREDICTION IN ATLANTIC SALMON

OLUMIDE VICTOR ONABANJO
European Master in Animal Breeding and Genetics

EUROPEAN MASTER IN ANIMAL BREEDING AND GENETICS (EMABG)

IMPUTATION TO AND USE OF WHOLE GENOME SEQUENCING FOR FINE MAPPING AND GENOMIC PREDICTION IN ATLANTIC SALMON

OLUMIDE VICTOR ONABANJO

MAY, 2023

Main supervisor: **PROF. THEO MEUWISSEN (NMBU)**

Co-supervisor(s): **PROF. ARMIN SCHMITT (UGOE)**
DR. BINYAM DAGNACHEW (NOFIMA)
DR. MUHAMMAD LUQMAN ASLAM (NOFIMA)

Co-funded by the Erasmus+ Programme of the European Union

DEDICATION

I dedicate this thesis write-up to the Glory of the Almighty God, who has kept me alive, hale, and healthy in His infinite mercy. All salutations and adorations are directed to Him for his immeasurable favor, grace, and guidance upon me.

ACKNOWLEDGEMENT

I acknowledge CMS-Edit project (funded by The Research Council of Norway under grant agreement number 294504) for providing the salmon whole genome sequencing data used for this study, Mowi ASA for providing the genotype and phenotype data (which is a subset of a complete dataset used in AquaIMPACT project funded by European Union's Horizon 2020 research grant agreement number 818367), and the Norwegian Research Infrastructure Services (NRIS) for providing me with the computational power needed for this research.

I sincerely appreciate my supervisors, Prof. Theo Meuwissen (NMBU), Prof. Armin Schmitt (UGOE), Dr. Binyam Dagnachew (NOFIMA), and Dr. Luqman Aslam (NOFIMA), for their support, patience, corrections, guidance, and critiques during this master thesis. I acknowledge that this thesis would not have been successful without their input.

Also, I appreciate the European Union for funding the Erasmus Mundus Joint Master Degree (EMJMD) program and the European Master in Animal Breeding and Genetics (EMABG) committee for finding me fit to be awarded the two-year scholarship that funded this master's degree. Special appreciation to Nathalie Schwaiger for her constant follow-up and support to all EMABG students. Furthermore, I would like to thank all the lecturers and professors at BOKU, NMBU, and UGOE who taught me during my master's study. Lastly, my sincere appreciation to my course advisors and coordinators, Ms. Stine Telneset (NMBU), Dr. Gareth Frank Difford (NMBU), and Dr. Birgit Zumbach (UGOE), for their help, direction, and guidance.

My forever gratitude goes to my parents, Mr. and Mrs. Onabanjo, and siblings (Olamide, Ayomide, and Oyindamola) for their support and encouragement. I sincerely appreciate fellow EMABG 2021 colleagues for making my time in the program worthwhile, Esther Taiwo, Tafara Kundai, Constance Noge, Amritha Veedu, Muhammad Aziz, Natasha Watson, and others.

Lastly, special appreciation goes to my Heart rob, Gowri Siva, for the unconditional love, emotional support, understanding, and encouragement during the period of this thesis.

Thank You, God bless you all.

OLUMIDE V. ONABANJO

Ås, 2023

ABSTRACT

Sea lice (*Lepeophtheirus salmonis*) infestation of Atlantic salmon (*Salmo salar*) is a significant challenge facing the Aquaculture industry. This parasite is known to be resistant to chemical control. Previous research that studied the genomic architecture of host resistance to sea lice using low and medium-density SNP panels did not identify any genome-wide significant QTL associated with the trait. Thus, it became imperative to study the genomic architecture of this trait using whole-genome sequencing (WGS) data. However, it is not cost-efficient to re-sequence thousands of individuals, hence genotype imputation. Therefore, this study aimed to estimate the imputation accuracy (with and without pedigree), perform imputation to WGS for target individuals, estimate heritabilities, perform association tests, and genomic prediction.

A 10-fold cross-validation method was adopted to estimate imputation accuracy (with and without the inclusion of pedigree information) of the reference individuals using FImpute3 software. After imputation accuracy was estimated, genotype imputation of the target population (3185 individuals of the 2017 year class) to whole genome sequencing was carried out without including pedigree information. The imputed genotype and array data of the target population were then used to estimate heritability, perform association tests and estimate the accuracies of genomic prediction for host resistance to sea lice.

The weighted average imputation accuracy (r) without pedigree was estimated to be ~0.85, while ~0.84 was estimated with pedigree. The heritability of host resistance to sea lice was estimated to be 0.21 and 0.22, based upon array and imputed data, respectively. The association test using array and imputed data did not identify any marker associated with sea lice resistance QTL at the genome-wide level. In contrast, one marker on chromosome 7 of the array data surpassed the chromosome-wide Bonferroni corrected threshold and thus was declared significant at the chromosome-wide level. Lastly, a 5-fold within-family cross-validation design was used to assess the accuracy of genomic prediction. The accuracy was estimated to be ~0.65 and ~0.64 for array and imputed data, respectively.

In conclusion, genotype imputation is a valuable tool that saves sequencing costs, and including pedigree information did not significantly improve the genotype imputation accuracy. The trait of interest is moderately heritable and polygenic. The genomic predictions using medium-density SNP genotyping array was equally good or better than using whole genome imputed data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

1.0 INTRODUCTION

The aquaculture industry worldwide is a rapidly growing industry with the potential to help meet the food requirements of the ever-growing human population, estimated to reach 9.8 billion by 2050 (FAO, 2016). Aquaculture emerged as a solution when the amount of wild catch in the world's fisheries threatened the ecosystem's balance. According to FAO (2019), aquaculture production has surpassed global capture fisheries since 2013. This progress can be attributed to the high fecundity, good feed conversion ratio, and growth rate of most aquaculture species.

The Norwegian Atlantic salmon (*Salmo salar*) is a leading farmed aquaculture species of significant economic importance. Norway is the world leader in the production of salmon, as it contributes over 50% of the global production (FAO, 2022). The Atlantic salmon industry is of particular interest because it is the country's largest non-oil export, generating employment and substantial foreign revenue (Myhre Jensen et al., 2020). The Norwegian Seafood Council reported that Norway exported 1.25 million tonnes of salmon valued at NOK 105.8 billion (over 9 billion euros) in 2022 (Norwegian SeaFood Council, 2023).

However, there persists a lingering challenge of ectoparasite infestation by sea louse (*Lepeophtheirus salmonis),* which causes substantial economic losses annually. Sea louse is a natural parasite that feeds on the blood and tissue of salmon (Barrett et al., 2020), which poses a significant challenge to the production, welfare, and profitability of salmon farming (Gharbi et al., 2015). Once infested, the host is predisposed to stress, anaemia, stunted growth, and many other viral and bacterial infections, which may eventually lead to death (Correa et al., 2017; Øverli et al., 2014). In 2021, Norway recorded a production loss of about 60 million salmon due to death, escapes, and rejections (Directory of Fisheries, 2021). Most of the deaths and rejections could be attributed to lice infestation. In 2009, Costello estimated the global loss of salmon due to sea lice infestation to be about 430 million US dollars worldwide (Costello, 2009).

To curtail this problem, some medicinal and non-medicinal methods were adopted. The extensive dependence on few medicinal options due to various environmental laws has resulted in sea lice building up resistance against these compounds (Aaen et al., 2015). In Norway, since 2008, there have been reports of sea lice resistance to compounds such as emamectin benzoate,

hydrogen peroxide, benzoyl urea, and pyrethroids (Aaen et al., 2015; Helgesen et al., 2014). However, since 2017 there has been an increase in the use of non-medicinal methods to reduce sea lice infestation. These methods include delousing laser, warm water dips, mechanical removal, removal by a soft brush, and plankton shielding skirts. Although safer for the environment, some of these methods are stressful for salmon, affect their welfare, and in some cases, increase post-treatment mortality rates (Myhre Jensen et al., 2020; Overton et al., 2019).

It has been observed that variations exist in the susceptibility of salmon to sea lice, which indicates the presence of additive genetic variance. This can be exploited by selective breeding for genetic improvement of this trait in the population (Tsai et al., 2016). To achieve this, the genomic architecture that confers sea lice resistance to some salmon samples needs to be studied through genome-wide association studies.

Although other researchers (Correa et al., 2017; Tsai et al., 2016) have used a varying number of markers (6k to 50k) to study the association and estimate genomic breeding values for sea lice resistance, this research differs from them in that it uses WGS markers which are denser than what was used by previous researchers.

1.1 Aims and Objectives

The aims and objectives of this research are:

- To determine the accuracy of WGS for genotype imputation (with and without pedigree information)
- To carry out genotype imputation for the target population (3,185 individuals) from 50k to WGS
- To estimate the heritability of host resistance to sea lice using the imputed genotype data and array data
- To carry out Genome Wide-Association Studies (GWAS) analysis using array and the imputed genotype data to detect QTLs associated with host resistance to sea lice
- To carry out genomic prediction and estimate the accuracy of the predicted genomic breeding values for sea lice resistance in salmon.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 Atlantic Salmon (*Salmo salar*)

The Atlantic salmon is a ray-finned fish belonging to the Salmonidae family. It is an anadromous fish that lives in both fresh and seawater. It is prominent in the northern Atlantic oceans and its tributary rivers. Due to its nutrient composition, high fecundity, and feed conversion ratio, the Atlantic salmon has become a very important aquaculture specie that is farmed intensively in Canada, Scotland, Chile, and Norway.

Source: (Norwegian Seafood Company, 2019)

Figure 1: Atlantic Salmon (*Salmo salar*)

2.1.1 History of Atlantic Salmon Farming in Norway

Wild salmon fishing, whaling, and sealing have been a part of Norwegian culture since immemorial. These activities led to the decline of the wild salmon population in Norwegian waters, consequently depressing the fishing communities' economy (Liu et al., 2011). Scientists came together in the 70s to solve this issue, breeding new salmon that performed better than their wild counterparts in size and docility (Simen Sætre and Kjetil S. Østli, 2021).

The salmon industry gained government support, and salmon farming became a large-scale commercial industry in the 80s. This industry has experienced significant growth ever since, with its production growing exponentially from less than 500 tonnes in the early 1970s to 1.5 million tonnes in 2021 (Directory of Fisheries, 2021). This massive salmon production output has positioned Norway as the world leader in the supply of Atlantic salmon (Liu et al., 2011).

2.1.2 Life-Cycle of Atlantic Salmon

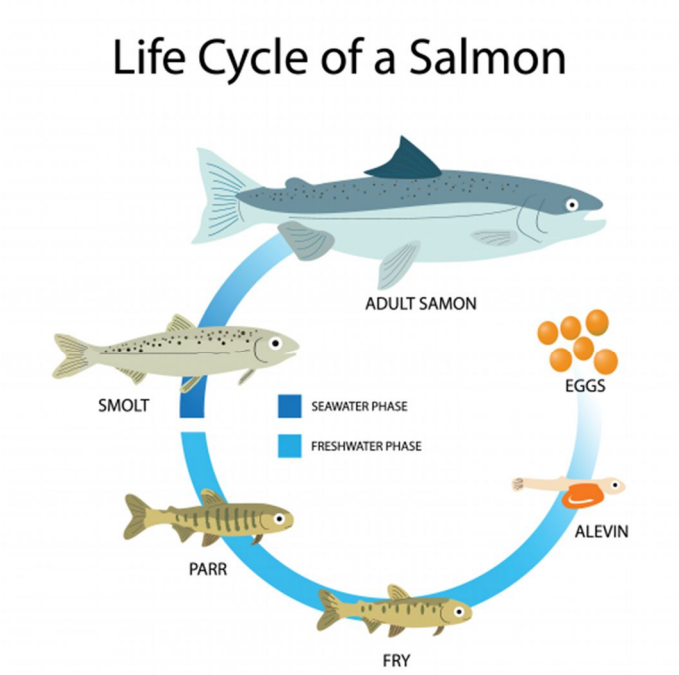
The life cycle of the wild Atlantic salmon, as shown in Figure 2, comprises six stages (egg, alevin, fry, parr, smolt, and adult) which take place in two phases, namely the freshwater phase and the marine phase. Atlantic salmon start their life in freshwater (river beds) as pea-sized orange eggs, which hatch after two to three months to become alevins. The alevins, which are partly transparent, feed on their attached yolk sac until they are ready to emerge from the riverbed about eight weeks or more post-hatching. Once they exhaust their yolk sac and develop fins, they are called fry (Salmon Facts, 2022).

The young fry swims to the river surface to feed on microscopic organisms and take up air. When the fry reaches a length of 5 to 8cm, and after they are over a year old, they transform into parr. Vertical parr marks, which help the fish to camouflage, appear. The parr remains in fresh water for about one to four years (depending on water temperature and food availability), feeding on aquatic insects. After the parr reaches about 12 to 24cm in length, they transform into a smolt. Then a silvery sheen replaces the parr mark. Also, they undergo smolting, characterized by the darkening of the edges of their pectoral and caudal fins. Furthermore, they undergo an internal change that prepares Smolt for life in the ocean (Marine Institute, 2022b).

Once the smolt migrates to the sea, they prey on other fish, such as eel and herring. They continue to feed and grow until they mature into adults. After living in the ocean for a year or more, adult salmon return to the river where it was born to spawn. Hormones control the migration period, and the adult salmon does not feed during its time in the river but survives on its fat reserves. An adult male salmon fertilizes the spawned egg. After spawning, the adult salmon are referred to as kelts. They are thin and weak because they do not feed during their time in freshwater. They migrate back to the ocean, where they feed and become strong again, although some do not survive. The eggs hatch after two to three months, and the life cycle begins again (Institute of Marine Research, 2020).

In summary, the wild Atlantic salmon life cycle spans about three to seven years. They spend their first two to three years in the river before migrating as smolts to the ocean. There, they spend about one to three years feeding and growing before returning to spawn in the river where they were born (Biologywise, 2023).

The life cycle of farmed salmon is similar to that of its wild counterparts, but it has been shortened to about four years with two phases (juvenile and grow-out phase). The juvenile phase (10 – 16 months) encompasses the stages from egg to smolt in land-based recirculatory aquaculture systems. During the grow-out phase, the smolts are transferred to sea pens, where they are fed for fourteen to twenty-four months till they reach maturity (3 to 8kg) and are harvested for processing (Artic Seafood Exports, 2023).



Source: (Inland Fisheries Ireland)

Figure 2: The Life-Cycle of Atlantic Salmon (*Salmo salar*)

2.1.3 Economic Importance of Atlantic Salmon

The Atlantic salmon is an aquaculture species of high economic importance that could help meet the animal source food needs of the growing human population. This potential can be attributed to its nutrient composition, high fecundity, and good feed conversion ratio. Farmed Atlantic salmon is Norway's most prominent non-oil export (Myhre Jensen et al., 2020). In 2022, Norway exported about 1.25 million tonnes of salmon valued at NOK 105.8 billion (Norwegian SeaFood Council, 2023), with France, Poland, and Denmark as the greatest importers of processed salmon. This high production level has positioned Norway as the world leader in salmon production, as the country contributes 50% of the global salmon production (FAO, 2022).

About 160 companies are operating over 1600 licenses in Norwegian waters. These companies provide about 9000 direct employment (Directory of Fisheries, 2021), excluding the indirect employment (machinery suppliers, feed producers, processing companies, transporting companies, etc.) provided along the salmon production food chain.

2.2 Whole-Genome Sequencing

Whole-genome sequencing refers to determining all DNA sequences in an organism's genome. Several techniques have been developed for this purpose and are classified into generations. There are First generation (Sanger and Maxam-Gilbert sequencing), Second generation or Next Generation (Roche, Illumina), and Third generation (Pacific Biosciences and Oxford Nanopore) sequencing.

Sanger sequencing (Sanger et al., 1977) is a popular First generation sequencing technique that Fredrick Sanger developed in the '70s. It was characterized by a manual, laborious, long-read (800 – 1000 bases), low-throughput sequencing technique that uses radiolabeled partially digested fragments and is also called the chain-termination method. It dominated the industry for three decades as it was the best technique at that time due to its high accuracy. Maxam-Gilbert (Maxam & Gilbert, 1977) also introduced his method, which was based on the chemical modification of DNA in the same year as Sanger.

The Next-generation (NGS) or second-generation sequencing technologies introduced between 2004 and 2006 dislodged the previous generation's technology and transformed biomedical research due to their high throughput, speed, and low cost (Mardis, 2013). It resulted in more output availability as this method processed millions of sequencing reactions from multiple samples in parallel (Tucker et al., 2009). However, a significant disadvantage is that the outputs are short-reads (50 – 300 bases), making mapping laborious and detecting structural variants difficult. Roche and Illumina are the primary providers of this technology. NGS's workflow includes library preparation, sequencing, and data analysis (Illumina).

The third generation of genome sequencing technology, such as Oxford Nanopore and Pacific Biosciences, boasts of long-read lengths that could reach 10kb (Hu et al., 2021). Although it is speculated to be less accurate when compared to the short reads, this long read length makes mapping and detection of structural variants easier. This sequencing technology has dramatically

reduced sequencing costs and time compared to NGS. In addition, these sequencing machines are very portable and affordable and do not require a complex library preparation process (Hu et al., 2021). These technologies have been used to successfully sequence the genome of humans and domestic species such as cattle, goat, chicken, pig, horse, and salmon.

2.2.1 Genotyping

Although advancement in technology and computational power has made genome sequencing very affordable, it is still not cost-efficient to sequence the whole genome of thousands of animals. Genotyping is simply the process of determining the genotype of an organism using denovo or known information from previous studies. The variants are observed when the DNA sequence of an organism is mapped against some reference genome.

Several genetic markers can be found in an organism's genome, which can be used to genotype individuals and populations. They include single nucleotide polymorphisms (SNPs), variable number tandem repeats (VNTRs), restriction fragment length polymorphisms (RFLPs), short tandem repeats (STRs), structural variants (SVs), and so on.

SNPs are by far the most popular and most widely studied type of genetic marker. This is because, unlike other markers, they are abundant in the genome, very informative, and their genotyping is automated (Wakeley et al., 2001). The SNP microarrays are a high throughput and very efficient way of genotyping hundreds of individuals with several thousand to millions of markers simultaneously. The outcrossing human population has over a million markers on its commercial genotyping array chip, while farm animals known to have a distinct family structure have a lower number of markers. Animal species such as sheep, goat, pig, horse, cattle, and chicken have cost-effective 50k SNP arrays (Meuwissen et al., 2013). Of these, chicken, salmon, and cattle have high-density chips of over 600,000 array SNPs because their genomes are well-studied compared to other farm animals.

2.3 Genotype Imputation

Over the years, there have been significant advances in sequencing and array genotyping technologies, but these technologies are still imperfect. Hence, NA (Not available) is assigned to markers that were not called due to low-quality control scores. Thus, genotype imputation is a primary pre-processing method used to infer the genotypes of such data points in preparation for

further analysis, such as genome-wide association studies (GWAS) analysis and genomic prediction (Meuwissen et al., 2001).

Also, imputation has helped to save costs associated with sequencing or high-density genotyping of several individuals within a population. This process is done by sequencing or high-density genotyping a few individuals (reference population) and using their genotype information to infer the genotypes at positions that were not genotyped for other individuals (target population) which were genotyped using low-density panel (Hickey et al., 2012). This method is advantageous in salmon aquaculture since they produce thousands of offspring, and it is not cost-efficient to genotype all individuals using high-density chips.

Several software programs have been developed for genotype imputation in various organisms (plants or animals), species, and populations. The popularly used ones include Beagle (Browning et al., 2018), FImpute (Sargolzaei et al., 2014), TASSEL (Bradbury et al., 2007), Impute (Howie et al., 2009), MiniMac (Das et al., 2016) to mention a few. Beagle (Browning et al., 2018), probably the most popular imputation software developed, was created to carry out genotype imputation in outbred human populations (Bradbury et al., 2007; Pook et al., 2019). In contrast, TASSEL (Bradbury et al., 2007) was developed to handle genotype imputation in fully homozygous plant lines (inbred lines). FImpute (Sargolzaei et al., 2014) was designed to handle genotype imputation in livestock, which, to some extent, are closely related. Therefore, it allows for the optional inclusion of pedigree information before carrying out genotype imputation. When pedigree information is included, the software depends on family relationships and linkage disequilibrium (LD) to impute missing genotypes. Otherwise, it depends entirely on linkage disequilibrium (Calus et al., 2014).

There are two major approaches that these software programs use to infer genotypes. Beagle, Impute, and MiniMac use the Hidden Markov Models approach (Baum & Petrie, 1966), while TASSEL and FImpute use the overlapping sliding window approach, which exploits relationships between reference and target individuals to infer genotypes.

2.3.1 Estimating Imputation Accuracies

Genotype imputation accuracies can be estimated either SNP or individual-wise, and several statistical approaches exist for estimating this accuracy. Some methods compare the imputed

genotype to the true genotype, while others do not (Ramnarine et al., 2015). Examples of imputation accuracy approaches that compare the imputed to true genotypes include correlation coefficient (r), squared correlation (r^2), concordance rate, error rate, and imputation quality score (Lin et al., 2010). In this case, the genotypes of SNPs are known, and some percentages are intentionally masked and imputed. The true genotypes are then compared to the imputation outcome across all animals using any of the above-stated metrics of imputation accuracy.

The Pearson correlation coefficient (r), which is the simplest form of calculating accuracy, directly compares the true and imputed genotype and returns the r value based on how linearly related the imputed and the true genotypes are, while squared correlation (r^2), also known as imputation reliability, is the square of the Pearson correlation coefficient (r) between true and imputed genotypes. The concordance rate is the proportion of correctly imputed loci, while the error rate is the opposite. The imputation quality score (IQS) has a maximum score of 1 and no theoretical minimum, and it is a concordance rate adjusted for chance (Lin et al., 2010).

However, in real situations, the true genotype of SNPs to be imputed are unknown, which is why we impute them (Stahl et al., 2021). Therefore, imputation software such as Beagle and MiniMac use statistics to determine the accuracy which they report. Numerous research have found this accuracy estimation without true genotypes to be very accurate and dependable; thus, they are widely used and accepted (Chanda et al., 2012; Ramnarine et al., 2015)

2.3.2 Factors Affecting Genotype Imputation Accuracy

Although genotype imputation methods are instrumental in data pre-processing and saving costs related to sequencing, it is essential to calculate the imputation accuracy as this would affect the outcomes of subsequent analysis (Deng et al., 2021). Factors affecting the accuracy of imputation include the proportion of genotypes to be imputed (Hickey et al., 2012; Zhang & Druet, 2010), the minor allele frequency of variants, the imputation method, the genetic distance between the reference and the target individuals (Carvalheiro et al., 2014), the number of individuals in the reference panel (Druet et al., 2010; Zhang & Druet, 2010), the sequencing coverage of reference panel, and the chromosomal position (Badke et al., 2013) amongst others.

2.4 Genome-Wide Association Studies (GWAS)

Genome-wide association studies (GWAS) use a statistical approach to map variants (from SNP microarrays or whole genome sequencing data) associated with traits of interest (The Wellcome Trust Case Control Consortium, 2007). In order to achieve this, the genotype of thousands to millions of variants and the phenotype (e.g., case and control) information of a reasonable number of individuals within a population should be available (Altshuler & Daly, 2007). Each genotype information is tested across many genomes to find those statistically associated with the traits or diseases of interest (Uffelmann et al., 2021). In humans, this method has assisted in identifying close to 200,000 SNPs associated with complex traits and diseases (Buniello et al., 2019).

The results of GWAS analysis can be used for various purposes, including understanding the underlying biology of a phenotype, predicting clinical risks, guiding drug development initiatives, and inferring potential causal relationships between risk factors and health outcomes (Uffelmann et al., 2021). Although GWAS have contributed significantly in explaining genotype-phenotype associations of various traits, it also has several limitations, including multiple testing burdens and spurious associations. In addition, GWAS analysis using only SNP variants may not pinpoint causal variants, may explain only a modest fraction of heritability, and may not identify all the genetic determinants of complex traits (Tam et al., 2019).

2.4.1 Multiple Testing Problem

The multiple testing problem is unavoidable in GWAS because thousands to millions of markers are simultaneously linearly regressed to the phenotypes. The high number of markers being tested increases the number of markers that could be found statistically significant by chance (false positives) and, therefore, must be corrected. In other words, increasing the number of markers to be tested increases the probability of committing a type I error. The probability of committing at least one type I error (accepting a non-significant test as significant) in a study is given as,

$$P(Type\ I\ error) = 1 - (1 - \alpha)^n$$

Where n is the number of independent tests carried out, and α is the chosen level of significance for one test. There are several methods for correcting this type of error in a GWAS analysis, and

the most commonly used methods are the Bonferroni correction and the false discovery rate (FDR).

The Bonferroni correction assumes that all markers are independent, although it is common knowledge that neighbouring SNPs on a chromosome tend to be inherited together and are therefore linked (The International HapMap Consortium, 2005). This correction method is said to be over-conservative, reducing the power of an experiment (that is, the probability of finding a significant QTL if it exists). It is expressed as,

$$alpha\ prime\ = \frac{\alpha}{n}$$

Where *alpha prime* is the genome-wide p-value, α is the chosen level of significance for one test, and n is the number of independent markers.

On the other hand, the false discovery rate (Benjamini & Hochberg, 1995) method allows the acceptance of a few false discoveries without eliminating the true discoveries. This method is carried out in two steps. First, the hypothesis is ranked then a cutoff is chosen along the rankings (Wei et al., 2009). It is expressed as

$$FDR = \frac{n * \ p - value}{rank}$$

2.4.2 Spurious Association

The spurious or biased association is another major limitation of GWAS caused by population stratification or population structure. Population stratification occurs when the population to be studied is mixed or heterogeneous, while one speaks of a population structure if a complex relationship exists among individuals. This problem can be detected by carrying out a principal component analysis (PCA) which detects population structures. The spurious association problem can be reduced using a linear mixed model (LMM), which accounts for genomic relationships (Kang et al., 2008). Also, it can be visually represented and detected by a quantile-quantile plot which compares the observed $-log(p - values)$ to that which are expected under the assumption of null hypothesis.

2.4.3 Fine Mapping

Since it is known that not all variants that are found statistically significant in a GWAS analysis are causal, a post-GWAS analysis that utilizes both statistical and functional methods is necessary to identify causal variants. Fine mapping analyzes a trait-associated region from a genome-wide association study to find the genetic variants most likely to have a causal impact on the trait under consideration (Schaid et al., 2018). To achieve this, the linkage disequilibrium structure and known genes associated with each region are examined. The chosen SNPs are then further assessed for their likely function using publicly accessible genomic annotation (Schaid et al., 2018).

For the purpose of prioritizing causal variations to explain association signals, various methods have been developed. They can be roughly divided into two groups: Bayesian approaches that assign a posterior probability of causality to each SNP and triaging variations based on p-values or linkage disequilibrium to the lead SNP (Spain & Barrett, 2015). The latter is an easier approach when compared to the former. In a Bayesian framework, each variant's evidence for association is assessed using a Bayes factor, which, under specific conditions, can be used to determine each variant's posterior probability of being causal for the connection in that region (Stephens & Balding, 2009). Both methods depend on raw genotype data that are not always available. Hence, several other approaches, such as CAVIARBF (Chen et al., 2015), FINEMAP (Benner et al., 2016), and PAINTOR (Kichaev et al., 2014), which make use of summary statistics from GWAS alone, have been developed.

2.5 Genomic Selection

Genomic selection is a form of marker-assisted selection (MAS) that simultaneously estimates the weights of thousands to millions of DNA markers (in this case, SNPs) to predict breeding values (Meuwissen et al., 2013). Unlike other MAS methods, which could only explain about 10% of total genetic variance due to their strict significance threshold, genomic selection uses the information of all SNPs, thereby explaining the majority of genetic variance (Goddard & Hayes, 2007). Also, genomic selection has provided a lasting solution to traditional breeding problems such as long generation intervals, slow genetic gain, and the high cost of keeping animals. This genomic selection method, introduced by Meuwissen et al., (2001), became popular and widely adopted after the advent of affordable genome-wide SNP chips in 2008.

To estimate genomic breeding values, the genotype and phenotype information of a group of animals regarded as the training population is required, while only the genotype information is needed for the selection population for which genomic breeding values would be predicted. Progenies with good genomic estimated breeding values are then selected from the selection population to serve as parents of the next generation. In the dairy industry, the adoption of genomic selection has doubled genetic gain by reducing the generation interval of bulls from around six years to less than two years (Pryce & Daetwyler, 2012). Genomic selection has also been widely used in poultry, pig, and salmon breeding.

2.5.1 Methods for Estimating Genomic Estimated Breeding Values

There are several methods of genomic selection used to estimate genomic breeding values. They can be broadly categorized as linear (Genomic BLUP) or non-linear (SNP-BLUP, Ridge regression, Bayes A, Bayes B, Bayes C, Bayes R, and Bayesian Lasso). Each method has pros and cons; they all work with different assumptions. The GBLUP is the same as the BLUP animal model except that it uses a genomic relationship matrix (G) which gives a more accurate relationship estimation than the pedigree-based relationship matrix (A), which is error-prone. This method assumes that each marker has an effect and explains an equal amount of genetic variance, which in reality is not the case (Meuwissen et al., 2016).

On the other hand, the non-linear models presuppose a prior distribution, leading to an uneven proportion of genetic variance being explained by various markers (Meuwissen et al., 2001). The Ridge regression and SNP-BLUP models assume that all SNP effects come from a normal distribution with minimal variance. Because the number of records is often significantly lower than the number of SNP effects to be estimated and the final estimates are heavily impacted by the prior, applying these models will produce very modest SNP effect estimates (Meuwissen et al., 2013).

The Bayesian methods of genomic selection, such as Bayes A, B, C, and R, are based on other assumptions of the distribution of SNP effects (Meuwissen et al., 2013). The Bayes A method allows the variance of SNP effects to differ for every SNP but assumes that all SNPs have effects. Bayes B assumes a t-distribution for SNPs with effects, resulting in some SNPs having huge effects (Meuwissen et al., 2001), while Bayes C assumes normal distribution and constant variance for SNPs with effects (Habier et al., 2011). In addition to allowing for some SNPs with

huge effects, such as those taken from the distribution with the greatest variance, Bayes R assumes a mixture of normal distributions for the effective SNPs (Erbe et al., 2012).

2.5.2 Accuracy of Genomic Estimated Breeding Values

In order to determine how well a model performed, it is necessary to estimate the accuracy of such a model. The correlation between the estimated and the true breeding value divided by the square root of heritability is the most commonly used metric for measuring prediction accuracy (Daetwyler et al., 2013). It is used in a case where we already have phenotypes and genotypes information of a large population, the population is divided into n-fold, and each fold is used to validate other animals in the population. This step is repeated until phenotypes are predicted for all animals in the population. The expected phenotype for each animal in the population is then compared to its true phenotype (Meuwissen et al., 2013). There are several factors affecting the accuracy of the genomic estimated breeding value. They include marker density, the level of linkage disequilibrium between the marker and QTL, heritability of the trait, the level of relatedness, homogeneity of the population, the size of the training population, and the estimation method.

2.5.3 Advantage of Using Whole Genome Sequencing Data for Genomic Selection

The key feature and distinction of genomic selection against other MAS methods is that it aims to include all genome-wide markers to explain more genetic variance (Goddard & Hayes, 2007). Hence, whole genome sequencing data gives a higher chance of capturing causative polymorphisms (Meuwissen et al., 2013). The benefits of using whole genome sequencing data in genomic selection are only visible if Bayesian methods (B, C, and R), which assume that not every SNP has an effect, are used. The accuracy of the Bayes B method in simulated WGS data was estimated to be between 83 and 97%, while GBLUP had just about 50% accuracy (Meuwissen et al., 2013). This difference in accuracy is because the GBLUP method assumes that all markers have equal effects, and it merely uses the WGS data to estimate genomic relationships between animals (Meuwissen et al., 2016).

The cost of computation, sequencing, and the large number of animals to sequence has been a stumbling block to the use of WGS data for genomic selection, but the availability of SNP chips,

long-read sequencing technology, imputation algorithm, and decreasing price of sequencing will further make it a possibility.

2.6 Sea Lice

Lepeophtheirus salmonis and Caligus elongatus are two known marine ectoparasitic copepods of the family Caligidae, posing a major economic challenge to salmon farming worldwide (Boxaspen, 2006). These parasites attach themselves to the skin and feed on their host's mucus, epidermal tissue, and blood, causing injury and diseases. The aquaculture industry worldwide is estimated to lose about 430 million US dollars annually to sea lice infestation and control (Costello, 2009).

2.6.1 History of Sea Lice Outbreak

In Norway, the sea lice outbreak first occurred during the 1960s, soon after the advent of open cage culture (Pike, A. W., and Wadsworth, S. L., 1999). The name “Lakselus” was used by fishermen to refer to the salmon louse. This term was said to be the first used in prints by Gisler in 1751 (Berland & Margolis, 1983).

2.6.2 Morphology of Sea Lice

In terms of size, the *Lepeophtheirus salmonis* is said to be double the size of *C. elongatus*. They have four body parts: the cephalothorax, the fourth (leg-bearing) segment, the genital complex, and the abdomen (Johnson & Albright, 1991). The cephalothorax covers all body segments up to the third leg-bearing segment, which resembles a broad shield. Each species has mouthpieces with an oral cone or siphon form that helps keep the parasite on the fish. In adult females, the genital complex makes up the majority of the body mass, and it grows to a size that is consistently bigger than that of adult males (Costello, 2006).

2.6.3 Life-Cycle and Reproduction

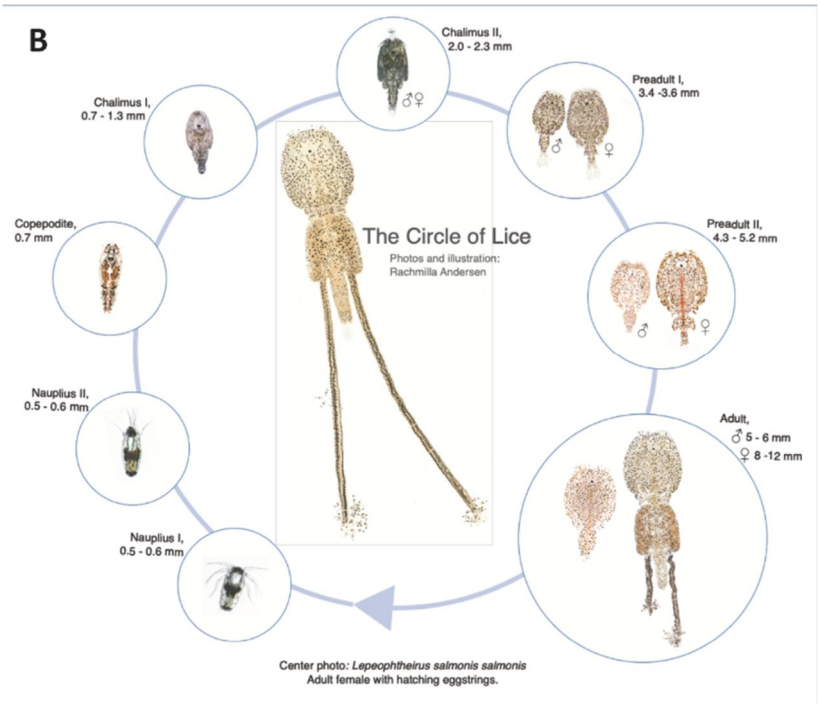
Sea lice (*Lepeophtheirus salmonis)* have a direct life cycle that occurs in about eight developmental stages (Hamre et al., 2013), as shown in Figure 3. The life cycle begins with the first two naupliar stages: free-swimming, non-feeding, and non-infectious. The naupliar larvae are from hatched egg strings produced by the adult female louse. The larvae at this stage drift with the ocean current since they are small (0.5 – 0.6mm). Also, this stage is temperature, light,

and salinity dependent as the larvae grow faster at higher temperatures and salinity (Marine Institute, 2022a). This naupliar stage goes through morphological changes and molts into the copepodid, an infective stage (Sea Lice Research Centre, 2023). Unless it comes across a suitable host, the copepodid stage is also free-swimming and not feeding as it lives off its fat reserve. Once in contact with a suitable host, attachment is made possible by using specially designed clawed appendages (Jacques Drolet, 2015). From that moment, the copepodid becomes a parasite that feeds on the host tissue, blood, and mucus.

Before the copepopid molts into the chalimus stage, the frontal filament, an attachment apparatus that binds the growing louse to hard structures like scales, cartilage, and fin rays on the host, develops (Jacques Drolet, 2015). This frontal filament gives the louse a good grip of the host and prevents it from falling off when it changes its exoskeleton to grow. Male and female louse differentiation is now microscopically conceivable at the chalimus II stage. Although a macroscopic distinction is impossible until the next stage, the female louse has a larger cephalothorax, the frontal body portion, than the male louse at this stage (Sea Lice Research Centre, 2023).

In the pre-adult stage of life, the frontal filament is shed off, and the louse can move freely on the host surface. Due to their larger size and ability to feed from a wider area, lice in the mobile stages inflict the most damage to the host. Furthermore, since they are mobile, they can move from one host to another (Jacques Drolet, 2015).

Reproduction occurs in the adult stage, as this is the stage of sexual maturity for female lice. In order to be the first to fertilize a particular female, males are often partnered up with pre-adult II females because they attain sexual maturity earlier and develop more quickly than females. The female lice bear the young in two protracted egg threads that are fastened to her vaginal area. The egg strings can have a maximum length of 50 mm and hold up to 700 eggs each. At 10°C, the female louse develops new egg threads every ten days, creating thousands of children after a few months (Sea Lice Research Centre, 2023). Then the cycle starts all over.



Source: (Contreras et al., 2020)

Figure 3: Life-Cycle and Morphology of *Lepeophtheirus salmonis*

2.6.4 Control of Sea Lice

Several methods have been adopted to control sea lice, including pesticides, physical treatments, functional feed, selective breeding, vaccination, fallowing and biological control. For simplicity, these methods are classified into medicinal/chemical or non-medicinal/chemical control methods.

2.6.4.1 Medicinal Control

Medicinal control involves using drugs and chemicals like hydrogen peroxide (Ron Tardiff, 2019). Medicinal control, which is majorly administered by bath treatment, has been widely used to control sea lice due to its effectiveness. Its extensive use over the years resulted in resistance by sea lice. In Norway, sea lice have been reported to be resistant since 2008 to compounds such as emamectin benzonate, hydrogen peroxide, benzoyl urea, and pyrethroids (Aaen et al., 2015; Helgesen et al., 2014). The pollution, resistance, and unintended adverse effect of these chemicals on other species living in the ocean necessitated more environmentally friendly and

safe sea lice control options. Thus, the shift to non-medicinal control options such as mechanical and thermal control.

2.6.4.2 Non-Medicinal Control

The non-medicinal control methods encompass mechanical, thermal, and biological controls. The mechanical methods of lice control require the fish to be pumped into a treatment system, then spray nozzles flush the sea lice off the fish (Overton et al., 2019). For the thermal delousing method, infested fish are placed in warm water for a few minutes to inactivate sea lice. Although these methods are safe, have no impact on non-target species, and are chemical-free, they can be stressful for the fish and increase mortality (Overton et al., 2019).

In the biological control method, fish such as Ballan wrasse (*Labrus bergylta*), which is not a host for sea lice, is reared with salmon in their cages. The job of this cleaner fish is to feed on sea lice that live on salmon (Ron Tardiff, 2019). In Norway, the use of wild wrasse expanded from 1.7 million fish in 2008 to 20 million in 2016 due to its effectiveness in controlling sea lice (Norwegian Directorate of Fisheries, 2017). To meet the demand, cleaner fish are now being farmed due to the recent industrial expansion of salmon. There have recently been welfare and sustainability concerns about using cleaner fish to control sea lice (Skiftesvik et al., 2014).

Other non-medicinal sea lice control methods include ultrasonic waves, laser-shooting robots, skirt barriers, freshwater baths, etc.

2.6.4.3 Selective Breeding

The resistance to chemicals and stress associated with mechanical and thermal methods necessitated seeking a favorable and cheaper alternative. In this regard, selective breeding is considered an affordable and effective alternative (Gharbi et al., 2015). This method uses quantitative genetics to select families resistant to sea lice to become future parents. This approach has been previously used in the salmon aquaculture industry against infectious pancreatic necrosis (IPN).

CHAPTER THREE

3.0 MATERIALS AND METHODS

3.1 Description of the Population

The data of the population studied for this research was provided by MOWI. It consists of samples from the year class 2017, comprising 3185 samples. Eighty-three sires and one hundred and eighty-two dams produced one hundred and ninety-one full-sib families with a median of 17sibs per family, as shown in Figure 4.

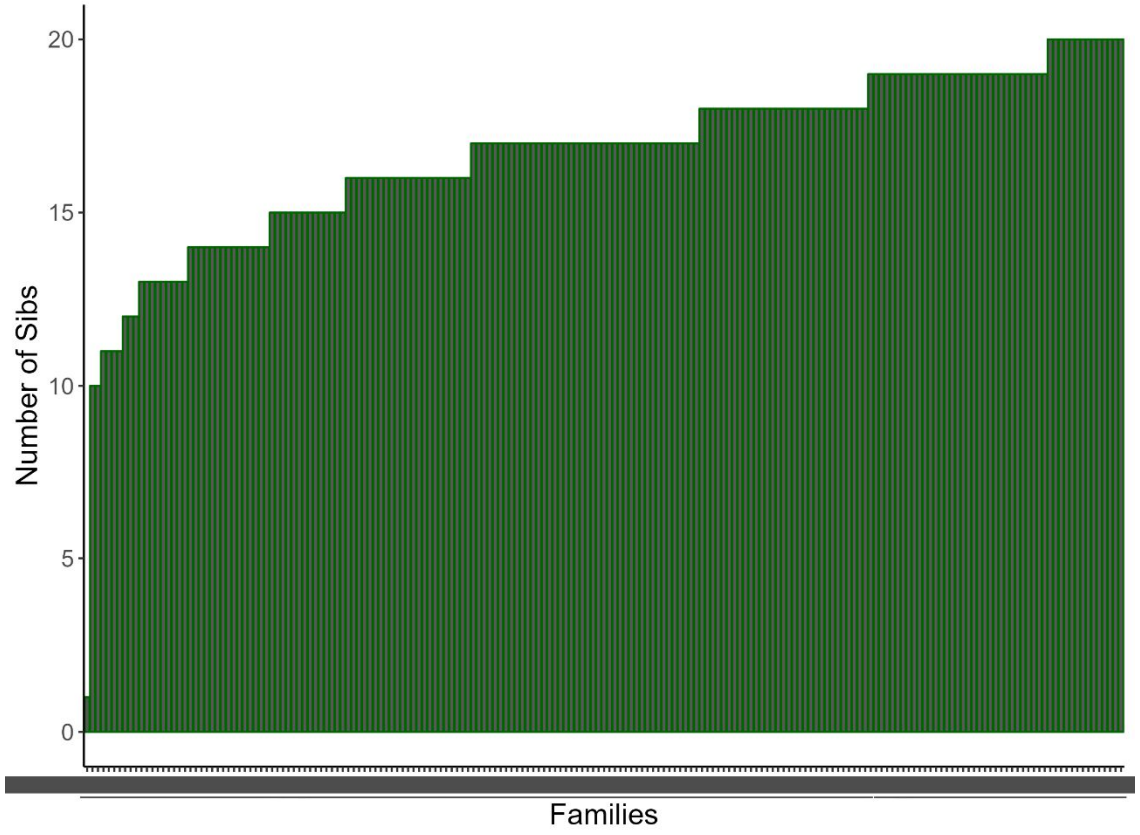


Figure 4: A Histogram showing the Frequency Distribution of Full-Sib Families in the 2017 Population

3.2 Phenotyping

The challenge test(s) was carried out in 4 tanks at Matre, Norway. The population (2017-year class) with an average weight of 109 grams was infested with ~45 copepodids per fish. The tank parameters were supervised and recorded, including water temperature, oxygen, and salinity.

After the infestation, regular monitoring was performed daily until most of the lice reached the chalimus I stage. A small fish sample was counted every 4-7 days post-lice infestation to evaluate the developmental stage of lice. The final counting of the complete set of fish was carried out by ten counters after ~85 days from the start of the challenge test when lice reached the chalimus III stage. The lice-counting process continued for approximately four days which was performed by anaesthetizing fish, counting lice, and recording body weight and length.

Figure 5 shows the frequency distribution of sea lice counted on the salmon samples, while Table 1 shows the descriptive statistics of the data. The population phenotyped consists of 3,185 fish samples, which were reduced to 2,935 after missing records were removed. Most samples had relatively low lice counts, while few had high counts resulting in a right-skewed distribution, as shown in Figure 5a. In other to normalize the phenotype distribution, a log transformation [$logLC = \log_e(sea\ lice\ count + 1)$] of this data was carried out as shown in Figure 5b. The transformation formula adds a constant value of 1 to all sea lice counts, allowing the transformation of zero (0) sea lice count if such exist.

Table 1: Descriptive statistics of phenotype data

	n	**Mean**	**Median**	**Min**	**Max**	**Var**	**Std. dev**
Lice count	2935	20.5	17	1	262	208.56	14.44
$log_e(lice\ count + 1)$	2935	2.89	2.89	0.69	5.57	0.35	0.59
Body weight (g)	2935	109.61	104.6	35	265	1055.01	32.48

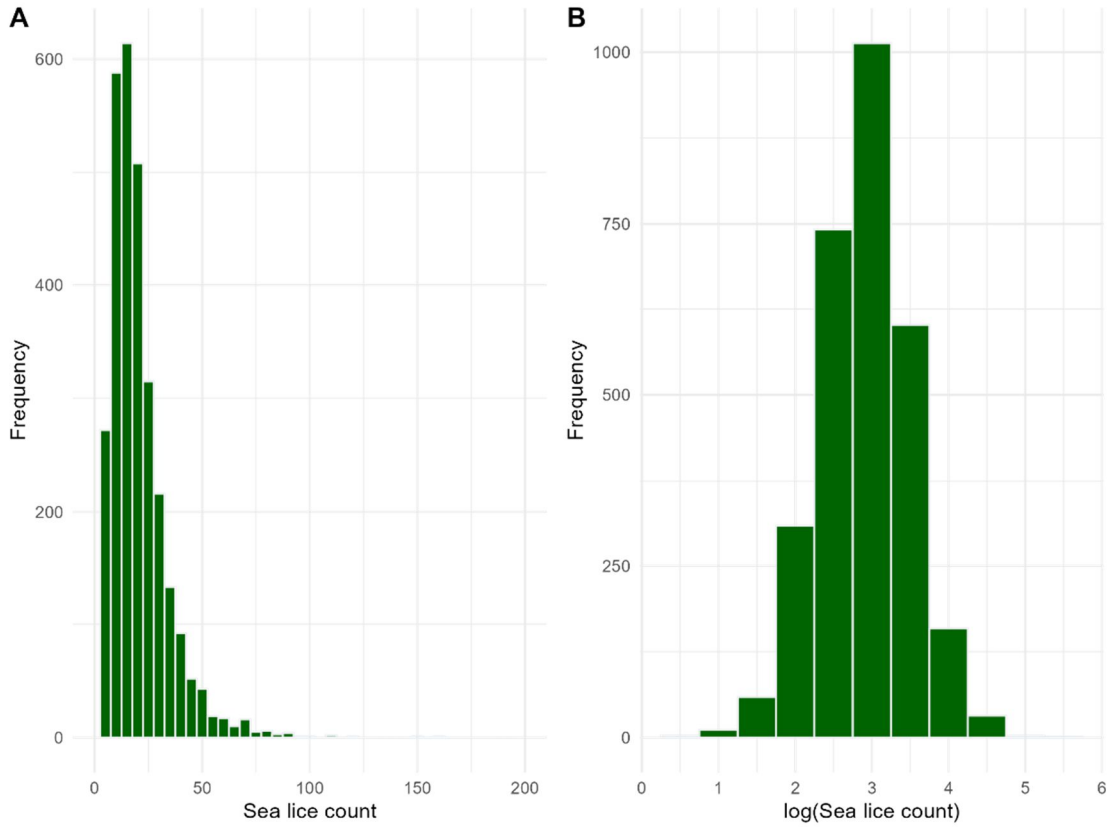


Figure 5: A Histogram showing the Frequency Distribution of a) Sea Lice Count (b) $Log1p$ of Sea Lice Counted on Salmon.

3.3 Sequencing

The whole-genome re-sequencing data was generated in the CMS-Edit project - funded by The Research Council of Norway under grant agreement number 294504. The variant call information (*.vcf.gz files) was made available for the current thesis. Briefly, 197 individuals covering siblings, parents, and relatives of the target individuals (lice count recorded) were available with whole genome re-sequence data. The whole-genome resequencing was performed using the BGISEQ platform with 150bp paired ends reads. The raw sequence reads were trimmed and filtered using Trimmomatic (Bolger et al., 2014). Subsequently, quality sequence data were aligned to the latest available Atlantic salmon reference genome sequence (assembly *Ssal_v3.1*) using BWA-MEM version 0.7.13-r1126 (Li, 2013), and then GATK (O'Connor & van der Auwera, 2017) pipeline was used for variant discovery and genotype calling. Table 2 shows the number of the whole-genome sequencing SNPs and the number of the overlapping genotyping array SNPs for each chromosome.

3.4 SNP Genotyping

The individuals recorded with lice count phenotype (target individuals) were genotyped using custom developed ~55K SNPs genotyping array (NOFSAL03, Affymetrix axiom array). The SNPs across sequence and the array data were searched for overlapping SNPs, which detected ~50K overlapping SNPs. The genetic map of the overlapping SNPs (~50K) between the whole-genome sequencing data and the genotypic array data is shown in Figure 6.

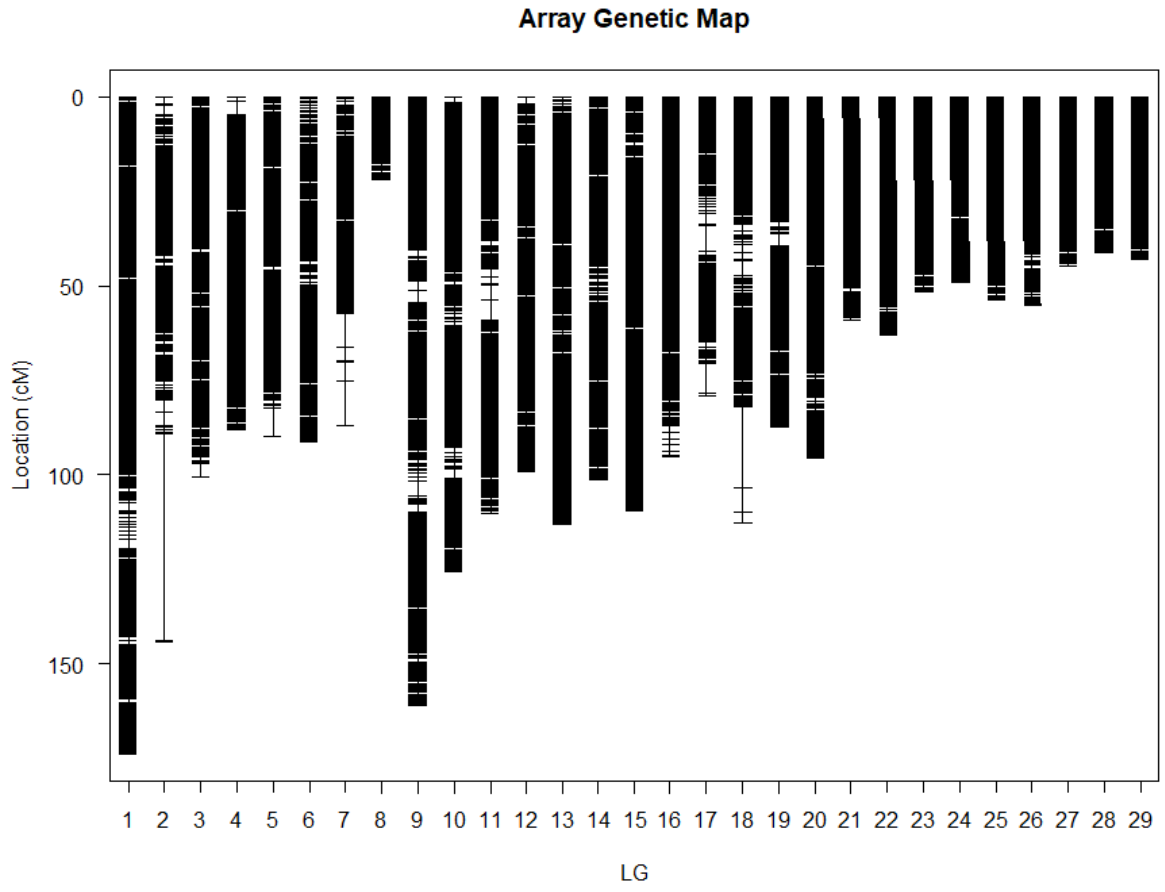


Figure 6: A Genetic Map Showing the Density of Markers on each Chromosome of the Array

Table 2: Summary of the Whole-Genome Sequencing and the Overlapping Array SNPs

Chromosome	Chromosome size (bp)	Number of Whole-Genome Sequencing SNPs	Number of Overlapping Genotyping Array SNPs
Ssa01	174,498,729	315,203	3,636
Ssa02	95,481,959	154,223	1,604
Ssa03	105,780,080	196,962	2,393
Ssa04	90,536,438	168,726	1,995
Ssa05	92,788,608	156,403	2,016
Ssa06	96,060,288	175,325	1,953
Ssa07	68,862,998	123,674	1,404
Ssa08	28,860,523	44,976	418
Ssa09	161,282,225	258,770	2,744
Ssa10	125,877,811	211,792	2,585
Ssa11	111,868,677	176,679	1,922
Ssa12	101,677,876	189,122	2,000
Ssa13	114,417,674	196,755	2,502
Ssa14	101,980,477	171,528	2,215
Ssa15	110,670,232	195,975	2,045
Ssa16	96,486,271	151,977	1,693
Ssa17	87,489,397	108,117	1,156
Ssa18	84,084,598	142,757	1,376
Ssa19	88,107,222	153,362	1,600
Ssa20	96,847,506	162,802	2,007
Ssa21	59,819,933	117,562	1,185
Ssa22	63,823,863	122,674	1,449
Ssa23	52,460,201	110,988	1,386
Ssa24	49,354,470	93,267	1,205
Ssa25	54,385,492	99,906	1,171
Ssa26	55,994,222	98,126	1,012
Ssa27	45,305,548	98,869	1,202
Ssa28	41,468,476	87,008	1,015
Ssa29	43,051,128	91,370	892
Total	**2,499,322,922**	**4,374,898**	**49781**

3.5 Genotype Imputation

This project used the FImpute3 software (Sargolzaei et al., 2014) to perform all genotype imputation. The reason for choosing this software over the widely used Beagle software is that it allows for the inclusion of pedigree information. The FImpute3 software uses an overlapping sliding window approach to efficiently exploit relationships or haplotype similarities between target and reference individuals. Firstly, the variant call format (VCF) file for sequence data was unzipped, converted to Plink's (Purcell et al., 2007) binary format using the "--make-bed" command option, and then to the FImpute3 input format. A population-based genotype imputation of the sequence data (reference population) was performed to impute data points with low-quality scores (missing genotypes) across all chromosomes. The imputed genotypes were then extracted and converted to Plink's binary format for further analysis. With this complete sequence data set, the animal- and SNP-based imputation accuracy was estimated for the reference population, with and without pedigree information.

3.5.1 Estimation of Imputation Accuracy

The imputation accuracy of the reference population (WGS data) was estimated with and without pedigree. This was done to assess the reference quality for the proposed reference-target whole genome sequencing imputation. Moreover, we were curious to find out if including pedigree information would improve the accuracy of genotype imputation. A 10-fold cross-validation method, as shown in Figure 7, was adopted to estimate these imputation accuracies. The reference population (197 individuals) was divided into ten folds which consisted of 20 individuals in each fold, except the last fold, which comprised 17 individuals.

The advantage of the n-fold cross-validation method over the random sampling method is that it allows all individuals to be used to train and validate in different iterations. The "--remove" command option in Plink (Purcell et al., 2007) was used to exclude the genotypes of each fold of validation individuals from the complete genotype file, while the training folds were retained in each iteration run. The validation set had only the ~50k SNP genotypes available on both sequence and array data. All other positions (~99%) were masked and imputed for. The data was converted into FImpute3 input format, and the training individuals with 4 million SNPs were used to impute for the ~3.95 million SNPs missing in the validation set in all iterations. The

imputed genotypes for all folds of the ten iterations were extracted, merged, and compared to the true genotypes.

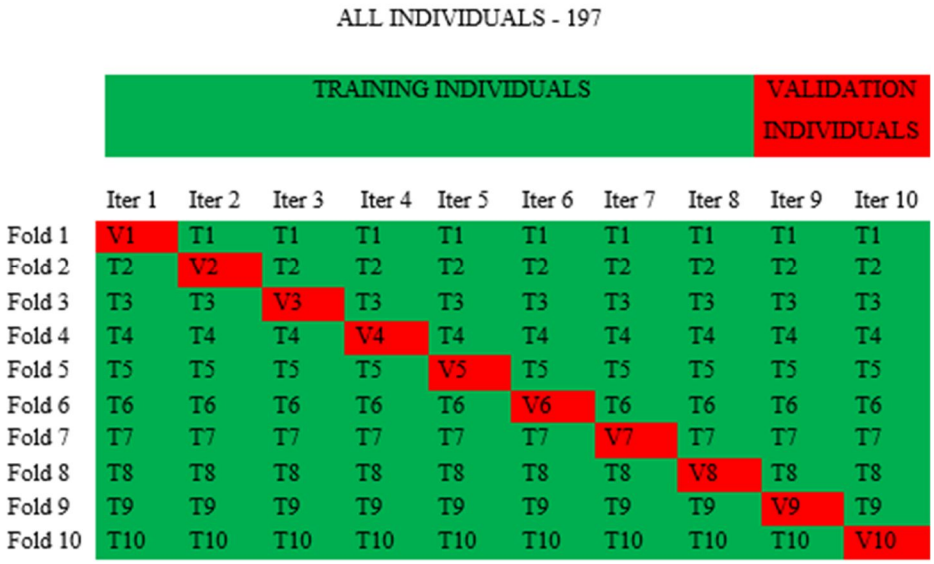


Figure 7: A Representation of The 10-Fold Cross-Validation used for Estimating the Imputation Accuracy

Pearson's correlation coefficient (r) was used to estimate the animal-based and SNP-based imputation accuracy, and the averages were reported per chromosome. The animal-based imputation accuracy is the Pearson's correlation coefficient (r) between the true genotypes for all markers of each animal to its imputed genotypes. In contrast, SNP-based accuracy is Pearson's correlation coefficient (r) between the true genotypes of each marker for all animals to its imputed genotypes. SNPs with poor imputation accuracy (r) < 0.6 were excluded from further analysis.

Pearson Correlation coefficient (r)

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

$r = correlation\ coefficient$

$x_i = values\ of\ the\ true\ genotypes$

$\bar{x} = mean\ of\ the\ true\ genotypes$

$y_i = values\ of\ the\ imputed\ genotypes$

$\bar{y} = mean\ of\ the\ imputed\ genotypes$

Since the chromosome size differs, the weighted average of all 29 chromosomes was calculated to give the genome-wide average SNP-based imputation accuracy.

Weighted Average (W)

$$W = \frac{\sum_{i=1}^{n} w_i X_I}{\sum_{i=1}^{n} w_i}$$

Where:

$W = weighted\ average$

$n = number\ of\ chromosomes\ to\ be\ averaged$

$w_i = weights\ applied\ to\ number\ of\ SNPs\ on\ each\ chromosome$

$X_I = number\ of\ SNPs\ on\ each\ chromosome\ to\ be\ averaged$

3.5.2 Reference-Target Imputation (50k to whole-genome sequencing)

Reference-target population genotype imputation was carried out to whole-genome sequencing. The imputed genotypes for the 3185 target population individuals were extracted and converted to Plink's raw format for genome-wide association studies (GWAS) and genomic prediction.

3.6 Population Structural Analysis

Before principal component analysis (PCA), quality control was carried out on both array and imputed data using plink1.9 (Purcell et al., 2007) to discard markers that had minor allele frequency (MAF) < 0.02 and those that deviated from Hardy-Weinberg equilibrium (HWE). PCA was then carried out on the quality-controlled imputed and array data to verify if any population structure exists in the data. All the individuals were plotted in R (R Core Team, 2021) for their PCA 1 and 2 values.

3.7 Estimation of Genetic Parameters

Genetic parameters were estimated using the GCTA software (Yang et al., 2011). The genomic estimates were computed using "--reml" command option of GCTA by implementing a univariate mixed animal model. The components of the univariate mixed animal model used are described below.

$$\boldsymbol{y = \mu + Xb + Zu + e}$$

Where:

y = Vector of observed phenotype ($log_e(sea\ lice\ count\ +\ 1)$)

μ = Overall mean of $log_e(sea\ lice\ count\ +\ 1)$

X and Z = assigned design matrices to the respective vectors b and u

b = Vector of fixed effects (Interaction between Tank*counter and body weight)

u = Vector of random additive genetic effects

e = Residuals

These fixed effects (Interaction between Tank*counter and body weight) used in the model were tested against the phenotype and confirmed to be significant. The narrow sense heritability of the trait of interest (sea lice resistance) was estimated using the formula.

$$h^2 = \frac{\sigma_a^2}{\sigma_p^2}$$

Where:

h^2 = Narrow sense heritability

σ_a^2 = Additive genetic variance

σ_p^2 = Phenotypic variance

3.8 Genome-Wide Association Studies (GWAS)

Genome-wide association analysis was conducted using the GCTA software (Yang et al., 2011). This software allows to detect SNPs that explain a substantial proportion of the phenotypic variance for a complex trait. The "--mlma" command option of GCTA initiated a mixed linear animal model. The model is described below.

$$\boldsymbol{y = \mu + Xb + Zu + M\alpha + e}$$

Where:

y = Vector of observed phenotype ($log_e(sea\ lice\ count\ +\ 1)$)

μ = Overall mean of $log_e(sea\ lice\ count\ +\ 1)$

$X, and\ Z$ = Incidence matrices to the respective vectors b and u, respectively

b = Vector of fixed effects (Interaction between tank*counter and body weight)

u = Polygenic effect

M = Incidence matrix for SNP containing marker genotype coded as 0, 1 or 2

α = Allelic substitution effect of each candidate SNP

e = Residuals

The Manhattan plot was used to visualize the $-log(p-value)$ against the chromosomal position of each SNP. In order to correct for multiple testing errors and avoid declaring non-significant SNPs as significant (false positives), Bonferroni (Genome-wide and chromosome-wide threshold) correction was computed with the formula below. Also, the false discovery rate (FDR) of the top ten most significant SNPs was estimated. The accepted FDR was fixed at ≤ 0.05, while for the Bonferroni correction, SNPs whose -$log(p-value)$ estimate surpassed the computed threshold were declared significant.

$$Genome-wide\ Bonferroni\ threshold = -log_{10}\left(\frac{0.05}{N_{snps}}\right)$$

$$Chromosome-wide\ Bonferroni\ threshold = -log_{10}\left(\frac{0.05 * N_{chromosomes}}{N_{snps}}\right)$$

$$FDR = \frac{N_{snps} * \ p-value}{rank}$$

Furthermore, a diagnostic quantile-quantile plot was used to compare the relationship between the observed p-value and the expected p-value under the null hypothesis of no association. Both plots were plotted using functions from the qqman (D. Turner, 2018) package in R. The genomic inflation factor (λ) of the qq-plot, which gives insight into the spurious association, was estimated using the formula

$$lambda(\lambda) = \frac{median(x^2)}{qchisq(p = 0.5, df = 1)}$$

3.9 Genomic Prediction

For genomic prediction, families with less than ten siblings and Individuals with no phenotypes (lice count and body weight) were excluded from this analysis. This exclusion reduced the sample size to 2875 individuals and 186 families. Firstly, I ordered the samples based on family identity (FID) and assigned them into five folds, each comprising 575 individuals. Then I ordered them back based on their individual identity (IID) to have the exact ordering as in the genotype file. A 5-fold within family cross-validation genomic predictions analysis was conducted using the Bayesian generalized linear regression (Pérez & los Campos, 2014) package (BGLR) in R. The Reproducing Kernel Hilbert Space (RKHS) model option was used in BGLR to estimate breeding values. At each iteration, the phenotypes of a fold were masked and assumed unknown (validation set), while those of other folds were not masked (training set). In this way, each fish would serve as training and validation at different times. Using the RKHS model, the genomic breeding value for each masked individual in the validation fold was predicted. The model used is the same as described in section 3.7.

The matrix form of the GBLUP (VanRaden, 2008) model is represented as

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z + G^{-1}}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

Where:

X and Z = incidence matrices

y = Vector of observed phenotype ($log_e(sea\ lice\ count\ +\ 1)$)

G = Genomic relationship matrix estimated based on covariance

$\hat{b}\ and\ \hat{u}$ = Vector of estimated fixed effect and estimated breeding values

λ = Variance ratio (var_e/var_u)

3.9.1 Accuracy of Genomic Prediction.

The prediction accuracy of genomic prediction was estimated by dividing the Pearson correlation between the estimated breeding value and adjusted phenotype by the square root of heritability. This accuracy was estimated for each of the five validation folds for the array and imputed data.

Also, the accuracy of all folds was estimated by extracting and merging the estimated breeding values of validation individuals for each validation fold, which makes up predicted breeding values for all samples. The predicted breeding values for all individuals are then correlated with their true adjusted phenotypes and divided by the square root of heritability. The standard error and accuracy of the folds were estimated and reported as the accuracy of the genomic prediction analysis.

$$Accuracy = \frac{cor(EBV, y_{adj})}{\sqrt{h^2}}$$

$$SE_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Where:

EBV = Estimated breeding values

y_{adj} = Phenotype adjusted for fixed effect

h^2 = heritability

SE_r = Standard error of the correlation

r^2 = correlation square

n = number of observations

CHAPTER FOUR

4.0 RESULTS

4.1. Imputation accuracy of Whole Genome Sequence (WGS) data

The chromosome-wise averages of imputation accuracies (animal- and SNP-based) with and without pedigree information are shown in Table 3. As seen in Figure 8, including the pedigree information in the imputation did not necessarily improve the chromosome-wise SNP-based average imputation accuracy (r). In this case, the most considerable difference between average imputation accuracy with and without pedigree was observed on chromosome 23, where the average accuracy with pedigree was higher by approximately 0.02 (2%). On the other hand, the average imputation accuracy without pedigree for chromosome 26 was higher than the average with pedigree. Although, for most chromosomes, the average imputation accuracy with pedigree was higher than without pedigree, the average differences were negligible.

Table 4 shows the number of SNPs per chromosome that met the individual SNP-wise imputation accuracy threshold $(r \geq 0.6)$ with and without the inclusion of the pedigree. Although the number of SNPs that met the threshold with pedigree (3176724 SNPs) exceeded those without pedigree (3141598 SNPs), the weighted average imputation accuracy without pedigree still had the highest weighted genome-wide average of approximately 0.85. Therefore, it was adopted for reference-target population imputation. Figure 9 shows the accuracies of the SNP-based imputation without pedigree information after excluding poorly imputed SNPs. Chromosome 8 was observed to have the lowest average (0.82), while chromosome 1 had the highest average imputation accuracy (0.86).

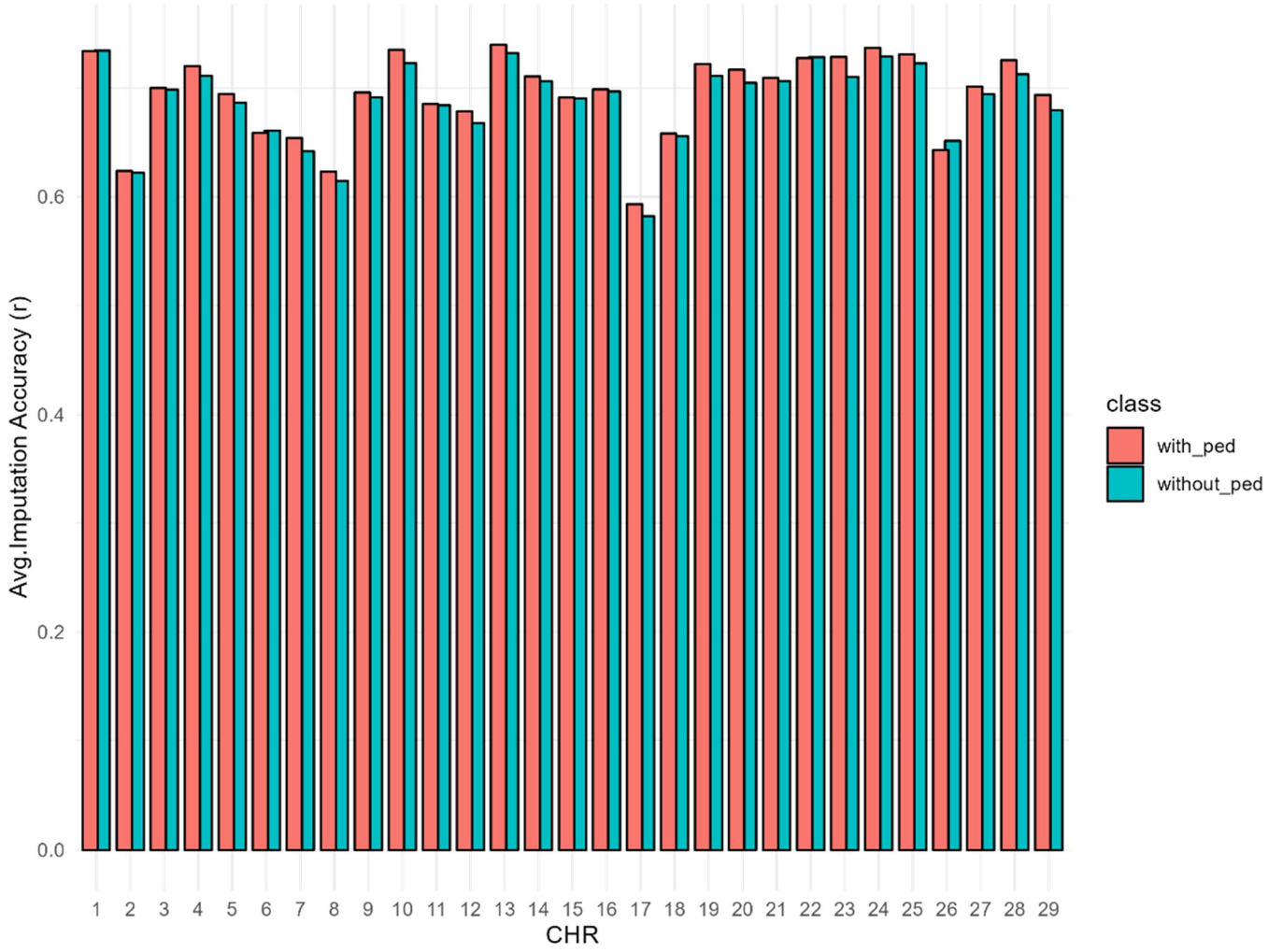


Figure 8: Bar Plot Comparing the Chromosome Averages of SNP-Based Imputation Accuracy (r) with and without Pedigree Information

Table 3: Average of SNPs and Animal-Based Imputation Accuracies (All SNPs)

Chromosomes	Avg. of Imputed SNPs Accuracy	Avg. of Imputed SNPs Accuracy (with pedigree)	Avg. of Animal-Based Accuracy	Avg. of Animal-Based Accuracy (with pedigree)
Ssa01	0.734	0.734	0.842	0.842
Ssa02	0.622	0.624	0.769	0.768
Ssa03	0.698	0.700	0.821	0.820
Ssa04	0.711	0.720	0.826	0.832
Ssa05	0.686	0.694	0.816	0.820
Ssa06	0.661	0.659	0.794	0.794
Ssa07	0.642	0.654	0.782	0.785
Ssa08	0.615	0.623	0.761	0.771
Ssa09	0.691	0.696	0.816	0.817
Ssa10	0.723	0.735	0.834	0.842
Ssa11	0.684	0.685	0.811	0.811
Ssa12	0.668	0.678	0.792	0.799
Ssa13	0.732	0.739	0.839	0.844
Ssa14	0.706	0.710	0.824	0.827
Ssa15	0.690	0.691	0.817	0.815
Ssa16	0.697	0.699	0.818	0.821
Ssa17	0.582	0.593	0.742	0.753
Ssa18	0.656	0.658	0.793	0.793
Ssa19	0.711	0.722	0.825	0.834
Ssa20	0.705	0.717	0.823	0.833
Ssa21	0.706	0.709	0.817	0.818
Ssa22	0.728	0.727	0.835	0.831
Ssa23	0.710	0.728	0.824	0.833
Ssa24	0.729	0.737	0.839	0.842
Ssa25	0.722	0.731	0.834	0.840
Ssa26	0.651	0.643	0.785	0.784
Ssa27	0.694	0.701	0.816	0.820
Ssa28	0.712	0.725	0.822	0.829
Ssa29	0.679	0.694	0.800	0.807
Weighted Average	**0.692**	**0.698**	**0.811**	**0.815**

Table 4: Averages of SNPs-Based Imputation Accuracy after Excluding Poorly Imputed SNPs

Chromosomes	Sequence	Array SNPs	Selected SNPs (r >= 0.6)	Selected SNPs (r >= 0.6) with pedigree	Average SNP accuracy (r >= 0.6)	Average SNP accuracy (r >= 0.6) with pedigree
Ssa01	315,203	3626	242209	241853	0.862	0.859
Ssa02	154,223	1586	96549	96866	0.837	0.827
Ssa03	196,962	2377	144037	144788	0.846	0.841
Ssa04	168,726	1993	126719	127886	0.850	0.851
Ssa05	156,403	2007	111493	112816	0.856	0.853
Ssa06	175,325	1936	119769	119666	0.832	0.826
Ssa07	123,674	1399	82372	84412	0.835	0.831
Ssa08	44,976	410	29014	29603	0.819	0.821
Ssa09	258,770	2738	187277	189755	0.842	0.836
Ssa10	211,792	2574	161101	164418	0.851	0.852
Ssa11	176,679	1913	125288	124784	0.852	0.847
Ssa12	189,122	1995	128315	131180	0.837	0.837
Ssa13	196,755	2499	149951	151568	0.862	0.862
Ssa14	171,528	2209	124349	125445	0.854	0.850
Ssa15	195,975	2044	140291	140565	0.841	0.834
Ssa16	151,977	1680	109846	110589	0.854	0.846
Ssa17	108,117	1154	64048	65661	0.840	0.835
Ssa18	142,757	1373	96174	96598	0.846	0.842
Ssa19	153,362	1599	112502	115062	0.855	0.856
Ssa20	162,802	2007	118256	120868	0.847	0.848
Ssa21	117,562	1184	86732	87858	0.837	0.830
Ssa22	122,674	1449	93358	93664	0.853	0.846
Ssa23	110,988	1386	82481	85395	0.846	0.845
Ssa24	93,267	1205	71355	72363	0.853	0.850
Ssa25	99,906	1170	75645	76774	0.853	0.851
Ssa26	98,126	1004	64531	62794	0.841	0.835
Ssa27	98,869	1202	70201	71100	0.848	0.845
Ssa28	87,008	1015	63665	66145	0.850	0.843
Ssa29	91,370	892	64070	66248	0.823	0.823
Total	**4374898**	**49626**	**3141598**	**3176724**	**0.848**	**0.844**

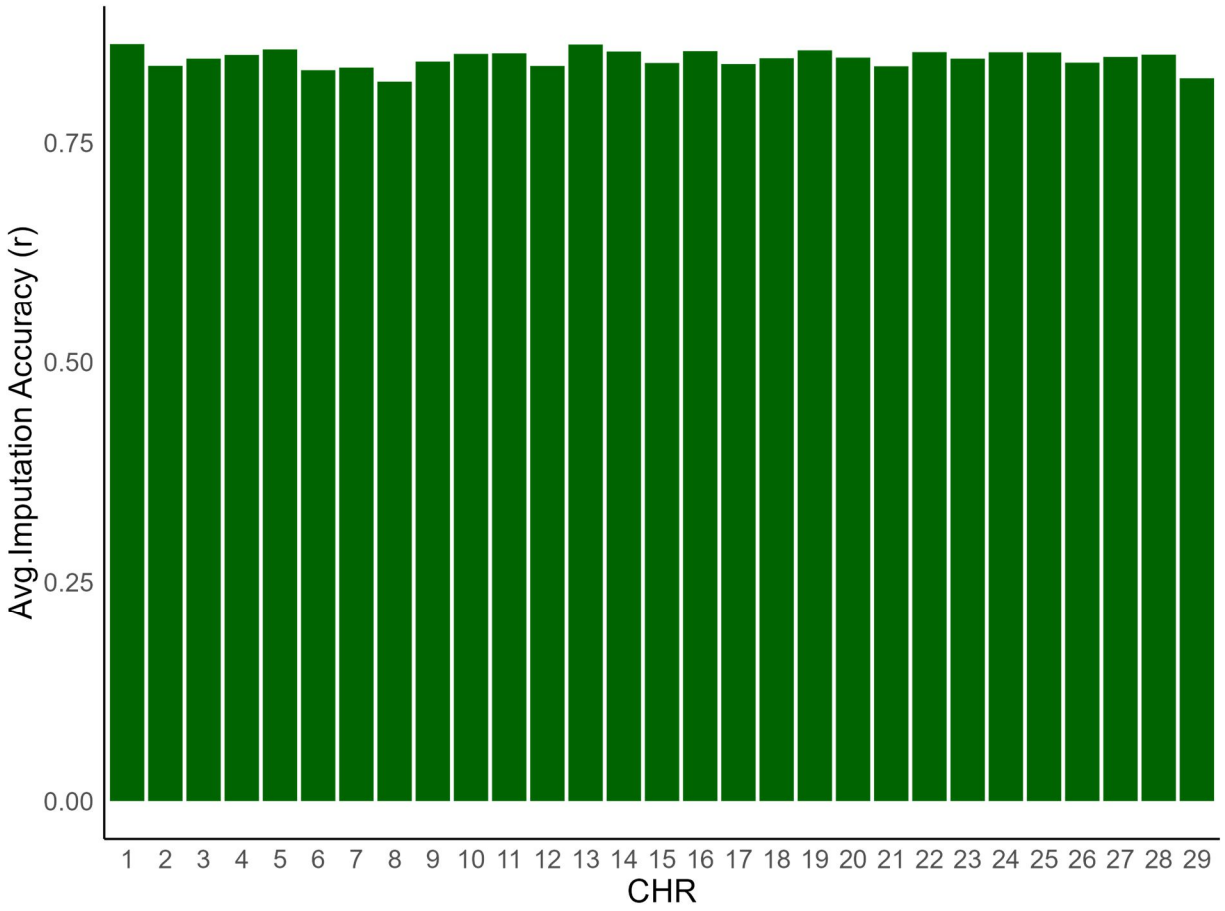


Figure 9: Bar plot showing the Chromosome-Wise Average Imputation Accuracies of SNPs used for the Reference-Target Population Imputation

4.2 Minor Allele Frequency (MAF) and Imputation Accuracy

The minor allele frequencies were divided into 25 bins from 0 to 0.5 at 0.02 intervals. The average imputation accuracies of SNPs at each MAF bin were estimated and plotted against their corresponding MAF, as shown in Figure 10. The average correlation (r) tends to increase with minor allele frequency. The least correlation was observed on chromosome 26 with a mean correlation of 0.71 at a MAF of 0.02, while the highest mean correlation was on chromosome 25 with 0.91 accuracy at 0.46 MAF. Other observed results fell between this minimum and maximum threshold.

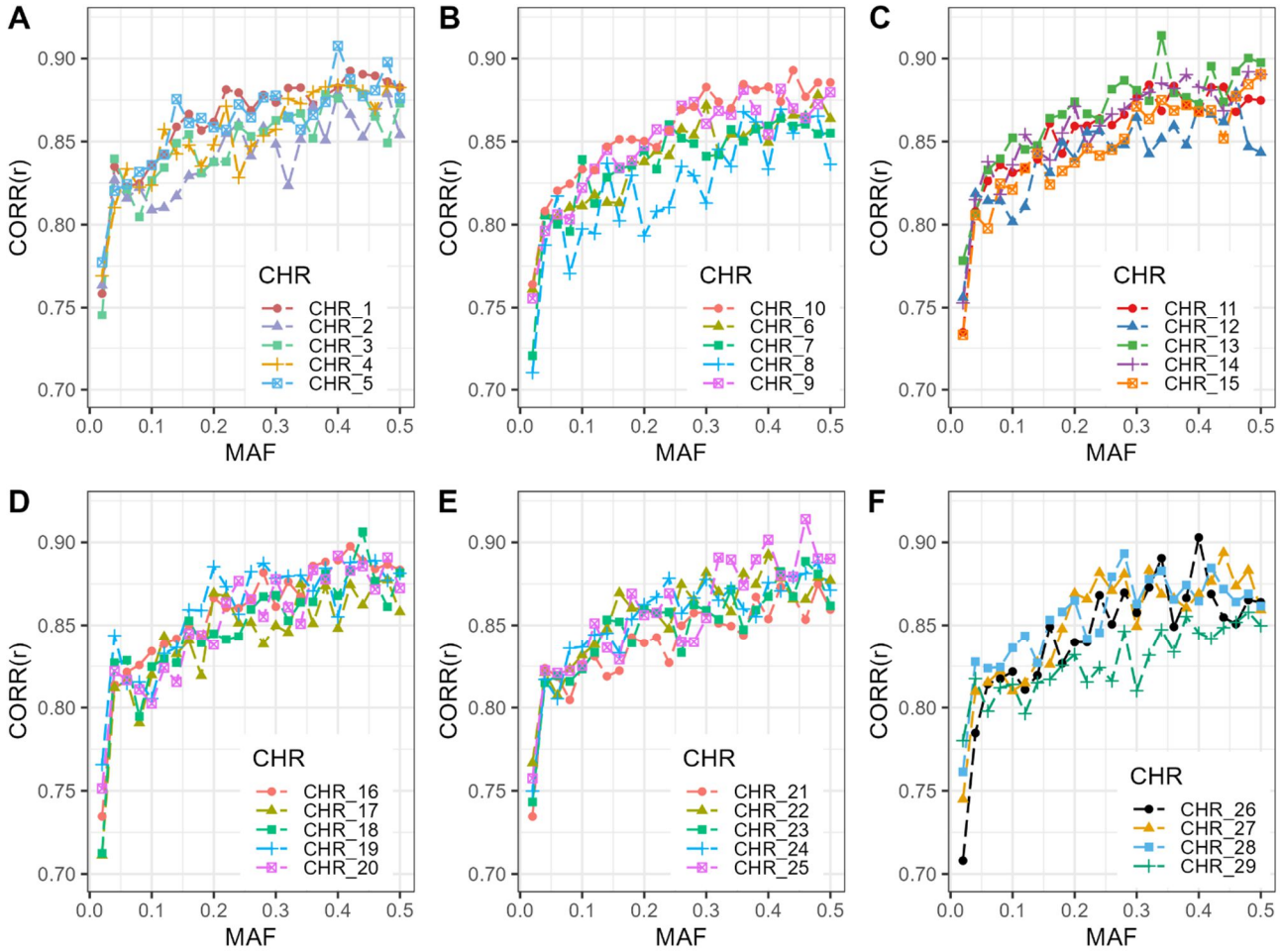


Figure 10: Plots showing Minor Allele Frequency (MAF) of Imputed SNPs against the Mean Correlation (r) of Imputation Accuracy for all Autosomal Chromosomes.

4.3 Population Structural Analysis

As shown in Figure 11, PCA 1 and 2 cumulatively explained 5.5% of the total genetic variation. It is debatable whether a distinct population structure exists in this population.

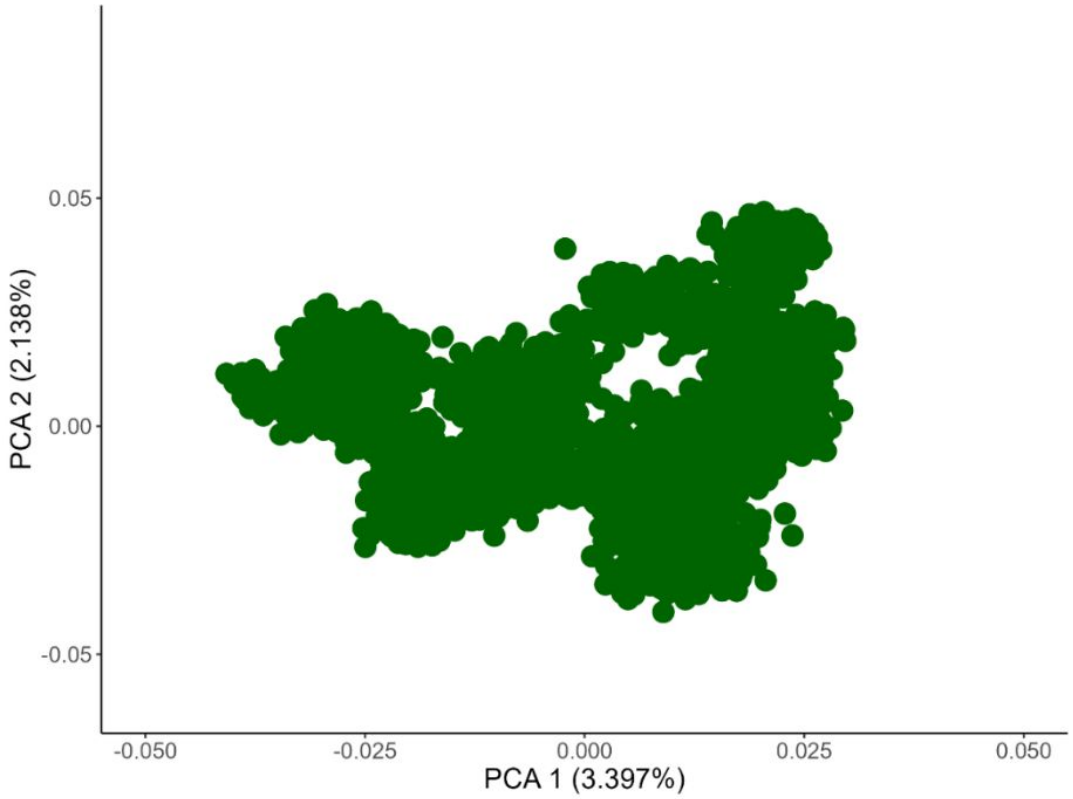


Figure 11: A Scatter Plot Showing the Population Structure of the Target Individuals

4.4 Estimation of Genetic Parameters

The obtained heritability for sea lice resistance in the studied population was estimated to be ~0.21 when Array SNPs data were used and ~0.22 when imputed SNPs data were used. The additive and residual variance estimates, alongside their standard errors, are also shown in Table 5.

Table 5: Estimates of Heritability and their Standard Errors for $log_e(Sea\ Lice\ Count\ + 1)$

Components	Imputed SNPs (3 million)	Array SNPs (50k)
$\boldsymbol{\sigma_a^2}$	0.061 ± 0.009	0.058 ± 0.008
$\boldsymbol{\sigma_e^2}$	0.218 ± 0.007	0.223 ± 0.007
$\boldsymbol{h^2}$	0.218 ± 0.028	0.206 ± 0.026

4.5 Genome-Wide Association Studies (GWAS)

The GWAS analysis results for array (50k) and imputed (~ 3 million) SNPs data showed that no SNP significantly affected the trait of interest. As seen in Figure 12, one SNP on chromosome 7 surpassed the chromosome-wide Bonferroni threshold (4.52) for the array data, but none reached the genome-wide Bonferroni threshold (5.99). On the other hand, none of the SNPs of the imputed data surpassed the Bonferroni chromosome-wide threshold of 6.26 nor the genome-wide threshold of 7.72. For the array data, QTL signals were observed on chromosomes 5, 7, 12, 16, 18, 22, and 25, while imputed data had QTL signals on chromosomes 1, 7, 12, 15, 16, and 24.

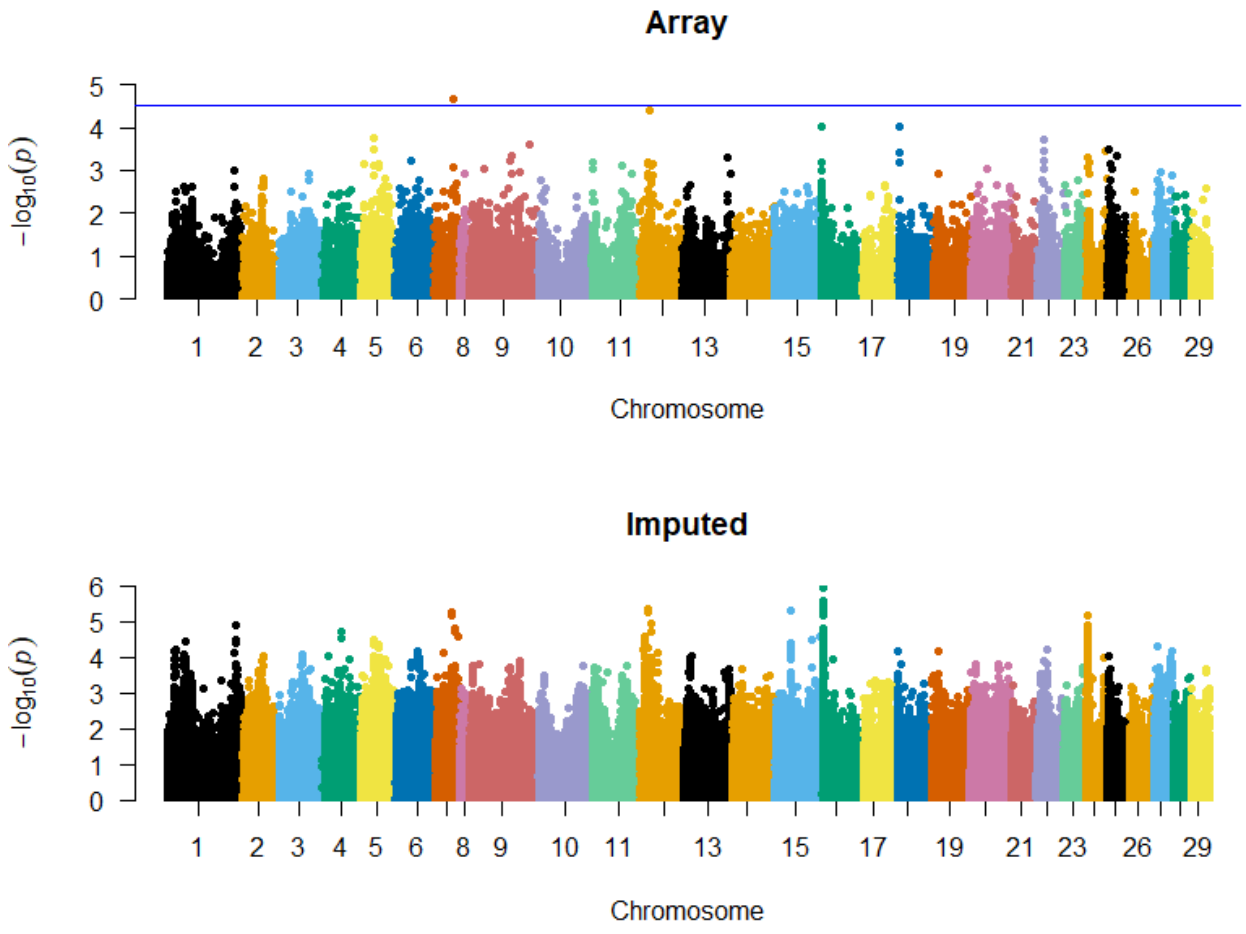


Figure 12: Manhattan Plot of the Array and Imputed Data GWAS showing the $-log(P-values)$ Distributed Across all Autosomal Chromosomes. The blue line is the chromosome-wide Bonferroni threshold.

The quantile-quantile plot of Figure 13 shows how close the observed p-values are to the expected p-values. It also confirms no significant associations since the observed p-values for the top SNPs were below the expected p-values. The genomic inflation factor (λ) values for the array and imputed SNPs were estimated to be 1.00 and 0.99, respectively. These values confirm the absence of spurious associations and significance.

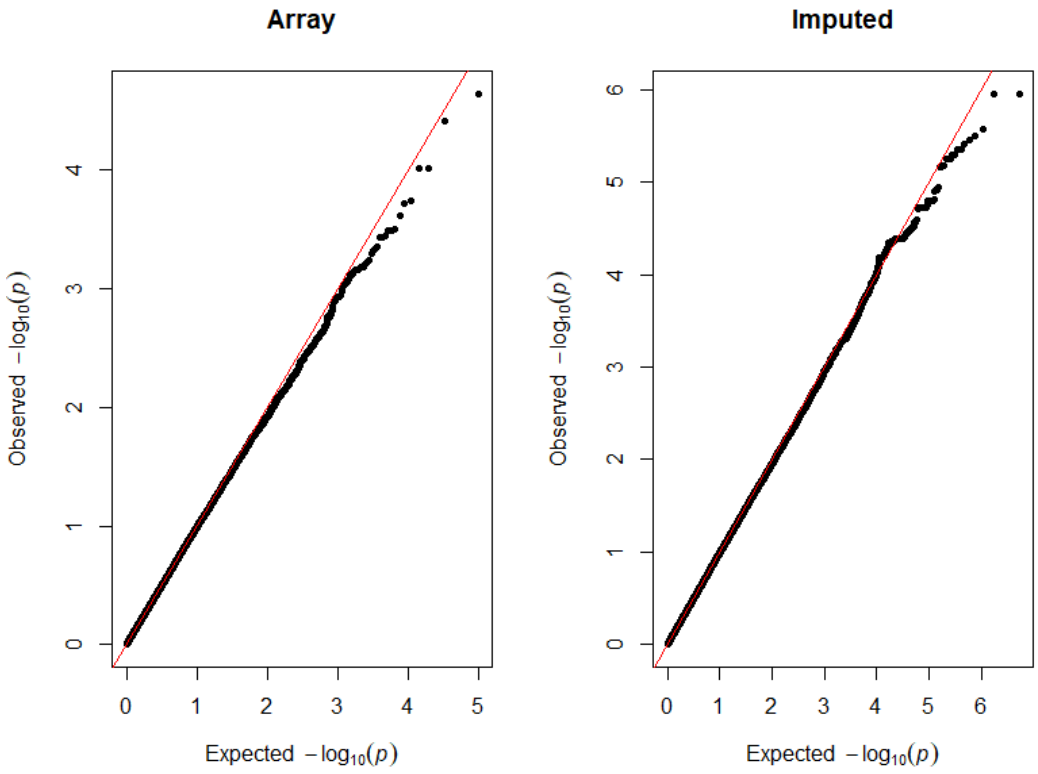


Figure 13: Quantile-Quantile Plot showing Observed against Expected $-log_{10}(p-values)$ for Array and Imputed data. The red diagonal line indicates the null hypothesis of no association.

The FDR of the top SNPs was estimated. For the array SNPs, FDR ranged from 0.926 to 1.000, while those of imputed data were all 1.000, as shown in Table 6. These values confirm that none of the array and imputed data were significant since they exceeded the accepted FDR threshold of 0.05.

Table 6: The FDR for the Top 10 SNPs According to the P-values of Array and Imputed data

Chr	SNP	A1	A2	Freq	b	se	p	FDR
Array Data								
Ssa07	HG993266.1_45156870	T	G	0.093	-0.127	0.030	2.244E-05	1.000
Ssa12	HG993271.1_26109621	T	G	0.285	0.105	0.025	3.812E-05	0.926
Ssa16	HG993275.1_2299421	A	G	0.393	-0.076	0.019	9.519E-05	1.000
Ssa18	HG993277.2_4439352	T	C	0.383	-0.068	0.018	9.543E-05	1.000
Ssa05	HG993264.1_31252801	C	T	0.074	0.139	0.037	1.800E-04	1.000
Ssa22	HG993281.1_16838003	G	T	0.479	0.069	0.018	1.906E-04	1.000
Ssa09	HG993268.2_140287766	T	G	0.474	0.070	0.019	2.431E-04	1.000
Ssa25	HG993284.1_5222628	C	T	0.385	0.063	0.018	3.141E-04	1.000
Ssa05	HG993264.1_32056685	T	C	0.235	-0.081	0.023	3.187E-04	1.000
Ssa05	HG993264.1_32056935	T	C	0.095	0.121	0.034	3.225E-04	1.000
Imputed Data								
Ssa16	HG993275.1_2670854	T	G	0.418	0.089	0.018	1.095E-06	1.000
Ssa16	HG993275.1_2652565	T	C	0.419	0.089	0.018	1.096E-06	1.000
Ssa16	HG993275.1_2650283	C	T	0.424	0.086	0.018	2.676E-06	1.000
Ssa16	HG993275.1_2695572	C	A	0.487	0.085	0.018	3.187E-06	1.000
Ssa16	HG993275.1_2666638	T	C	0.424	0.085	0.018	3.535E-06	1.000
Ssa16	HG993275.1_2695854	A	G	0.487	0.084	0.018	3.841E-06	1.000
Ssa16	HG993275.1_2698184	G	C	0.424	0.084	0.018	4.375E-06	1.000
Ssa12	HG993271.1_20073710	T	C	0.287	0.083	0.018	4.399E-06	1.000
Ssa15	HG993274.1_35023984	A	G	0.085	0.144	0.032	4.988E-06	1.000
Ssa16	HG993275.1_2671264	G	T	0.424	0.084	0.018	5.002E-06	1.000

4.6 Accuracy of Genomic Prediction

The accuracy of genomic prediction for each of the 5-fold cross-validation and all folds of the array and imputed data are shown in Table 7. For the array and the imputed data, individuals in the validation fold three had the best prediction accuracy of 0.721 and 0.715, respectively. In comparison, the prediction accuracy for fold-five individuals was the lowest, with a prediction accuracy of 0.578 and 0.569, respectively. Overall, using more SNPs (imputed data) with the GBLUP method did not improve the accuracy of genomic prediction.

Table 7: Accuracy of Genomic Prediction and their Standard Error

	Fold 1±SE_r	**Fold 2±SE_r**	**Fold 3±SE_r**	**Fold 4±SE_r**	**Fold 5±SE_r**	**All Folds±SE_r**
Array (50k)	0.676±0.018	0.654±0.018	0.722±0.018	0.605±0.018	0.578±0.018	0.645±0.018
Imputed (3 million)	0.676±0.018	0.663±0.018	0.715±0.018	0.587±0.018	0.569±0.018	0.641±0.018

CHAPTER FIVE

5.0 DISCUSSION

5.1 Genotype Imputation Accuracies

This research reported a weighted average genome-wide imputation accuracy (r) of 0.84 when pedigree information was included and 0.85 without pedigree information. Manousi (2021) assessed the effect of the new Atlantic salmon genome assembly on imputation accuracy. Using the new and old genome assemblies, the researcher compared imputation accuracies with Beagle (Browning & Browning, 2009) and Fimpute3 (Sargolzaei et al., 2014). An imputation accuracy (r^2) of 0.83 was reported for Fimpute3 (with pedigree). This accuracy would translate to a Pearson correlation coefficient (r) value of 0.91 which is higher than the 0.84 (with pedigree) found here. The modest reference population size (n = 197) used in this study and high imputation proportion (a factor of 80) when compared to Manousi (2021) study (1300 samples, imputation from 44k to 400k, which translates to a factor of x10) could explain the differences in accuracies reported.

Yoshida et al., (2018) reported an imputation accuracy (r) ranging between 0.74 to 0.98, having tested different imputation scenarios. They performed genotype imputation to 50k SNPs using Fimpute2.2 software (Sargolzaei et al., 2014) with varying SNP densities (500, 3k, 6k) and varying numbers of reference and validation animals in the Atlantic salmon population. Our findings agree with the imputation accuracy range found by Yoshida et al., (2018), although the sample size and number of SNPs used were lower.

Furthermore, Kijas et al., (2017) used a multi-generation reference population of Tasmanian Atlantic salmon to carry out imputation from 5k to 78k. They reported a high genotype imputation accuracy of between 0.89 – 0.97, while Tsai et al., (2017) reported an imputation accuracy (r) of 0.62 to 0.90 in UK-farmed salmon.

The imputation accuracies reported in all these salmon studies were relatively high, further confirming the relevance of genotype imputation in saving costs relating to high-density genotyping or re-sequencing of a large number of animals in the aquaculture industry.

5.2 Relationship between MAF and Imputation Accuracy

The relationship between MAF and imputation accuracy was observed by dividing SNPs into 25 bins according to their MAF. The imputation accuracy is the average correlation (r) between the true and the imputed genotypes for SNPs belonging to a particular bin. Imputation accuracy increased with an increase in minor allele frequency (MAF), which corresponds with the findings of Tsai et al., (2017), Jiang et al., (2022), and Pausch et al., (2013), who reported an increasing imputation accuracy for known variants.

Although it was observed that imputation accuracy slightly dropped at some bins as MAF increased, this could be due to the low number of SNPs in those bins resulting in sampling errors

5.3 Genetic Parameters

The heritability of sea lice resistance in the population of Norwegian Atlantic salmon studied in this research was estimated to be 0.21 and 0.22 for array and imputed data, respectively. These findings are consistent with the reports of the recent research of Tsairidou et al., (2020), where a heritability estimate of 0.19 for sea lice resistance using a high-density SNP panel was reported. Also, a previous study by Tsai et al., (2016) reported heritability estimates of 0.22 and 0.33 for sea lice resistance after studying two populations of salmon.

Gjerde et al., (2011) reported a heritability estimate of 0.33 for SalmoBreed population, while Gharbi et al., (2015) reported a heritability of 0.30 for the same trait in a Scottish salmon population. Furthermore, Rochus et al., (2018) reported an estimated heritability of 0.29 when lice count phenotype data were log-transformed and 0.17 when it was not. Some low heritabilities have also been reported for host resistance to sea lice. Correa et al., (2017) reported an estimated heritability of 0.12, while Kjetså et al., (2020) and Odegård et al., (2014) estimated the heritability of sea lice resistance to be 0.14.

The differences observed in the heritability estimated and reported by various researchers could be due to the species of salmonoid and sea lice studied, phenotype transformation, the difference in population or year class, the type of model used, and the type of challenge test (land-based or sea cages) carried out and pedigree versus genomic estimates. All heritability estimates reported in these various studies fall in the low to moderate heritability range and therefore suggest that the trait of interest can be improved by selective breeding.

5.4 Genome-Wide Association Studies (GWAS)

Findings of the association test using the array (50k) and imputed (3 million) SNPs data in this research confirm the polygenic nature of sea lice resistance trait. Although no SNP was identified as significant at the genome-wide level for both scenarios, one SNP on chromosome 7 of the array was observed to have a chromosome-wide significance. Array and Imputed data had strong signals on chromosomes 7, 12, and 16. If these regions are studied, they might harbor putative genes that affect sea lice resistance. Our result of no significant genome-wide QTL for sea lice resistance agrees with the findings of Correa et al., (2017) and Tsai et al., (2016), respectively.

On the other hand, Cáceres et al., (2022), Robledo et al., (2019), and Rochus et al., (2018) reported to have found significant QTLs associated with host resistance to sea lice. Cáceres et al., (2022) performed a meta-GWAS analysis for sea lice (*Caligus rogercresseyi*) load in Atlantic salmon using over 6000 samples from four-year classes and reported the detection of highly associated regions on chromosomes 3 and 12. He concluded that the high experimental power due to the combination of several-year classes is advantageous in identifying these QTLs.

Robledo et al., (2019) also found three single QTLs significant for host resistance to sea lice (*Caligus rogercresseyi*) and explain about 4% of genetic variation. Rochus et al., (2018) used the forward multiple linear regression and a mixed linear model and detected QTLs on different chromosomes. The mixed linear model identified two QTLs located on chromosome 1 and 23, respectively, while the other model identified 70 SNPs, many of which may be due to not correcting for population structure.

The differences in reports could be due to the species of sea lice studied, as those who found association studied *Caligus rogercresseyi*, the population of salmon studied, population structure, and the type of challenge test (land-based or sea cages). Also, the genomic inflation values (λ) of these studies were not reported; therefore, the associations reported could have been due to spurious associations.

5.5 Accuracy of Genomic Prediction

This study reported the accuracy of a 5-fold within-family cross-validation scheme for genomic prediction. The assessment of each fold showed that for the array and the imputed data, individuals in validation fold three had the best prediction accuracy of 0.722 and 0.715, respectively. In contrast, individuals in fold five had the lowest prediction accuracy of 0.578 and 0.569. The variation observed in the prediction accuracy across folds is due to differences in the number of sibs per family. Some families had more sibs in the training fold than others; therefore, the breeding value of their sibs in the validation set is predicted more accurately. This finding agrees with the conclusion of Fraslin et al., (2022) when they studied the impact of the genetic relationship between training and validation population in Atlantic salmon. Fraslin et al., (2022) concluded that a close genetic relationship between training and validation individuals enhances the accuracy of genomic prediction.

The accuracy of genomic prediction for all validation with array (~50k) and imputed (3 million) data was estimated to be 0.645 ± 0.018 and 0.641 ± 0.018, respectively. Our finding agrees with Kjetså et al., (2020), who reported a prediction accuracy of 0.671 and 0.669 for sea lice resistance when 215k and 750k SNPs were used. Although, it is essential to note that the heritability estimate reported by Kjetså et al., (2020) was way below what was estimated in this study. The massive difference in the SNP density did not improve the prediction accuracy of the GBLUP model, thus corresponding to the report of Tsai et al., (2016), where they stated that medium SNP densities are sufficient to achieve maximum genomic prediction accuracy.

The potential of high SNP densities (sequencing data) can be maximized to improve the accuracy of genomic prediction. The Bayesian variables (Bayes B, C, etc.) models should be adopted to achieve this. This is because, unlike the GBLUP method, which assumes that all markers have equal effects with constant variance across the genome, Bayesian methods work with a prior assumption that few markers have significant effects while others have no effect. This assumption is more realistic; thus, the Bayesian variables methods may outperform the GBLUP method when using a high-density marker.

Tsai et al., (2017) carried out 5-fold cross-validation and reported an estimated genomic prediction accuracy of 0.58 and 0.60 for sea lice resistance when using imputed and true genotypes. Also, Fraslin et al., (2022) reported an estimated prediction accuracy of 0.49 (2014

population) and 0.39 (2010 population) for sea lice count. Tsai et al., (2017) and Fraslin et al., (2022) estimated the accuracy by dividing the correlation between the predicted genomic breeding values and the phenotype by the square root of heritability. In contrast, this study estimated accuracy by dividing the correlation between predicted genomic breeding values and the adjusted phenotypes by the square root of heritability. This formula using the adjusted phenotype for fixed effects rather than the unadjusted phenotype is better because the unadjusted phenotype consists of fixed, random effects of animals and residuals. Therefore, estimating accuracy by calculating the correlation between estimated breeding values and phenotypes gives a lower accuracy estimate. Using the adjusted phenotype gives a better accuracy as it gets us closer to estimating the effect of markers.

CHAPTER SIX

6.0 CONCLUSION AND RECOMMENDATION

6.1 Conclusion

In conclusion, this study's accuracy of genotype imputation was relatively good, but including pedigree information in genotype imputation did not significantly improve its imputation accuracy. Host resistance to sea lice is a moderately heritable trait that can be improved with selective breeding. The high signals observed across the chromosomes with no significant associated QTL detected confirms the polygenic nature of host resistance to sea lice. Lastly, the 50k SNP data for this study was sufficient to conduct GWAS analysis and accurately predict genomic breeding values.

6.2 Recommendations

A meta-analysis combining salmon populations should be used to increase the power of GWAS analysis. Bayes B genomic prediction method may improve the prediction accuracy of imputed data.

References

Aaen, S. M., Helgesen, K. O [Kari Olli], Bakke, M. J., Kaur, K., & Horsberg, T. E [Tor Einar] (2015). Drug resistance in sea lice: A threat to salmonid aquaculture. *Trends in Parasitology*, *31*(2), 72–81. https://doi.org/10.1016/j.pt.2014.12.006

Altshuler, D., & Daly, M. (2007). Guilt beyond a reasonable doubt. *Nature Genetics*, *39*(7), 813–815. https://doi.org/10.1038/ng0707-813

Artic Seafood Exports. (2023). *ATLANTIC SALMON LIFE CYCLE*. https://arcticseafoodexport.com/atlantic-salmon-life-cycle/

Badke, Y. M., Bates, R. O., Ernst, C. W., Schwab, C., Fix, J., van Tassell, C. P., & Steibel, J. P. (2013). Methods of tagSNP selection and other variables affecting imputation accuracy in swine. *BMC Genetics*, *14*, 8. https://doi.org/10.1186/1471-2156-14-8

Barrett, L. T., Oppedal, F., Robinson, N., & Dempster, T. (2020). Prevention not cure: a review of methods to avoid sea lice infestations in salmon aquaculture. *Reviews in Aquaculture*, *12*(4), 2527–2543. https://doi.org/10.1111/raq.12456

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benner, C., Spencer, C. C. A., Havulinna, A. S., Salomaa, V., Ripatti, S., & Pirinen, M. (2016). Finemap: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, *32*(10), 1493–1501. https://doi.org/10.1093/bioinformatics/btw018

Berland, B., & Margolis, L. (1983). The early history of 'Lakselus' and some nomenclatural questions relating to copepod parasites of salmon. *Sarsia*, *68*(4), 281–288. https://doi.org/10.1080/00364827.1983.10420582

Biologywise. (2023). *Salmon Life Cycle*. https://biologywise.com/salmon-life-cycle

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boxaspen, K. (2006). A review of the biology and genetics of sea lice. *ICES Journal of Marine Science*, *63*(7), 1304–1316. https://doi.org/10.1016/j.icesjms.2006.04.017

Bradbury, P. J., Zhang, Z [Zhiwu], Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). Tassel: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics*, *103*(3), 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., . . . Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005-D1012. https://doi.org/10.1093/nar/gky1120

Cáceres, P., Lopéz, P., Garcia, B., Cichero, D., Ødegård, J [J.], Moen, T [T.], & Yáñez, J. M [J. M.]. (2022). *Meta-analysis of GWAS for sea lice load in Atlantic salmon.* https://doi.org/10.1101/2022.09.28.509902

Calus, M. P. L [M. P. L.], Bouwman, A. C., Hickey, J. M [J. M.], Veerkamp, R. F., & Mulder, H. A. (2014). Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: A review of livestock applications. *Animal : An International Journal of Animal Bioscience*, *8*(11), 1743–1753. https://doi.org/10.1017/S1751731114001803

Carvalheiro, R., Boison, S. A., Neves, H. H. R., Sargolzaei, M., Schenkel, F. S., Utsunomiya, Y. T., O'Brien, A. M. P., Sölkner, J., McEwan, J. C., van Tassell, C. P., Sonstegard, T. S., & Garcia, J. F. (2014). Accuracy of genotype imputation in Nelore cattle. *Genetics Selection Evolution*, *46*(1), 69. https://doi.org/10.1186/s12711-014-0069-1

Chanda, P., Yuhki, N., Li, M., Bader, J. S., Hartz, A., Boerwinkle, E., Kao, W. H. L., & Arking, D. E. (2012). Comprehensive evaluation of imputation performance in African Americans. *Journal of Human Genetics*, *57*(7), 411–421. https://doi.org/10.1038/jhg.2012.43

Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A., & Schaid, D. J. (2015). Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics*, *200*(3), 719–736. https://doi.org/10.1534/genetics.115.176107

Contreras, M., Karlsen, M., Villar, M., Olsen, R. H., Leknes, L. M., Furevik, A., Yttredal, K. L., Tartor, H., Grove, S., Alberdi, P., Brudeseth, B., & La Fuente, J. de (2020). Vaccination with Ectoparasite Proteins Involved in Midgut Function and Blood Digestion Reduces Salmon Louse Infestations. *Vaccines*, *8*(1). https://doi.org/10.3390/vaccines8010032

Correa, K., Lhorente, J. P., Bassini, L., López, M. E., Di Genova, A., Maass, A., Davidson, W. S., & Yáñez, J. M [José M.] (2017). Genome wide association study for resistance to Caligus rogercresseyi in Atlantic salmon (Salmo salar L.) using a 50K SNP genotyping array. *Aquaculture*, *472*, 61–65. https://doi.org/10.1016/j.aquaculture.2016.04.008

Costello, M. J. (2006). Ecology of sea lice parasitic on farmed and wild fish. *Trends in Parasitology*, *22*(10), 475–483. https://doi.org/10.1016/j.pt.2006.08.006

Costello, M. J. (2009). The global economic cost of sea lice to the salmonid farming industry. *Journal of Fish Diseases*, *32*(1), 115–118. https://doi.org/10.1111/j.1365-2761.2008.01011.x

D. Turner, S. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Journal of Open Source Software*, *3*(25), 731. https://doi.org/10.21105/joss.00731

Daetwyler, H. D [Hans D.], Calus, M. P. L [Mario P. L.], Pong-Wong, R., los Campos, G. de, & Hickey, J. M [John M.] (2013). Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics*, *193*(2), 347–365. https://doi.org/10.1534/genetics.112.147983

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., . . . Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Deng, T., Zhang, P., Garrick, D., Gao, H [Huijiang], Wang, L., & Zhao, F. (2021). Comparison of Genotype Imputation for SNP Array and Low-Coverage Whole-Genome Sequencing Data. *Frontiers in Genetics*, *12*, 704118. https://doi.org/10.3389/fgene.2021.704118

Directory of Fisheries (2021). Key figures from Norwegian Aquaculture Industry 2021.

Druet, T., Schrooten, C., & Roos, A. P. W. de (2010). Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *Journal of Dairy Science*, *93*(11), 5443–5454. https://doi.org/10.3168/jds.2010-3255

Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., Mason, B. A., & Goddard, M. E [M. E.] (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, *95*(7), 4114–4129. https://doi.org/10.3168/jds.2011-5019

FAO (2016). The State of Food and Agriculture (SOFA): Climate change, agriculture and food security. https://www.fao.org/3/i6030e/i6030e.pdf

FAO (2019). FAO Aquaculture Newsletter. No. 60 (August). Rome. https://www.fao.org/documents/card/en/c/ca5223en/

FAO. (2022). *The State of World Fisheries and Aquaculture 2022*. FAO. https://doi.org/10.4060/cc0461en

Fraslin, C., Yáñez, J. M [José M.], Robledo, D., & Houston, R. D. (2022). The impact of genetic relationship between training and validation populations on genomic prediction accuracy in Atlantic salmon. *Aquaculture Reports*, *23*, 101033. https://doi.org/10.1016/j.aqrep.2022.101033

Gharbi, K., Matthews, L., Bron, J., Roberts, R., Tinch, A., & Stear, M. (2015). The control of sea lice in Atlantic salmon by selective breeding. *Journal of the Royal Society, Interface*, *12*(110), 574. https://doi.org/10.1098/rsif.2015.0574

Gjerde, B., Ødegård, J [Jørgen], & Thorland, I. (2011). Estimates of genetic variation in the susceptibility of Atlantic salmon (Salmo salar) to the salmon louse Lepeophtheirus salmonis. *Aquaculture*, *314*(1-4), 66–72. https://doi.org/10.1016/j.aquaculture.2011.01.026

Goddard, M. E [M. E.], & Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics = Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie*, *124*(6), 323–330. https://doi.org/10.1111/j.1439-0388.2007.00702.x

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*, 186. https://doi.org/10.1186/1471-2105-12-186

Hamre, L. A., Eichner, C., Caipang, C. M. A., Dalvin, S. T., Bron, J. E., Nilsen, F., Boxshall, G., & Skern-Mauritzen, R. (2013). The Salmon Louse Lepeophtheirus salmonis (Copepoda: Caligidae) life cycle has only two Chalimus stages. *PLoS ONE*, *8*(9), e73539. https://doi.org/10.1371/journal.pone.0073539

Helgesen, K. O [K. O.], Bravo, S., Sevatdal, S [S.], Mendoza, J., & Horsberg, T. E [T. E.] (2014). Deltamethrin resistance in the sea louse Caligus rogercresseyi (Boxhall and Bravo) in Chile: Bioassay results and usage data for antiparasitic agents with references to Norwegian conditions. *Journal of Fish Diseases*, *37*(10), 877–890. https://doi.org/10.1111/jfd.12223

Hickey, J. M [John M.], Crossa, J., Babu, R., & los Campos, G. de (2012). Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. *Crop Science*, *52*(2), 654–663. https://doi.org/10.2135/cropsci2011.07.0358

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, *5*(6), e1000529. https://doi.org/10.1371/journal.pgen.1000529

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, *82*(11), 801–811. https://doi.org/10.1016/j.humimm.2021.02.012

Illumina. *Next-Generation Sequencing for Beginners.* https://www.illumina.com/science/technology/next-generation-sequencing/beginners.html

Inland Fisheries Ireland. *Atlantic salmon (Salmo salar)*. https://www.fisheriesireland.ie/species/atlantic-salmon-salmo-salar

Institute of Marine Research. (2020). *Salmon – Atlantic*. https://www.hi.no/en/hi/temasider/species/salmon--atlantic

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

Jacques Drolet. (2015). *Sea LICE RISK REDUCTION PROGRAM.* https://doi.org/10.13140/RG.2.1.2254.5766

Jiang, Y., Song, H., Gao, H [Hongding], Zhang, Q., & Ding, X. (2022). Exploring the optimal strategy of imputation from SNP array to whole-genome sequencing data in farm animals. *Frontiers in Genetics*, *13*, 963654. https://doi.org/10.3389/fgene.2022.963654

Johnson, S. C., & Albright, L. J. (1991). The developmental stages of Lepeophtheirus salmonis (Krøyer, 1837) (Copepoda: Caligidae). *Canadian Journal of Zoology*, *69*(4), 929–950. https://doi.org/10.1139/z91-138

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*(3), 1709–1723. https://doi.org/10.1534/genetics.107.080101

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., & Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics*, *10*(10), e1004722. https://doi.org/10.1371/journal.pgen.1004722

Kijas, J., Elliot, N., Kube, P., Evans, B., Botwright, N., King, H., Primmer, C. R., & Verbyla, K. (2017). Diversity and linkage disequilibrium in farmed Tasmanian Atlantic salmon. *Animal Genetics*, *48*(2), 237–241. https://doi.org/10.1111/age.12513

Kjetså, M. H., Ødegård, J [J.], & Meuwissen, T [T.H.E.] (2020). Accuracy of genomic prediction of host resistance to salmon lice in Atlantic salmon (Salmo salar) using imputed high-density genotypes. *Aquaculture*, *526*, 735415. https://doi.org/10.1016/j.aquaculture.2020.735415

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. http://arxiv.org/pdf/1303.3997v2

Lin, P., Hartz, S. M., Zhang, Z [Zhehao], Saccone, S. F., Wang, J., Tischfield, J. A., Edenberg, H. J., Kramer, J. R., M Goate, A., Bierut, L. J., & Rice, J. P. (2010). A new statistic to evaluate imputation reliability. *PLoS ONE*, *5*(3), e9697. https://doi.org/10.1371/journal.pone.0009697

Liu, Y., Olaf Olaussen, J., & Skonhoft, A. (2011). Wild and farmed salmon in Norway—A review. *Marine Policy*, *35*(3), 413–418. https://doi.org/10.1016/j.marpol.2010.11.007

Manousi, D. (2021). Assessing the effects of the new Atlantic salmon (Salmo salar) genome assembly on imputation accuracy. https://nmbu.brage.unit.no/nmbu-xmlui/handle/11250/2772899

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry (Palo Alto, Calif.)*, *6*, 287–303. https://doi.org/10.1146/annurev-anchem-062012-092628

Marine Institute. (2022a). *Life cycle of the Salmon Louse*. https://www.marine.ie/site-area/areas-activity/aquaculture/sea-lice/life-cycle-salmon-louse

Marine Institute. (2022b). *Salmon Life Cycle*. https://www.marine.ie/site-area/areas-activity/fisheries-ecosystems/salmon-life-cycle

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(2), 560–564. https://doi.org/10.1073/pnas.74.2.560

Meuwissen, T., Hayes, B. J., & Goddard, M. E [M. E.] (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Meuwissen, T., Hayes, B., & Goddard, M [Mike] (2013). Accelerating improvement of livestock with genomic selection. *Annual Review of Animal Biosciences*, *1*, 221–237. https://doi.org/10.1146/annurev-animal-031412-103705

Meuwissen, T., Hayes, B., & Goddard, M [Mike] (2016). Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers*, *6*(1), 6–14. https://doi.org/10.2527/af.2016-0002

Myhre Jensen, E., Horsberg, T. E [Tor Einar], Sevatdal, S [Sigmund], & Helgesen, K. O [Kari Olli] (2020). Trends in de-lousing of Norwegian farmed salmon from 2000–2019—Consumption of medicines, salmon louse resistance and non-medicinal control methods. *PLOS ONE*, *15*(10), e0240894. https://doi.org/10.1371/journal.pone.0240894

Norwegian Directorate of Fisheries. (2017). *Sale of Farmed Cleaner Fish 2012–2016*. https://www.fiskeridir.no/English/Aquaculture/Statistics/Cleanerfish-Lumpfish-and-Wrasse

Norwegian Seafood Company. (2019). *Atlantic Salmon – Salmo Salar*. https://norseaco.no/atlantic-salmon/

Norwegian SeaFood Council. (2023). *Norway's seafood exports worth NOK 151.4 billion in 2022*. https://en.seafood.no/news-and-media/news-archive/norways-seafood-exports-worth-nok-151.4-billion-in-2022/

O'Connor, B. D., & van der Auwera, G. (2017). *Genomics analysis with Spark, Docker, and clouds: A guide to big data tools for genomics research* (First edition). O'Reilly. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2420556

Odegård, J., Moen, T [Thomas], Santi, N., Korsvoll, S. A., Kjøglum, S., & Meuwissen, T. H. E. (2014). Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). *Frontiers in Genetics*, *5*, 402. https://doi.org/10.3389/fgene.2014.00402

Øverli, Ø., Nordgreen, J., Mejdell, C. M., Janczak, A. M., Kittilsen, S., Johansen, I. B., & Horsberg, T. E [Tor E.] (2014). Ectoparasitic sea lice (Lepeophtheirus salmonis) affect behavior and brain serotonergic activity in Atlantic salmon (Salmo salar L.): Perspectives on animal welfare. *Physiology & Behavior*, *132*, 44–50. https://doi.org/10.1016/j.physbeh.2014.04.031

Overton, K., Dempster, T., Oppedal, F., Kristiansen, T. S., Gismervik, K., & Stien, L. H. (2019). Salmon lice treatments and salmon mortality in Norwegian aquaculture: a review. *Reviews in Aquaculture*, *11*(4), 1398–1417. https://doi.org/10.1111/raq.12299

Pausch, H., Aigner, B., Emmerling, R., Edel, C., Götz, K.-U., & Fries, R. (2013). Imputation of high-density genotypes in the Fleckvieh cattle population. *Genetics Selection Evolution*, *45*(1), 3. https://doi.org/10.1186/1297-9686-45-3

Pérez, P., & los Campos, G. de (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, *198*(2), 483–495. https://doi.org/10.1534/genetics.114.164442

Pike, A. W., and Wadsworth, S. L. (Ed.). (1999). *Advances in Parasitology*. *Sealice on Salmonids: Their Biology and Control*. Advances in Parasitology Volume 44. Elsevier.

Pook, T., Mayer, M., Geibel, J., Weigend, S., Cavero, D., Schoen, C. C., & Simianer, H. (2019). *Improving imputation quality in BEAGLE for crop and livestock data.* https://doi.org/10.1101/577338

Pryce, J. E., & Daetwyler, H. D [H. D.] (2012). Designing dairy cattle breeding schemes under genomic selection: a review of international research. *Animal Production Science*, *52*(3), 107. https://doi.org/10.1071/AN11098

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., Bakker, P. I. W. de, Daly, M. J., & Sham, P. C. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

R Core Team. (2021). *R: A Language and Environment for Statistical Computing* (Version 4.1.2-foss-2021b) [Computer software]. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ramnarine, S., Zhang, J., Chen, L.-S., Culverhouse, R., Duan, W., Hancock, D. B., Hartz, S. M., Johnson, E. O., Olfson, E., Schwantes-An, T.-H., & Saccone, N. L. (2015). When Does Choice of Accuracy Measure Alter Imputation Accuracy Assessments? *PLoS ONE*, *10*(10), e0137601. https://doi.org/10.1371/journal.pone.0137601

Robledo, D., Gutiérrez, A. P., Barría, A., Lhorente, J. P., Houston, R. D., & Yáñez, J. M [José M.] (2019). Discovery and Functional Annotation of Quantitative Trait Loci Affecting Resistance to Sea Lice in Atlantic Salmon. *Frontiers in Genetics*, *10*, 56. https://doi.org/10.3389/fgene.2019.00056

Rochus, C. M., Holborn, M. K., Ang, K. P., Elliott, J. A. K., Glebe, B. D., Leadbeater, S., Tosh, J. J., & Boulding, E. G. (2018). Genome-wide association analysis of salmon lice (Lepeophtheirus salmonis) resistance in a North American Atlantic salmon population. *Aquaculture Research*, *49*(3), 1329–1338. https://doi.org/10.1111/are.13592

Ron Tardiff. (2019). *The current state of sea lice management*. https://www.asc-aqua.org/the-current-state-of-sea-lice-management/

Salmon Facts. (2022). *The Life Cycle of Wild Atlantic Salmon*. https://salmonfacts.org/life-cycle-of-wild-atlantic-salmon/?utm_content=cmp-true

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sargolzaei, M., Chesnais, J. P., & Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*, *15*(1), 478. https://doi.org/10.1186/1471-2164-15-478

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews. Genetics*, *19*(8), 491–504. https://doi.org/10.1038/s41576-018-0016-z

Sea Lice Research Centre. (2023). *THE ATLANTIC SALMON LOUSE*. https://slrc.w.uib.no/about-sea-lice/the-atlantic-salmon-louse/

Simen Sætre and Kjetil S. Østli. (2021). *The New Fish. The Global History of Salmon Farming.* https://norla.no/nb/books/1286-the-new-fish-the-global-history-of-salmon-farming

Skiftesvik, A. B., Blom, G., Agnalt, A.-L., Durif, C. M., Browman, H. I., Bjelland, R. M., Harkestad, L. S., Farestveit, E., Paulsen, O. I., Fauske, M., Havelin, T., Johnsen, K., & Mortensen, S. (2014). Wrasse (Labridae) as cleaner fish in salmonid aquaculture – The Hardangerfjord as a case study. *Marine Biology Research*, *10*(3), 289–300. https://doi.org/10.1080/17451000.2013.810760

Spain, S. L., & Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human Molecular Genetics*, *24*(R1), R111-9. https://doi.org/10.1093/hmg/ddv260

Stahl, K., Gola, D., & König, I. R. (2021). Assessment of Imputation Quality: Comparison of Phasing and Imputation Algorithms in Real Data. *Frontiers in Genetics*, *12*, 724037. https://doi.org/10.3389/fgene.2021.724037

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews. Genetics*, *10*(10), 681–690. https://doi.org/10.1038/nrg2615

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews. Genetics*, *20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1

Tsai, H.-Y., Hamilton, A., Tinch, A. E., Guy, D. R., Bron, J. E., Taggart, J. B., Gharbi, K., Stear, M., Matika, O., Pong-Wong, R., Bishop, S. C., & Houston, R. D. (2016). Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations. *Genetics, Selection, Evolution : GSE*, *48*(1), 47. https://doi.org/10.1186/s12711-016-0226-9

Tsai, H.-Y., Matika, O., Edwards, S. M., Antolín-Sánchez, R., Hamilton, A., Guy, D. R., Tinch, A. E., Gharbi, K., Stear, M. J., Taggart, J. B., Bron, J. E., Hickey, J. M [John M.],

& Houston, R. D. (2017). Genotype Imputation To Improve the Cost-Efficiency of Genomic Selection in Farmed Atlantic Salmon. *G3 (Bethesda, Md.)*, *7*(4), 1377–1383. https://doi.org/10.1534/g3.117.040717

Tsairidou, S., Hamilton, A., Robledo, D., Bron, J. E., & Houston, R. D. (2020). Optimizing Low-Cost Genotyping and Imputation Strategies for Genomic Selection in Atlantic Salmon. *G3 (Bethesda, Md.)*, *10*(2), 581–590. https://doi.org/10.1534/g3.119.400800

Tucker, T., Marra, M., & Friedman, J. M. (2009). Massively parallel sequencing: The next big thing in genetic medicine. *American Journal of Human Genetics*, *85*(2), 142–154. https://doi.org/10.1016/j.ajhg.2009.06.022

Uffelmann, E., Huang, Q. Q., Munung, N. S., Vries, J. de, Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1). https://doi.org/10.1038/s43586-021-00056-9

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Wakeley, J., Nielsen, R., Liu-Cordero, S. N., & Ardlie, K. (2001). The discovery of single-nucleotide polymorphisms--and inferences about human demographic history. *The American Journal of Human Genetics*, *69*(6), 1332–1347. https://doi.org/10.1086/324521

Wei, Z., Sun, W., Wang, K., & Hakonarson, H. (2009). Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics (Oxford, England)*, *25*(21), 2802–2808. https://doi.org/10.1093/bioinformatics/btp476

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678. https://doi.org/10.1038/nature05911

Yang, J., Lee, S. H., Goddard, M. E [Michael E.], & Visscher, P. M. (2011). Gcta: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yoshida, G. M., Carvalheiro, R., Lhorente, J. P., Correa, K., Figueroa, R., Houston, R. D., & Yáñez, J. M [José M.] (2018). Accuracy of genotype imputation and genomic predictions in a two-generation farmed Atlantic salmon population using high-density and low-

density SNP panels. *Aquaculture*, *491*, 147–154. https://doi.org/10.1016/j.aquaculture.2018.03.004

Zhang, Z [Z.], & Druet, T. (2010). Marker imputation with low-density marker panels in Dutch Holstein cattle. *Journal of Dairy Science*, *93*(11), 5487–5494. https://doi.org/10.3168/jds.2010-3501

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway