THE UNIVERSITY
of EDINBURGH

Genomic basis of growth traits and host resistance against sea lice (*L. Salmonis*) in Atlantic salmon (*S. Salar*)

Hsin-Yuan Tsai

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

The University of Edinburgh

2017

Declaration

I declare that this thesis was composed by me, and the works contained herein is my own except where explicitly stated otherwise in the text. In chapters 2, 3, 4, 5 and 6, where data were generated or obtained from collaborative and online resources, and these have been clearly specified. My contributions to the thesis have been clearly stated in each chapter.

Hsin Yuan Tsai
Date: Feb/2017

Acknowledgements

First of all, I would like to express my sincere thanks to Ross Houston, Chris Haley and Steve Bishop, for your supervision, and supports during my PhD. I also appreciate the thesis committee members, Alan Archibald and Paula Brunton, for their suggestion at the committee meetings. Without your help, I do not think I can finish this degree.

Secondly, I thank to Jose Yanez, Jeffrey Schoenebeck and Pam Wiener, for their correction to the thesis, and arguments during the viva. Your suggestions helped to improve the thesis.

Thirdly, I thank to all people at the Landcatch, Roslin Institute, Kings Building and other institutes for your generous assistance during the PhD and my project, and people in the author list in my published manuscripts. Too much people I have to express my appreciation, therefore, please accept my apologies that I cannot detail all of your names here.

List of Publications

Peer-Reviewed

Tsai H.Y., Robledo, D., Lowe, N.R., Bekaert, M., Taggart, J.B., Bron, J.E., Houston, R.D., 2016. Construction and annotation of a high density SNP linkage map of the Atlantic salmon (Salmo salar) genome. **G3: Genes Genomes Genet.** 6, 2173–2179. (Based on **Chapter 2**)

Tsai H.Y., Hamilton, A., Guy, D.R., Tinch, A.E., Bishop, S.C., Houston, R.D., 2015. The genetic architecture of growth and fillet traits in farmed Atlantic salmon (Salmo salar). **BMC Genet.** 16, 51. (Based on **Chapter 3**)

Tsai H.Y., Hamilton, A., Tinch, A.E., Guy, D.R., Gharbi, K., Stear, M.J., Oswald, M., Bishop, S.C., Houston, R.D., 2015. Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. **BMC Genomics** 16, 969. (Based on **Chapter 4**)

Tsai H.Y., Hamilton, A., Guy, D.R., Tinch, A.E., Bishop, S.C., Houston, R.D., 2015. Verification of SNPs associated with growth traits in two populations of farmed atlantic salmon. **Int. J. Mol. Sci.** 17, 1–10. (Based on **Chapter 5**)

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. **Genet. Sel. Evol.** 48, 1–11. (Based on **Chapter 6**)

Conference Proceedings

Tsai H.Y., Bishop S.C., Houston R.D. The genetic regulation of growth and sea lice resistance in farmed Atlantic salmon (S. Salar). **Hendrix Genetics Academics, Eindhoven, Netherlands.** May 2015.

Tsai H.Y., Hamilton A., Guy D.R., Tinch A.E., Bishop S.C., & Houston R.D. Genome wide association study identifies SNPs associated with early growth performance in Atlantic salmon (S. Salar). **The International Symposium on Genetics in Aquaculture XII, Santiago de Compostela, Spain.** June 2015.

Tsai H.Y., A. Hamilton, D.R. Guy, A.E. Tinch, K. Gharbi, M.J. Stear, O. Matika, S.C. Bishop, R.D. Houston. GWAS and genomic selection of growth traits and host resistance to sea lice in Atlantic salmon. **British Society of Animal Sciences, Edinburgh, U.K.** October 2015.

Tsai H.Y., A. Hamilton, D.R. Guy, A.E. Tinch, J.E. Bron, K. Gharbi, M.J. Stear, O. Matika, S.C. Bishop, R.D. Houston. Genomic Prediction of Host Resistance to Sea Lice (L. salmonis) in Atlantic Salmon (S. salar). **International Plant and Animal Genome XXIV, San Diego, CA, U.S.A.** January 2016.

Tsai H.Y., A. Hamilton, D.R. Guy, A.E. Tinch, J.E. Bron, J.B. Taggart, K. Gharbi, M.J. Stear, O. Matika, R. Pong-Wong, J.M. Hickey, S.C. Bishop, R.D. Houston. Genomic selection for breeding improved resistance to sea lice in Farmed Atlantic Salmon. **Aquaculture Europe 2016, Edinburgh, U.K.** September 2016.

Table of Contents

Table and Figure List

Abstract

Background

Atlantic salmon (*Salmo Salar*) is a key aquaculture species in several countries. Since its critical role in economic sector and scientific research, this species has been relatively extensively investigated, in comparison with other farmed and wild aquatic species. However, the genetic components associated with growth and fillet-related traits are lack consistency, and the issue of sea louse disease in both wild and famed salmon is still unsolved.

Objectives

Overall aim of this project was to understand the genetic basis of growth-related traits and host resistance to sea lice using three large commercial farmed salmon populations. Specifically, the method of quantitative trait loci (QTL) mapping, genome-wide association study (GWAS), and genomic prediction (GS) were utilized to dissect the genetic architectures associated with traits of interest in our experimental populations. Prior to this, linkage mapping was performed to construct a high-density linkage map for Atlantic salmon.

Results

Linkage map

A linkage map was firstly constructed underlying a SNP array containing 132 K validated SNPs. 96,396 SNPs were successfully assigned to 29 chromosomes that correspond to the linkage group number of European Atlantic salmon. 6.5 % of unassigned contigs, which was equal to 1 % of recent whole genome reference assembly (GCA_000233375.4) anchored to exist chromosomes by referring to linkage mapping result.

Genetic components associated with growth traits

Heritabilities of growth-related traits were about 0.5 to 0.6 in adult and juvenile farmed salmon. The QTL mapping and GWAS suggested the growth-related traits are likely a polygenic genetic architecture with no major QTL segregating. The prediction accuracy estimated by genomic prediction showed that approximately

5,000 SNP markers could achieve the highest accuracy in body weight and length in juvenile salmon within population.

Genetic components associated with lice resistance

The heritability of lice resistance was 0.22 to 0.33 using pedigree and genetic relationship matrices respectively. GWAS indicated that the host resistance to sea lice was likely polygenic with no individual SNP surpassed the genome-wide significance threshold. Genomic prediction showed that about 5 to 10 K SNPs was able to achieve the asymptote of accuracy in closely related animals, while the greatest advantage of genomic prediction was observed in non-sibling test within population.

Conclusions

As the growth-related traits and lice resistance are both likely polygenic and population-specific, the genomic prediction is an efficient approach to capture the genetic variances of the traits in selection candidates in experimental population, especially for traits with low heritability such as flesh colour and lice resistance. Family-based selection method is the better choice than mass selection to accumulate the genetic effects in corresponding SNP platform. Given the high cost of genotyping and field data collection, the genotyping-by-sequencing and genotype imputation are likely the way to make significant improvements in relevant research.

Chapter 1

General Introduction

1.1 Overview

Selective breeding programs applied in livestock and crop species have continuously improved industry productivity for decades. However, only approximately 9 % of aquaculture production is derived from selective breeding [1]. The use of molecular genetic information in breeding programs has transformed livestock and crop breeding, and is at a relatively formative stage for a few aquaculture species. With the advances in nucleic acid sequencing technology and bioinformatics, a large number of studies have been conducted to identify the quantitative trait loci (QTLs) associated with phenotypes with economical and biological importance in livestock and aquaculture species [2]. Due to the high economic value and scientific interest of Atlantic salmon [3], the primary goal of this thesis was to apply several molecular and quantitative genetic approaches to investigate the genetic architecture and improvement of different traits of economic interest in commercial farmed salmon populations.

1.2 Atlantic Salmon

1.2.1 Atlantic Salmon (*Salmo salar*) and Farming

Atlantic salmon (*Salmo salar*) is an anadromous species in the family *Salmonidae*, which was initially found in the northern Atlantic Ocean. With the migration of human beings and activities, Atlantic salmon were also documented in northern Pacific areas. Several distinct life stages are observed in this species, from eggs hatch to juveniles, which takes roughly one to three years in natural freshwater environment. At the smolt stage, salmon typically start a long distance migration, from native rivers to the ocean. While growing up as the grilse phase in the ocean, fish groups typically return to the same freshwater tributary where they were hatched, for mating and spawning (Figure 1-1).

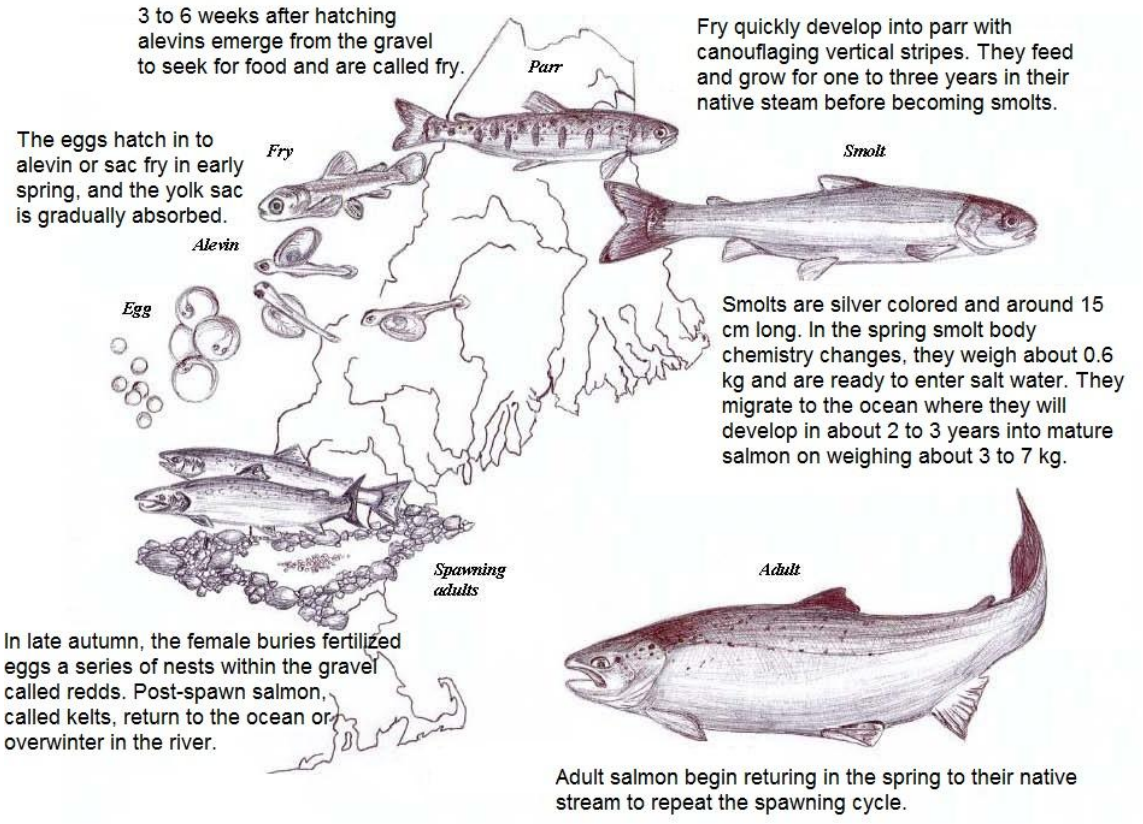


Figure 1-1. Life cycle of Atlantic salmon (*Salmo Salar*). The illustration was adapted from http://www.nefsc.noaa.gov/press_release/2011/SciSpot/SS1107/

For Atlantic salmon farming, the industry is still relatively young compared to terrestrial animal farming such as poultry and livestock, with the first organized salmon breeding programs established in the early 1970s in Norway. Data reported by Food and Agriculture Organization (F.A.O.) indicate that the major producing countries of Atlantic salmon are Norway, Chile, Canada, and Scotland, as well as minor numbers in Ireland, New Zealand, Australia, and the United States. Recently, aquaculture was recognized as the fastest growing source of animal protein production in the world (Table 1-1). Atlantic salmon production is a major contributor to global aquaculture, and the current consumption of farmed salmon is nearly three times higher than 1980s [4].

Table 1-1. The annual production of terrestrial and aquatic farmed animals from 2011 to 2013. This table was adapted from Gjedrem et al. [5].

Species	**Annual improvement in growth (%)**	2011 (million tons)	2012 (million tons)	2013 (million tons)
Aquaculture	**5.7**	62.7	66.3	69.6
Pig meat	**2.5**	109.2	112.7	114.6
Poultry meat	**2.0**	102.6	104.9	106.8
Beef meat	**0.2**	67.3	67.4	67.5
Fishing	**-1.8**	93.5	90.6	90.1

1.2.2 Genetic Basis of Growth Performance in Salmon

For most farmed animal breeding programs, growth rate is the major criterion due to its economic importance, which is directly related to the profits of the industry. A higher growth rate is associated with a good feed conversion rate and shorter time to harvest. Selective breeding has been implemented to enhance the genetic improvement in growth performance and fillet traits in several aquaculture species since 1970s [1, 6]. A review written in 1980s showed that the growth traits are heritable, the heritabilities of body weight and length at early stage (~ 6 months) of farmed Atlantic salmon were estimated at approximately 0.15 to 0.17 using pedigree information, and about 0.45 at 3.5 year post-hatching [7].

Benefitting from the availability of genomic resources in salmon and improvement of computing methods, trait heritabilities can be estimated using genomic-based methods in addition to pedigree methods. At present, the heritabilities of growth-related traits have been estimated at 0.5 to 0.6 in both young and adult fish by fitting genomic information in the animal model [8], and the genetic gain obtained for body weight in farmed

Atlantic salmon in a breeding program is notably high, at 12.7 % per generation [5]. The growth rate of individuals is considered as a part of the complex regulatory processes which are typically regulated by multiple controlling networks involving several genes and metabolic pathways. So far, there is no major or consistent QTL reported associated with fillet production and growth-related traits. As such, the current consensus is that growth traits are highly heritable, and the genetic gain is generationally increased by long term genetic improvement, but the traits have a highly polygenic architecture [9].

1.2.2.1 Genetic Basis of Sea Lice (*Lepeophtheirus salmonis*) Resistance in Farmed Salmon

Salmon lice have caused large negative effects on both wild and farmed salmon for many years. There are two major species of sea louse impacting on farmed salmon in European areas, namely *Lepeophtheirus salmonis* (Kröyer, 1837) and *Caligus elongatus* (von Nordmann, 1832) (Figure 1-2), of which *L. salmonis* is the primary problem for commercial production. *Caligus rogercresseyi* is the major parasite of salmon in Chile [10]. Controlling outbreaks of these parasites is essential for the economics of salmon farming, and also from an animal welfare and environmental perspective.

The first literature documenting the sea louse as a parasitic copepod was based on wild Atlantic salmon [11], and farmed salmon in Norway in 1960s [12, 13]. While in 1970s, sea louse was soon addressed as a major parasitic threat to commercial salmon farming (*e.g.* net cage salmon farming) [14], and has persisted and worsened in recent years [15, 16].

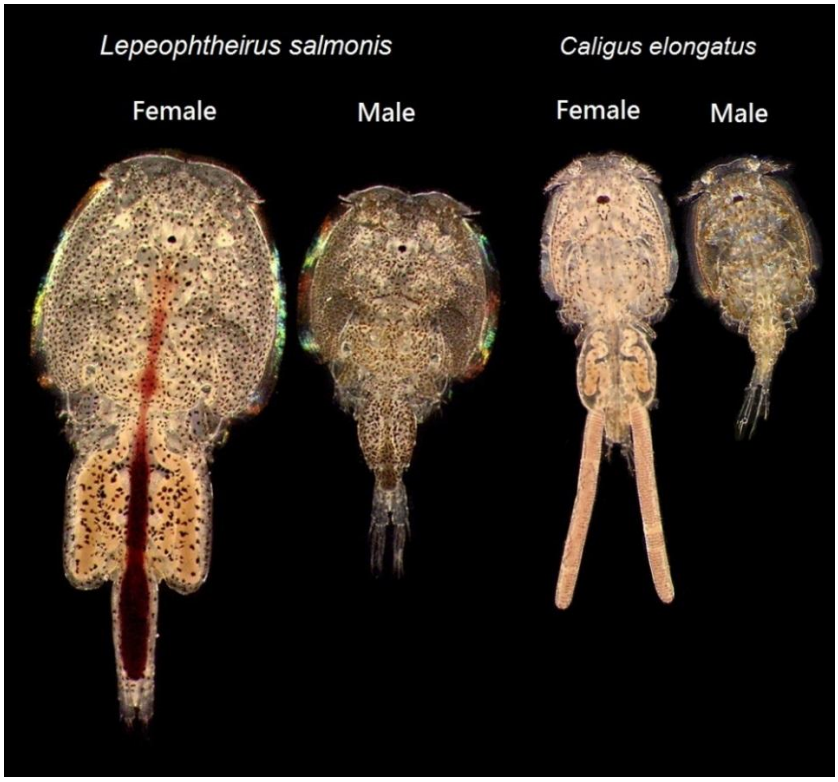


Figure 1-2. Outward appearance of Lepeophtheirus salmonis and Caligus elongatus. Credits: Lars Hamre, Sea Lice Research Centre, UIB. The illustration was adapted from http://www.slrc.no/about-sea-lice/what-is-a-sea-louse/

Sea lice infection can be diagnosed from the surface of fish skin; the symptoms include skin damage, osmotic imbalance, and increased susceptibility to other infections (*e.g.* secondary bacterial or fungal infection), as a result it can induce host immune suppression, slower host growth rate and even death [17]. All of these can reduce the gross production by approximately 5 %, while combining with relevant treatment cost, sea lice have caused about £25 million financial losses in Scottish salmon industry annually [18].

Several approaches have been proposed and implemented to control sea lice disease. Currently, chemotherapeutant treatment is the primary way that the farmers combat outbreaks of sea lice, and the most widespread chemical drug for sea lice control is to use SLICE® (emamectin benzoate), an avermectin drugs used as in-feed manner [13].

However, frequent chemical treatment may result in certain environmental and human health implications, and the emergence of resistance amongst the parasites. As such, the usage of chemical drug has been regulated in most salmon farming countries. For example, in the U.K., the Scottish salmon producers are authorized to administer the chemical treatments with prescribed discharge consent annually for specific chemotherapeutants. However, chemotherapautants alone are often insufficient to control lice outbreaks, and alternative control measures are required to combat this major issue for both authorities and industries.

Encouragingly, studies conducted in both sea-cage and experimental tank environment challenge trials demonstrate that the heritability of sea lice resistance was about 0.2 to 0.3 [15, 19–21], implying that selective breeding can contribute to sustainable sea lice control. As such, breeding for improving host resistance to sea lice in farmed salmon has become an increasingly important component of sea lice disease control [20, 22], and is likely to help the industry reduce the usage and cost of chemical treatments.

1.3 Atlantic Salmon Genomic Resources and Linkage Mapping

Genetic and physical maps of the genomes of farmed animal species are essential tools for mapping and utilizing genetic variation underpinning traits of economic importance. Rather than physical maps, which the marker position is determined by the specific physical distance along the corresponding chromosomes, a linkage map is constructed based on the recombination frequencies between the molecular markers during the crossover of homologous chromosomes. At present, linkage maps are available in around fifty fish species, and most of them are aquaculture species with economic importance, although the density of genetic maps is typically lower than most terrestrial livestock and commercial crops [23]. Construction of a high density linkage map can improve the quality of *de novo* genome sequence assembly, and also assist in high

resolution mapping of QTLs associated with phenotypes of importance due to their commercial value or scientific interest [24–26]. Benefitting from the advances of next-generation sequencing technologies and latest genotyping approaches (*e.g.* genotyping-by-sequencing (GBS) and restriction site-associated DNA (RAD) sequencing), medium to high density genetic maps for aquatic species with commercial importance are increasingly available (*e.g.* Rainbow trout [27], Atlantic salmon [3, 26], Channel catfish [28], Turbot [29], Nile tilapia [30] and European seabass [31]). Amongst all aquaculture species, the genomic resources of Atlantic salmon are the most extensive, due to its role in both economic and scientific interest. Several sparse to high genetic maps were developed using different forms of molecular markers, including amplified fragment length polymorphisms (AFLPs), microsatellites, and SNPs [32–34]. In the past five years, large amount of SNPs in Atlantic salmon were discovered [35, 36], making high density genetic map become increasingly available [25, 26]. More recently, the whole genome assemblies of Atlantic salmon and Rainbow trout have been published [3, 27], offering valuable genomic resources for a wide range of evolutionary and aquaculture genetic research in different disciplines.

1.4 Application of Genomics to Selective Breeding of Salmon

Traditional breeding programs were based on the phenotypic performances of the selection candidates consisting of the genetic and environmental component [37]. However, there are several traits that cannot easily be measured on selection candidates themselves, such as disease resistance, fillet quality, and growth / survival in commercial environments. As such, breeders utilize the candidates' pedigrees, measuring relevant traits on individuals related to the candidates, in order to select the candidates for mating and breeding. However, selection relying on pedigree records still limits the accuracy of selection as it is not possible to distinguish between offspring from the same full-sibling

family, *e.g.* one can only utilize between-family genetic variation and not within-family genetic variation.

In a typical commercial salmon aquaculture breeding program, the high fecundity of the species results in large nuclear families, comprising thousands of offspring. Therefore, using within-family genetic variation is critical for maximizing genetic improvement for key traits. With recent advances in molecular genetics, information at the DNA level (*e.g.* molecular markers) have recently been exploited in the breeding programs, in addition to pedigree methods. By selecting directly for favorable alleles within families, breeders can significantly improve genetic gain, which is known as marker-assisted selection (MAS). Applying molecular markers in the breeding program offers the potential to distinguish between full-sibs more efficiently, and can also be applied to traits that cannot be measured directly on the candidates.

Large numbers of molecular markers have become available for certain aquaculture species in recent years, expedited by the development of high-throughput sequencing technology and bioinformatics tools. When these markers are genotyped for animals with performance trait records, it is possible to detect and map genes or genomic regions associated with target traits. For Atlantic salmon, due to its high economic value for aquaculture, which have been better studied than other salmonid species, and has a more extensive genomic toolbox [3]. Previous studies have indicated that growth and host resistance to disease are heritable traits [38], and several have investigated the QTLs or QTNs associated with growth performance [8, 39], flesh colour [9] and disease resistance [38, 40–42] in salmonid species. However, with the exception of the major QTL affecting resistance to the IPN virus, there is typically a lack of consistency for QTL results across studies and populations (*e.g.* QTLs associated with body weight).

With the availability of high density SNP arrays [25, 26, 35] and recent reference genome assemblies [3, 27], it is now relatively straight forward to gather extensive genome-wide SNP genotypes in salmonids. These have enabled genome-wide

association studies (GWAS), whereby the association between SNPs dispersed across the genome, and traits of interest are tested at a population-wide level. To date, GWAS has been employed to understand the genetic basis of traits related to disease and growth performance in Atlantic salmon [16, 43–45]. The results of these studies highlight the genetic architecture of these production traits, and identify individual SNPs that may explain a small proportion of the underlying genetic variation. However, the application of individual SNPs associated with polygenic traits is unlikely to be of great importance in marker-assisted selection, due to the very small proportion of genetic variation explained.

Genomic selection approaches are alternative means of utilizing genome-wide markers in modern selective breeding programs, especially for those phenotypes controlled by many QTLs of small effect. Rather than testing the significance of all individual SNP by using GWAS, genomic prediction uses information from all markers to estimate genetic merit of a selection candidate, aiming to increase the genetic gains per generation via prediction of genomic estimated breeding values [46]. Since many factors have been reported to affect the prediction accuracy, such as homogeneity of population [47–49] and relatedness between validation and training population [50, 51], several prediction methods have been proposed to improve the prediction accuracy in genomic prediction under different types of breeding programs (*e.g.* reviewed by [52]). However, the optimal models to use for genomic prediction are still under debate. Currently, most genomic prediction studies conducted in aquaculture species used simulated population data [53], but relatively few have used experimental data, partly due to high cost [22]. The cost of genotyping is still a barrier for industries and researchers to extensively study and apply genomic prediction and GWAS to aquaculture species. However, early results indicate that genomic prediction is a reliable way to predict the breeding values of traits in selection candidate in both simulation studies and experimental population in aquaculture research [22, 53]. As such, genomic selection is likely to be of critical importance to aquaculture breeding and, meanwhile, maximizing prediction accuracy with minimal cost for SNP genotyping is an important goal.

1.5 Outline of Thesis

The overall aim of this thesis was to develop and utilize genomic approaches to understand the genetic factors associated with growth traits and host resistance to sea lice in farmed Atlantic salmon. Specifically, the objectives of each chapter are listed below, along with the corresponding manuscripts in each chapter, respectively.

In chapter 2, to provide a genomic resource for trait mapping and genomic selection, I constructed a salmon linkage map comprising approximately 100 K SNP markers across the entire salmon genome, and integrated the results with the updated reference genome assembly (Genbank Accession GCA_000233375.4, [3]). 622 individual salmon from 62 nuclear families were genotyped using developed Affymetrix SNP array [35]. 111,908 SNPs passing through the quality control (QC) were retained in the linkage mapping analysis. The Lep-Map2 software [54] was used to assign SNPs to 29 linkage groups that correspond to the karyotype of European Atlantic salmon, and to estimate the most likely order of those QC-SNPs. Several of the previously unmapped reference genome contigs were anchored to possible chromosome / region by referring to the linkage mapping result. The recombination pattern of male and female salmon across the entire genome was also compared to investigate the difference between sexes.

In chapter 3, I mapped QTL associated with performance and quality traits in a large commercial salmon population. The fish were approximately 3 years post-hatching when measured. The heritabilities of 12 traits recorded at harvest processing were estimated and compared. Due to the large disparity in recombination rate between male and female salmon [55], a two-step approach was employed to efficiently perform the QTL mapping. Firstly, the sire-based mapping was applied with sparse SNPs (2 to 3 SNPs per chromosome, [26]) to detect the putative QTLs associated with traits of interest. Secondly, the candidate chromosomes / QTLs that were detected in sire-base stage were

confirmed and fined the genomic position using denser SNP platform (10 SNPs per candidate chromosome, [26]) in dam-based mapping.

In chapter 4, I applied a GWAS to evaluate the genetic association of individual SNPs on a high density SNP array [35] containing approximately 132 K SNPs, with body weight (g) and length (mm) in juvenile Atlantic salmon. The pedigreed fish population was around 1 year post-hatching (n = 622 including 534 offspring, 28 sires, and 60 dams). The heritabilities of traits were estimated by genomic data and pedigree information respectively. Candidate genes harbouring significant SNPs evaluated by GWAS were identified. Secondly, I also performed genomic prediction with ascending marker densities (0.5 K, 1 K, 5 K, 10 K, 20 K, 33 K and 112 K) to assess the utility of genomic prediction for both growth traits using best linear unbiased prediction (BLUP) fitting genomic (GBLUP) and pedigree relationship matrix (PBLUP) respectively.

In chapter 5, I extended results from chapter 3 and 4 to verify SNPs associated with growth in salmon. In this chapter, I addressed the concern that the significant SNPs identified in GWAS are likely to contain a mix of true associations and false positives in those polygenic growth traits. Thus, I selected two candidate SNP markers from genome-wide significant QTL regions ($p < 0.05$, chapter 3) and fourteen nominal significance SNPs identified by GWAS ($p < 0.001$, chapter 4) to verify these SNPs' effects on growth traits in a separate commercial population. The genes harbouring the significant SNPs were identified by alignment to the salmon assembly reference genome (GCA_000233375.4, [3]), and were discussed in the context of their potential role in underpinning genetic variation in salmon growth.

In chapter 6, I investigated the genetic architecture of host resistance to sea louse (*Lepeophtheirus salmonis*), and tested genomic prediction approaches for this trait. Two pedigreed populations were sampled from 2007 (n = 621 comprising 531 offspring, 30 sires and 60 dams) and 2010 (n = 874 comprising 588 offspring, 98 sires, and 188 dams) year groups, and genotyped with 132 K and 33 K genome-wide distributed SNP

platforms respectively. The heritability of resistance was estimated by genomic and pedigree relationship matrices respectively. The genetic architecture was investigated using a GWAS. To predict the breeding values of traits in individual underlying different marker densities, all SNPs (33 K) were randomly chosen to construct ascending low-marker-density genomic relationship matrices (20 K, 10 K, 5 K, 1 K, and 0.5 K). Five-fold cross validation analyses were employed in four scenarios where the degree of relationship between the training and validation sets varied [(i) within population random selection, (ii) within population full-sibling, (iii) within population non-sibling, and (iv) across two populations].

1.6 References

1. Gjedrem T, Robinson N, Rye M: The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. *Aquaculture* 2012, 350-353:117–129.

2. Yáñez JM, Newman S, Houston RD: Genomics in aquaculture to better understand species biology and accelerate genetic progress. *Front Genet* 2015, 6(April):128.

3. Lien S, Koop BF, Sandve SR, Miller JR, Matthew P, Leong JS, Minkley DR, Zimin A, Grammes F, Grove H, Gjuvsland A, Walenz B, Hermansen RA, Schalburg K Von, Rondeau EB, Genova A Di, Samy JKA, Vik JO: The Atlantic salmon genome provides insights into rediploidization. *Nature* 2016, 533:200–205.

4. F.A.O: *FAO Yearbook.Fishery and Aquaculture Statistics.* Rome; 2014.

5. Gjedrem T, Robinson N: Advances by selective reeding for aquatic species: a review. *Agric Sci* 2014, 05:1152–1158.

6. Gjedrem T: Disease Resistant Fish and Shellfish Are within Reach: A Review. *J Mar Sci Eng* 2015, 3:146–153.

7. Trygve Gjedrem: Genetic variation in quantitative traits and selective breeding in fish

and shellfish. *Aquaculture* 1983, 33:51–72.

8. Houston RD, Bishop SC, Hamilton A, Guy DR, Tinch AE, Taggart JB, Derayat A, McAndrew BJ, Haley CS: Detection of QTL affecting harvest traits in a commercial Atlantic salmon population. *Anim Genet* 2009, 40:753–5.

9. Baranski M, Moen T, Våge DI: Mapping of quantitative trait loci for flesh colour and growth traits in Atlantic salmon (Salmo salar). *Genet Sel Evol* 2010, 42:17.

10. Boxshall, GA; Bravo S: On the identity of the common Caligus (Copepoda : Siphonostomatoida : Caligidae) from salmonid netpen systems in southern Chile. *Contrib TO Zool* 2000, 69:137–146.

11. White HC: "Sea lice" (Lepeophtheirus) and death of salmon. *J Fish Res Board Canada* 1940, 5:172–175.

12. Hastein T, Bergsjo T: The salmon lice Lepeophtheirus salmonis as the cause of disease in farmed salmonids. *Riv Ital di Piscic e Ittiopatologia* 1976, 11:3–5.

13. Frenzl B: Understanding key factors associated with the infection of farmed Atlantic salmon by the salmon louse Lepeophtheirus Salmonis. PhD Degree Thesis. University of Stirling; 2014.

14. Heuch PA, Bjørn PA, Finstad B, Holst JC, Asplin L, Nilsen F: A review of the Norwegian "National Action Plan Against Salmon Lice on Salmonids": The effect on wild salmonids. *Aquaculture* 2005, 246:79–92.

15. Gharbi K, Matthews L, Bron J, Roberts R, Tinch A, Stear MJ: The control of sea lice in Atlantic salmon by selective breeding. *J R Soc Interface* 2015, 12:0574.

16. Correa K, Lhorente JP, Bassini L, López ME, Di Génova A, Maass A, Davidson WS, Yáñez JM: Genome wide association study for resistance to Caligus rogercresseyi in Atlantic salmon (Salmo salar L.) using a 50K SNP genotyping array. *Aquaculture* 2016. In Press.

17. Frazer LN, Morton A, Krkosek M: Critical thresholds in sea lice epidemics: evidence, sensitivity and subcritical estimation. *Proc R Soc B Biol Sci* 2012, 279:1950–1958.

18. Costello MJ: The global economic cost of sea lice to the salmonid farming industry. *J Fish Dis* 2009, 32:115–118.

19. Kolstad K, Heuch PA, Gjerde B, Gjedrem T, Salte R: Genetic variation in resistance of Atlantic salmon (Salmo salar) to the salmon louse Lepeophtheirus salmonis. *Aquaculture* 2005, 247:145–151.

20. Gjerde B, Ø degård J, Thorland I: Estimates of genetic variation in the susceptibility of Atlantic salmon (Salmo salar) to the salmon louse Lepeophtheirus salmonis. *Aquaculture* 2011, 314:66–72.

21. Houston, R.D., Bishop, S.C., Guy, D.R., Tinch A.E., Taggart, J.B., Bron, J.E., Downing, A., Stear, M.J., Gharbi, K., Hamilton A: Genome Wide Association Analysis for Resistance to Sea Lice in Atlantic Salmon: Application of a Dense SNP Array. In *Proc 10th World Congr Genet Appl to Livest Prod*. Vancouver; 2014:10–12.

22. Odegård J, Moen T, Santi N, Korsvoll SA, Kjøglum S, Meuwissen THE: Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). *Front Genet* 2014, 5(November):402.

23. Yue GH: Recent advances of genome mapping and marker-assisted selection in aquaculture. *Fish Fish* 2014, 15:376–96.

24. Wang L, Wan ZY, Bai B, Huang SQ, Chua E, Lee M, Pang HY, Wen YF, Liu P, Liu F, Sun F, Lin G, Ye BQ, Yue GH: Construction of a high-density linkage map and fine mapping of QTL for growth in Asian seabass. *Sci Rep* 2015, 5:16358.

25. Gonen S, Lowe NR, Cezard T, Gharbi K, Bishop SC, Houston RD: Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing. *BMC Genomics* 2014, 15:166.

26. Lien S, Gidskehaug L, Moen T, Hayes BJ, Berg PR, Davidson WS, Omholt SW, Kent MP: A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genomics* 2011, 12:615.

27. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C, Labadie K, Alberti A, Aury J-M, Louis A, Dehais P, Bardou P, Montfort J, Klopp C, Cabau C, Gaspin C, Thorgaard GH, Boussaha M, Quillet E, Guyomard R, Galiana D, Bobe J, Volff J-N, Genêt C, Wincker P, Jaillon O, Roest Crollius H, Guiguen Y: The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun* 2014, 5:3657.

28. Liu S, Li Y, Qin Z, Geng X, Bao L, Kaltenboeck L, Kucuktas H, Dunham R, Liu Z: High-density interspecific genetic linkage mapping provides insights into genomic incompatibility between channel catfish and blue catfish. *Anim Genet* 2016, 47:81–90.

29. Wang W, Hu Y, Ma Y, Xu L, Guan J, Kong J: High-density genetic linkage mapping in turbot (Scophthalmus maximus L) based on SNP markers and major sex- and growth-related regions detection. *PLoS One* 2015, 10:1–16.

30. Guyon R, Rakotomanga M, Azzouzi N, Coutanceau JP, Bonillo C, D'Cotta H, Pepey E, Soler L, Rodier-Goud M, D'Hont A, Conte M a, van Bers NEM, Penman DJ, Hitte C, Crooijmans RPM a, Kocher TD, Ozouf-Costaz C, Baroiller JF, Galibert F: A high-resolution map of the Nile tilapia genome: a resource for studying cichlids and other percomorphs. *BMC Genomics* 2012, 13:222.

31. Palaiokostas C, Bekaert M, Taggart JB, Gharbi K, McAndrew BJ, Chatain B, Penman DJ, Vandeputte M: A new SNP-based vision of the genetics of sex determination in European sea bass (Dicentrarchus labrax). *Genet Sel Evol* 2015, 47:68.

32. Gilbey J, Verspoor E, McLay A, Houlihan D: A microsatellite linkage map for Atlantic salmon (Salmo salar). *Anim Genet* 2004, 35:98–105.

33. Moen T, Hoyheim B, Munck H, Gomez-Raya L: A linkage map of Atlantic salmon (Salmo salar) reveals an uncommonly large difference in recombination rate between the sexes. *Anim Genet* 2004, 35:81–92.

34. Moen T, Hayes B, Baranski M, Berg PR, Kjøglum S, Koop BF, Davidson WS, Omholt SW, Lien S: A linkage map of the Atlantic salmon (Salmo salar) based on EST-derived SNP markers. *BMC Genomics* 2008, 9:223.

35. Houston RD, Taggart JB, Cézard T, Bekaert M, Lowe NR, Downing A, Talbot R, Bishop SC, Archibald AL, Bron JE, Penman DJ, Davassi A, Brew F, Tinch AE, Gharbi K, Hamilton A: Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). *BMC Genomics* 2014, 15:90.

36. Yáñez JM, Naswa S, López ME, Bassini L, Correa K, Gilbey J, Bernatchez L, Norris A, Neira R, Lhorente JP, Schnable PS, Newman S, Mileham A, Deeb N, Di Genova A, Maass A: Genomewide single nucleotide polymorphism discovery in Atlantic salmon (Salmo salar): validation in wild and farmed American and European populations. *Mol Ecol Resour* 2016, 16:1002–11.

37. Falconer DS, Mackay TFC: *Introduction to Quantitative Genetics*. 4th edition. Harlow, Essex, UK.: Longmans Green; 1996.

38. Yáñez JM, Houston RD, Newman S: Genetics and genomics of disease resistance in salmonid species. *Front Genet* 2014, 5(November):415.

39. Boulding EG, Culling M, Glebe B, Berg PR, Lien S, Moen T: Conservation genomics of Atlantic salmon: SNPs associated with QTLs for adaptive traits in parr from four trans-Atlantic backcrosses. *Heredity (Edinb)* 2008, 101:381–91.

40. Houston RD, Haley CS, Hamilton A, Guy DR, Mota-Velasco JC, Gheyas AA, Tinch AE, Taggart JB, Bron JE, Starkey WG, McAndrew BJ, Verner-Jeffreys DW, Paley RK, Rimmer GSE, Tew IJ, Bishop SC: The susceptibility of Atlantic salmon fry to freshwater infectious pancreatic necrosis is largely explained by a major QTL. *Heredity*

(Edinb) 2010, 105:318–27.

41. Gheyas AA, Houston RD, Mota-Velasco JC, Guy DR, Tinch AE, Haley CS, Woolliams JA: Segregation of infectious pancreatic necrosis resistance QTL in the early life cycle of Atlantic Salmon (Salmo salar). *Anim Genet* 2010, 41:531–536.

42. Gonen S, Baranski M, Thorland I, Norris A, Grove H, Arnesen P, Bakke H, Lien S, Bishop SC, Houston RD: Mapping and validation of a major QTL affecting resistance to pancreas disease (salmonid alphavirus) in Atlantic salmon (Salmo salar). *Heredity (Edinb)* 2015, 115:1–10.

43. Sodeland M, Gaarder M, Moen T, Thomassen M, Kjøglum S, Kent M, Lien S: Genome-wide association testing reveals quantitative trait loci for fillet texture and fat content in Atlantic salmon. *Aquaculture* 2013, 408-409:169–174.

44. Gutierrez AP, Lubieniecki KP, Davidson EA, Lien S, Kent MP, Fukui S, Withler RE, Swift B, Davidson WS: Genetic mapping of quantitative trait loci (QTL) for body-weight in Atlantic salmon (Salmo salar) using a 6.5K SNP array. *Aquaculture* 2012, 358-359:61–70.

45. Correa K, Lhorente JP, López ME, Bassini L, Naswa S, Deeb N, Di Genova A, Maass A, Davidson WS, Yáñez JM: Genome-wide association analysis reveals loci associated with resistance against Piscirickettsia salmonis in two Atlantic salmon (Salmo salar L.) chromosomes. *BMC Genomics* 2015, 16:1–9.

46. Meuwissen THE, Hayes BJ, Goddard ME: Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 2001, 157:1819–1829.

47. Hayes BJ, Bowman PJ, Chamberlain AC, Goddard ME: Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* 2009, 92(2):433-43.

48. de Roos APW, Hayes BJ, Spelman RJ, Goddard ME: Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* 2008, 179(3):1503-12.

49. Ibáñẽz-Escriche N, Fernando RL, Toosi A, Dekkers JCM: Genomic selection of purebreds for crossbred performance. *Genet Sel Evol* 2009, 41:1–10.

50. Clark SA, Hickey JM, Daetwyler HD, van der Werf JHJ: The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet Sel Evol* 2012, 44:1–9.

51. Habier D, Fernando RL, Dekkers JCM: The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 2007, 177:2389–2397.

52. Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G, Hickey JM: Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics* 2013, 193:347–365.

53. Sonesson AK, Meuwissen THE: Testing strategies for genomic selection in aquaculture breeding programs. *Genet Sel Evol* 2009, 41:37.

54. Rastas P, Calboli FCF, Guo B, Shikano T, Merilä J: Construction of ultra-dense linkage maps with Lep-MAP2: stickleback F2 recombinant crosses as an example. *Genome Biol Evol* 2016. In Press.

55. Hayes BJ, Gjuvsland A, Omholt S: Power of QTL mapping experiments in commercial Atlantic salmon populations, exploiting linkage and linkage disequilibrium and effect of limited recombination in males. *Heredity (Edinb)* 2006, 97:19–26.

Chapter 2

Construction and annotation of a high density SNP linkage map of the Atlantic salmon (*Salmo salar*) genome

Hsin Y. Tsai, Diego Robledo, Natalie R. Lowe, Michael Bekaert, John B Taggart,

James E Bron and Ross D Houston

DOI: 10.1534/g3.116.029009

RDH, JEB, and JBT conceived and designed the study; NRL performed laboratory experiments; HYT (linkage mapping), DR (transcriptome analysis), and MB analysed data; HYT, DR, and RDH wrote the manuscript.

This chapter has been published in G3 at the following URL:

http://www.g3journal.org/content/6/7/2173.full

2.0 Introduction

Linkage maps and a reference genome sequence are now available for Atlantic salmon. However, these two resources have not yet been integrated, and the major objective of this chapter was to build up a high density genetic map using a public available Atlantic salmon SNP array containing 132 K markers. Based on this map, an additional objective of the project was to align and integrate the results with recent high quality Atlantic salmon reference genome assembly. Having both a linkage map and a physical map of a species' genome is advantageous for modern genetic analysis, which will help to refine QTL mapping (chapter 3), identify the loci of interest using GWAS (chapter 4), perform the association analysis and putative gene identification (chapter 5), and more recent widely applied method in aquaculture breeding schemes, to build up the platform of SNPs for genomic selection (chapter 6). In addition, the integrated genetic map allows characterization of the recombination landscape of the Atlantic salmon genome, and comparison between the sexes.

2.1 Abstract

Background

High density linkage maps are useful tools for fine-scale mapping of quantitative trait loci, and characterisation of the recombination landscape of a species' genome. Genomic resources for Atlantic salmon (*Salmo salar*) include a well-assembled reference genome and high density SNP arrays. Our aim was to create a high density linkage map, and to align it with the reference genome assembly.

Results

Over 96 K SNPs were mapped and ordered on the 29 salmon linkage groups using a pedigreed population comprising 622 fish from 60 nuclear families, all genotyped with the 'ssalar01' high density SNP array. The number of SNPs per group showed a high positive correlation with physical chromosome length (r = 0.95). While the order of markers on the genetic and physical maps was generally consistent, areas of discrepancy were identified. Approximately 6.5 % of the previously unmapped reference genome sequence was assigned to chromosomes using the linkage map. Male recombination rate was lower than females across the vast majority of the genome, but with a notable peak in sub-telomeric regions. Finally, using RNA-Seq data to annotate the reference genome, the mapped SNPs were categorised according to their predicted function, including annotation of ~ 2.5 K putative non-synonymous variants.

Conclusions

The highest density SNP linkage map for any salmonid species has been created, annotated, and integrated with the Atlantic salmon reference genome assembly. This map highlights the marked heterochiasmy of salmon, and provides a useful resource for salmonid genetics and genomics research.

2.2 Chapter Introduction

Linkage maps are valuable tools for the investigation of the genetic basis of complex traits in farmed animal species. For several decades, linkage maps have enabled the mapping of quantitative trait loci (QTL), and formed the basis of attempts at positional cloning of these QTL in both terrestrial [1] and aquatic farmed species [2]. High throughput sequencing technologies have now expedited the discovery of millions of single nucleotide polymorphism (SNP) markers [3]. These SNPs form the basis of modern, high-resolution genetics studies, and underpin genomic selection for faster genetic improvement in terrestrial livestock and, laterally, aquaculture breeding programmes [4–8]. Scoring of genome-wide SNPs in large populations is achieved either through genotyping by sequencing [9], or by creation and application of SNP arrays (*e.g.* [10]). High density linkage maps based on these SNP datasets can aid in high resolution mapping of loci underpinning complex traits in farmed animals (*e.g.* [11, 12]), improvements in assembly of reference sequences [13], and knowledge of the recombination landscape of the genome (*e.g.* [14, 15]).

Reference genome assemblies are now available for several aquaculture species, including Atlantic salmon [16, 17]. Once anchored and annotated, these genome assemblies provide invaluable physical maps of the genome. Due to a recent whole genome duplication, and the relatively high frequency of long and diverse repeat elements [16–18], assembly of the Atlantic salmon genome has been challenging, with ~ 22 % of the current assembly (NCBI GCA_000233375.4) yet to be assigned to chromosome. Salmonid species exhibit marked heterochiasmy, with males showing very low recombination rates across much of the genome, but with much higher recombination rates thought to occur in telomeric regions (*e.g*. [19–23]). This phenomenon may be related to the pairing and recombination between homeologous regions of the genome, particularly in males [24–26]. Several high density SNP arrays exist for Atlantic salmon [10, 27], and integrated linkage maps based on those arrays would facilitate detailed interrogation of the unusual recombination landscape. Further,

while the high density SNP arrays have been applied for GWAS and genomic prediction [7, 28, 29], such studies would be enhanced by annotation of the SNPs according to their genomic position, nearby genes, and their predicted effects.

Therefore, the purposes of this study were: (i) to construct a linkage map of the SNPs contained on the publicly available high density Affymetrix Atlantic salmon SNP array 'ssalar01' [10]; (ii) to align and compare the linkage map to the latest Atlantic salmon reference genome assembly (Genbank assembly accession GCA_000233375.4); (iii) to assign previously unmapped reference genome contigs and genes to chromosomes; (iv) to investigate and compare patterns of male and female recombination across the genome; and (v) to annotate the SNPs according to their position relative to putative genes, including prediction of variant effects.

2.3 Methods

2.3.1 Animals

The population used for the linkage analysis was a subset of those described in Gharbi et al. [30], purchased from Landcatch Natural Selection (LNS, Ormsary, UK). The juvenile fish used in the current study were from the 2007 year group of the LNS broodstock and were from 60 full sibling families (28 sires and 60 dams) comprising at least six progeny per family. The trial (which focussed on resistance to sea lice) was performed by Marine Environmental Research Laboratory (Machrihanish, UK) and under approval of ethics review committee in the University of Stirling (Stirling, UK). Full details of the trial and the population used have been described previously [28, 30, 31].

2.3.2 SNP Array Genotyping

Genomic DNA from each sample was extracted (Qiagen, Crawley, UK) and genotyped for the 'ssalar01' Affymetrix Axiom SNP array containing ~ 132 K validated SNPs. Details of the creation and testing of the SNP array are given in Houston et al. [10]. Details of the quality control filtering of the genotypes are given in Tsai et al. [28]. Briefly, the Plink software was used to filter the validated SNPs by removing individuals and SNPs with excessive (> 1 %) Mendelian errors, and SNPs with minor allele frequency (MAF) < 0.05 in this dataset. In total, 111,908 SNPs were retained for 622 fish (534 offspring, 28 sires and 60 dams). Details of all the SNP markers are available at dbSNP [32] (NCBI ss# 947429275 - 947844429.).

2.3.3 Linkage Analysis

Lep-Map2 [33] was used to construct the linkage maps. The 'Filtering' function was applied to the initial input dataset, with 'MAFLimit' set at 0.05 (consistent with filtering described above), and 'dataTolerance' set at 0.001 to remove markers exhibiting significant segregation distortion. The 'SeparateChromosomes' function was applied to cluster markers into linkage groups, with the LOD threshold of 36 applied (chosen because this is the level at which 29 groups were formed, consistent with the expected karyotype of European Atlantic salmon). The function 'JoinSingles' was applied to assign additional single SNPs to existing linkage groups. Subsequently, the function 'OrderMarkers' was applied to estimate the marker order within each linkage group. Using parallelised computing, this step was repeated several times to assess consistency of marker order between replicates. Sex-specific linkage maps were generated because of the known difference in recombination rate between male and female Atlantic salmon [20, 23, 34, 35]. To compare the genetic and physical maps, the flanking sequence for each SNP locus (35 bp either side) was aligned with the Atlantic salmon reference genome assembly (Genbank assembly GCA_000233375.4) [16], and only complete and exact matches to the reference genome (e-value = 3×10^{-29}) were retained. In cases where the SNP flanking sequence aligned exactly with > 1 genomic region, the

alignment corresponding to the chromosome that was consistent with the linkage mapping of the SNP was retained.

2.3.4 RNA Sequencing

Atlantic salmon fry samples from two different families from the Scottish breeding nucleus of Landcatch Natural Selection Ltd were selected for RNA sequencing, corresponding to families 'B' and 'S' in Houston et al. [36]. Full details of the library preparation and sequencing are given in Houston et al. [10] (although for the current study, only two of the three families previously sequenced were used for assembling the transcriptome. This was because the third family 'C' had large variation in sequence coverage between samples). Briefly, a total of 48 individual fry were homogenised in 5 mL TRI Reagent (Sigma, USA) using a Polytron mechanical homogeniser (Kinemetica, Switzerland). The RNA was isolated from 1 mL of the homogenate, using 0.5 vol. RNA precipitation solution (1.2 mol/L sodium chloride; 0.8 mol/L sodium citrate sesquihydrate) and 0.5 vol. isopropanol. Following re-suspension in nuclease-free water, the RNA was purified using the RNeasy Mini kit (Qiagen, UK). The RNA integrity numbers from the Bioanalyzer 2100 (Agilent, USA) were all over 9.9. Thereafter, the Illumina Truseq RNA Sample Preparation kit v1 protocol was followed directly, using 4 µg of RNA per sample as starting material. Libraries were checked for quality and quantified using the Bioanalyzer 2100 (Agilent, USA), before being sequenced in barcoded pools of 12 individual fish on the Illumina Hiseq 2000 instrument (100 base paired-end sequencing, v3 chemistry) and all sequence data were deposited in the European Nucleotide Archive under accession number ERP003968.

2.3.5 Transcriptome Assembly

The quality of the sequencing output was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/; version 0.11.2). Quality filtering and removal of residual adaptor sequences was conducted on read pairs using

Trimmomatic v.0.32 [37]. Specifically, residual Illumina specific adaptors were clipped from the reads, leading and trailing bases with a Phred score less than 15 were removed, and the read trimmed if a sliding window average Phred score over four bases was less than 20. Only paired-end reads where both sequences had a length greater than 36 bases post-filtering were retained. The most recent salmon genome assembly (ICSASG_v2, ncbi assembly GCA_000233375.4) was used as a reference for read mapping. Filtered reads were mapped to the genome using Tophat2 v. 2.0.12 [38] that leverages the short read aligner Bowtie2 v.2.2.3 [39], allowing a maximum of two mismatches. Using Cuffdiff v.2.2.1 [40], the aligned reads were merged into a transcriptome assembly. The transcriptome was annotated against NCBI's non-redundant protein and nucleic acid databases using local Blast v.2.3.0+ [41] with a cut-off e-value of 10^{-5}. The completeness of the salmon transcriptome was evaluated using Blast searches with a cut-off e-value of 10^{-25} against a set of 248 core eukaryotic genes [42].

2.3.6 SNP Annotation

For every gene, the most highly expressed transcript variant was selected to identify candidate coding regions using Transdecoder v.2.0.1 (http://transdecoder.sourceforge.net/). Open Reading Frames (ORF) were predicted for every transcript, requiring a minimum of 100 amino acids (to reduce the number of potential false positives). All the predicted proteins were aligned against the manually curated UniRef90 database using local Blast v.2.3.0+ [41] with a cut-off e-value of 10^{-5}, discarding ORFs without positive matches. Finally, the longest ORF was selected as the canonical protein for each transcript. The final set of coding regions was used to build a genome annotation file which was used to predict the functional significance of all the SNPs on the 'ssalar01' SNP array using SnpEff v.4.2 [43].

2.4 Results and Discussion

2.4.1 Linkage Map Construction

A pedigreed population of 622 individual Atlantic salmon (534 offspring, 28 sires and 60 dams) were successfully genotyped using the high density Affymetrix SNP array 'ssalar01' [10]. SNPs were assigned to putative linkage groups and then ordered on each linkage group using Lep-Map2 [33]. A total of 111,908 SNPs were retained following QC filtering, of which 96,396 (86 %) were assigned and ordered on the 29 linkage groups (which correspond to the karyotype of European Atlantic salmon). The number of SNPs per chromosome varied from 1128 to 6080, and was positively correlated with the number of SNPs per chromosome in previously published Atlantic salmon SNP linkage maps of Lien et al. [20] (r = 0.94), and Gonen et al. [23] (r = 0.87). The flanking sequences of the SNPs on the linkage map were aligned to the salmon reference genome assembly (GCA_000233375.4) to determine their putative physical position (Additional File 1). There was a high positive correlation between the genetic map position and the reference sequence position of the SNPs (Table 2-1), and the number of SNPs per chromosome was dependent on chromosome sequence length (Figure 2-1). SNP density for the successfully genotyped and mapped markers from the 'ssalar1' array is relatively constant across the genome, with an average of 1 SNP per ~ 23 kb in the assembled chromosomes, and 1 SNP per 0.05 cM (male) and 0.07 cM (female) in the full linkage map.

Figure 2-1. Comparison of the number of SNPs in corresponding chromosomes and physical length retrieving from recent reference assembly (Genbank assembly reference GCA_000233375.4). The correlation was approximately 0.95.

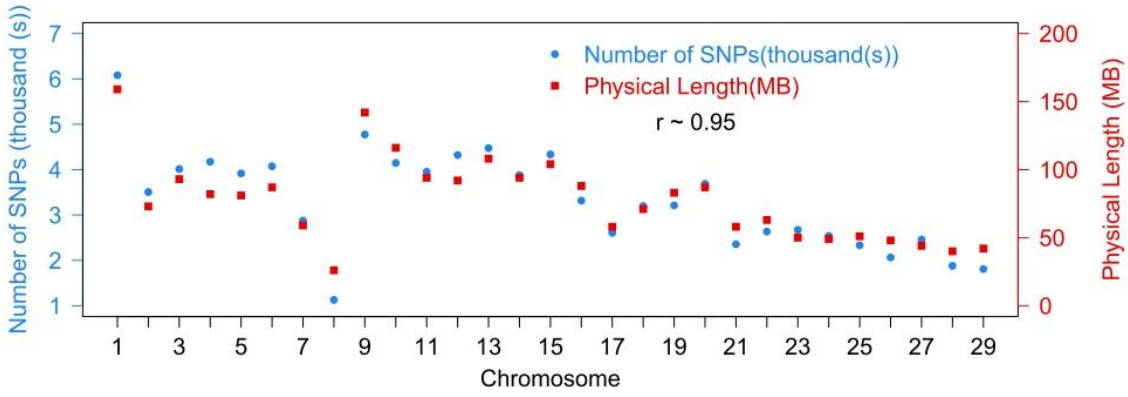


Table 2-1. The characteristics of the physical and genetic maps of the 29 Atlantic salmon (pairs of) chromosomes (Genbank reference GCA_000233375.4).

				Male		Female	
Chr.	SNPs	Physical Length (MB)*	Physical Length of Unassigned Contigs (MB)*	Max (cM)	Correlation[$]	Max (cM)	Correlation[$]
1	6,080	159	1.6	428.8	0.97	551.3	0.98
2	3,506	73	3.1	173.5	0.80	404.4	0.85
3	4,013	93	2.2	332.2	0.84	467.7	0.96
4	4,173	82	1.1	156.6	0.82	183.6	0.95
5	3,916	81	1.9	274.4	0.91	529.9	0.93
6	4,073	87	2.3	264.2	0.88	689.1	0.89
7	2,875	59	1.2	183.7	0.85	249.0	0.97
8	1,128	26	0.6	181.6	0.87	326.4	0.97
9	4,774	142	1.7	278.8	0.77	392.2	0.81
10	4,146	116	0.9	82.83	0.79	166.8	0.97
11	3,953	94	2.8	166.2	0.79	291.0	0.81
12	4,321	92	2.6	95.65	0.80	239.5	0.80
13	4,472	108	1.3	178	0.62	213.8	0.91
14	3,878	94	1.4	96.4	0.73	123.5	0.92
15	4,335	104	1.9	77.34	0.64	136.9	0.91
16	3,316	88	2.3	141.9	0.80	137.7	0.90
17	2,607	58	2.0	171.2	0.90	307.2	0.96
18	3,196	71	1.4	91.68	0.85	105.9	0.92
19	3,210	83	1.5	74.49	0.76	103.2	0.90
20	3,687	87	1.5	96.52	0.82	112.5	0.93
21	2,355	58	0.7	93.2	0.80	159.1	0.84
22	2,634	63	0.4	73.64	0.74	78.0	0.88
23	2,670	50	0.6	77.53	0.65	84.4	0.96
24	2,538	49	0.3	379	0.91	458.2	0.97
25	2,332	51	0.7	147	0.92	175.3	0.96
26	2,063	48	2.2	166.2	0.92	161.8	0.95
27	2,458	44	0.4	73.31	0.72	72.6	0.91
28	1,878	40	0.7	143.1	0.94	156.0	0.99
29	1,809	42	0.6	70.24	0.73	76.4	0.88
Total	96,396	2,242	41.9	4769	-	7153.2	-
Avg	3,324	77	1.4	164.5	0.81	246.7	0.92

*: The physical length is taken from the latest Atlantic salmon genome assembly [Genbank reference GCA_000233375.4 [16]], and 'unassigned contigs' are those that were unplaced on the reference assembly but mapped to the chromosome in the linkage map.

$: The correlation between the genetic distance of SNPs (cM) on the linkage map and the physical distance (bp) according to the reference genome assembly.

The most recent Atlantic salmon reference genome assembly (GCA_000233375.4) contains 2,240 MB of sequence contigs anchored to chromosomes (78 % of total assembly), and 647 MB of contigs that are not yet assigned to chromosome (22 % of total assembly). Linkage mapping was used extensively to orientate reference genome contigs and scaffolds, and identify putative misassemblies in the recently-published salmon genome paper [17]. However, those linkage maps are unpublished. In the current study, a total of 4,581 previously unassigned contigs comprising 41.9 MB of sequence were tentatively mapped to the 29 salmon chromosomes (Table 2-1, Additional File 2). While additional experiments would be required to confirm the correct position of these genome contigs, this linkage map has enabled an additional ~ 1 % of the entire reference genome assembly to be tentatively mapped to chromosomes, corresponding to ~ 6.5 % of the previously unassigned genome assembly. These contigs were spread across all 29 chromosome pairs (Table 2-1; details given in Additional File 1). Novel potentially misassembled regions were also identified in the reference sequence via regions of discordance between the linkage and physical maps, an example of which is between ~ 11.5 MB and 11.8 MB on Chromosome 26 (Additional File 3).

There were substantial differences in the patterns of recombination between the sexes. The female linkage map covered 7,153 cM (ranging from 72.6 to 689.0 cM per chromosome) whereas the male linkage map covered 4,769 cM (ranging from 70.2 to 428.8 cM per chromosome) (Table 2-1). Overall, the female map was ~ 1.5 × longer than the male map, consistent with previous Atlantic salmon SNP linkage maps [20, 23]. The pattern of recombination across the genome was notably different between the sexes, with female recombination rates being higher across much of the genome, except for some subtelomeric regions where male recombination was substantially higher (*e.g.* Figure 2-2). This phenomenon has been observed in several previous salmonid linkage maps [19–21, 23, 44], but the availability of the reference genome enables a more detailed investigation. Therefore, linkage and physical maps were aligned and a proxy of

recombination rate (number of centimorgans per megabase) was estimated at regular intervals on each chromosome, with each interval corresponding to 2 % of the total chromosome's physical length. The average recombination rate for each corresponding interval on the 29 chromosomes was calculated and graphed against the distance from the nearest telomere (Figure 2-3). The results highlight the phenomenon of markedly high male recombination in some sub-telomeric regions, on average ~ 10 × higher than regions of the genome nearer the middle of the chromosome (Figure 2-3).

Figure 2-2. A comparison between genetic and physical maps of a representative chromosome (Chr 22), reflecting the recombination pattern difference between males and females. Details of genetic distance and physical distance for all mapped loci are given in Additional File 1.

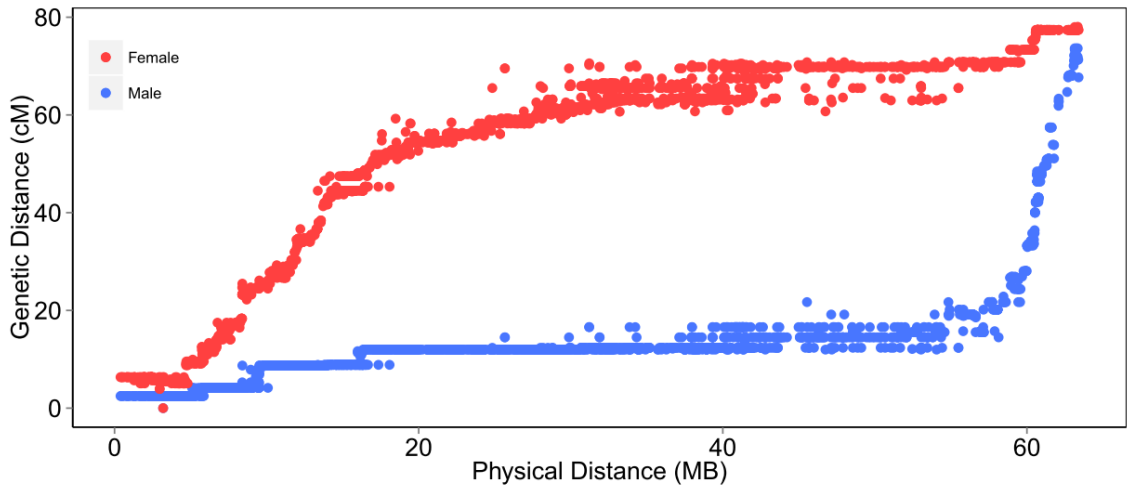
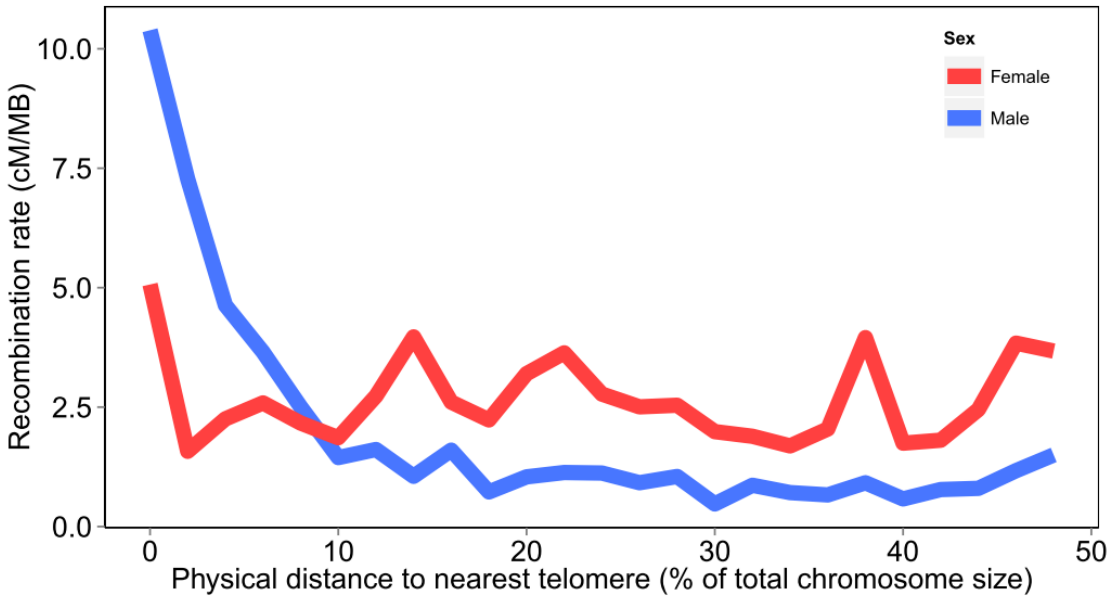


Figure 2-3. A comparison of male and female recombination level (cM / Mb) graphed according to physical distance from the nearest chromosome end (expressed as a percentage of total chromosome size in megabases).



2.4.2 Transcriptome Assembly and Annotation

To annotate the mapped SNPs and predict their function according to their position relative to putative genes, an annotated reference transcriptome was created. RNA-seq of 48 individual salmon fry yielded 927 M raw paired-end sequence reads, of which 93 % remained after trimming and filtering. Filtered reads were aligned to the most recent Atlantic salmon reference genome assembly (GCA_000233375.4; 82.2 % concordant pair alignment) to generate a reference transcriptome. The alignment resolved 202,009 putative transcripts corresponding to 65,803 putative genes, consisting of 36,846 single transcript genes and 28,957 multi-transcript genes (Table 2-2, Additional File 4). The average length of the transcripts was 4,127 bp with an N50 of 5,710, an N90 of 2,323 and > 90 % of transcripts longer than 500 bp. The assembled transcripts were annotated using BLASTx and BLASTn searches against the NCBI non-redundant protein and nucleic acid databases respectively. Of the 65,804 total putative genes, 58,416 (88.8 %)

showed significant similarity to known proteins, while an additional 2,732 (4.2 %) showed significant similarity to nucleotide entries in the NCBI non-redundant nucleotide database (Additional file 5). The proportion of unannotated genes was higher for the shorter transcript sequences (Additional file 6), but all transcripts were retained (since a relevant minimum size threshold was not apparent). The completeness of the transcriptome was evaluated against a set of 248 core eukaryotic genes described in Parra et al. [42] ; 247 of these genes were found in our transcriptome (BLASTn e-value < $E10^{-25}$), 222 of which had at least 90 % coverage, and 153 of which were fully covered. A total of 53,950 identified genes were located within chromosomes on the Atlantic salmon genome assembly, while the remaining 11,853 were aligned to unassigned contigs. Of these 11,853 genes, 1,647 (13.9 %) were located in contigs assigned to chromosomes using the linkage map of the current study (Table 2-1; Additional file 7).

Table 2-2. Summary statistics for the Atlantic salmon RNA-seq transcriptome assembly.

Transcriptome assembly details	**Number**
Transcripts	202,009
Genes	65,803
Single transcript genes	36,846
Multi-transcript genes	28,957
Genes in assembled chromosomes	53,950
Genes in unassigned contigs	11,853
Average transcript length	4,127
N50	5,710
N90	2,323
Transcripts > 500 bp	195,224
Genes annotated using protein database	58,416
Genes annotated using DNA database	2,732

2.4.3 SNP Annotation

The RNA-seq based transcriptome described above was used to predict open reading frames and protein sequences in order to annotate the SNPs present on the 'ssalar01' array (Table 2-3, Additional file 8). A total of 106,424 SNPs (95 %) matched a single genome location, while 2,857 SNPs matched two different genomic positions, related in part to the salmonid specific genomic duplication. An additional 880 SNPs mapped to 3 or more genome locations, indicative of repetitive elements or protein domains. It should be noted that filtering of SNPs during the design process for the array would have removed the majority of SNPs mapping to two or more locations [10]. The tentative annotation of all SNPs is given (Additional file 6), but only those mapping to unique genomic regions are described below. Of these 106,424 unique SNPs, 48,842 (45.9 %) were located in putative genes, with the remainder mapping to intergenic regions. Of the genic SNPs, the majority were in putative intronic regions (34,534 – 70.7 %), although 483 of these were associated with splicing regions and therefore have a higher likelihood of being functionally relevant. The remaining genic SNPs were mapped to putative UTRs (8,091), with a larger amount of SNPs in the 3' UTR than expected (6,224 vs 1,867 5'UTR); and to putative exons (5,856). A total of 2,465 putative non-synonymous SNPs were identified, in addition to 39 SNPs predicted to cause gain / loss of start / stop codons, which have a high likelihood of functional consequences (Additional File 8). As an example, a premature stop codon was found in phospholipase D, an enzyme which produces the signal molecule phosphatidic acid which is also a precursor for the biosynthesis of many other lipids [45]. The distribution of the SNP functional categories across the 29 chromosome pairs is given in Table 2-4. It is important to note that these predicted SNP effects will contain a proportion of false positives due to inevitable errors in the predicted structure of the genes. Nonetheless, their annotation combined with their linkage and physical mapping provides a valuable resource for users of the high density 'ssalar01' array in particular, and for salmonid genomics researchers in general.

Table 2-3. Predicted numbers, location and effect of the mapped SNPs according to their position on the annotated reference genome.

Summary of annotated SNPs					
Intergenic	57,582				
Genic	48,842	UTR	8,091	5’	1,867
				3’	6,224
		Intron	34,534	Splice region	483
				Non splice region	34,051
		Exon	5,856	Synonymous	3,352
				Non-synonymous	2,465
				Gain or loss of start / stop codon	39

Table 2-4. Number of predicted genes and functional categories of SNPs split according to chromosome.

Genes and SNPs per chromosome					
Chromosome	Genes	Exonic SNPs	Intronic SNPs	UTR SNPs	Intergenic SNPs
1	3,507	181	877	206	4,717
2	2,711	222	1,116	284	1,630
3	2,741	225	1,209	312	2,026
4	2,255	246	1,301	309	2,066
5	2,286	220	1,184	299	2,030
6	2,441	217	1,286	312	2,006
7	1,526	152	928	192	1,455
8	875	44	335	67	525
9	3,062	244	1,415	374	2,563
10	2,568	217	1,341	300	2,140
11	2,308	162	1,168	249	2,207
12	2,672	268	1,398	349	2,088
13	2,524	276	1,516	328	2,181
14	2,343	236	1,154	314	2,034
15	2,400	271	1,415	294	2,138
16	2,205	193	1,003	253	1,721
17	1,770	144	744	206	1,307
18	1,767	142	1,041	205	1,654
19	1,694	125	1,013	203	1,743
20	2,072	211	1,093	257	1,830
21	1,056	129	700	160	1,252
22	1,398	153	811	189	1,416
23	1,138	142	863	192	1,400
24	1,040	146	860	187	1,238
25	1,032	113	585	133	1,431
26	1,372	102	606	128	1,082
27	1,096	129	828	195	1,221
28	912	92	593	147	992
29	821	88	598	120	937
Total	55,592	5,090	28,981	6,764	51,030
Avg	1,917	176	999	233	1,760

2.5 Chapter Conclusions

A linkage map comprising > 96 K SNPs from the ‘ssalar01’ array was created, annotated and integrated with the reference genome assembly. This represents the highest density SNP linkage map for any salmonid species. Alignment of the linkage and physical maps revealed good agreement between genetic map, and the mapping allowed a further circa 1 % of the salmon reference genome assembly to be tentatively assigned to chromosomes. Marked heterochiasmy was observed, with male recombination rate substantially lower than females across much of the genome, but with a notably high level in some sub-telomeric regions. Finally, the mapped SNPs were annotated and categorised according to their predicted function. The map will be another useful resource for salmonid genomics research.

2.6 Conclusions

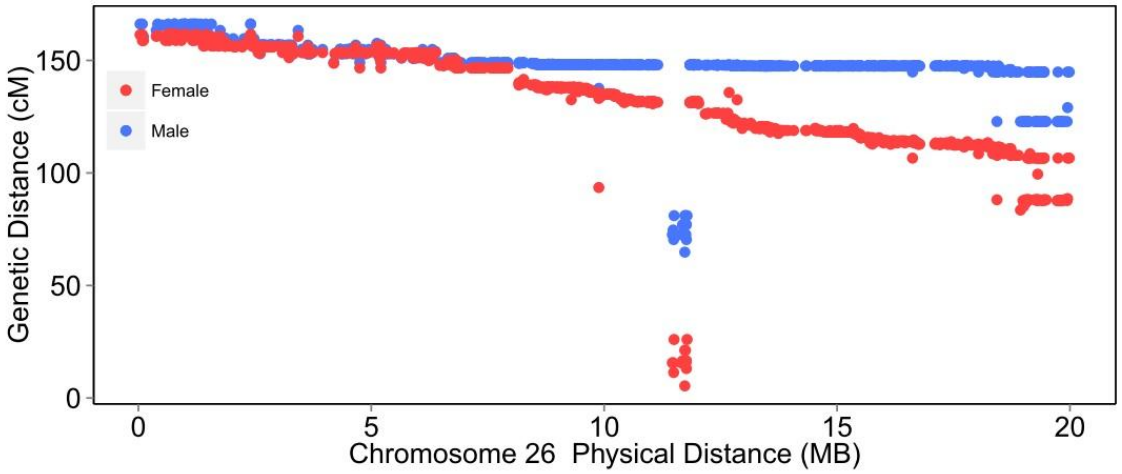
In this chapter, I successfully constructed a high density genetic map comprising approximately 100 K SNPs across 29 linkage groups. The results characterized the unique recombination landscape in male and female salmon, and also assigned 6.5 % of previously unmapped reference genome to genomic regions in corresponding chromosomes. This new map provides valuable genetic information to investigate the loci or QTLs associated with target traits that are studied in later chapters. Specifically, the linkage map integrated with the physical map will help to accurately select markers for QTL mapping, loci and gene identification associated with the traits studied, and SNP panel construction used for routine genomic selection breeding program. The outcomes of these analyses will be discussed in later chapters, including chapter 3 to chapter 6.

2.7 Additional Files

Additional file 1. Details of the linkage map of the 29 chromosomes. The file includes SNP IDs, their position on the linkage map (cM), their position on the reference genome (contig ID and position in bp), the flanking sequence and the two SNP alleles. There are separate sheets for each of the 29 chromosomes. Link: https://goo.gl/LoNajr

Additional file 2. Unassigned genome contigs. The position of each of the previously unassigned reference genome contigs (Genbank assembly reference GCA_000233375.4, [16]) on the 29 linkage groups. Link: https://goo.gl/9Jq1Si

Additional file 3. Potential misassembly in reference genome. A graph of the linkage map versus the physical map for the first 20 MB of chromosome 26, highlighting a potential misassembly of a region between 11.5 and 11.8 MB.

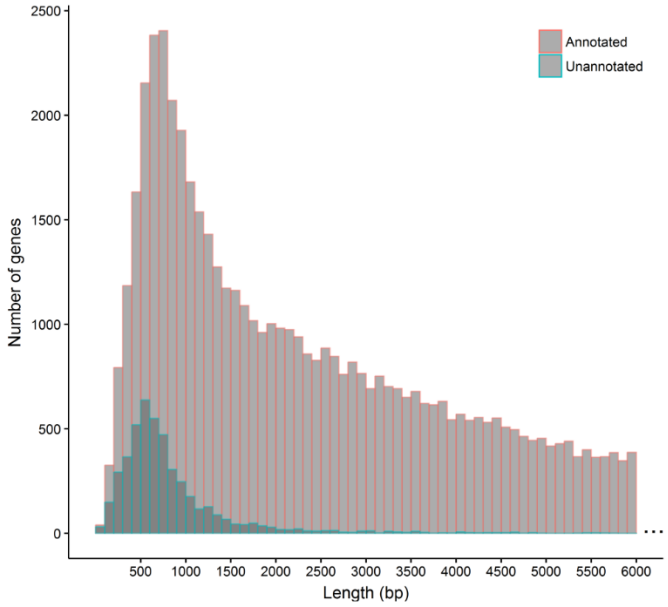


Additional file 4. Reference transcriptome sequence (fasta). All the sequences of the Atlantic salmon transcriptome in fasta format. Link: https://goo.gl/StCveN

Additional file 5. Reference transcriptome annotation. The position of every putative gene in the genome (chromosome or scaffold, start and end positions, and DNA strand),

length in base pairs and annotation against NCBI's databases (description of the best match, e-value and similarity) are shown. Link: https://goo.gl/jgHkat

Additional file 6. Length distribution of annotated and unannotated genes. Length distribution (bin width = 100bp) for the annotated and unannotated genes of the transcriptome.



Additional file 7. Number of previously unmapped genes assigned to chromosomes using the linkage map. Number of unmapped genes (placed in previously unassigned genome contigs), assigned to each Atlantic salmon chromosome using the linkage map. Link: https://goo.gl/89fjDL

Additional file 8. SNP annotation. The position of each SNP in the Atlantic salmon genome, the SNP-array ID, the genomic and the alternative variants, the effects of the alternative variant and, if applicable, the affected protein, gene and its annotation are shown. In some cases two different genes / proteins are affected by the SNP due to them being overlapping and transcribed from different DNA strands. SNPs aligning with more

than one region of the genome assembly are given in separate worksheets. Link: https://goo.gl/kyvSOe

2.8 Acknowledgements

The authors gratefully acknowledge (i) Mike Stear, Karim Gharbi, Bill Roy, Matt Tinsley, Alastair Hamilton, Derrick Guy and Alan Tinch for provision of the data and samples from the population used for map construction; (ii) David Verner-Jeffreys, Richard Paley, Georgina Rimmer and Ian Tew at the Centre for Environment Fisheries and Aquaculture Science (Cefas) for contribution to the disease challenge experiment from which the RNA-Seq samples were derived; and (iii) Lep-Map 2 software developer, Pasi Rastas, for the technical suggestions for the linkage map construction. RNA sequencing and SNP array genotyping were carried out by Edinburgh Genomics, The University of Edinburgh. Edinburgh Genomics is partly supported through core grants from NERC (R8/H10/56), MRC (MR/K001744/1) and BBSRC (BB/J004243/1). This research was supported by Biotechnology and Biological Sciences Research Council (BBSRC) grants (BB/H022007/1, BB/F002750/1, BB/F001959/1) awarded to The Roslin Institute and University of Stirling, and by BBSRC Institute Strategic Funding Grants to The Roslin Institute (BB/J004235/1, BB/J004324/1, BB/J004243/1). DR was funded by a postgraduate grant from Fundación Barrié. The authors also acknowledge the support of the MASTS pooling initiative (The Marine Alliance for Science and Technology for Scotland) in the completion of this study. MASTS is funded by the Scottish Funding Council (grant reference HR09011) and contributing institutions. The authors declare that they have no competing interests.

2.9 References

1. Goddard ME, Hayes BJ: Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* 2009, 10:381–91.

2. Danzmann RG, Gharbi K: Gene mapping in fishes: a means to an end. *Genetica* 2001, 111:3–23.

3. Liu ZJ: *Next Generation Sequencing and Whole Genome Selection in Aquaculture*. Blackwell Publishing Ltd., Des Moines, IA., 2010.

4. Meuwissen THE, Hayes BJ, Goddard ME: Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 2001, 157:1819-1829.

5. Goddard ME, Hayes BJ, Meuwissen TH: Genomic selection in livestock populations. *Genet Res* 2010, 92:413–421.

6. Sonesson AK: Genomic Selection for Aquaculture: Principles and Procedures. In *Next Gener Seq Whole Genome Sel Aquac*; 2010:151–163.

7. Correa K, Lhorente JP, López ME, Bassini L, Naswa S, Deeb N, Di Genova A, Maass A, Davidson WS, Yáñez JM: Genome-wide association analysis reveals loci associated with resistance against Piscirickettsia salmonis in two Atlantic salmon (Salmo salar L.) chromosomes. *BMC Genomics* 2015, 16:854.

8. Yáñez JM, Newman S, Houston RD: Genomics in aquaculture to better understand species biology and accelerate genetic progress. *Front Genet* 2015, 6(April):128.

9. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML: Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat*

Rev Genet 2011, 12:499–510.

10. Houston RD, Taggart JB, Cézard T, Bekaert M, Lowe NR, Downing A, Talbot R, Bishop SC, Archibald AL, Bron JE, Penman DJ, Davassi A, Brew F, Tinch AE, Gharbi K, Hamilton A: Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). *BMC Genomics* 2014, 15:90.

11. Shi Y, Wang S, Gu Z, Lv J, Zhan X, Yu C, Bao Z, Wang A: High-density single nucleotide polymorphisms linkage and quantitative trait locus mapping of the pearl oyster, Pinctada fucata martensii Dunker. *Aquaculture* 2014, 434:376–384.

12. Wang L, Wan ZY, Bai B, Huang SQ, Chua E, Lee M, Pang HY, Wen YF, Liu P, Liu F, Sun F, Lin G, Ye BQ, Yue GH: Construction of a high-density linkage map and fine mapping of QTL for growth in Asian seabass. *Sci Rep* 2015, 5:16358.

13. Fierst JL: Using linkage maps to correct and scaffold de novo genome assemblies: Methods, challenges, and computational tools. *Front Genet* 2015(June), 6:220.

14. Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens H-J, Crooijmans RPMA, Besnier F, Lathrop M, Muir WM, Wong GK-S, Gut I, Andersson L: A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Res* 2009, 19 (3):510–519.

15. Tortereau F, Servin B, Frantz L, Megens H-J, Milan D, Rohrer G, Wiedmann R, Beever J, Archibald AL, Schook LB, Groenen MAM: A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics* 2012, 13:12.

16. Davidson WS, Koop BF, Jones SJM, Iturra P, Vidal R, Maass A, Jonassen I, Lien S, Omholt SW: Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome Biol*

2010, 11:403.

17. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, Grammes F, Grove H, Gjuvsland A, Walenz B, Hermansen RA, von Schalburg K, Rondeau EB, Di Genova A, Samy JKA, Olav Vik J, Vigeland MD, Caler L, Grimholt U, Jentoft S, Inge Våge D, de Jong P, Moen T, Baranski M, Palti Y, Smith DR, et al.: The Atlantic salmon genome provides insights into rediploidization. *Nature* 2016, 533(7602):200-5.

18. de Boer JG, Yazawa R, Davidson WS, Koop BF: Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics* 2007, 8:422.

19. Sakamoto T, Danzmann RG, Gharbi K, Howard P, Ozaki A, Khoo SK, Woram RA, Okamoto N, Ferguson MM, Holm L-E, Guyomard R, Hoyheim B: A Microsatellite Linkage Map of Rainbow Trout (Oncorhynchus mykiss) Characterized by Large Sex-Specific Differences in Recombination Rates. *Genetics* 2000, 155:1331–1345.

20. Lien S, Gidskehaug L, Moen T, Hayes BJ, Berg PR, Davidson WS, Omholt SW, Kent MP: A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genomics* 2011, 12:615.

21. Miller MR, Brunelli JP, Wheeler PA, Liu S, Rexroad 3rd CE, Palti Y, Doe CQ, Thorgaard GH: A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Mol Ecol* 2011, 21:237-249.

22. Brieuc MSO, Waters CD, Seeb JE, Naish K a: A dense linkage map for Chinook salmon (Oncorhynchus tshawytscha) reveals variable chromosomal divergence after an ancestral whole genome duplication event. *G3 (Bethesda)* 2014, 4:447–60.

23. Gonen S, Lowe NR, Cezard T, Gharbi K, Bishop SC, Houston RD: Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing. *BMC Genomics* 2014, 15:166.

24. Allendorf FW, Bassham S, Cresko WA, Limborg MT, Seeb LW, Seeb JE: Effects of Crossovers Between Homeologs on Inheritance and Population Genomics in Polyploid-Derived Salmonid Fishes. *J Hered* 2015, 106 (3):217–227.

25. Wright Jr. JE, Johnson K, Hollister A, May B: Meiotic models to explain classical linkage, pseudolinkage, and chromosome pairing in tetraploid derivative salmonid genomes. *Isozymes Curr Top Biol Med Res* 1983, 10:239–260.

26. Allendorf FW: Tetraploidy and the evolution of salmonid fishes, pp.55-93 in *Tetraploidy and the Evolution of Salmonid Fishes*, edited by B.J. Turner, Plenum Press, New York. 1984.

27. Yáñez JM, Naswa S, López ME, Bassini L, Correa K, Gilbey J, Bernatchez L, Norris A, Neira R, Lhorente JP, Schnable PS, Newman S, Mileham A, Deeb N, Di Genova A, Maass A: Genome-wide single nucleotide polymorphism (SNP) discovery in Atlantic salmon (*Salmo salar*): validation in wild and farmed American and European populations. *Mol Ecol Resour* 2016, 16(4):1002–1011.

28. Tsai HY, Hamilton A, Tinch AE, Guy DR, Gharbi K, Stear MJ, Oswald M, Bishop SC, Houston RD: Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. *BMC Genomics* 2015, 16:969.

29. Odegård J, Moen T, Santi N, Korsvoll SA, Kjøglum S, Meuwissen THE: Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). *Front Genet*

2014, 5(November):402.

30. Gharbi K, Matthews L, Bron J, Roberts R, Tinch A, Stear MJ: The control of sea lice in Atlantic salmon by selective breeding. *J R Soc Interface* 2015, 12:0574.

31. Housto RD, Bisho SC, Gu DR, Tinc AE, Taggar JB, Bro JE, Downin A, Stea MJ, Gharb K, Hamilto A: Genome Wide Association Analysis for Resistance to Sea Lice in Atlantic Salmon: Application of a Dense SNP Array. *Proc 10th World Congr Genet Appl to Livest Prod* 2014:10–12.

32. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001, 29 (1):308–311.

33. Rastas P, Calboli FCF, Guo B, Shikano T, Merilä J: Construction of ultra-dense linkage maps with Lep-MAP2: stickleback F2 recombinant crosses as an example. *Genome Biol Evol* 2016. In Press.

34. Gilbey J, Verspoor E, McLay A, Houlihan D: A microsatellite linkage map for Atlantic salmon (Salmo salar). *Anim Genet* 2004, 35:98–105.

35. Moen T, Hoyheim B, Munck H, Gomez-Raya L: A linkage map of Atlantic salmon (Salmo salar) reveals an uncommonly large difference in recombination rate between the sexes. *Anim Genet* 2004, 35:81–92.

36. Houston RD, Haley CS, Hamilton A, Guy DR, Mota-Velasco JC, Gheyas AA, Tinch AE, Taggart JB, Bron JE, Starkey WG, McAndrew BJ, Verner-Jeffreys DW, Paley RK, Rimmer GSE, Tew IJ, Bishop SC: The susceptibility of Atlantic salmon fry to freshwater infectious pancreatic necrosis is largely explained by a major QTL. *Heredity (Edinb)* 2010, 105:318–27.

37. Bolger AM, Lohse M, Usadel B: Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma* 2014, 30 (15):2114–2120.

38. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013, 14:1–13.

39. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012, 9:357–9.

40. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 2012, 7:562–578.

41. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25 (17):3389–3402.

42. Parra G, Bradnam K, Korf I: CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinforma* 2007, 23 (9):1061–1067.

43. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. *Fly (Austin)* 2012, 6:80–92.

44. Seeb JE, Pascal CE, Grau ED, Seeb LW, Templin WD, Harkins T, Roberts SB: Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids. *Mol Ecol Resour* 2011, 11:335–48.

45. McDermott M, Wakelam MJO, Morris AJ: Phospholipase D. *Biochem Cell Biol* 2004, 82:225–253.

Chapter 3

The genetic architecture of growth and fillet traits in farmed Atlantic salmon (*Salmo salar*)

Tsai, H.Y., Hamilton, A., Guy, D.R., Tinch, A.E., Bishop, S.C., Houston, R.D.

DOI: 10.1186/s12863-015-0215-y

RDH and SCB proposed and designed the experiments. Tissue samples and phenotype measurements were supported by AET, DRG and AH. Data analyses were performed by HYT and RDH. Manuscript was written by HYT and RDH.

This chapter has been published in BMC Genetics at the following URL:
http://www.biomedcentral.com/1471-2156/16/51

3.0 Introduction

Growth traits are likely to be polygenic in nature, but highly heritable, compared with disease resistance traits. Previous literature indicates that QTLs affecting growth phenotypes in salmon vary between different populations. The main objectives of this chapter were to characterize the genetic basis of growth traits in a large (n = ~5000) commercial farmed salmon population containing 198 full-sibling families, and to estimate the heritabilities of growth and muscle-related traits, and to map the QTLs using two-stage QTL mapping analysis. The advantages of a two-stage QTL mapping applied in salmon breeding is primarily because of the large recombination ratio difference between male and female salmon, which has been characterized in their linkage maps described in chapter 2. Initially, the lower recombination rate in male salmon can provide relatively higher power and lower genotyping cost to identify the potential QTLs and linkage groups of interest in sire-based analysis. Subsequently, the QTLs of target will be confirmed using additional molecular markers aiming to aid the QTL mapping resolution using dam-based analysis. Finally, the findings of this chapter can help to understand the genetic basis of growth traits, not only in fish production via salmon farming, but also in evolutionary and conservation biology in salmonids, as the QTL mapping results will be compared with other closed species (*e.g.* rainbow trout (*Oncorhynchus mykiss*)).

3.1 Abstract

Background

Performance and quality traits such as harvest weight, fillet weight and flesh color are of economic importance to the Atlantic salmon aquaculture industry. The genetic factors underlying these traits are of scientific and commercial interest. However, such traits are typically polygenic in nature, with the number and size of QTL likely to vary between studies and populations. The aim of this study was to investigate the genetic basis of several growth and fillet traits measured at harvest in a large farmed salmon population by using SNP markers. Due to the marked heterochiasmy in salmonids, an efficient two-stage mapping approach was applied whereby QTL were detected using a sire-based linkage analysis, a sparse SNP marker map and exploiting low rates of recombination, while a subsequent dam-based analysis focused on the significant chromosomes with a denser map to confirm QTL and estimate their position.

Results

The harvest traits all showed significant heritability, ranging from 0.05 for fillet yield up to 0.53 for the weight traits. In the sire-based analysis, 1,695 offspring with trait records and their twenty sires were successfully genotyped for the SNPs on the sparse map. Chromosomes 13, 18, 19 and 20 were shown to harbor genome-wide significant QTL affecting several growth-related traits. The QTL on chr. 13, 18 and 20 were detected in the dam-based analysis using 512 offspring from ten dams and explained approximately 6 – 7 % of the within-family variation in these traits.

Conclusions

We have detected several QTL affecting economically important complex traits in a commercial salmon population. Overall, the results suggest that the traits are relatively polygenic and that QTL tend to be pleiotropic (affecting the weight of several components of the harvested fish). Comparison of QTL regions across studies suggests that harvest trait QTL tend to be relatively population-specific. Therefore, the application of marker or genomic selection for improvement in these traits is likely to be

most effective when the discovery population is closely related to the selection candidates (*e.g.* within-family genomic selection).

3.2 Chapter Introduction

Traditional selective breeding has rapidly improved economically important traits in aquaculture species, such as growth and disease resistance in aquaculture species [1]. Atlantic salmon have been more extensively studied than most other aquaculture species due to its high economic value and the significant scientific interest in salmonid species [2]. However, the genetic factors affecting some complex traits of economic importance, such as size, morphology and composition, are not yet well known. The limitations to detecting and defining these genetic factors may include a previous lack of genomic resources, the polygenic nature of the traits in question, and the relatively recent whole genome duplication (*e.g.* [3, 4]) in the salmonid lineage.

Genomic resources for salmonids are rich in comparison to most aquacultural species [5]. Benefitting from the development of next generation sequencing (*e.g.* [6]), abundant genetic markers have been discovered in most salmonid species (*e.g.* [7–10]). Many other genomic resources and salmonid-specific databases are available, *e.g.* the Genomics Research on All Salmon (GRASP, http://web.uvic.ca/grasp/) and SalmonDB (http://salmondb.cmm.uchile.cl/). Furthermore, the genomes of rainbow trout [3] and Atlantic salmon [2] have been sequenced and assembled, which provide reference sequences for genomic studies of these and other salmonid species [11].

Understanding the genetic basis of phenotypic variation is a fundamental goal of biological research. Quantitative genetic analysis has been widely used to apportion variation in the traits of interest into genetic and environmental factors [12]. A further goal is to ascertain the genetic architecture of these traits, and quantitative trait loci (QTL) mapping is useful for this purpose. This approach has been widely applied in most farmed animal and plant species to improve genetic breeding programs [13–16]. To date, QTL mapping relating to the growth performance of farmed salmonid species have been undertaken in Atlantic salmon [17–21], Coho salmon [22, 23], Arctic char [24], Chinook salmon [25] and Rainbow trout [26, 27]. The loci associated with these

apparently polygenic growth traits tend to vary between studies, which may reflect population differences or gene by environment interaction.

Traits of economic interest in aquaculture species include those pertaining to the efficient production of high quality fillets. As such, overall growth rate is important, alongside the relative proportion of particular components of the fish (fillet, guts, and head, etc.). Ultimately, fillet weight is a key economic trait, and variation in this characteristic significantly depends on the proliferation and composition of white and red muscle. Muscle cell development and proliferation are part of a complex regulatory process and intricately linked with the development of the skeleton. These processes are typically controlled by networks involving many genes and biological pathways [28]. As such, a polygenic architecture of variation in this trait may be expected. Previous studies have shown that the less desirable parts of Atlantic salmon (*e.g.* head weight and vertebral weight) have a significant positive correlation with desirable traits such as harvest and fillet yields [29]. By detecting and selecting haplotypes at specific QTL, it may be possible to improve the proportion of fillet within the fish for any given growth rate (albeit caution should also be applied to ensure overall wellbeing and robustness of the fish).

The objective of this study was to detect and characterize QTL affecting growth and fillet characteristics in farmed salmon, using SNP markers genotyped in several large families reared under commercial aquaculture conditions. Due to the lower recombination rate observed throughout much of the genome in male salmon, compared to female salmon [30], the efficiency of QTL detection is increased by using a two stage analysis. In this strategy, QTL are first detected in a sire analysis using few markers per chromosome, and the chromosomes harbouring significant QTL are then genotyped for additional markers and analysed using dam mapping parents. Here, we use this approach with the overall target of improving understanding of the genetic regulation of growth and fillet characteristics in Atlantic salmon, and providing candidate regions for

potential application in marker-based selection to capture within-family variation in these traits.

3.3 Methods

3.3.1 Animals and Phenotype Measurement

A commercial salmon population comprising 198 full-sibling families derived from 136 sires and 198 dams (Landcatch Natural Selection, Ormsary, UK) was utilized in this experiment. Details of this population have been previously published [31–33]. Briefly, approximately 5,000 fish were harvested at ~ 3 years of age and measured for overall and component weight traits: harvest weight (kg), gutted weight (kg), deheaded weight (kg), fillet weight (kg), gutted yield (%), fillet yield (%), head weight (kg), gut weight (kg), body waste weight (kg) and total waste weight (kg), fat percentage [% as estimated using a Torry Fatmeter (Distell Ltd, Aberdeen, Scotland)]; and fillet color [assessed visually using the Roche SalmoFan scale (Hoffmann-La Roche, U.K.), ranging from 20 (Yellow) to 34 (Red)]. Details of trait measurements at harvest are given in Powell et al. [29]. A fin clip sample of each fish was retained for DNA extraction. All animals were reared and harvested in accordance with all relevant national and EU legislation concerning health and welfare. Landcatch are accredited participants in the RSPCA Freedom Foods standard, the Scottish Salmon Producers Organization Code of Good Practice, and the EU Code-EFABAR Code of Good Practice for Farm Animal Breeding and Reproduction Organizations. The traits of fat percentage and gut weight were $\log_{10}$ transformed to approximate a normal distribution. Two generation pedigree records were available for all fish and the sex of the offspring was not observable at harvest and processing. Heritability estimates for some of the traits have been estimated previously in the larger population from which the QTL families were sampled [31, 32]. For gut, head, waste and total waste weight, the polygenic heritability was estimated in this larger population using a simple animal model, $Y_{ij} = \mu + A_i + e_{ij}$, where Y_{ij} is the trait value

measured in the individual i, μ is the overall mean value of the trait, A_i is the additive genetic effect of the individual based on the pedigree information and e_{ij} is the residual error. The heritability for each of the traits was estimated using the above model, and the procedure was described in Tsai et al. [32].

3.3.2 SNP Marker Selection and Genotyping

To account for the large differences in recombination rate between male and female salmon, a two-stage QTL detection and mapping strategy was employed [30, 34]. Stage 1 used sire mapping parents (low recombination), with few markers per chromosome to detect chromosomes containing putative QTL. Stage 2 used dam mapping parents, with a denser marker coverage, to confirm QTL on significant chromosomes and estimate QTL position. For stage 1, the twenty sires in the population with the most progeny were chosen for analysis (total n = 1,695). The sparse panel of SNP markers described in Gonen et al. [35], largely taken from Moen et al. [36], were provided to LGC Genomics (Herts, U.K.) for the design of Kompetitive Allele Specific PCR (KASP) assays (see details at http://www.lgcgroup.com/products/kasp-genotyping-chemistry/kasp-technical-resources/#.VVUKo_1waM8) for genotyping. From these, a total of 51 informative SNPs, with one to three SNPs per chromosome, were genotyped in all 1,695 offspring (Table S1). Stage 2 aimed to confirm the QTL detected in stage 1 and to estimate their position on the chromosome. Therefore, stage 2 focused on three putative QTL-containing chromosomes (chr. 13, 18, and 20) detected in stage 1. Thirty additional segregating SNP markers (Table S1) [9] were chosen to be positioned at regular (~ 10 cM) intervals across the candidate chromosome according to published linkage maps. As such, it was anticipated that this marker density would be sufficient to estimate approximate position of QTL on chromosomes. These SNPs were selected for genotyping in the ten dams with the largest number of offspring. A total of ten, eight and eight informative SNPs from chr. 13, 18 and 20 respectively were genotyped in the 512 offspring of these dams (which were a subset of the offspring genotyped offspring in stage 1).

3.3.3 Linkage and QTL Mapping

Sex-specific genetic maps were constructed using Crimap version 2.4 [37]. The 'prepare' option was used to create the input files (markers had previously been assigned to linkage groups based on a LOD score of > 4.0), followed by the 'build' option to estimate marker order, and 'fixed' option to estimate the map distance between the markers. Where relevant, the 'flipsn' option was used to test different order permutations and determine the most likely marker order.

For both sire and dam based QTL detection, a two stage linear regression-based linkage analysis was performed using the GridQTL software [38]. The conditional probability of inheriting a particular haplotype from the sire or dam was inferred from the marker genotypes in all offspring, at 1 cM intervals. Subsequently, the trait value was regressed on the probability that a particular haplotype allele was inherited from the sire (stage 1) or the dam (stage 2). At each genomic location, the model containing a single QTL is compared to a model with no QTL resulting in an F Ratio statistic. The chromosome-wide significance thresholds for each trait were computed by permutation using 10,000 iterations per chromosome. With 29 chromosomes, the expected number of false positive was 1.45 at the 5 % significance threshold, and 0.29 at the 1 % significance threshold per genome scan respectively. The genome-wide thresholds were determined by applying the Bonferroni correction [39] to 29 independent chromosomes. In addition, in the stage 2 dam-based analysis, the confidence intervals for the QTL were estimated using bootstrapping with 10,000 permutations. In order to estimate the size of the effect of the significant QTL on the traits, the within-family variation explained by the QTL (PVE) was calculated using the following equation: $h^2_{QTL} = 4[1-(MSE_{full} / MSE_{reduced})]$ for sire-based analysis. For the dam-based analysis, because the dams were nested within sires (full-sibling families), the estimated equation was revised to $h^2_{QTL} = 2[1-(MSE_{full} / MSE_{reduced})]$, where the MSE_{full} is the mean square error of the performed model including the QTL and $MSE_{reduced}$ is the model including the family mean only.

For traits related to the component weights of the fish, the QTL analyses were repeated including harvest weight as a covariate in the analysis. This was done to assess and distinguish QTL associated with an overall growth effect on the fish, versus QTL associated with proportional growth of specific components (*e.g.* fillet and waste, etc.).

3.4 Results

Trait records of 1,695 offspring derived from twenty sire families were obtained from a larger dataset of ~ 5,000 salmon measured at harvest (~ 3 years old). The heritability of the weight traits was significant and consistent with previous estimated (h^2 = 0.52 to 0.53). For the traits not previously analysed in this population (*i.e.* gut, head, waste and total waste weight) the heritabilities ranged from 0.15 to 0.32. Summary statistics from the QTL-mapping offspring and population-wide estimates of heritability for these traits are given in Table 3-1-a. The weight traits showed a high phenotypic and genetic correlation (Table 3-1-b) and fitting overall harvest weight as a covariate in the animal model reduced the estimated h^2 for the component traits to 0.02 – 0.05 (although these were still significantly different from zero).

Table 3-1-a. Summary statistics and heritabilies for the phenotypes used in this study.

Trait	Sample Size†	Mean (SD)	Heritability (SE) [29]
Harvest Weight	1524	2.57 (0.63)	0.52 (0.05)
Gutted Weight	1616	2.35 (0.58)	0.53 (0.05)
Gutted Yield	1447	0.92 (0.02)	0.04 (0.01)
Deheaded Weight	1604	2.06 (0.52)	0.52 (0.05)
Fillet Weight	1516	1.70 (0.42)	0.53 (0.05)
Fillet Yield	1363	0.66 (0.04)	0.05 (0.02)
Fat Percentage	1679	12.2 (5.58)	0.18 (0.03)
Fillet Colour	1322	29.0 (0.73)	0.14 (0.03)
Head Weight	1475	0.32 (0.08)	0.21 (0.03)*
Gut Weight	1447	0.42 (0.08)	0.30 (0.04)*
Body Waste Weight	1426	0.33 (0.12)	0.15 (0.02)*
Total Waste Weight	1422	0.65 (0.17)	0.32 (0.04)*

Gut weight (kg) = harvest weight - gutted weight; Head weight (kg) = gutted weight - deheaded weight.

Waste weight (kg) = deheaded weight - fillet weight (weight of vertebrae and caudal fin); Total waste weight (kg) = head weight + body waste weight.

*: The heritability was estimated in this study and the used population was the same as Tsai et al. [32].

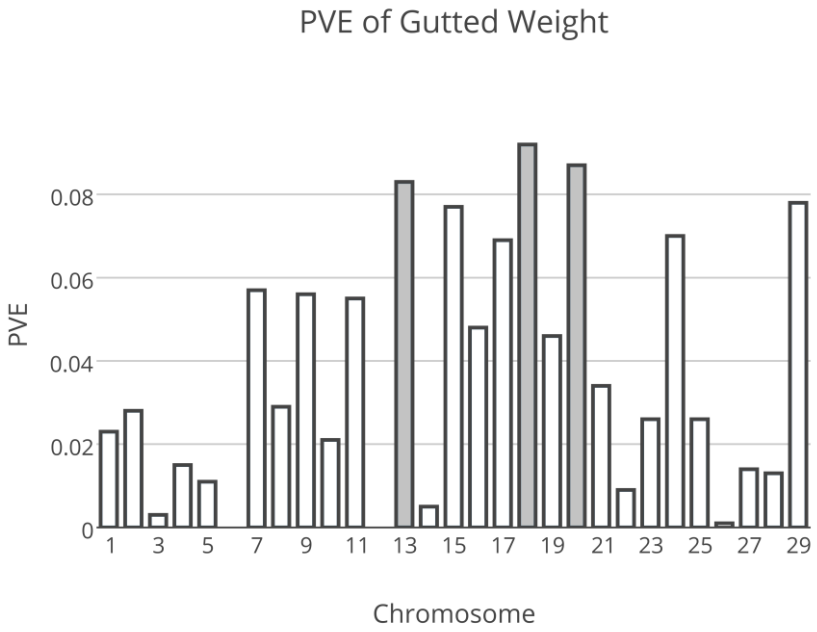
†: Only the number of individuals used in the calculation is shown, after removal of missing data.

In total, 51 SNP markers dispersed over all 29 chromosomes were successfully genotyped in the parents and offspring. In the sire-based QTL mapping analysis, a total of 13 chromosomes showed suggestive evidence for a QTL (chromosome-wide $p < 0.05$), while four chromosomes showed a significant effect on growth-related traits at the genome-wide leve1 (chr. 13, 18, 19, and 20; Table 3-2, Figure 3-1).

The QTL typically affected several of the weight measurements and, given the high phenotypic correlations between these traits (r ~ 0.97-1.00), it is plausible that these results reflect single pleiotropic QTL on each chromosome, rather than distinct linked QTL. Interestingly, when harvest weight was fitted as a covariate (as a proxy for an overall measure of growth), the QTL affecting the component traits on chr. 18, 19 and 20 were no longer significant, suggesting these QTL affect overall growth of the fish. In contrast, on chr. 13, most of the QTL effects for the component traits remained after fitting the covariate, suggesting putative proportional differences in the growth of components of the fish. In addition, four new QTLs (chr. 12, 22, 23, and 25) reached chromosome-wide significance in the sire-based analysis with the inclusion of harvest weight as a covariate in the analysis (Table 3-2). The proportion of within-family phenotypic variance explained (PVE) varied between 8 and 10 % for the genome-wide significant QTL in the sire-based analysis, suggesting QTL of moderate but not large effect in this population.

Three of the genome-wide significant QTL in the sire-based analysis (chr. 13, 18, and 20) were tested in a dam-based analysis using 512 offspring from ten dams, and a denser SNP marker map of the significant chromosomes (Table S1). The genome-wide significant QTL affecting gutted, deheaded and total waste weight on chr. 20 was confirmed in the dam-based analysis and mapped to a best estimated position of 21, 19 and 14 cM respectively, although the 95 % confidence intervals encompassed the entire linkage map for this chromosome (Table 3-3). The evidence for QTL on chr. 13 and 18 was not as strong in the dam-based analysis, with only gut weight (chr. 13) and gutted weight (chr. 18) showing chromosome-wide significance (in the analysis with harvest weight included as a covariate).

Figure 3-1. The distribution of PVE according to chromosome in the sire-based analysis for the representative weight trait of gutted weight. Gray represents the chromosome showing genome-wide significance ($p < 0.05$) in sire-based analysis. Chromosome 20 also showed chr-wide significance in dam-based analysis ($p < 0.05$).



For the chr. 20 QTL, there were three sires and three dams segregating for a QTL affecting at least one weight trait, and the average size of the allelic substitution effect for deheaded weight of the salmon in segregating parents was consistent across all segregating parents, with an average effect of 620 grams (Table 3-4).

Table 3-1-b. Genetic and phenotypic correlation of traits using in this study.

Genetic / Phenotypic	Harvest Weight	Gutted Weight	Gutted Yield	Deheaded Weight	Fillet Weight	Fillet Yield	Fat Percentage	Fillet Colour	Gut Weight	Head Weight	Body Waste Weight	Total Waste Weight
Harvest Weight	-	1.00	0.16	1.00	1.00	0.35	0.84	-0.17	-0.96	0.97	1.00	0.98
Gutted Weight	1.00	-	0.19	1.00	1.00	0.33	0.83	-0.20	-0.95	0.98	0.99	0.99
Gutted Yield	-0.02	0.06	-	0.19	0.20	0.53	0.05	-0.27	0.13	0.08	0.06	0.09
Deheaded Weight	0.98	0.98	0.06	-	0.99	0.37	0.83	-0.19	-0.95	0.97	1.00	0.98
Fillet Weight	0.97	0.97	0.05	0.97	-	0.41	0.82	-0.20	-0.95	0.95	1.00	0.98
Fillet Yield	0.02	0.06	0.31	0.08	0.27	-	0.21	-0.15	-0.21	0.09	0.23	0.19
Fat Percentage	0.41	0.41	0.04	0.41	0.42	0.07	-	-0.19	-0.82	0.76	0.84	0.80
Fillet Colour	-0.08	-0.08	-0.02	-0.07	-0.08	0.03	-0.06	-	0.10	-0.24	-0.13	-0.12
Gut Weight	-0.77	-0.72	0.56	-0.71	-0.72	0.12	-0.30	0.05	-	-0.94	-0.99	-0.96
Head Weight	0.61	0.62	0.04	0.47	0.92	-0.09	0.21	-0.11	-0.45	-	0.99	1.00
Body Waste Weight	0.62	0.61	0.09	0.63	0.41	-0.65	0.25	-0.04	-0.42	0.63	-	1.00
Total Waste Weight	0.83	0.83	0.08	0.83	0.67	-0.48	0.31	-0.07	-0.59	0.88	0.93	-

Table 3-2. Results of sire-based QTL mapping analysis and proportion of phenotypic variance explained by each chromosome.

	Harvest Weight		Fillet Weight				Gutted Weight				Deheaded Weight				Fillet Yield			
					covariate[$]				covariate[$]				covariate[$]				covariate[$]	
Chr	F ratio	PVE	F ratio	PVE	F ratio	PVE	F ratio	PVE	F ratio	PVE	F ratio	PVE	F ratio	PVE	F ratio	PVE	F ratio	PVE
1	-	-	1.88*	0.047	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	1.96*	0	-	-	-	-	-	-	1.90*	n.a.†	1.89*	n.a.†
7	-	-	-	-	-	-	2.03*	0.057	-	-	2.07*	0.06	1.79*	0.037	-	-	-	-
9	-	-	2.04*	0.055	-	-	2.12*	0.056	-	-	2.26*	0.062	-	-	-	-	-	-
10	-	-	2.08*	0.053	-	-	-	-	-	-	-	-	-	-	-	-	-	-
11	2.10*	0.065	2.27*	0.077	2.26**	0.071	1.96*	0.055	-	-	1.91*	0.052	-	-	2.32*	n.a.†	2.31**	n.a.†
12	-	-	-	-	-	-	-	-	-	-	-	-	2.05*	0.078	-	-	-	-
13	2.49**	0.074	2.53**	0.075	1.90*	0.07	2.78**	0.083	-	-	2.67**	0.08	1.81*	0.034	2.37*	n.a.†	2.37**	n.a.†
15	2.12*	0.076	-	-	-	-	2.24*	0.077	-	-	2.43*	0.091	-	-	-	-	-	-
16	2.23*	0.06	-	-	-	-	2.06*	0.048	-	-	2.19*	0.055	-	-	-	-	-	-
17	2.22*	0.082	2.47*	0.096	-	-	2.10*	0.069	-	-	2.37*	0.085	-	-	-	-	-	-
18	2.59**	0.083	2.76**	0.092	-	-	2.89**	0.092	-	-	2.82**	0.089	-	-	-	-	-	-
19	2.00*	0.05	2.62**	0.078	-	-	2.00*	0.046	-	-	2.00*	0.049	-	-	-	-	-	-
20	2.66**	0.09	2.42*	0.077	-	-	2.76**	0.087	-	-	2.76**	0.09	-	-	-	-	-	-
21	-	-	-	-	-	-	1.72*	0.034	1.89*	0.114	-	-	-	-	-	-	-	-
22	-	-	-	-	-	-	-	-	-	-	-	-	1.95*	0.063	-	-	-	-
23	-	-	-	-	1.86*	0.033	-	-	-	-	-	-	-	-	-	-	-	-
24	2.52*	0.068	2.58*	0.07	-	-	2.66*	0.07	2.45*	n.a.†	2.84*	0.079	2.56*	0.066	-	-	-	-
27	1.77*	0.039	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
28	1.88*	0.043	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
29	-	-	1.97*	0.06	-	-	2.31*	0.078	-	-	2.32*	0.079	-	-	-	-	-	-

Table 3-2. Continued.

	Head Weight				Gut Weight				Body Waste Weight		SalmoFan Scale		Total Waste Weight			
			covariate[$]				covariate[$]								covariate[$]	
Chr	F ratio	PVE	F ratio	PVE	F ratio	PVE	F ratio	PVE	F ratio	PVE	F ratio	PVE	F ratio	PVE	F ratio	PVE
1	1.76*	0.07	-	-	-	-	-	-	-	-	-	-	1.80*	0.068	-	-
3	-	-	-	-	-	-	-	-	-	-	2.30*	0.088	-	-	3.20**	0.154
6	2.12*	0.074	-	-	-	-	-	-	-	-	-	-	2.05*	0.078	-	-
7	-	-	-	-	1.73*	0.044	1.72*	0.048	2.01*	0.063	-	-	2.49*	0.075	-	-
9	2.05*	0.07	-	-	-	-	-	-	-	-	-	-	-	-	-	-
11	1.96*	n.a.†	-	-	-	-	-	-	-	-	-	-	-	-	2.00*	0.099
13	2.72**	0.072	-	-	-	-	-	-	-	-	-	-	2.44*	0.046	1.91*	0.052
16	-	-	-	-	1.98*	0.044	-	-	-	-	-	-	-	-	-	-
17	-	-	-	-	2.16*	0.093	-	-	-	-	2.11*	0.085	-	-	-	-
18	2.11*	0.069	-	-	-	-	-	-	-	-	-	-	1.88*	0.047	-	-
19	2.49**	0.068	-	-	-	-	-	-	-	-	-	-	2.22*	0.033	-	-
20	2.02*	0.067	-	-	2.19**	0.063	-	-	-	-	-	-	1.99*	0.048	-	-
21	-	-	1.72*	n.a.†	-	-	-	-	1.85*	0.03	-	-	1.76*	0.046	-	-
22	-	-	1.90*	0.4	-	-	-	-	-	-	-	-	-	-	-	-
24	2.24*	0.068	-	-	3.03**	0.094	3.13**	0.126	-	-	-	-	-	-	-	-
25	-	-	1.97*	0.4	-	-	-	-	-	-	2.04*	0.067	-	-	-	-

*: chromosome-wide significance at $p<0.05$; **: genome-wide significance at $p<0.05$; PVE: proportion of phenotypic variance for half-sib analysis

†: Due to the MSE_{full} value being equal to $MSE_{reduced}$.

$: Harvest weight was fitted as covariate.

Table 3-3. Results of dam-based QTL mapping analysis and proportion of phenotypic variance explained for significant trait/chromosome combinations.

Chr	Trait	Dam F-ratio	PVE	Average QTL position (cM)	95% C.I. for QTL Position (cM)
20	**Gutted Weight**	2.48*	0.06	20.8	0.0 - 43.0
	Deheaded Weight	2.71*	0.07	19.4	
	Total Waste Weight	2.35*	0.06	14.0	
	Body Waste Weight	2.18*	0.06	12.4	0.0 - 40.0
13†	**Gut Weight**	2.49*	0.07	42.0	0.0 - 64.0
18†	**Gutted Weight**	2.62*	0.07	20.2	0.0 - 39.0

*:chromosome-wide significance at $p < 0.05$; †: QTLs found in the analysis fitting harvest weight as covariate.

PVE: proportion of phenotypic variance for full-sib analysis.

Table 3-4. The QTL effect on growth traits and associated absolute T values in segregating individual parents for the significant QTL at chr. 20.

Sire-based analysis	Traits	QTL effect estimate (SE)* (g)	Absolute T value
J9L2M0088	Harvest Weight	-580 (170)	3.43
	Fillet Weight	-430 (110)	4.01
	Gutted Weight	-650 (150)	4.41
	Deheaded Weight	-580 (130)	4.41
	Head Weight	-80 (20)	3.61
	Total Waste Weight	-140 (50)	2.98
J9L2M0091	Harvest Weight	-360 (160)	2.19
J9L3M3080	Total Waste Weight	170 (70)	2.29
Dam-based analysis			
J9L2F0144	Gutted Weight	570 (170)	3.36
	Deheaded Weight	480 (150)	3.28
	Total Waste Weight	140 (50)	2.53
J9L2F1295	Gutted Weight	590 (250)	2.35
	Deheaded Weight	480 (200)	2.35
	Body Waste Weight	200 (60)	3.56
	Total Waste Weight	270 (80)	3.52

J9L2F0695	Deheaded Weight	-940 (400)	2.32

* The sign + or – is arbitrary when compared across families but indicates the direction of the allelic effect within families (*e.g.* an allele decreasing harvest weight in sire J9L2M0088 also decreased fillet, gutted, deheaded, head and total waste weight.)

3.5 Discussion

In this study, the genetic basis and architecture of growth-related traits was investigated in a large commercial population of Atlantic salmon using a two-stage QTL mapping approach. All traits measured showed significant evidence for heritability and significant weight-related QTLs were observed on chr. 13, 18, 19 and 20 in the sire-based analysis. These QTL typically affected several of the weight measurements taken at harvest, which reflects the high positive correlation between these traits and suggests that their effect is related to overall size of the fish. However, the QTL on chr. 13 may have effects on the weight of components of the fish independent of an overall growth effect, as indicated by an analysis including harvest weight as a covariate.

A QTL affecting several of the growth-related traits on chr. 20 was confirmed in the dam-based analysis. This chromosome has previously been shown to harbor QTL affecting body weight of Atlantic salmon at younger age (10 months; [16]). However, a comparison of the QTL detected in the current study with those observed in previous studies (Table 3-5) shows that, even amongst populations of salmon measured at similar age, QTL tend to be rather population-specific. This may reflect differing underlying quantitative trait nucleotide affecting growth of the populations, genotype by environment interaction, or simply that a proportion of QTL identified in most studies are likely to be 'false positives'. The weight traits measured at harvest had high positive genetic and phenotypic correlations (r ~ 0.97-1.00 in phenotypic and ~ 0.99-1.00 in genetic correlation), and this is generally reflected in the QTL results, because individual QTL tended to affect the weight of several components of the fish. This is a phenomenon observed in several other studies (*e.g.* [18]) and suggests that improvement of the growth of all components of the fish in breeding programs can be made by simply measuring overall harvest weight. This will improve harvest weight, the most important trait, but is likely to also improve potentially undesirable traits such as gut weight. Achieving different rates of gain in individual components of the fish using QTL or conventional family-based selection is likely to be challenging and may require more

Table 3-5. Comparison of harvest weight QTL chromosomes in Atlantic salmon from this and previous studies.

	Gutierrez et al. [17]		Baranski et al. [18]	Houston et al. [19]	This study	
					Sire	Dam
Age	**~27 months**	**~38 months**	**~36 months**	**~30 months**	**~36 months**	
Chr						
1		C	C			
2	G	C		C/G		
3				C		
4			G			
5	C	C	G			
6				C		
7			C			
8	C			C		
9	C					
10	C		G			
11			C		C	
12						
13			C	C/G	G	
14						
15	C	C			C	
16		C	C		C	
17					C	
18			C		G	C
19	C				C	
20					G	C
21	C			C		
22			C			

23			C		
24					C
25		C	C		
26		G			
27		C			C
28					C
29	C		C		

C: chromosome-wide significance; G: genome-wide significance

detailed or accurate measures of these component traits. However, the existence of QTL affecting fillet weight seemingly independent of overall harvest weight (*e.g.* chr. 11) suggests that there are potentially some genes affecting component traits partially independently of harvest weight that could be targets for further study.

Atlantic salmon are closely related to rainbow trout and previous studies in trout have reported several QTLs affecting body mass [25, 40-41]. There was some overlap between these QTL and the genome-wide significant QTL identified in the current study, in particular for body mass QTL mapped to trout chromosomes 1q and 16q/12p [26], chr. 9p and 21p [42] and chr. 16q [40], which correspond to chr. 13 and 18 in salmon. In addition, corresponding QTL regions showing chromosome-wide significance with body weight were also discovered between Chinook salmon (chr. 25) [25] and Atlantic salmon (chr. 28) (this study). These results raise the possibility that some of the QTL affecting complex growth traits may be conserved across salmonid species. However, clearly some overlap between studies will occur by chance and the likelihood of the underlying QTL being common in both species will become more apparent with further studies and a finer mapping resolution. The confidence intervals associated with the QTL in the current study were large which precluded the meaningful identification of potential underlying candidate genes. However, known candidate genes explaining a small percentage of variation in growth in this population (*myostatin* [31] and *IGF1* [32]) do not coincide with the QTL identified here.

The size of the QTL effects in the current study was typically around 5-9 % and 6-7 % of the within-family phenotypic variance in the sire and dam-based analysis respectively. While this may be an overestimate due to the Beavis effect [41], it is certainly plausible that markers linked to these QTL may be of use in selective breeding programs. However, the confidence intervals were large and this indicates that while the two-stage mapping approach employed appears to be effective at detecting QTL, the fine mapping to smaller chromosome regions in the dam analysis may benefit from additional markers. The results of this and other studies support the hypothesis that complex traits such as weight are polygenic, which may reflect the involvement of diverse regulation pathways related to energy balance, muscle cell proliferation and skeletal growth. The fact that the proportion of variation explained by the QTL is smaller than in previous studies (*e.g.* [19]) is probably due to the large sample size of the current study (*i.e.* ~ 1700 offspring for the sire-based analysis), and hence potentially more reliable estimates of QTL effect size [41]. Further, the two-step approach provided some degree of within-study validation for the detected QTL on chr. 18. The traits of most commercial interest in salmon production, such as fillet weight were affected by the QTL on chr. 13, 18, 19, and 20 (genome-wide significance) in the sire-based analysis. Notably, except chr. 19 in sire-based analysis - for which further study may be merited - all of these QTL regions showed a significant effect on gutted weight and deheaded weight.

No QTL affecting fat content were detected in our study. Interestingly, components of fat content of salmon, such as n-3 long chain polyunsaturated acid, are highly heritable [43]**.** Therefore, perhaps more consideration could be given to the investigation of the genetic architecture of the specific components of the fat content of the fillet, as opposed to a more crude overall measure of fat levels. Naturally, this refinement of phenotype would incur a greater cost. In addition, only three QTL (chr. 3, 17 and 25) were shown to affect fillet colour at the chromosome-wide significance level. Chromosomes 3 and 26 have previously been suggested to harbor QTL associated with fillet colour traits [18]. The heritability of this trait is relatively low in this study (h^2 ~ 0.1 - 0.2) when compared

with weight related traits in Atlantic salmon [44], although recently published studies have given higher heritabilities [45] and fillet color has been suggested to show a significant association with a single locus SCAR marker [46]. It has also been suggested that fillet colour is positively correlated with overall body weight in farmed Coho salmon ($r \sim 0.4 \pm 0.5$) [46] and Atlantic salmon ($r \sim 0.49$) [47]. This may be related to the inclusion of dietary additives such as astaxanthin, canthaxanthin and carotenoid, which are included in feed to enhanced fillet pigmentation [48]. As such, protein / muscle gains may be accompanied by an associated increase in colour additives. However, we did not observe a correlation between harvest weight and fillet colour in our study. In part, this may be due to a lack of fillet colour variation observed in the population (coefficient of variation ~ 0.025). Of the putative colour QTL in the current study, only chr. 17 showed some evidence for an effect on growth-related traits, while chr. 3 and 26 were associated with fillet colour independent of the other traits measured. Given the economic importance of this trait, further study of these putative QTL and other aspects of the genetic regulation of colour are merited.

Marker-assisted selection has been applied in the salmon aquaculture industry for several traits, the foremost example being resistance to the Infectious Pancreatic Necrosis virus [34, 49-50]. However, the genetic architecture of resistance to this disease was unusually monogenic, with a single QTL explaining most of the genetic variation. For more typical complex traits such as growth or fillet component traits, the optimal use of markers in selective breeding programs has yet to be established. Clearly, the advantages of using markers in selection for aquaculture are maximal where the traits are difficult or impossible to measure on the selection candidates themselves, and some of the harvest traits fall into this category. However, due to the lack of large-effect QTL and the putative population-specificity of those QTL, it is unlikely that QTL-targeted, across population marker-assisted selection will be a highly effective tool for breeding. With the recent development of high density SNP arrays (*e.g.* [7]), genomic selection may be a more effective (albeit expensive) means of capturing variation at QTL of small effect, but is likely to be the most effective when the training and selection population

are closely related. Within family genomic selection using lower marker density may be a more cost-effective method of capturing the within-family genetic variation associated with QTLs that are relatively population-specific [51]. The large full-sibling family sizes and routine sib-testing in salmon breeding schemes makes such approaches feasible and powerful.

3.6 Chapter Conclusions

This study investigated the genetic basis of traits measured at harvest in a large commercial population of Atlantic salmon. The traits showed significant heritability and four genome-wide significant QTL were identified on chr. 13, 18, 19 and 20. The QTL on chr. 20 had relatively large effects on several weight-related traits that were consistent in the sire and dam analysis. The abundant putative QTLs provide a broad view of the genetic architecture of body weight and component traits in salmon. It is likely that weight traits in salmon are controlled by a finite number of partially population-specific loci of moderate-effect, in addition to a large polygenic component. These factors should be accounted for when considering the optimal methods of applying genomic markers in selective breeding programs.

3.7 Conclusions

In this chapter, the heritabilities of weight-related traits were approximately 0.5. Multiple QTLs were detected associating with growth-related traits. The chromosome 13, 18, and 20 were shown to harbor significant QTL in both sire and dam-based analysis, explaining 6 to 7 % of within-family variation in growth traits in the population. Unfortunately, the QTL affecting growth traits is still lack of consistency between previous and current studies in farmed and wild Atlantic salmon. In addition, there was little evidence for colocation of QTL on synetic regions of chromosomes from evolutionary closely related species such as rainbow trout. However, this chapter offers certain valuable indicative genomic regions which potentially can be confirmed by GWAS in chapter 4, and reflects the fact that the genomic selection may be an appropriate approach for selective breeding in salmon, especially for polygenic traits – of which growth is an example.

3.8 Additional Files

Supplementary Table 1: Details of the SNP markers used in this study. Link: https://goo.gl/q9IjHt

3.9 Acknowledgements

The authors acknowledge funding from the Biotechnology and Biological Sciences Research Council (BBSRC) via grant number BB/H022007/1 and an Institute Strategic Funding Grant to The Roslin Institute, and the Studying Funding from Ministry of Education, Taiwan. We also thank Oswald Matika for assistance with the data analysis.

3.10 References

1. Gjedrem T, Robinson N, Rye M: The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. *Aquaculture* 2012, 350-353:117–129.

2. Davidson WS, Koop BF, Jones SJM, Iturra P, Vidal R, Maass A, Jonassen I, Lien S, Omholt SW: Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome Biol* 2010, 11:403.

3. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C, Labadie K, Alberti A, Aury J-M, Louis A, Dehais P, Bardou P, Montfort J, Klopp C, Cabau C, Gaspin C, Thorgaard GH, Boussaha M, Quillet E, Guyomard R, Galiana D, Bobe J, Volff J-N, Genêt C, Wincker P, Jaillon O, Roest Crollius H, Guiguen Y: The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun* 2014, 5:3657.

4. Macqueen DJ, Johnston IA: A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from spe. *Proc R Soc B* 2014, 281:20132881.

5. Yáñez JM, Houston RD, Newman S: Genetics and genomics of disease resistance in salmonid species. *Front Genet* 2014, 5(November):415.

6. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA: Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 2008, 3:e3376.

7. Houston RD, Taggart JB, Cézard T, Bekaert M, Lowe NR, Downing A, Talbot R, Bishop SC, Archibald AL, Bron JE, Penman DJ, Davassi A, Brew F, Tinch AE, Gharbi

K, Hamilton A: Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). *BMC Genomics* 2014, 15:90.

8. Palti Y, Gao G, Liu S, Kent MP, Lien S, Miller MR, Rexroad CE, Moen T: The Development and Characterization of a 57K SNP Array for Rainbow Trout. *Mol Ecol Resour* 2015;15(3):662-72.

9. Lien S, Gidskehaug L, Moen T, Hayes BJ, Berg PR, Davidson WS, Omholt SW, Kent MP: A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genomics* 2011, 12:615.

10. Gonen S, Lowe NR, Cezard T, Gharbi K, Bishop SC, Houston RD: Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing. *BMC Genomics* 2014, 15:166.

11. Danzmann RG, Cairney M, Davidson WS, Ferguson MM, Gharbi K, Guyomard R, Holm L, Leder E, Okamoto N, Ozaki A, Iii CER, Sakamoto T, Taggart JB, Woram RA: A comparative analysis of the rainbow trout genome with 2 other species of fish (Arctic charr and Atlantic salmon) within the tetraploid derivative Salmonidae family (subfamily : Salmoninae). *Genome* 2005, 1051:1037–1051.

12. Weller JI: *Quantitative Trait Loci Analysis in Animals, Second Edition.* London: CABI Publishing; 2009.

13. Bishop SC, Woolliams JA: Genomics and disease resistance studies in livestock. *Livest Sci* 2014;166:190–8.

14. Mittler R, Blumwald E: Genetic engineering for modern agriculture: challenges and perspectives. *Annu Rev Plant Biol* 2010, 61:443–62.

15. Khatkar MS, Thomson PC, Tammen I, Raadsma HW: Quantitative trait loci mapping in dairy cattle: review and meta-analysis. *Genet Sel Evol* 2004, 36:163–90.

16. Yue GH: Recent advances of genome mapping and marker-assisted selection in aquaculture. *Fish Fish* 2014;15(3):376–96.

17. Gutierrez AP, Lubieniecki KP, Davidson EA, Lien S, Kent MP, Fukui S, Withler RE, Swift B, Davidson WS: Genetic mapping of quantitative trait loci (QTL) for body-weight in Atlantic salmon (Salmo salar) using a 6.5K SNP array. *Aquaculture* 2012, 358-359:61–70.

18. Baranski M, Moen T, Våge DI: Mapping of quantitative trait loci for flesh colour and growth traits in Atlantic salmon (Salmo salar). *Genet Sel Evol* 2010, 42:17.

19. Houston RD, Bishop SC, Hamilton A, Guy DR, Tinch AE, Taggart JB, Derayat A, McAndrew BJ, Haley CS: Detection of QTL affecting harvest traits in a commercial Atlantic salmon population. *Anim Genet* 2009, 40:753–5.

20. Boulding EG, Culling M, Glebe B, Berg PR, Lien S, Moen T: Conservation genomics of Atlantic salmon : SNPs associated with QTLs for adaptive traits in parr from four trans-Atlantic backcrosses. *Heredity (Edinb)* 2008;101:381–91.

21. Reid DP, Szanto A, Glebe B, Danzmann RG, Ferguson MM: QTL for body weight and condition factor in Atlantic salmon (Salmo salar): comparative analysis with rainbow trout (Oncorhynchus mykiss) and Arctic charr (Salvelinus alpinus). *Heredity (Edinb)* 2005;94:166–72.

22. McClelland EK, Naish KA: Quantitative trait locus analysis of hatch timing, weight, length and growth rate in coho salmon, Oncorhynchus kisutch. *Heredity (Edinb)* 2010, 105:562–73.

23. O'Malley KG, McClelland EK, Naish KA: Clock genes localize to quantitative trait loci for stage-specific growth in juvenile coho salmon, Oncorhynchus kisutch. *J Hered* 2010, 101:628–32.

24. Küttner E, Moghadam HK, Skúlason S, Danzmann RG, Ferguson MM: Genetic architecture of body weight, condition factor and age of sexual maturation in Icelandic Arctic charr (Salvelinus alpinus). *Mol Genet Genomics* 2011, 286:67–79.

25. Everett M V, Seeb JE: Detection and mapping of QTL for temperature tolerance and body size in Chinook salmon (Oncorhynchus tshawytscha) using genotyping by sequencing. *Evol Appl* 2014;7:480–92.

26. Wringe BF, Devlin RH, Ferguson MM, Moghadam HK, Sakhrani D, Danzmann RG: Growth-related quantitative trait loci in domestic and wild rainbow trout (Oncorhynchus mykiss). *BMC Genet* 2010, 11:63.

27. Haidle L, Janssen JE, Gharbi K, Moghadam HK, Ferguson MM, Danzmann RG: Determination of quantitative trait loci (QTL) for early maturation in rainbow trout (Oncorhynchus mykiss). *Mar Biotechnol (NY)* 2008, 10:579–92.

28. Shi J, Luo L, Eash J, Ibebunjo C, Glass DJ: The SCF-Fbxo40 complex induces IRS1 ubiquitination in skeletal muscle, limiting IGF1 signaling. *Dev Cell* 2011, 21:835–47.

29. Powell J, White I, Guy D, Brotherstone S: Genetic parameters of production traits in Atlantic salmon (Salmo salar). *Aquaculture* 2008, 274:225–231.

30. Hayes BJ, Gjuvsland A, Omholt S: Power of QTL mapping experiments in commercial Atlantic salmon populations, exploiting linkage and linkage disequilibrium and effect of limited recombination in males. *Heredity (Edinb)* 2006, 97:19–26.

31. Peñaloza C, Hamilton A, Guy DR, Bishop SC, Houston RD: A SNP in the 5' flanking region of the myostatin-1b gene is associated with harvest traits in Atlantic salmon (Salmo salar). *BMC Genet* 2013, 14:112.

32. Tsai HY, Hamilton A, Guy DR, Houston RD: Single nucleotide polymorphisms in the insulin-like growth factor 1 (IGF1) gene are associated with growth-related traits in farmed Atlantic salmon. *Anim Genet* 2014, 45:709–15.

33. Guy DR: Genetic Resistance to Infectious Pancreatic Necrosis virus in Pedigreed Atlantic salmon (Salmo salar). PhD thesis: University of Edinburgh; 2011:184.

34. Houston RD, Haley CS, Hamilton A, Guy DR, Tinch AE, Taggart JB, McAndrew BJ, Bishop SC: Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar). *Genetics* 2008, 178:1109–15.

35. Gonen S, Baranski M, Thorland I, Norris A, Grove H, Arnesen P, Bakke H, Lien S, Bishop SC, Houston RD: Mapping and validation of a major QTL affecting the resistance to pancreas disease (salmonid alphavirus) in Atlantic salmon (Salmo salar). *Heredity* 2015(Nov), 115(5):405-14.

36. Moen T, Hayes B, Baranski M, Berg PR, Kjøglum S, Koop BF, Davidson WS, Omholt SW, Lien S: A linkage map of the Atlantic salmon (Salmo salar) based on EST-derived SNP markers. *BMC Genomics* 2008, 9:223.

37. Green P, Falls K, Crooks S: *Documentation for CRI-MAP, Version 2.4.* USA: Washington School of MedicineSt Louis, MO; 1990:63110.

38. Seaton G, Hernandez J, Grunchec J, White I, Allen J, Koning D, Wei W, Berry D, Haley C, Knott S: (http://www.ngs.ac.uk). 2006.

39. Knott SA, Marklund L, Haley CS, Andersson K, Davies W, Ellegren H, Fredholm M, Hansson I, Hoyheim B, Lundström K, Moller M, Andersson L: Multiple marker

mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics* 1998, 149:1069–80.

40. O'Malley KG: Quantitative Trait Loci for Spawning Date and Body Weight in Rainbow Trout: Testing for Conserved Effects Across Ancestrally Duplicated Chromosomes. *J Hered* 2003, 94:273–284.

41. Beavis W: The power and deceit of QTL experiments: lessons from comparative QTL studies. In *Proceedings of the Forty-Ninth Annual Corn & Sorghum Industry Research Conference.* Washington, DC.: American Seed Trade Association.; 1994:250–266.

42. Martyniuk CJ, Perry GML, Mogahadam HK, Ferguson MM, Danzmann RG: The genetic architecture of correlations among growth-related traits and male age at maturation in rainbow trout. *J Fish Biol* 2003, 63:746–764.

43. Leaver MJ, Taggart JB, Villeneuve L, Bron JE, Guy DR, Bishop SC, Houston RD, Matika O, Tocher DR: Heritability and mechanisms of n-3 long chain polyunsaturated fatty acid deposition in the flesh of Atlantic salmon. *Comp Biochem Physiol Part D Genomics Proteomics* 2011, 6:62–9.

44. Norris AT, Cunningham EP: Estimates of phenotypic and genetic parameters for flesh colour traits in farmed Atlantic salmon based on multiple trait animal model. *Livest Prod Sci* 2004, 89:209–222.

45. Odegård J, Moen T, Santi N, Korsvoll S a, Kjøglum S, Meuwissen THE: Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). *Front Genet* 2014, 5(November):402.

46. Araneda C, Neira R, Iturra P: Identification of a dominant SCAR marker associated with colour traits in Coho salmon (Oncorhynchus kisutch). *Aquaculture* 2005, 247:67–73.

47. Quinton CD, McMillan I, Glebe BD: Development of an Atlantic salmon (Salmo salar) genetic improvement program: Genetic parameters of harvest body weight and carcass quality traits estimated with animal models. *Aquaculture* 2005, 247:211–217.

48. Albert G.J. Tacon: *The Nutrition and Feeding of Farmed Fish and Shrimp - a Training Manual.* Rome: FAO 1987.

49. Moen T, Baranski M, Sonesson AK, Kjøglum S: Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar): population-level associations between markers and trait. *BMC Genomics* 2009, 10:368.

50. Houston RD, Haley CS, Hamilton A, Guy DR, Mota-Velasco JC, Gheyas AA, Tinch AE, Taggart JB, Bron JE, Starkey WG, McAndrew BJ, Verner-Jeffreys DW, Paley RK, Rimmer GSE, Tew IJ, Bishop SC: The susceptibility of Atlantic salmon fry to freshwater infectious pancreatic necrosis is largely explained by a major QTL. *Heredity (Edinb)* 2010, 105:318–27.

51. Lillehammer M, Meuwissen THE, Sonesson AK: A low-marker density implementation of genomic selection in aquaculture using within-family genomic breeding values. *Genet Sel Evol* 2013, 45:39.

Chapter 4

Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array

Tsai, H.Y., Hamilton, A., Tinch, A.E., Guy, D.R., Gharbi, K.,
Stear, M.J., Oswald, M., Bishop, S.C., Houston, R.D.
DOI: 10.1186/s12864-015-2117-9

RDH and SCB designed the overall study. MJS and KG were responsible for collecting the trait data. AET, AH, and DRG provided pedigree and trait data. HYT, OM and RDH analysed the data. HYT and RDH wrote the manuscript.

This chapter has been published in BMC Genomics at the following URL:
http://www.biomedcentral.com/1471-2164/16/969

4.0 Introduction

A high density genetic map was constructed in chapter 2. A QTL mapping analysis was performed in chapter 3 to study the QTLs associating with growth traits in a 3-year-old salmon population. In this chapter, I apply the genome-wide genetic markers to estimate the significance level of individual SNP using genome-wide association analysis, and to predict the estimated genetic merits of individuals using genomic and pedigree-based prediction. Both methods aim to investigate the genetic architecture of complex growth traits in a 1-year-old salmon population. In chapter 3, chromosomes 13, 18, and 20 were shown to be associated with growth traits in adult salmon. The GWAS conducted in this chapter can help to verify the results identified in chapter 3 using a different methodology with a more comprehensive coverage of genome-wide markers.

4.1 Abstract

Background

The genetic architecture of complex traits in farmed animal populations is of interest from a scientific and practical perspective. The use of genetic markers to predict the genetic merit (breeding values) of individuals is commonplace in modern farm animal breeding schemes. Recently, high density SNP arrays have become available for Atlantic salmon, which facilitates genomic prediction and association studies using genome-wide markers and economically important traits. The aims of this study were (i) to use a high density SNP array to investigate the genetic architecture of weight and length in juvenile Atlantic salmon; (ii) to assess the utility of genomic prediction for these traits, including testing different marker densities; (iii) to identify potential candidate genes underpinning variation in early growth.

Results

A pedigreed population of farmed Atlantic salmon (n = 622) were measured for weight and length traits at one year of age, and genotyped for 111,908 segregating SNP markers using a high density SNP array. The heritability of both traits was estimated using pedigree and genomic relationship matrices, and was comparable at around 0.5 and 0.6 respectively. The results of the GWA analysis pointed to a polygenic genetic architecture, with no SNPs surpassing the genome-wide significance threshold, and one SNP associated with length at the chromosome-wide level. SNPs surpassing an arbitrary threshold of significance ($P < 0.005$, ~ top 0.5 % of markers) were aligned to an Atlantic salmon reference transcriptome, identifying 109 SNPs in transcribed regions that were annotated by alignment to human, mouse and zebrafish protein databases. Prediction of breeding values was more accurate when applying genomic (GBLUP) than pedigree (PBLUP) relationship matrices (accuracy ~ 0.7 and 0.58 respectively) and 5,000 SNPs were sufficient for obtaining this accuracy increase over PBLUP in this specific population.

Conclusions

The high density SNP array can effectively capture the additive genetic variation in complex traits. However, the traits of weight and length both appear to be very polygenic with only one SNP surpassing the chromosome-wide threshold. Genomic prediction using the array is effective, leading to an improvement in accuracy compared to pedigree methods, and this improvement can be achieved with only a small subset of the markers in this population. The results have practical relevance for genomic selection in salmon and may also provide insight into variation in the identified genes underpinning body growth and development in salmonid species.

4.2 Chapter Introduction

Atlantic salmon (*Salmo salar*), an anadromous species found primarily in the northern Atlantic Ocean, is widely known for its importance in both wild fishing and aquaculture. According to statistics from the Food and Agriculture Organization (FAO), the estimated global economic value of this species in 2010 was approximately $7.8 billion [1]. Atlantic salmon is also a model for genomic studies of salmonid species with extensive genomic resources and a recent availability of an assembled reference genome sequence [2]. Atlantic salmon breeding programs are the most advanced of all aquaculture species and routinely incorporate genomic information to construct pedigrees, and to improve selection accuracy via marker-assisted or genomic selection [3].

Genome-wide association studies (GWAS) are employed to assess the association between DNA sequence variants (typically SNPs) dispersed throughout the genome and complex traits of interest. To date, abundant GWAS have been conducted on human [4] and terrestrial livestock species [5, 6], resulting in the discovery of several genes and underlying mutations affecting traits of medical and economic importance. However, despite the contribution of GWAS to terrestrial livestock and human medical research, relatively few GWAS have been undertaken in aquaculture species to date, and have typically utilized relatively sparse SNP chips [7–9]. Recently, a high density publicly available SNP chip containing ~132 K verified SNP markers was developed [10] and gives the opportunity to apply GWAS at a resolution previously not possible in salmon.
Commercially important traits for salmon farming such as growth and disease resistance are a major focus for scientific research, with several QTL mapping studies performed for growth performance (*e.g.* [11–13]) and disease resistance (summarized in [14]). Studies of the genetic basis of growth related traits using QTL linkage mapping identified chromosome regions of interest; however, there is a lack of consistency between the location of the QTL in different populations [11, 13, 15]. Potentially, GWAS may be able to address some of the drawbacks of QTL mapping, such as the

possible omission of QTL due to inadequate marker density [16]. Additionally, since GWAS detects SNPs in population-wide linkage disequilibrium with QTL affecting the trait, the potential for applying these markers directly in selective breeding is greater. While single marker-assisted selection is of limited value for polygenic traits, genomic estimated breeding values (GEBVs) can be calculated for candidate breeding animals using marker data, even in the absence of trait and / or pedigree information [17]. Studies using simulated data have shown the accuracy of prediction of breeding values using genomic data was significantly higher than using pedigree records alone [18, 19]. Few studies of genomic prediction using real data have been performed in aquaculture species, although one recent analysis of a recently admixed farmed Atlantic salmon population suggests that a genomic prediction approach can be effective at improving the accuracy of selection compared to pedigree records alone [20].

The objectives of this study were (i) to use the high density (~ 132 K) SNP array to estimate genetic parameters for weight and length of juvenile farmed salmon and compare to those based on pedigree; (ii) to detect individual SNPs / chromosomes associated with these traits; (iii) to estimate breeding values and prediction accuracy for the two traits by applying the pedigree and the genomic relationship matrix across different marker densities; (iv) to identify putative growth candidate genes by annotating the most significant markers from transcribed regions of the genome.

4.3 Methods

4.3.1 Ethics Statement

All animals were reared in accordance with the U.K. Home Office regulations regarding the use of animals in experiments. The trial was carried out at the Marine Environmental Research Laboratory (Machrihanish, UK) and approved by the ethical review committee in the University of Stirling (Stirling, UK). Fish were purchased from Landcatch which are accredited participants in the RSPCA Freedom Foods standard, the Scottish Salmon

Producers Organization Code of Good Practice, and the EU Code-EFABAR Code of Good Practice for Farm Animal Breeding and Reproduction Organizations.

4.3.2 Animal and Phenotype Measurement

The population used in the current study was a subset of those described in Gharbi et al. [33]. Briefly, eggs from the 2007 cohort of Landcatch Natural Selection (LNS, Ormsary, UK) broodstock fish were hatched and reared in separate family tanks in freshwater. At the post-smolt stage, fish were transferred to sea water environment (Machrihanish, UK). The one-year-old post-hatch fish from 62 full sibling families were PIT-tagged and transferred to a single tank. All fish were measured for body weight (g) and body length (mm). Parents and offspring of families represented by a minimum of 6 fish in the population (712 fish from 61 full sibling families) were selected for genotyping. The PIT tags were used to assign offspring to parents and construct the pedigree.

4.3.3 SNP Array Genotyping

DNA from the 712 fish was extracted using the DNeasy-96 tissue DNA extraction kits (Qiagen, Crawley, UK) and then genotyped for the Affymetrix Axiom SNP array containing ~ 132 K validated SNPs [10] (http://www.affymetrix.com/support/technical/datasheets/axiom_salmon_genotyping_array_datasheet.pdf). Starting with these validated SNPs, filtering of SNP data was performed using the Plink software [34] to remove individuals and SNPs with excessive (> 1 %) Mendelian errors and SNPs with minor allele frequency (MAF) < 0.05 in this dataset. A total of 111,908 remaining SNPs were retained for 622 fish (534 offspring, 28 sires and 60 dams). The phenotypic sex of the offspring was unknown and, therefore, the Y-specific probes on the array were used to predict the genetic sex of the fish based on the putative sex determining gene [35], as described in Houston et al. [10].

4.3.4 Statistical Analysis

4.3.4.1 Heritability Estimation

Genetic parameters for the weight and length traits were tested fitting animal as a random effect. The estimation was performed using a REML analysis assuming the following model:

$\mathbf{y = Xb + Zu + e}$ *(1)*

where **y** is the observed trait, **b** is the fixed effect of sex, **u** is the vector of additive genetic effects, **e** is the residual error and **X** and **Z** the corresponding incidence matrices for fixed effects and additive effects, respectively. The covariance structure for the genetic effect was calculated either using pedigree (**A**) or genomic (**G**) information (*i.e.* $\mathbf{u} \sim N(0, A\sigma_a^2)$ or $N(0, G\sigma_a^2)$). Hence, the narrow sense of heritability was estimated by the additive genetic variance and total phenotypic variance, equaling to:

$h^2_a = \sigma^2_a / \sigma^2_p$ *(2)*

where σ^2_a is the additive genetic variance and σ^2_p is the total phenotypic variance which is a sum of $\sigma^2_a + \sigma^2_e$.

The analysis was implemented using the ASReml 3.0 software [36]. The genomic relationship required for the analysis was calculated using the Genabel 'R' package [37] and method of VanRaden [38], and then inverted applying the standard 'R' function.

4.3.4.2 Genome-wide Association Study

The GWAS was performed using the two-step GRAMMAR method implemented in Genabel [37]. Firstly, the trait data were corrected for the fixed effect and polygenic effects (fitting the genomic relationship matrix) using model (1) above. Secondly, the association between the individual SNPs and the residuals from model (1) was applied using the 'mmscore' method [39]. The genome-wide and chromosome-wide significance thresholds were determined by Bonferroni correction (0.05 / N), where N represents the number of QC-filtered SNPs across the entire genome (genome-wide) and on each chromosome (chromosome-wide).

Subsequently the allelic substitution effects of SNPs from the GWA analysis surpassing an arbitrary relaxed threshold (P < 0.005, ~ top 0.5 % of all markers) were estimated using ASReml 3.0 [36] fitting the mixed model (1) as previously described plus the SNP as the fixed effects.

The SNP additive effect (α) was calculated as half the difference between the predicted phenotypic means of the two homozygotes, (AA-BB)/2, and the dominance effect (δ) was AB – [(AA+BB)/2], where the AB represents the predicted phenotypic mean of the heterozygote. The proportion of genetic variance explained (PVE) by the SNP was estimated using the following equation:

$$PVE = [2pq\ (\alpha + \delta\ (q - p))^2]/V_A \qquad (3)$$

where α and δ are the additive and dominance effect respectively, the p is the frequency of the major allele and q is the frequency of the minor allele, and V_A is the total additive genetic variance of the trait obtained when no SNP effects are included in the model. For certain markers containing two genotypes, the dominance effect (δ) was not included in the equation (Appendix 1).

4.3.4.3 Genomic Prediction

Estimated breeding values were obtained using the pedigree relationship (PBLUP) or the genomic one (GBLUP). These predictions were compared in terms of their accuracy to predict an unknown phenotype. In order to do so, a five-fold cross validation analysis was performed using the individuals with genotype data and phenotypes in both traits.

The individuals were randomly divided into five non-overlapping subsets (*i.e.* each subset contains one fifth of the data corresponding to ~ 106 individuals). One subset of data was then used as a validation set and the reminder of the data [four fifths (n ~ 425)] was used as the training population. The phenotype recorded in the validation population was then masked and breeding values were estimated using ASReml 3.0 assuming model (1). Accuracy was calculated as the correlation between the predicted EBVs of the validation set and the actual phenotypes divided by the square root of the heritability

[~ $r(y_1,y_2)/h$] using all individuals. The whole procedure was independently replicated five times and average accuracy values were calculated.

4.3.4.4 Comparison of Different SNP Densities

We compared the use of different SNP marker densities of 0.5 K, 1 K, 5 K, 10 K, 20 K, 33 K, and 112 K (full dataset) for GEBV calculation. Firstly, as part of a pipeline for designing a lower density SNP genotyping platform for routine genomic evaluations, a subset of ~ 33 K SNPs were selected from the ~ 132 K array as follows: (i) only polymorphic high resolution SNPs were retained as defined using Affymetrix software, (ii) only one SNP per genome contig in the salmon genome assembly was retained (NCBI Assembly GCA_000233375.1), (iii) removed one of every pair of SNPs with $r^2 >$ 0.65 based on the Landcatch Natural Selection samples from the test plate of samples as described in Houston et al. [10], (iv) removed any remaining SNPs with a MAF < 0.1 and 'missingness' > 0.03 in the above samples and (v) added any SNPs not included above that reached a nominal significance threshold in a genome-wide association analysis for disease resistance (data not shown). From this reduced set of ~ 33 K SNPs, further subsets were taken at random to create SNP densities of 0.5 K, 1 K, 5 K, 10 K, and 20 K markers.

4.3.4.5 Putative Gene Identification

Based on the result of the GWA analysis, the SNPs surpassing the relaxed significance threshold (P < 0.005 in model (1), ~ top 0.5 % of markers) were chosen to identify those located within or proximal to genes. Firstly, the flanking sequence of all the significant markers were aligned (using blastn) with an Atlantic salmon fry transcriptome database from RNA-seq of salmon fry in a separate study in which a large proportion of the SNPs on the array were discovered (described in Houston et al. [10]). Only markers whose flanking sequences exactly matched exactly with reference transcriptome database at the SNP position were selected. These transcripts were used to align (using blastx) with

human (*Homo sapiens*), mouse (*Mus musculus*), and zebrafish (*Danio rerio*) peptide reference database respectively (downloaded from http://www.ensembl.org/index.html; May 2014), from which a stringent criterion of e-value ≃ 0 were used as evidence for homology. Secondly, for each unique peptide in each of the species, the corresponding gene id, associated gene name, chromosome position, and gene ontology (GO) were retrieved from ensembl biomart database (retrieved from http://www.ensembl.org/biomart; *Jun.* 2014) respectively. The corresponding chromosome of SNP markers were identified by aligning the marker and its flanking sequence with salmon reference genome sequence (AKGD00000000.4) and existing LG mapping [10].

4.4 Results

4.4.1 Summary Statistics and Heritability

The final dataset used for the GWAS consisted of ~ 112 K QC-filtered SNPs successfully genotyped in 622 fish (from 61 full sibling families) with weight and length measurements taken approximately 1 year post-hatching. Sex of the fish was predicted based on the Y-specific probes on the SNP array (as described in Houston et al. [10]) and the population was evenly split between males and females (1:1.03). The weight and length traits were highly correlated at the phenotypic and genetic level (r ~ 0.96 in both). The overall heritability for both traits, as estimated by the genomic relationship matrix was ~ 0.6, compared to ~ 0.5 using the pedigree relationship matrix (Table 4-1).

Table 4-1. The heritability and summary statistics of the weight and length phenotypes.

	Weight (g)	Length (mm)
Mean (std dev)	112.0 (24.0)	214.1 (16.1)
Heritability*(std err):		
G-matrix	0.60 (0.07)	0.61 (0.07)
A-matrix	0.48 (0.10)	0.51 (0.11)

*: Heritability was estimated by the genomic relationship matrix (G-matrix) and pedigree-based relationship matrix (A-matrix) respectively.

4.4.1 Genome-wide Association Analysis

To determine which individual SNPs were associated with weight and length, a GWAS was performed on all markers. No SNPs reached the genome-wide significance level (using the stringent Bonferroni correction), whereas one SNP mapping to chromosome 17 surpassed the chromosome-wide significance level for length and was estimated to explain ~ 7 % of the additive genetic variation (Table 4-2). 684 of the 111,908 SNPs surpassed an (arbitrary) relaxed threshold [nominal $P < 0.005$ from model (1)] and were used for determining putative candidate genes (Appendix 2 and Table 4-3). The p-value, allele frequency, additive and dominance effect, and proportion of additive genetic variance due to the SNP for the top ten markers for weight and length are given in Table 4-2. Full lists of the SNPs surpassing the relaxed threshold are given in Appendix 1. The proportion of genetic variation explained by the top ten markers ranged between 0.003 to 0.12. Approximately 40 K SNPs had been assigned to corresponding chromosome using sire-based linkage mapping performed by Crimap software [21] as described in Houston et al. [10] and using the reference genome sequence (AKGD00000000.4). The quantile-quantile (Q-Q) plots generated using model (1) in the GWA analysis for weight and length are given in Appendix 3.

Table 4-2. The p-value, allele frequency, additive (α) and dominance (δ) effect, and proportion of additive genetic variance explained for the top ten SNP markers associated with weight and length.

Weight							
Marker	**P-value**	**Allele Frequency**		**Additive effect (s.e.)**	**Dominance effect (s.e.)**	**PVE**	**Chromosome**
		p	q				(Unknown: n/a)
*AX87944147	2.8 e-05	0.69	0.31	4.97 (1.88)	8.76 (2.09)	0.003	n/a
*AX87934338	6.4 e-05	0.61	0.39	7.22 (2.00)	3.22 (2.08)	0.08	16
AX87992121	9.5 e-05	0.54	0.46	7.55 (1.97)	0.18 (2.11)	0.08	n/a
AX87888225	1.0 e-04	0.94	0.06	7.00 (6.28)	23.83 (6.66)	0.06	n/a
AX87943138	1.2 e-04	0.69	0.31	8.34 (2.07)	2.65 (2.29)	0.10	21
AX88223695	1.2 e-04	0.80	0.20	3.32 (2.76)	16.54 (3.02)	0.04	28
AX87959413	1.3 e-04	0.58	0.42	7.34 (1.81)	3.61 (1.96)	0.08	28
AX88127533	1.4 e-04	0.59	0.41	7.43 (1.84)	2.71 (1.98)	0.07	28
*AX87963258	1.4 e-04	0.57	0.43	5.80 (1.47)	2.00 (2.04)	0.05	17
AX88282141	1.5 e-04	0.56	0.44	6.68 (1.77)	0.56 (1.96)	0.07	21

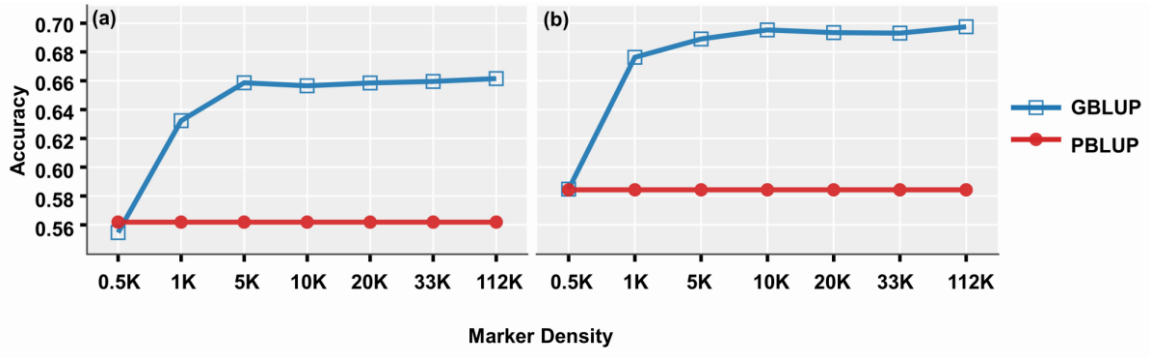
Length							
Marker	**P-value**	**Allele Frequency**		**Additive effect (s.e.)**	**Dominance effect (s.e.)**	**PVE**	**Chromosome**
		p	q				(Unknown: n/a)
***AX87963258**	**1.7 e-05**	**0.57**	**0.43**	**4.42 (0.99)**	**1.27 (1.37)**	**0.07**	17
AX88141678	5.3 e-05	0.77	0.23	6.84 (1.88)	1.74 (1.98)	0.07	5
*AX87944147	5.4 e-05	0.69	0.31	3.19 (1.27)	5.77 (1.40)	0.003	n/a
*AX87934338	7.3 e-05	0.61	0.39	4.91 (1.34)	1.71 (1.40)	0.08	16
AX87959512	9.1 e-05	0.68	0.32	5.46 (1.48)	0.21 (1.55)	0.08	20
AX88083269	1.0 e-04	0.59	0.41	4.76 (1.16)	1.99 (1.40)	0.08	n/a
AX88089073	1.6 e-04	0.70	0.30	4.77 (1.62)	1.07 (1.65)	0.05	20
AX88048182	1.6 e-04	0.78	0.22	6.65 (1.88)	1.96 (2.00)	0.12	5
AX88267406	1.6 e-04	0.78	0.22	6.65 (1.88)	1.96 (2.00)	0.12	5
AX88287764	1.7 e-04	0.85	0.15	3.33 (3.38)	12.33 (3.47)	0.04	n/a

Bold: chromosome-wide significance ($p < 0.05$).
*: SNP appears in the lists for both traits.

4.4.3 Genomic Prediction within Population

The use of the SNP markers for genomic prediction (GBLUP) of the weight and length traits was assessed and compared to the equivalent model using the pedigree to define relationships between the animals (PBLUP) using a five-fold cross validation design. The accuracy of the GBLUP model was approximately 20 % higher than PBLUP for both traits when using all markers in the model, reaching a value of approximately 0.7 within this population. Interestingly, while the prediction accuracy was improved by approximately 20 % with increased marker density from 0.5 K to 5 K SNPs, there was very little or no improvement in accuracy of prediction with increased marker density beyond this level. At the lowest marker density analyzed (0.5 K), the accuracy of GBLUP and PBLUP had the similar performance in both traits (Figure 4-1). However, it should be noted that the training and validation populations used for this analysis will contain closely related animals.

Figure 4-1. The estimated prediction accuracy of the (a) length and (b) weight traits when applying GBLUP and PBLUP across different marker densities (from 0.5 K to 112 K SNPs).



4.4.4 Putative Gene Identification

A large proportion of the SNPs on the 132 K Axiom array were derived from an RNA-Seq experiment and, therefore, are likely to be located within genes. 109 of the 684 SNPs surpassing a nominal significance threshold were matched with salmon fry transcriptome data using blastn alignment. From these 109 transcripts, twelve, seven, and fifteen corresponding unique peptides were identified from human, mouse, and zebrafish database respectively using strict alignment criteria ($E \simeq 0$). Five genes were identified in all reference species, while ten, seven, and two genes were detected specifically in the zebrafish, human, and mouse databases respectively. Details including the associated gene name, putative chromosome in Atlantic salmon, gene ontology (GO), transcript id and gene id are given in Appendix 2. Summaries of the identified genes are given in Table 4-3 while the effects associated with these genetic markers are given in Appendix 1.

The single marker that surpassed the chromosome-wide significance level for length (and also appears to have similar association with weight; Table 4-1) was annotated as Retinoic acid-induced protein 2 (*RAI2*; Table 4-3). Retinoic acid is a critical regulator of development, cellular growth, and differentiation [22] although the specific role of this RA induced gene is unknown.

Table 4-3. Summary of the putative homologous genes associated with SNPs surpassing the relaxed threshold (P < 0.005), the associated SNP name and predicted chromosome location on the salmon genome. The details of corresponding transcript id and SNP effect are given in Appendix 1 and 2.

Marker ID	Gene	Chromosome*	Reference Species
AX88089073	*POMT1*	20	Human / Mouse / Zebrafish
AX87884170	*MYH9*	03	Human / Mouse / Zebrafish
AX88052896	*GAPDH (GAPDHS)*	05	Human / Mouse / Zebrafish
AX87900517	*NOTCH3*	06	Human / Mouse / Zebrafish
AX88070408	*WDR35*	01	Human / Mouse / Zebrafish
AX88276725	*WDR35*	01	Human / Mouse / Zebrafish
AX88067081	*AGRN*	15	Human / Mouse / Zebrafish
AX87963258**	***RAI2***	**17**	**Human / Mouse**
AX87914686	*KNDC1*	01	Human / Mouse
AX87934385	*TXNRD3*	12	Human / Mouse
AX87906812	*ARHGEF7*	16 / 17	Human / Zebrafish
AX88009559	*DLG5*	01	Human / Zebrafish
AX87895800	*KLHL42*	17	Human / Zebrafish
AX87913460	*GUCY2F*	13	Human
AX87934385	*TXNRD1*	12	Zebrafish
AX88060914	*MYO18AB*	20	Zebrafish
AX87883353	*SYTL5*	21	Zebrafish
AX87913460	*GC3*	13	Zebrafish
AX88168740	*SI:CH211-181D7.1*	03	Zebrafish
AX88009559	*DLG5A*	01	Zebrafish
AX88254864	*PGBD4(5 OF 8)*	02	Zebrafish
AX88049616	*PGBD4(5 OF 8)*	02	Zebrafish

*: Corresponding chromosome was based on the Atlantic salmon reference genome (AKGD00000000.4) and the chromosome assignments given in Houston et al. [10], see methods for additional details.
**: Chromosome-wide significance.
AGRN: agrin; ***ARHGEF7***: Rho guanine nucleotide exchange factor (GEF) 7; ***GAPDH***: Glyceraldehyde-3-Phosphate Dehydrogenase; ***DLG5***: Discs, Large Homolog 5 (Drosophila); ***RAI2***: Retinoic acid-induced protein 2; ***KNDC1***: Kinase Non-Catalytic C-Lobe Domain (KIND) Containing 1; ***GUCY2F***: Guanylate Cyclase 2F, Retinal; ***POMT1***: Protein-O-Mannosyltransferase 1; ***GC3***: guanylate cyclase 2D, membrane (retina-specific); ***KLHL42***: kelch-like family member 42; ***TXNRD1***: Thioredoxin Reductase 1; ***TXNRD3***: Thioredoxin Reductase 3; ***WDR35***: WD repeat domain 35; ***MYH9***: myosin, heavy chain 9, non-muscle; ***NOTCH3***: notch 3; ***MYO18AB***: myo18ab; ***SYTL5***: synaptotagmin-like 5.

4.5 Discussion

4.5.1 Genome-wide Association Study

A high density SNP array [10] was applied to estimate genetic parameters and map SNPs associated with early growth of farmed salmon, as reflected by weight and length measurements at 1 year of age. The estimates of trait heritability when using the genomic relationship matrix was comparable but slightly higher than the equivalent model using the pedigree relationships (~ 0.6 vs ~ 0.5). While no SNPs surpassed the stringent genome-wide significance threshold, one SNP surpassed the chromosome-wide threshold for length ($p < 0.05$). Therefore, the GWAS results suggest that early growth in salmon is highly heritable but with a polygenic architecture and no evidence for major QTL. Based on previous linkage mapping and the current salmon reference genome assembly (AKGD00000000.4), the individual SNPs with the lowest P value for the growth traits were located on chr. 5, 16, 17, 20, 21 and 28. QTL associated with growth traits have been reported on the same chromosomes in some (but not all) previous studies in Atlantic salmon (*e.g.* [11–13, 15]). The proportion of variance explained (PVE) by each individual marker was relatively small (up to 12 %) for the growth traits. The data in the current study support previous studies suggesting that there is a lack of consistent, cross-population, major QTL affecting growth in Atlantic salmon. Previous studies have performed GWA analyses to identify genetic variants associated with complex traits such as flesh texture, fat content and sexual maturation by using a ~ 6 K SNP array in farmed Atlantic salmon [7, 9]. In the current study, we used a substantially higher density of SNPs (~ 112 K), which may have facilitated the detection of QTL in regions not covered by previous lower density SNP platforms.

4.5.2 Assessment of Genomic Prediction

Due to practical and financial limitations, such as the previous lack of a high density genotyping platform, relatively few studies into genomic prediction have been undertaken using real data in aquaculture species. A recent study by Odegard et al. [20] showed prediction accuracies of 0.34 and 0.36 for the traits of sea lice resistance and fillet colour respectively when using PBLUP, whereas GBLUP improved the accuracies by 32 % to 51 % for sea lice resistance, and up to 22 % for fillet colour. Previous studies applying simulated data have also indicated that GBLUP would have significantly higher accuracy compared to the equivalent model using pedigree records in the typical half / full-sibling family structure of salmon breeding programs (*e.g.* [20, 23]). Our results also show that the BLUP model applying genomic data had higher accuracy than using pedigree information for both the weight and length traits, with an improvement of approximately 20 % to values close to 0.7. This is promising for the application of genomic prediction within salmon breeding programs, where it may be most effective for traits evaluated in siblings or other close relations of the selection candidates.

It is also noteworthy that ~ 5000 high quality informative SNPs are sufficient to achieve this increase in prediction accuracy in this population. Genotyping and field data collection are costly and the relative advantage of using SNP data in selection depends on these costs versus the value of the extra improvement in the traits of interest. Therefore, while the high density SNP array is more than adequate for within-population genomic prediction, the use of a cheaper and lower density SNP platform is likely to be most cost-effective. The cost-benefit is also likely to be most favourable for traits with high economic value and that cannot be measured directly on the selection candidates (*e.g.* disease resistance or fillet quality traits). However, it is important to note that (i) this is a relatively small dataset for assessing genomic prediction and (ii) the training and validation population will contain closely related animals. As such, genomic prediction in this dataset will be based on both linkage and linkage disequilibrium information, which is likely to result in increased accuracy of prediction and reduced need for high

density markers compared to scenarios where the training and validation populations are distantly related to each other. It is plausible that with more distant relationships between the populations, a higher marker density and larger sample size would be required to achieve improvements in selection accuracy over traditional BLUP. Further, the high levels of linkage disequilibrium will result in greater power to detect QTL via GWAS, but lower resolution to estimate QTL position. Simulation studies are generally consistent with the results based on real data presented in the current study: Vela-Avitua et al. [23] reported that the prediction accuracy using sparse genomic data was equivalent or lower than using the classical pedigree model with sparse markers (10 – 20 SNPs / M) across traits with different heritabilities ($h^2 \sim 0.1$, 0.3 and 0.8), while Hickey et al. [24] demonstrated that increasing marker density above ~ 10 K results in little or no improvement in prediction accuracy in maize populations, while the results of Gorjanc et al. [25] also show only minor increases in accuracy above this level in simulated livestock datasets. Finally, Odegard et al. [20] detected little increase in accuracy with increases in marker density above 22 K for fillet colour or lice resistance in a commercial salmon population.

4.5.3 Putative Gene Identification

The single SNP exceeding the chromosome-wide significance level for length was mapped to chr.17, and its predicted location is within the retinoic acid-induced protein 2 gene (*RAI2*). Although the function of *RAI2* is not yet clear, this gene is suggested to be involved in the control of cellular growth and embryo development [26]. Retinoic acid is well established as a key regulator of growth and differentiation in early life [22], and is involved in the regulation of bone formation and mineralization in salmon [27]. Therefore, *RAI2* can be considered both a positional and a biological candidate for an effect on regulation of growth in juvenile salmon. Genes associated with the other markers discovered surpassing the arbitrary relaxed significance threshold ($P < 0.005$) were also identified by aligning with human, mouse, and zebrafish databases (Table 4-3). Amongst these was a SNP in *POMT1* (Protein-O-Mannosyltransferase 1) which

produces the *POMT* enzyme complex, dysregulation of which can contribute to the formation of abnormal basement membranes, which can lead to muscular dystrophy [28]. Interestingly, the *AGRN* (agrin) gene also appears to have a key regulatory role in basement membranes of neuromuscular junctions, and is involved in the inhibition, storage, activation of varied growth factors [29], clustering of voltage-gated sodium channels, and G-protein coupled acetylcholine receptor signaling pathway [30], all of which are essential for fundamental cell development. In addition, *NOTCH3* (notch 3) and the *NOTCH3* receptor have critical roles in the development and maintenance of vascular smooth muscle cells [31, 32]. Finally, genes associated with ATP binding and motor activity, such as *MYH9* (myosin, heavy chain 9) and *MYO18AB* were also identified amongst the nominally significant markers. While a proportion of the nominally significant SNPs (and therefore the genes identified) will clearly be false positives, highlighting these genes provides the opportunity to cross-reference with future studies to identify with higher confidence the putative candidates underlying growth in salmon.

4.6 Chapter Conclusions

The results of the current study show that early growth traits are highly heritable in farmed Atlantic salmon, and that the heritability can be estimated by using either the genomic or the pedigree relationship matrix. The GWA analysis showed that there are likely to be small effect QTL on several chromosomes, but there was no evidence for major QTL and these traits appear to be highly polygenic in nature. A SNP in the retinoic acid-induced protein 2 gene on chromosome 17 reached chromosome-wide significance, and is a plausible positional and functional candidate gene. Other genes identified from nominally significant SNPs will be useful for cross-referencing with similar studies in salmon and may form candidates for follow up studies to assess their function in regulation of growth in salmon. For breeding value prediction using genomic and pedigree data, GBLUP had better accuracy than PBLUP in general with accuracy of ~ 0.7 attained for early growth traits using GBLUP in this population. As few as 5 K SNPs gives close to maximal accuracy within population, suggesting that only moderate marker density is likely to be suitable for GS breeding programs for similar highly heritable but polygenic traits where the discovery populations have close relationships with the selection candidates. However, it is important to note that increased marker density is likely to be advantageous, alongside larger sample size, when attempting to predict genomic breeding values in more distantly related animals.

4.7 Conclusions

The heritabilities of body weight and length in juvenile fish were around 0.5 to 0.6 estimated by genomic and pedigree-based method. GWAS suggested both growth traits are controlled by multiple loci, with no SNP surpassing the stringent genome-wide significance threshold. A SNP reached chromosome-wide significance was located on chromosome 17 and its putative gene is *RAI2*. In general, the QTLs associated with growth traits in this study were not notably consistent with the linkage-based study in chapter 3. This may reflect a combination of (i) different lifecycle stages, (ii) the polygenic nature of the traits, and (iii) the different genetic backgrounds. However, genomic prediction applying genomic BLUP showed that approximately 5 K SNPs was able to obtain the highest prediction accuracy in both traits in within population. Those accuracies were superior to the pedigree-based method, highlighting that genomic selection is likely able to make a significant contribution to salmon breeding. The significant markers identified in this chapter will be used for verification of SNP-trait association in chapter 5.

4.8 Additional Files

Appendix 1. Summary of significant markers, their p-values, allele frequencies, additive and dominance effects, and proportion of genetic variance due to the SNPs, for weight and length respectively. Link: https://goo.gl/OBsCQn

Appendix 2. List of identified putative gene name, chromosome position, gene ontology (GO), transcript id and gene id of three reference species databases. Link: https://goo.gl/SqkOsT

Appendix 3. The Q-Q plots of weight and length generated in GWA analysis. Link: https://goo.gl/WFPzfg

4.9 Acknowledgements

The authors would like to thank Edinburgh Genomics for SNP array genotyping, Bill Roy and Matt Tinsley at the Marine Environmental Research Laboratory and James Bron from the University of Stirling for contributions to trait and sample collection, and Ricardo Pong-Wong for the advice with the methods in statistical analysis. Funding for the study was received from the Technology Strategy Board grant (TP 5771-40299), by the UK Biotechnology and Biological Sciences Research Council (BBSRC) grants (BB/H022007/1, BB/F002750/1, BB/F001959/1), by a BBSRC Institute Strategic Funding Grant to The Roslin Institute.

4.10 References

1. F.A.O: *FAO Yearbook. Fishery and Aquaculture Statistics.* F.A.O. Rome; 2010.

2. Davidson WS, Koop BF, Jones SJM, Iturra P, Vidal R, Maass A, Jonassen I, Lien S, Omholt SW: Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome Biol* 2010, 11:403.

3. Yáñez JM, Newman S, Houston RD: Genomics in aquaculture to better understand species biology and accelerate genetic progress. *Front Genet* 2015, 6(April):128.

4. Hardy J, Singleton A: Genomewide association studies and human disease. *N Engl J Med* 2009, 360:1759–68.

5. Riggio V, Matika O, Pong-Wong R, Stear MJ, Bishop SC: Genome-wide association and regional heritability mapping to identify loci underlying variation in nematode resistance and body weight in Scottish Blackface lambs. *Heredity (Edinb)* 2013, 110:420–9.

6. Bermingham ML, Bishop SC, Woolliams JA, Pong-Wong R, Allen AR, McBride SH, Ryder JJ, Wright DM, Skuce RA, McDowell SW, Glass EJ: Genome-wide association study identifies novel loci associated with resistance to bovine tuberculosis. *Heredity (Edinb)* 2014, 112:543–51.

7. Sodeland M, Gaarder M, Moen T, Thomassen M, Kjøglum S, Kent M, Lien S: Genome-wide association testing reveals quantitative trait loci for fillet texture and fat content in Atlantic salmon. *Aquaculture* 2013, 408-409:169–174.

8. Yue GH: Recent advances of genome mapping and marker-assisted selection in aquaculture. *Fish Fish* 2014, 15:376–96.

9. Gutierrez AP, Yáñez JM, Fukui S, Swift B, Davidson WS: Genome-Wide Association Study (GWAS) for Growth Rate and Age at Sexual Maturation in Atlantic Salmon (Salmo salar). *PLoS One* 2015, 10:e0119730.

10. Houston RD, Taggart JB, Cézard T, Bekaert M, Lowe NR, Downing A, Talbot R, Bishop SC, Archibald AL, Bron JE, Penman DJ, Davassi A, Brew F, Tinch AE, Gharbi K, Hamilton A: Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). *BMC Genomics* 2014, 15:90.

11. Gutierrez AP, Lubieniecki KP, Davidson EA, Lien S, Kent MP, Fukui S, Withler RE, Swift B, Davidson WS: Genetic mapping of quantitative trait loci (QTL) for body-weight in Atlantic salmon (Salmo salar) using a 6.5K SNP array. *Aquaculture* 2012, 358-359:61–70.

12. Baranski M, Moen T, Våge DI: Mapping of quantitative trait loci for flesh colour and growth traits in Atlantic salmon (Salmo salar). *Genet Sel Evol* 2010, 42:17.

13. Houston RD, Bishop SC, Hamilton A, Guy DR, Tinch AE, Taggart JB, Derayat A, McAndrew BJ, Haley CS: Detection of QTL affecting harvest traits in a commercial Atlantic salmon population. *Anim Genet* 2009, 40:753–5.

14. Yáñez JM, Houston RD, Newman S: Genetics and genomics of disease resistance in salmonid species. *Front Genet* 2014, 5(November):415.

15. Tsai HY, Hamilton A, Guy DR, Tinch AE, Bishop SC, Houston RD: The genetic architecture of growth and fillet traits in farmed Atlantic salmon (Salmo salar). *BMC Genet* 2015, 16:51.

16. Mackay TFC, Stone EA, Ayroles JF: The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet* 2009, 10:565–77.

17. Meuwissen THE, Hayes BJ, Goddard ME: Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 2001, 157:1819–1829.

18. Jonas E, de Koning D-J: Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. *Front Genet* 2015, 6(February):49.

19. Sonesson AK, Meuwissen TH: Testing strategies for genomic selection in aquaculture breeding programs. *Genet Sel Evol* 2009, 41:37.

20. Odegård J, Moen T, Santi N, Korsvoll SA, Kjøglum S, Meuwissen THE: Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). *Front Genet* 2014, 5(November):402.

21. Green P, Falls K, Crooks S: *Documentation for CRI-MAP, Version 2.4.* USA: Washington School of MedicineSt Louis, MO; 1990:63110.

22. Niederreither K, Dollé P: Retinoic acid in development: towards an integrated view. *Nat Rev Genet* 2008, 9:541–553.

23. Vela-Avitúa S, Meuwissen TH, Luan T, Ø degård J: Accuracy of genomic selection for a sib-evaluated trait using identity-by-state and identity-by-descent relationships. *Genet Sel Evol* 2015, 47:9.

24. Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM, Grondona M, Zambelli A, Windhausen VS, Mathews K, Gorjanc G: Evaluation of Genomic

Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation. *Crop Sci* 2014, 54:1476.

25. Gorjanc G, Cleveland MA, Houston RD, Hickey JM: Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genet Sel Evol* 2015, 47:12.

26. Walpole SM, Hiriyana KT, Nicolaou A, Bingham EL, Durham J, Vaudin M, Ross MT, Yates JR, Sieving PA, Trump D: Identification and characterization of the human homologue (RAI2) of a mouse retinoic acid-induced gene in Xp22. *Genomics* 1999, 55:275–283.

27. Ø rnsrud R, Lock EJ, Glover CN, Flik G: Retinoic acid cross-talk with calcitriol activity in Atlantic salmon (Salmo salar). *J Endocrinol* 2009, 202:473–482.

28. Manya H, Chiba A, Yoshida A, Wang X, Chiba Y, Jigami Y, Margolis RU, Endo T: Demonstration of mammalian protein O-mannosyltransferase activity: coexpression of POMT1 and POMT2 required for enzymatic activity. *Proc Natl Acad Sci U S A* 2004, 101:500–5.

29. Groffen AJ, Buskens CA, van Kuppevelt TH, Veerkamp JH, Monnens LA, van den Heuvel LP: Primary structure and high expression of human agrin in basement membranes of adult lung and kidney. *Eur J Biochem* 1998, 254:123–8.

30. Groffen A. J, Ruegg MA, Dijkman H, van de Velden TJ, Buskens CA, van den Born J, Assmann KJ, Monnens LA, Veerkamp JH, van den Heuvel LP: Agrin Is a Major Heparan Sulfate Proteoglycan in the Human Glomerular Basement Membrane. *J Histochem Cytochem* 1998, 46:19–27.

31. Wang T, Baron M, Trump D: An overview of Notch3 function in vascular smooth muscle cells. *Prog Biophys Mol Biol* 2008, 96:499–509.

32. Wang Q, Zhao N, Kennard S, Lilly B: Notch2 and notch3 function together to regulate vascular smooth muscle development. *PLoS One* 2012, 7:1–8.

33. Gharbi K, Matthews L, Bron J, Roberts R, Tinch A, Stear M: The control of sea lice in Atlantic salmon by selective breeding. *J. R. Soc. Interface* 2015, 12:0574.

34. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M AR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007, 81:559–75.

35. Yano A, Guyomard R, Nicol B, Jouanno E, Quillet E, Klopp C, Cabau C, Bouchez O, Fostier A, Guiguen Y: An immune-related gene evolved into the master sex-determining gene in rainbow trout, Oncorhynchus mykiss. *Curr Biol* 2012, 22:1423–8.

36. Gilmour, A.R., Gogel, B. J., Cullis, B.R., Thompson R: ASReml User Guide. 4th edition. Hemel Hempstead, UK: VSN International Ltd; 2014(January).

37. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 2007, 23:1294–6.

38. VanRaden PM: Efficient methods to compute genomic predictions. *J Dairy Sci* 2008, 91:4414–23.

39. Chen W-M, Abecasis GR: Family-based association tests for genomewide association scans. *Am J Hum Genet* 2007, 81:913–2.

Chapter 5

Verification of SNPs Associated with Growth Traits in Two Populations of Farmed Atlantic Salmon

Tsai, H.Y., Hamilton, A., Guy, D.R., Tinch, A.E., Bishop, S.C., Houston, R.D.

DOI:10.3390/ijms17010005

The experiments were designed by RDH and SCB. Tissue samples and phenotype measurements were provided by AET, DRG and AH. Data analyses were performed by HYT. The manuscript was written by HYT and RDH.

This chapter has been published in Int. J. Mol. Sci. at the following URL: http://www.mdpi.com/1422-0067/17/1/5

5.0 Introduction

In chapter 3 and 4, I discovered several QTLs and SNPs significantly associated with growth and weight-related traits in adult and juvenile salmon population. Due to relatively low sample size in the association analysis, genetic architecture and population differences (*e.g.* linkage disequilibrium, environments, ages, and sex), the previous results may contain the mix of true associations and false positives. In this chapter, I select sixteen SNPs based on the results of the front chapters to verify the genetic effect of identified SNPs with a separate population. The sixteen SNPs will be selected from top 0.5 % significant marker identified by GWAS in chapter 4, and QTLs targeted in chapter 3, which are listed in the method section in this chapter. Identifying the association between SNPs and phenotypes of interest across different populations can offer more robust evidence in the association analysis. When the association is confirmed, the putative gene within the vicinity of the significant SNPs will be identified, which can help to study the potential function of significant genes, with implications for both wild and farmed Atlantic salmon.

5.1 Abstract

Understanding the relationship between genetic variants and traits of economic importance in aquaculture species is pertinent to selective breeding programmes. High-throughput sequencing technologies have enabled the discovery of large numbers of SNPs in Atlantic salmon, and high density SNP arrays now exist. A previous genome-wide association study (GWAS) using a high density SNP array (132 K SNPs) has revealed the polygenic nature of early growth traits in salmon, but has also identified candidate SNPs showing suggestive associations with these traits. The aim of this study was to test the association of the candidate growth-associated SNPs in a separate population of farmed Atlantic salmon to verify their effects. Identifying SNP-trait associations in two populations provides evidence that the associations are true and robust. Using a large cohort (N = 1152), we successfully genotyped eight candidate SNPs from the previous GWAS, two of which were significantly associated with several growth and fillet traits measured at harvest. The genes proximal to these SNPs were identified by alignment to the salmon reference genome and are discussed in the context of their potential role in underpinning genetic variation in salmon growth.

5.2 Chapter Introduction

The Food and Agriculture Organization (FAO) reported that the worldwide production of farmed finfish was approximately 66.6 million tonnes from 2011 to 2012, an increase of 26 % compared with 2008 to 2009 [1]. The demands for high quality animal proteins are continuously expanding due to global economic development and human population increase. Aquaculture has a major role in fulfilling the increased requirement of protein consumption, and the continuous improvement of farming scale, sustainability and efficiency is required. Selective breeding for key production traits (such as feed efficiency and disease resistance) in finfish and shellfish species is an essential component of this improvement. However, aquaculture breeding schemes are generally fewer and less developed than terrestrial livestock and plants [2,3]. Gjedrem et al. [4] indicated that less than 10 % of aquaculture production was based on genetically-improved stock. Notably, the annual genetic gain in selective breeding programmes of aquaculture species is typically higher than that of farmed terrestrial species [4], highlighting that genetic improvement of the key economic traits can be readily achieved.

The development of high throughput sequencing technologies has expedited the discovery of millions of genome-wide SNPs, particularly for salmonid species, which have high economic values; *e.g.*, Atlantic salmon [5,6], rainbow trout [7,8] and sockeye salmon [9]. To date, for Atlantic salmon, traits, such as fillet colour, sexual maturation and fat percentage, have been initially studied using genome-wide association (GWA) analyses using an SNP array with approximately 6 K markers [10,11]. Additionally, GWAS for host resistance to sea lice [12], host resistance to *Piscirickettsia salmonis* [13] and early growth traits [14] have been performed using higher density SNP chips (50 or 132 K SNPs). Around 70 to 100 million years ago, the ancestor of modern salmonids underwent a whole genome duplication (WGD) event [15,16], which was followed by extensive modifications of both the genome and transcriptome and is still under the process of returning to diploidy [17,18]. The relics of the duplicated genomes

generated by WGD complicate the discovery and interpretation of genomic variation, partly due to the difficulty in distinguishing true segregating polymorphism from paralogous variation [17]. Nonetheless, the vast majority of SNPs discovered to date in salmonid species segregate in a diploid manner [19].

The heritability of growth traits, such as body weight and length, in Atlantic salmon is moderate to high (*e.g.* [10,20,21]); but these complex traits are usually considered highly polygenic, and the underlying physiological basis for growth is likely to involve networks of many interacting genes. Typically, functional networks regulating growth-related traits involve hundreds of candidate genes [22,23]. Detecting and investigating the function of each individual gene within such complex networks is practically unfeasible. However, clues to the possible roles of particular candidate genes can be determined by associating genomic variation within or close to the gene with phenotypic variation in the trait of interest on a population scale. Herein lies the potential of GWAS to inform the underlying biology of the trait in question, in addition to providing potential markers for selective breeding programmes. Several previous studies of the association between candidate gene polymorphisms and phenotypic variation in salmon populations have focused on well-known candidates with previously-demonstrated physiological roles in the trait of interest (*e.g.* [19,24,25]). With the advent of high density and high throughput genotyping assays, GWAS and subsequent alignment to a reference genome [26] can identify positional candidate genes in a more systematic manner. However, with all association studies, it is important to assess the robustness of any putative significant result by testing the association between the SNP and the trait in a separate population / study. Therefore, the aims of this study were (i) to test the association of a subset of the most significant SNPs associated with weight and length of juvenile salmon [14] in another large cohort of fish and (ii) to identify and discuss putative candidate genes proximal to the SNPs that may directly contribute to variation in the growth phenotypes.

5.3 Methods

5.3.1 Animals

The GWAS used to identify the SNPs with putative association with growth in commercial salmon populations was based on the 2007 year group population of the Landcatch Natural Selection (LNS; Ormsary, UK) broodstock that were measured for weight and length at the end of the freshwater period (~ 1 year old; "Population 1") [14]. To test the candidate SNPs in a new population, 1152 individuals were randomly selected from a larger population (n ~ 5000) comprising the 1999 year group of LNS broodstock that were measured for weight and other fillet traits at harvest ("Population 2"). The 1152 genotyped fish were across 191 full sibling families from 131 sires and 185 dams. The phenotypes were measured by LNS at harvest (approximately 3 years old), including overall harvest weight (kg), gutted weight (kg), deheaded weight (kg), fillet weight (kg), head weight (kg), gut weight (kg), body waste weight (kg) and total waste weight (kg), fat percentage (% as estimated using a Torry Fatmeter (Distell Ltd., Aberdeen, Scotland)) and fillet colour (assessed visually using the Roche SalmoFan scale (Hoffmann-La Roche, West Sussex, UK), ranging from 20 (Yellow) to 34 (Red)). The body waste weight was calculated as deheaded weight minus fillet weight (weight of vertebrae and caudal fin), and total waste weight was by head weight plus body waste weight. Details of the population and phenotype measurement are given in Tsai et al. and Peñaloza et al. [20,25]. An adipose fin tissue sample of each individual was clipped and retained for DNA extraction using DNeasy-96 tissue DNA extraction kits (Qiagen, Crawley, UK).

All animals were reared in accordance with all relevant national and EU legislation concerning health and welfare. Landcatch is an accredited participant in the RSPCA (Royal Society for the Prevention of Cruelty to Animals) Freedom Foods standard, the Scottish Salmon Producers Organization Code of Good Practice and the EU Code-EFABAR (http://www.responsiblebreeding.eu/) Code of Good Practice for Farm Animal Breeding and Reproduction Organizations.

5.3.2 SNP Selection and Genotyping

The candidate SNPs were selected based on two relevant studies [14,20]. Firstly, a GWA analysis was performed in Population 1 to select the candidate markers for genotyping [14], and a proportion of the SNPs surpassing a nominal significance ($p \sim 10^{-3}$) were selected. Secondly, chromosome 20 was identified as containing loci affecting growth and fillet-related traits in Population 2 [20]. Therefore, two SNPs with nominally significant association with weight and length ($p \sim 10^{-2}$) [14] from this QTL region were also included in the shortlist for further investigation. The details of candidate SNPs are given in Table S1. In total, sixteen candidate SNPs were selected for assay design and genotyping in Population 2, of which eight were successfully genotyped and showed segregation. Candidate SNP markers and their flanking sequences were provided to LGC Genomics (Herts, UK) for the design of "kompetitive allele-specific PCR (KASP)" assays (see KASP technique details at [38]) for genotyping with 1152 offspring in Population 2.

5.3.3 Statistical Analysis

5.3.3.1 Heritability Estimation and SNP Associations

The heritability of the traits was calculated as described previously [20]. The simple animal model (Model (1)) was used to estimate the additive genetic effect of each SNP genotype (G):

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{G} + \boldsymbol{A} + \boldsymbol{e} \quad (1)$$

where $\boldsymbol{Y}$ represents the observed phenotype, $\boldsymbol{\mu}$ is the overall mean of the trait, $\boldsymbol{G}$ is the fixed effect of the SNP genotype, $\boldsymbol{A}$ is the additive genetic effect and $\boldsymbol{e}$ is the residual error.

For estimating heritability, the equivalent model was used, but without the SNP effect (***G***) using the model:

$$h^2_a = \sigma^2_a / \sigma^2_p \quad (2)$$

where σ^2_a is the additive genetic variance and σ^2_p is the total phenotypic variance. The analysis was performed by ASReml 3.0 [39].

5.3.3.2 Allelic Substitution Assessment

The allelic substitution effects of informative SNPs were estimated using Model (1) performed by ASReml 3.0 [39]. The SNP genotype was fitted as the fixed effect in the analysis. The additive effect of the candidate marker was calculated as the difference of the predicted phenotypic means of two homozygotes divided by two, which was given as (AA − BB)/2, and the dominance effect was AB − ((AA + BB)/2), where the AB represents the predicted phenotypic means from heterozygote and AA or BB are from homozygote in the statistical analysis. The proportion of genetic variance due to SNP (PVE) was also estimated, by the following equation [40]:

$$\text{PVE} = [2pq\,(\alpha + \delta(q - p))^2]/V_A \quad (3)$$

where α and δ are the additive and dominance effect, respectively, p is the frequency of the most frequent allele, q is the frequency of the minor allele and V_A is the total additive genetic variance of the trait obtained when no SNP effects are included in the model.

5.3.3.3 Candidate Gene Identification

To identify candidate genes near the significant SNPs, the flanking sequence was aligned to the Atlantic salmon reference genome assembly (GCA_000233375.4), and the corresponding genome contig and position of the SNPs were noted. Approximately 20 kb of sequence surrounding the SNPs were repeat masked (retrieved from [41]), and a BlastX analysis was used to detect putative genes within the vicinity of the SNPs.

5.4 Results

5.4.1 Heritability Estimation

The population used in the analysis was a random subset of a larger population (Population 2) measured for overall and component weight traits, colour and fat content. Heritabilities of fillet-related traits were moderate to high (0.52 to 0.53), whereas the waste weights (*e.g.* head weight) were approximately 0.3. The heritability of fat percentage and fillet colour was slightly lower (0.14 to 0.18). The phenotypic and genetic correlations were high for all of the weight-related traits ($r \sim 0.96$ to 0.99), but with little correlation between weight traits and fillet colour ($r = -0.08$). A summary of the heritability estimation and general statistics are given in Table 5-1, and they were consistent with estimates made on the larger population analysed previously [19].

5.4.2 Association between SNPs and Traits of Interest

Based on the results of the 2007 year group (Population 1) GWA analysis, 16 nominally significant SNPs were selected for genotyping in the 1999 year group (Population 2). These SNPs were chosen from QTL regions on chromosomes 16, 21 and 28 for weight and chromosomes 5, 16, 17 and 20 for length (Figures 5-1 and 5-2). Assays failed for six SNPs, and two more were monomorphic (details of selected markers were tabulated in Table S1). Of the remaining eight successfully genotyped SNPs, two were significantly associated with several growth traits (Table 5-2).

Table 5-1. The summary statistics and heritability estimates for the harvest traits.

Traits	Mean (SD)	Heritability (SE)
Harvest weight (kg)	2.65 (0.72)	0.52 (0.05)
Head weight (kg)	0.30 (0.12)	0.21 (0.03)
Body waste weight (kg)	0.34 (0.15)	0.15 (0.02)
Total waste weight (kg)	0.67 (0.21)	0.32 (0.04)
Gutted weight (kg)	2.42 (0.65)	0.53 (0.05)
Deheaded weight (kg)	2.11 (0.57)	0.52 (0.05)
Fillet weight (kg)	1.76 (0.48)	0.53 (0.05)
Fat percentage (%)	13.2 (5.98)	0.18 (0.03)
Fillet colour (20–34)	28.9 (0.74)	0.14 (0.03)
Gut weight (kg)	0.22 (0.08)	0.30 (0.04)

The SNP AX88270804 was significantly associated ($p < 0.05$) with most of the fillet and waste traits, including a suggestive association with fat content ($p < 0.1$). The adenine allele corresponds to higher trait means for the carcass weight and fatness traits. The SNP AX88141678 was associated with overall harvest weight, head weight and gutted weight ($p < 0.05$). At this SNP, the adenine allele was also associated with higher trait means for the carcass and overall weight traits. The estimation of the additive genetic variation explained by the SNPs indicated that AX88270804 explained a small percentage of the overall variation in fillet traits (~ 1 %), waste traits (2 % to 3 %) and fat percentage (4 %). The SNP AX88141678 explained approximately 1 % of the additive genetic variation in the weight-related traits (Table 5-2). To account for variation in the overall size of the fish when analysing component traits, Model (1) was preformed, including harvest weight as a covariate. In this analysis, most of the SNP-trait associations were no longer significant, but SNP AX88270804 showed an association with body waste weight and total waste weight.

Figure 5-1. The Manhattan plot of body weight in the GWAS of Population 1 [14]. The Bonferroni genome-wide significance threshold is $p \sim 4.50 \times 10^{-7}$.

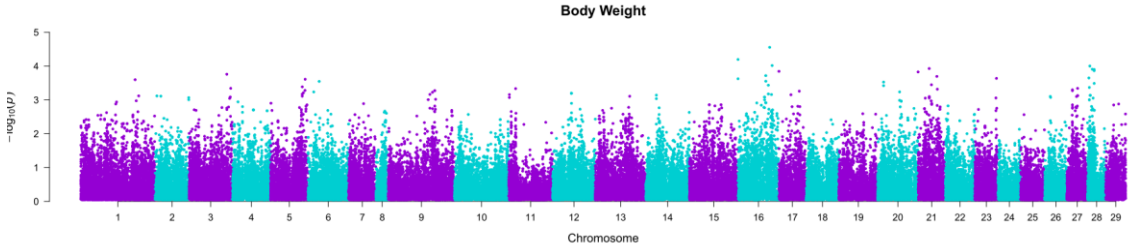
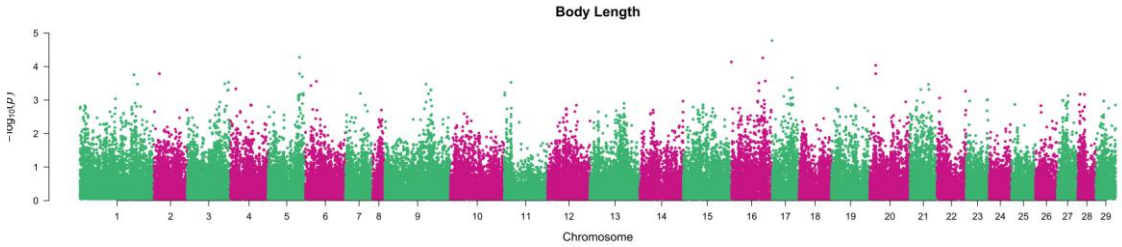


Figure 5-2. The Manhattan plot of body length in the GWAS of Population 1 [14]. The Bonferroni genome-wide significance threshold is $p \sim 4.50 \times 10^{-7}$.



5.4.3 QTL Regions Characterization and Putative Gene Identification

The corresponding flanking sequences for the two significant SNPs were aligned with the reference genome (assembly GCA_000233375.4), and the putative genes proximal to the SNPs were identified, indicating that the loci AX88141678 (chr. 5) and AX88270804 (chr. 16) were located within *MEP1A* (meprin A subunit beta-like) and *PCNT* (pericentrin), respectively. AX88270804 was located in an exon (non-synonymous), whereas AX88141678 was in a non-coding region (Table 5-2). The details of all SNPs tested in the current study are given in Table S1. The main results of the GWA analysis in Population 1 are given in Tsai et al. [14]. However, due to the recent availability of a chromosome-anchored reference genome sequence assembly for

Atlantic salmon (GCA_000233375.4), we used BlastN to align the flanking sequence of the SNPs on the array with the assembly to identify their putative chromosome and position. This information was used to draw Manhattan plots to view the QTL regions from which the candidate SNPs were chosen (Figures 5-1 and 5-2).

Table 5-2. Results of the association analysis including the predicted mean value (and standard error) and proportion of additive genetic variance due to SNP (PVE) for each trait and genotype class.

Traits	AX88141678 (Gene: *MEP1A*)				AX88270804 (Gene: *PCNT*)			
	A/A	A/G	G/G	PVE (%)	A/A	A/G	G/G	PVE (%)
# of fish	*n* = 651	*n* = 436	*n* = 52		*n* = 281	*n* = 581	*n* = 265	
Harvest weight	2.59 (0.04)	2.63 (0.05)	2.33 (0.1)**	0.3	2.66 (0.05)	2.60 (0.04)	2.50 (0.06)*	1
Head weight	0.30 (0.01)	0.30 (0.01)	0.26 (0.02)**	1.3	0.31 (0.01)	0.3 (0.01)	0.28 (0.01)**	1
Body waste weight	0.33 (0.01)	0.35 (0.01)	0.32 (0.02)	0.1	0.34 (0.01)	0.35 (0.01)	0.31 (0.01)**	3
Total waste weight	0.65 (0.01)	0.67 (0.01)	0.61 (0.03)	0	0.66 (0.02)	0.67 (0.01)	0.61 (0.02)**	2
Gutted weight	2.37 (0.04)	2.39 (0.04)	2.19 (0.09)*	0.2	2.41 (0.05)	2.38 (0.04)	2.28 (0.05) **	1
Deheaded weight	2.07 (0.03)	2.10 (0.03)	1.94 (0.08)	0.05	2.10 (0.04)	2.09 (0.03)	1.99 (0.04) *	1
Fillet weight	1.71 (0.03)	1.76 (0.03)	1.59 (0.07)**	0	1.76 (0.04)	1.72 (0.03)	1.67 (0.04)	1
Fat percentage	13.19 (0.27)	13.12 (0.31)	12.41 (0.84)	0.4	13.65 (0.39)	13.17 (0.28)	12.45 (0.4) *	4
Fillet colour	28.96 (0.04)	28.90 (0.05)	29.03 (0.12)	0.1	28.98 (0.06)	28.90 (0.04)	28.97 (0.06)	0.02
Gut weight	0.21 (0)	0.22 (0)	0.20 (0.01)	0.01	0.22 (0.01)	0.22 (0)	0.20 (0.01) **	3

* Overall SNP $p < 0.1$; ** overall SNP $p < 0.05$.

5.5 Discussions

Abundant SNPs discovered by modern sequencing technologies and bioinformatics tools have allowed us to better understand the association between genomic variation and production traits in aquatic species [27]. In a recent study, we applied a high density SNP array (~ 132 K) [6] to identify candidate markers associated with weight and length traits in a farmed salmon population measured at one year of age [14]. To test a subset of promising SNPs from the previous study in a different population, we successfully genotyped eight SNPs in a population of 1152 salmon with growth and harvest-related traits measured at three years of age. Two SNPs were found to be significantly

associated with several growth and harvest traits in the second population, implying that these SNPs are linked to QTL with effects on growth at multiple stages of the salmon production cycle. For the remaining six SNPs where no significant association was detected, this may reflect false positives in the initial study or false negatives in the current study. Alternatively, SNPs may have specific lifecycle stage-specific effects on growth that were not observed in both studies due to the difference in age at which the salmon were measured (one and three years respectively). While only weight and length were measured in the GWAS [14], there were eight fillet- and carcass-related traits measured in the current study. Therefore, for the two SNPs that were validated in the current study, the use of these additional measurements helps to determine a more specific growth phenotype associated with the SNP effects. For example, the SNP AX88270804 was associated with fat percentage in the current study, which indicates that the faster growth associated with the favourable allele also leads to increased fat content of the fish.

Alignment of the SNP flanking sequences with the Atlantic salmon reference genome predicted that AX88270804 was a synonymous exonic SNP within the *PCNT* gene and showed a significant association with several muscle and skeletal growth traits ($p < 0.05$) in Population 2 (current study) and growth traits ($p \sim 10^{-4}$) in Population 1 [14]. The SNP explained between 1 % and 4 % of the genetic variation in various harvest traits. In humans, the *PCNT* gene encodes the centrosome protein pericentrin, which contributes to the organisation of the mitotic spindle for the segregation of the chromosomes during cell division, thus influencing cell cycle progression. Mitotic centrosome dysfunction caused by pericentrin mutations can be expected to cause disturbances in cell division and is known to result in seriously stunted growth of the body and brain [28,29]. Interestingly, the SNP within the *PCNT* gene in salmon also has a suggestive association with fat percentage, explaining approximately 4 % of the genetic variation. As expected, the allele associated with faster growth is also associated with increased fatness (Table 5-2). Major mutations in the *PCNT* gene in humans also affect adipocyte differentiation and can result in dyslipidemia as part of a wider insulin

resistance syndrome. The fact that *PCNT* function is necessary for normal growth and lipid regulation in humans raises the possibility that further minor genetic variation within and around the gene may contribute to phenotypic variation in these traits. However, the role of the pericentrin in salmonid species has not yet been established.

The SNP AX88141678 was found in the intronic region of the *MEP1A* gene, which encodes meprin A subunit alpha. Meprins are zinc metalloendopeptidases that are predominantly found in kidney and intestinal brush border membranes in mammals and are known to play a role in protein metabolism [30]. Like *PCNT*, little is known about the function of *MEP1A* in Atlantic salmon, but interestingly, diet manipulation in another salmonid species (rainbow trout (*O. mykiss*)) has been shown to result in marked expression changes of *MEP1A* in the intestine [31]. In addition, *MEP1A* expression was shown to differ between domesticated and wild brook char (*Salvelinus fontinalis*) and its putative effect on growth factors was postulated to be the underlying mechanism for the higher expression in selected fish [32]. Therefore, while the association with growth traits may be due to variation in nearby candidate genes, the association of an SNP within the *MEP1A* gene and growth traits and its postulated functional connection to the growth traits raise the possibility that the causative effect underlying this association may be mediated via the *MEP1A* gene itself. It is worth noting that the genotype means for the SNP suggest an overdominance effect, which may explain why the additive variation explained is very small (Table 5-2).

Loci AX88141678 and AX88270804 were mapped to chr. 5 and chr. 16 using sire-based linkage mapping, respectively [6], and alignment with the reference genome assembly. A recent quantitative trait loci (QTL) mapping study by our group [20] in the same population as the current study showed that chr. 16 harbours loci affecting several growth traits with chromosome-wide significance in a sire-based analysis, although no QTL were detected on chr. 5. To date, there is a lack of consistency between the locations of the QTL affecting growth traits in different studies and commercial salmon populations [20,33–35]; therefore, the growth traits are considered to be regulated by

population-specific and polygenic factors. Further, while the association between the *PCNT* and *MEP1A* candidate gene polymorphisms and growth-related phenotypes measured in two different populations of salmon is encouraging, the direction of the allelic effects between the two studies was generally not consistent (see Table S2). For both SNPs in the current study, fish carrying two copies of the adenine allele had better growth performance than other genotypes, whereas in Tsai et al. [14], this genotype was associated with lower weight and length values. This may be due to opposing effects in different lifecycles and environments (freshwater *versus* seawater). A genotype by environment interaction has been shown to be evident for the direction of association of individual SNPs (*e.g.* [36]). Alternatively, these SNPs may be marking QTL some distance away, and the relationship between marker and QTL may vary from population to population. The QTL regions identified in the GWAS cover a relatively large region of the chromosomes (Figures 5-1 and 5-2). As such, while identifying chromosomal regions and putative genes harbouring variation contributing to growth phenotypes in salmon is of biological interest, it is unlikely that specific marker-assisted selection for these individual loci will be of high value, in particular for growth traits, which are directly measurable on the selection candidates themselves. This is particularly the case because genomic prediction using relatively few genome-wide markers can lead to very accurate prediction of breeding values for complex traits, such as growth (*e.g.* accuracy ~ 0.7 for juvenile weight and length in [14]). Therefore, genomic selection-based breeding schemes are likely to be increasingly utilised for the improvement of polygenic traits as genotyping technology becomes more affordable [14,37], especially for those traits with high economic value and that are difficult to be visualized (*e.g.* milk yield in dairy and fillet weight in fish).

5.6 Chapter Conclusions

In genome-wide association studies of complex and polygenic traits, the significant SNPs identified are likely to contain a mix of true associations and false positives. Therefore, verification of GWAS findings in a separate population is an important validation step, and SNP associations identified in more than one population are more likely to be reflecting real QTL. We identified two (out of eight successfully genotyped) SNPs that showed an association with growth traits in two different populations, and two different lifecycle stages, in Atlantic salmon. The SNPs are within the pericentrin and meprin alpha genes, which both have potentially relevant functional connections to the growth and harvest traits studied. Further investigation of these candidate genes may be merited to identify putative causative variation.

5.7 Conclusions

The heritabilities of growth traits are highly heritable, and comparable with the findings in chapter 3. Eight SNPs were successfully genotyped, and two of which showed significant associations with several weight-related traits across two populations. Loci AX88141678 and AX88270804 were mapped to chromosome 5 and 16, and the putative genes were *MEP1A* and *PCNT*, respectively. In comparison with results in chapter 3 and chapter 4, the linkage groups or genes showing significant association with targeted traits only have a little in common (*e.g.* chromosome 16 was identified in sire-based analysis only in chapter 3), implying the growth traits are regulated by population-specific and polygenic factors. The results highlight that the effects of significant SNPs may be varying in different populations and studies.

5.8 Additional Files

Supplementary Table 1. The markers selected from GWA analysis for verification. Link: http://goo.gl/F9oDB5

Supplementary Table 2. The marker ID, p-value estimated by the GWA analysis, putative gene and flanking sequence of two informative SNPs. The results were given in Tsai et al. [14]. Link: http://goo.gl/F9oDB5

5.9 Acknowledgements

The authors acknowledge funding from the Biotechnology and Biological Sciences Research Council (BBSRC) via Grant Number BB/H022007/1 and by Institute Strategic Funding Grants to the Roslin Institute (BB/J004235/1 and BB/J004324/1). Hsin Y. Tsai is supported by the Studying Funding from the Ministry of Education, Taiwan.

5.10 References

1. FAO. *Yearbook. Fishery and Aquaculture Statistics*; F.A.O.: Rome, Italy, 2010.

2. Yáñez, J.M.; Newman, S.; Houston, R.D. Genomics in aquaculture to better understand species biology and accelerate genetic progress. *Front. Genet.* 2015(6): 128.

3. Bishop, S.C.; Woolliams, J.A. Genomics and disease resistance studies in livestock. *Livest. Sci.* 2014, 166:190–198.

4. Gjedrem, T.; Robinson, N.; Rye, M. The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. *Aquaculture* 2012, 350-353:117–129.

5. Lien, S.; Gidskehaug, L.; Moen, T.; Hayes, B.J.; Berg, P.R.; Davidson, W.S.; Omholt, S.W.; Kent, M.P. A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genom.* 2011, 12:615.

6. Houston, R.D.; Taggart, J.B.; Cézard, T.; Bekaert, M.; Lowe, N.R.; Downing, A.; Talbot, R.; Bishop, S.C.; Archibald, A.L.; Bron, J.E.; *et al*. Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). *BMC Genom.* 2014, 15:90.

7. Palti, Y.; Gao, G.; Liu, S.; Kent, M.P.; Lien, S.; Miller, M.R.; Rexroad, C.E.; Moen, T. The Development and Characterization of a 57K SNP Array for Rainbow Trout. *Mol. Ecol. Resour.* 2015, 15:662–672.

8. Sánchez, C.C.; Smith, T.P.L.; Wiedmann, R.T.; Vallejo, R.L.; Salem, M.; Yao, J.; Rexroad, C.E. Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genom.* 2009, 10:559.

9. Everett, M.V; Seeb, J.E. Detection and mapping of QTL for temperature tolerance and body size in Chinook salmon (Oncorhynchus tshawytscha) using genotyping by sequencing. *Evol. Appl.* 2014, 7:480–492.

10. Gutierrez, A.P.; Yáñez, J.M.; Fukui, S.; Swift, B.; Davidson, W.S. Genome-Wide Association Study (GWAS) for Growth Rate and Age at Sexual Maturation in Atlantic Salmon (Salmo salar). *PLoS ONE* 2015, 10:e0119730.

11. Sodeland, M.; Gaarder, M.; Moen, T.; Thomassen, M.; Kjøglum, S.; Kent, M.; Lien, S. Genome-wide association testing reveals quantitative trait loci for fillet texture and fat content in Atlantic salmon. *Aquaculture* 2013, 408-409:169–174.

12. Houston, R.D.; Bishop, S.C.; Guy, D.R.; Tinch A.E.; Taggart, J.B.; Bron, J.E.; Downing, A.; Stear, M.J.; Gharbi, K.; Hamilton, A. Genome Wide Association Analysis for Resistance to Sea Lice in Atlantic Salmon: Application of a Dense SNP Array. In Proceedings of the 10th World Congress on Genetics Applied to Livestock Production, Vancouver, BC, Canada, 17–22 August 2014; pp. 10–12.

13. Correa, K.; Lhorente, J.; Lopez, M.; Bassini, L.; Naswa, S.; Deeb, N.; di Genova, A.; Maas, A.; Davidson, W.; Yáñez, J. Genome-wide association analysis reveals loci associated with resistance against Piscirickettsia salmonis in two Atlantic salmon (Salmo salar L.) chromosomes. *BMC Genom.* 2015, 16:854.

14. Tsai, H.Y.; Hamilton, A.; Tinch, A.E.; Guy, D.R.; Gharbi, K.; Stear, M.J.; Oswald, M.; Bishop, S.C.; Houston, R.D. Genome wide association and genomic prediction

for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. *BMC Genom.* 2015, 16:969.

15. Macqueen, D.J.; Johnston, I.A. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. R. Soc. B* 2014, 281:1778.

16. Berthelot, C.; Brunet, F.; Chalopin, D.; Juanchich, A.; Bernard, M.; Noël, B.; Bento, P.; Da Silva, C.; Labadie, K.; Alberti, A.; *et al*. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat. Commun.* 2014, 5:3657.

17. Gidskehaug, L.; Kent, M.; Hayes, B.J.; Lien, S. Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array. *Bioinformatics* 2011, 27:303–310.

18. Wolfe, K.H. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2001, 2:333–241.

19. Tsai, H.Y.; Hamilton, A.; Guy, D.R.; Houston, R.D. Single nucleotide polymorphisms in the insulin-like growth factor 1 (IGF1) gene are associated with growth-related traits in farmed Atlantic salmon. *Anim. Genet.* 2014, 45:709–715.

20. Tsai, H.Y.; Hamilton, A.; Guy, D.R.; Tinch, A.E.; Bishop, S.C.; Houston, R.D. The genetic architecture of growth and fillet traits in farmed Atlantic salmon (Salmo salar). *BMC Genet.* 2015, 16:51.

21. Powell, J.; White, I.; Guy, D.; Brotherstone, S. Genetic parameters of production traits in Atlantic salmon (Salmo salar). *Aquaculture* 2008, 274:225–231.

22. Mackay, T.F.C.; Stone, E.A.; Ayroles, J.F. The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* 2009, 10:565–77.

23. Warde-Farley, D.; Donaldson, S.L.; Comes, O.; Zuberi, K.; Badrawi, R.; Chao, P.; Franz, M.; Grouios, C.; Kazi, F.; Lopes, C.T.; *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 2010, 38:W214–W220.

24. Moghadam, H.K.; Poissant, J.; Fotherby, H.; Haidle, L.; Ferguson, M.M.; Danzmann, R.G. Quantitative trait loci for body weight, condition factor and age at sexual maturation in Arctic charr (Salvelinus alpinus): comparative analysis with rainbow trout (Oncorhynchus mykiss) and Atlantic salmon (Salmo salar). *Mol. Genet. Genom.* 2007, 277:647–661.

25. Peñaloza, C.; Hamilton, A.; Guy, D.R.; Bishop, S.C.; Houston, R.D. A SNP in the 5' flanking region of the myostatin-1b gene is associated with harvest traits in Atlantic salmon (Salmo salar). *BMC Genet.* 2013, 14:112.

26. Davidson, W.S.; Koop, B.F.; Jones, S.J.M.; Iturra, P.; Vidal, R.; Maass, A.; Jonassen, I.; Lien, S.; Omholt, S. W. Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome Biol.* 2010, 11:403.

27. Yáñez, J.M.; Houston, R.D.; Newman, S. Genetics and genomics of disease resistance in salmonid species. *Front. Genet.* 2014, 5:415.

28. Rauch, A.; Thiel, C.T.; Schindler, D.; Wick, U.; Crow, Y.J.; Ekici, A.B.; van Essen, A.J.; Goecke, T.O.; Al-Gazali, L.; Chrzanowska, K.H; *et al*. Mutations in the Pericentrin (PCNT) Gene Cause Primordial Dwarfism. *Science* 2008, 319:816–819.

29. Müller, T.; Rumpel, E.; Hradetzky, S.; Bollig, F.; Wegner, H.; Blumenthal, A.; Greinacher, A.; Endlich, K.; Endlich, N. Non-muscle myosin IIA is required for the development of the zebrafish glomerulus. *Kidney Int.* 2011, 80:1055–1063.

30. Sterchi, E.E.; Stöcker, W.; Bond, J.S. Meprins, membrane-bound and secreted astacin metalloproteinases. *Mol. Asp. Med.* 2010, 29:309–328.

31. Kirchner, S.; McDaniel, N.K.; Sugiura, S.H.; Soteropoulos, P.; Tian, B.; Fletcher, J.W.; Ferraris, R.P. Salmonid microarrays identify intestinal genes that reliably monitor P deficiency in rainbow trout aquaculture. *Anim. Genet.* 2007, 38: 319–331.

32. Sauvage, C.; Derôme, N.; Normandeau, E.; St-Cyr, J.; Audet, C.; Bernatchez, L. Fast transcriptional responses to domestication in the brook charr Salvelinus fontinalis. *Genetics* 2010, 185:105–112.

33. Gutierrez, A.P.; Lubieniecki, K.P.; Davidson, E.A.; Lien, S.; Kent, M.P.; Fukui, S.; Withler, R.E.; Swift, B.; Davidson, W.S. Genetic mapping of quantitative trait loci (QTL) for body-weight in Atlantic salmon (Salmo salar) using a 6.5K SNP array. *Aquaculture* 2012, 358-359:61–70.

34. Baranski, M.; Moen, T.; Våge, D.I. Mapping of quantitative trait loci for flesh colour and growth traits in Atlantic salmon (Salmo salar). *Genet. Sel. Evol.* 2010, 42:17.

35. Houston, R.D.; Bishop, S.C.; Hamilton, A.; Guy, D.R.; Tinch, A.E.; Taggart, J.B.; Derayat, A.; McAndrew, B.J.; Haley, C.S. Detection of QTL affecting harvest traits in a commercial Atlantic salmon population. *Anim. Genet.* 2009, 40:753–755.

36. Bastiaansen, J.W.M.; Bovenhuis, H.; Lopes, M.S.; Silv, F.F.; Megens, H.J.; Calu, M.P.L. SNP Effects Depend on Genetic and Environmental Context. In Proceedings

of the 10th World Congress on Genetics Applied to Livestock Production, Vancouver, BC, Canada, 17–22 August 2014; pp. 356–362.

37. Odegård, J.; Moen, T.; Santi, N.; Korsvoll, S.A.; Kjøglum, S.; Meuwissen, T.H.E. Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). *Front. Genet.* 2014, 5:402.

38. LGC Genomics KASP (kompetitive allele-specific PCR) assays. Available online: http://www.lgcgroup.com/products/kasp-genotyping-chemistry/kasp-technicalresources/#.VWC3TU9VhXt (accessed on 01 November 2015).

39. Gilmour, A.R.; Gogel, B.J.; Cullis, B.R.; Thompson, R. *ASReml User Guide*, 4th ed.; VSN International Ltd: Hemel Hempstead, UK, 2014.

40. Falconer, D.S.; Mackay, T.F.C. *Introduction to Quantitative Genetics*, 4th ed.; Longmans Green: Harlow, Essex, UK, 1996.

41. Centre for biomedical research, RepeatMasker, University of Victoria. Available online: http://lucy.ceh.uvic.ca/repeatmasker/cbr_repeatmasker.py (accessed on 30 August 2015).

Chapter 6

Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations

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.

DOI: 10.1186/s12711-016-0226-9

RDH and SCB were responsible for the overall experimental design; JEB, JBT, MJS and KG designed experiments and collected data; AET, AH, and DRG provided phenotype and pedigree data; HYT, OM, RPW and RDH analysed data; HYT and RDH prepared the manuscript.

This chapter has been published in Genetic Selection Evolution at the following URL:

http://gsejournal.biomedcentral.com/articles/10.1186/s12711-016-0226-9

6.0 Introduction

In previous chapters (chapter 3 to chapter 5), I have applied QTL mapping, GWAS, and genomic prediction to uncover the genetic basis of growth traits. My results showed that the prediction accuracies of both body weight and length were high. Because the growth traits are highly heritable and relatively easier to be measured, the genomic prediction applied with growth trait is likely less useful in comparison with disease traits. The most commercially important trait is sea lice (*L. salmonis*) to date, which is thought as the most critical parasitic disease in salmon farming in European countries. According to statistics, sea lice have caused approximately £25 million financial losses annually in Scotland. As such, genomic prediction is likely able to provide more contributions especially for disease resistance traits in salmon breeding. In this chapter, I aim to understand the genetic architecture of host resistance to sea lice by using GWAS and genomic prediction. Specifically, the objectives are to estimate the heritability of lice resistance trait, and to assess the prediction breeding values of individuals using pedigree and genomic-based approach, and to identify the single SNP associated with lice resistance trait by performing GWAS. The locations of identified loci in the corresponding linkage map were identified by using data generated in chapter 2.

6.1 Abstract

Background

Sea lice have significant negative economic and welfare impacts on marine Atlantic salmon farming. Since host resistance to sea lice has a substantial genetic component, selective breeding can contribute to control of lice. Genomic selection uses genome-wide marker information to predict breeding values, and can achieve markedly higher accuracy than pedigree-based methods. Our aim was to assess the genetic architecture of host resistance to sea lice, and test the utility of genomic prediction of breeding values. Individual lice counts were measured in challenge experiments using two large Atlantic salmon post-smolt populations from a commercial breeding programme, which had genotypes for ~ 33 K single nucleotide polymorphisms (SNPs). The specific objectives were to: (i) estimate the heritability of host resistance; (ii) assess its genetic architecture by performing a genome-wide association study (GWAS); (iii) assess the accuracy of predicted breeding values using varying SNP densities (0.5 to 33 K) and compare it to that of pedigree-based prediction; (iv) evaluate the accuracy of prediction in closely and distantly related animals.

Results

Heritability of host resistance was significant (0.22 to 0.33) in both populations using either pedigree or genomic relationship matrices. The GWAS suggested that lice resistance is a polygenic trait, and no genome-wide significant quantitative trait loci (QTL) were identified. Based on cross-validation analysis, genomic predictions were more accurate than pedigree-based predictions for both populations. Although prediction accuracies were highest when closely-related animals were used in the training and validation sets, the benefit of having genomic versus pedigree-based predictions within a population increased as the relationships between training and validation sets decreased. Prediction accuracy reached an asymptote with a SNP density of ~ 5 K within populations, although higher SNP density was advantageous for cross-population

prediction.

Conclusions

Host resistance to sea lice in farmed Atlantic salmon has a significant genetic component. Phenotypes relating to host resistance can be predicted with moderate to high accuracy within populations, with a major advantage of genomic over pedigree-based methods, even at relatively sparse SNP densities. Prediction accuracies across populations were low, but improved with higher marker densities. Genomic selection can contribute to lice control in salmon farming.

6.2 Chapter Introduction

Genomic selection (GS) involves the prediction of individual breeding values for complex traits by combining statistical methods with genome-wide single nucleotide polymorphism (SNP) data. Relationships between SNPs and traits of interest are first determined within a reference (or training) population, and then they are used to identify selection candidates with high genetic merit in the absence of phenotype records [1, 2]. The feasibility of GS schemes depends on the availability of a high-quality SNP genotyping platform and on extensive trait records collected in the reference populations. Due to the increased availability of high-density SNP chips and the development of genotyping-by-sequencing for several economically important livestock and aquaculture species (*e.g.* [3–7]), GS has become a widely used approach, particularly for traits of economic and welfare importance (*e.g.* disease resistance). The accuracy of predicted breeding values based on genomic data is expected to be substantially higher than that based on pedigree records alone, but depends on many variables, including the genetic architecture of the trait, SNP density, sample size, and the degree of relationship between the reference and validation sets [8, 9].

In Atlantic salmon farming, ectoparasitic copepods, commonly known as sea lice (specifically *Lepeophtheirus salmonis* in Europe and *Caligus rogercresseyi* in Chile), are the primary threat to sustainable production, and have a negative economic, animal welfare, and environmental impact. Symptoms of *L. salmonis* infection include skin lesions, osmotic imbalance, and increased susceptibility to other infections as a result of host immune suppression and skin damage [10]. Frequent chemical treatments are required to control louse infections on commercial farms and result in large annual costs, potential environmental damage, and a high prevalence of drug-resistant lice [10, 11]. However, there is encouraging evidence from challenge trials that revealed heritabilities of approximately 0.2 to 0.3 for lice resistance, as measured by counts of lice on the fish (*e.g.* [11–14]), highlighting host genetic variation in resistance to lice. Therefore, selective breeding to improve host resistance to lice in farmed salmon populations is an

increasingly important component of disease control [9, 11]. Given the importance of the sea lice issue to the salmon industry, this trait is also a high priority candidate for GS to accelerate the production of stocks with increased resistance.

The quantitative genetic models that underpin GS can be broadly split into two categories based on the assumptions that underlie the genetic architecture of the trait. The first category assumes an even distribution of the genetic variance across the genome and includes genomic best linear unbiased prediction (GBLUP) methods. The second category allows for heterogeneity in the contribution of the markers to the genetic variance, which is typically modelled using Bayesian methods (*e.g.* [15]). While the Bayesian methods (*e.g.* Bayes B) are generally more accurate than GBLUP on simulated data, particularly when the number of quantitative trait loci (QTL) that underlie the genetic variance is small [8], prediction accuracy using 'experimental' data in livestock breeding schemes is often very similar with either of these two methods [16]. Genomic prediction using these models relies both on capturing linkage disequilibrium (LD) between SNPs and QTL and on accurate estimates of realised genetic relationships between individuals [9, 17]. In typical farm animal populations, prediction accuracy depends largely on the latter [18], but the persistency of prediction accuracy across generations and between unrelated populations depends on the LD between SNPs and QTL [2, 9, 17]. For most commercial aquaculture breeding programmes, the availability of large full-sib families facilitates extensive trait measurements on individuals that are closely related to the selection candidates. Therefore, within-population genomic prediction will capitalise on realised genetic relationships, and the role of LD between SNPs and QTL may be less crucial [9, 18]. However, for salmon with a discrete 3-or 4-year generation interval, accuracy of prediction across adjacent year groups with limited genetic connectivity between them will depend more on LD, and is likely more challenging.

Family-based selective breeding programmes for Atlantic salmon have traditionally focused on economically important traits that can be easily measured on the selection candidates (*e.g.* growth) and on traits that can be measured on close relatives (*e.g.* full

and half siblings), such as disease resistance and processing traits. Studies of GS in aquaculture using both simulated and 'experimental' data have suggested that genomic prediction can result in more accurate breeding values than traditional pedigree-based approaches (*e.g.* [9, 19–21]). However, the cost-efficiency of GS is critical; both high-density SNP arrays and extensive collection of trait data can be prohibitively expensive for routine genomic evaluations. Therefore, knowledge of the optimal design of reference populations and of the required SNP density is important, as well as quantification of the benefit that can be expected from the implementation of GS.

The objectives of this study were to (i) estimate the heritability of host resistance to sea lice using both genomic and pedigree-based methods, (ii) analyse the genetic architecture of host resistance by performing a GWAS, (iii) assess the accuracy of genomic prediction using various SNP densities up to 33 K SNPs and compare it to that of pedigree-based prediction, and (iv) test genomic prediction accuracies in closely and more distantly related reference and validation populations.

6.3 Methods

6.3.1 Animal and Challenge Experiment

The animals used in the study were taken from a commercial Atlantic salmon breeding programme (Landcatch, UK). Due to the four-year generation interval, the breeding program consists of four sub-populations (referred to as year groups), two of which were studied. Full details for population I (2007 year group, n = 624) were previously described in Tsai et al. [21]. Briefly, this population consisted of 531 genotyped offspring with complete phenotype and genotype information, derived from 61 nucleus families (30 sires and 59 dams). The families in population I were reared in separate tanks until approximately 9 months post-hatch, at which time they were mixed. Population II (2010 year group, n = 874) comprised 151 families (98 sires and 188

dams), with 588 offspring that were phenotyped and genotyped. The families in population II were mixed at first feeding and reared in a single common tank. The lice challenge trials were conducted at the Marine Environmental Research Laboratory (Machrihanish, UK) in 2007 and 2010, respectively. The challenge protocols were similar for both populations; the fish (approximately 1 year post-hatching) were challenged in a single tank with a moderate dose of copepodid larvae (90 to 96 larvae per fish) and then monitored daily until most lice had moulted into chalimus I. Sampling and measurements began on day 7 post-challenge and lasted one and 4.5 days for populations I and II, respectively (for population I, lice counts were shown to be stable between 7 and 17 days post-challenge [11]). Prior to lice counting, fish were euthanized with benzocaine as described in Gharbi et al. [11]. Phenotypes including weight (g), length (mm), and sea lice count [number of sea lice per fish, measured using a stereo-microscope (Olympus SZ-40)] were recorded for each fish. An adipose fin clip was collected and stored in ethanol for DNA extraction. For population I, pedigree information for each individual was traced by using passive integrated transponder (PIT) tags. For population II, a standard parentage assignment panel of 108 SNPs was screened on a Sequenom platform (DNA LandMarks Inc., Canada) to construct the pedigree.

All animals were reared in accordance with relevant national and EU legislation concerning health and welfare. The challenge experiment was performed by the Marine Environmental Research Laboratory (Machrihanish, UK) under approval of the ethics review committee of the University of Stirling (Stirling, UK) and according to Home Office license requirements. Landcatch are accredited participants in the RSPCA Freedom Foods standard, the Scottish Salmon Producers Organization Code of Good Practice, and the EU Code-EFABAR Code of Good Practice for Farm Animal Breeding and Reproduction Organizations.

6.3.2 SNP Genotyping

DNA was extracted from fin tissue samples using the DNeasy 96 tissue DNA extraction kit (Qiagen, UK). Population I was genotyped with an Affymetrix Axiom SNP array that included ~ 132 K SNPs [22] and population II was genotyped with the custom Affymetrix Axiom ~ 35 K array described in Tsai et al. [21]. This 35 K array is used for routine genomic evaluations and includes a subset of high-quality SNPs of the 132 K array that were selected based on having a good distribution throughout the genome and minimal LD between pairs of SNPs [21]. Sex of the fish was predicted by using the Y-specific probes on the 132 K array, as described by Houston et al. [22]. Filtering of SNP data was performed using the Plink software [23], excluding SNPs with Mendelian errors, with a minor allele frequency (MAF) lower than 0.1 and with a proportion of missing genotypes greater than 0.03. Finally, approximately 33 K SNPs were retained for analyses in both populations.

6.3.3 Genetic Parameters for Lice Resistance

6.3.3.1 Data Normalization

The raw data for lice counts showed a positively skewed distribution [See Additional file 1 Figure S1], thus to normalize this distribution, we transformed the data using a previously applied approach that also accounts for an approximation of the surface area of the fish [13]:

$$\log_e \mathrm{LD} = \log_e\left((\mathrm{LC}+1)/(\mathrm{BW})^{\wedge 2/3}\right), \quad (1)$$

where LC is the number of lice counted on the fish (plus one to avoid a computation error since some fish may have zero lice), $(\mathrm{BW})^{\wedge 2/3}$ is an approximation of the whole

surface of the skin of each individual, where BW represents the body weight (g) at the time of the sea lice challenges. A moderate correlation of 0.35 was found between body surface and lice count.

6.3.3.2 Estimation of Genetic parameters

The heritability of host resistance to sea lice count (and of weight and length traits) was estimated using both genomic and pedigree-based analyses for the two populations. Only fish with complete phenotype and genotype records were included, resulting in 531 and 588 fish in populations I and II, respectively. Heritabilities were estimated by ASReml 3.0 [24] using genomic and pedigree-based relationship matrices (**G**-matrix and **A**-matrix, respectively) with the following mixed model:

$$\mathbf{Y} = \mu + \mathbf{Xb} + \mathbf{Za} + \mathbf{e}, \quad (2)$$

where **Y** is a vector of observed phenotypes, μ is the overall mean of phenotype records, **b** is the vector of fixed effects, **a** is a vector of additive genetic effects distributed as $\sim N(0, \mathbf{G}\sigma_a^2)$ or $N(0, \mathbf{A}\sigma_a^2)$ where σ_a^2 is the additive (genetic) variance, **G** and **A** are the genomic and pedigree relationship matrices, respectively. **X** and **Z** are the corresponding incidence matrices for fixed and additive effects, respectively, and **e** is a vector of residuals. If the SNPs applying sex as the fixed effect did not surpass the genome-wide significance threshold (Bonferroni correction (0.05/N), where N represents the number of QC-filtered SNPs across the entire genome), it was omitted from subsequent analyses. The genomic relationship matrix was constructed by the Genabel R package [25] using the method of VanRaden [26] and then inverted by applying a standard R function (https://www.r-project.org/). Narrow sense heritability was estimated as the ratio of

additive genetic variance to total phenotypic variance. Phenotypic correlations between traits were estimated using ASReml 3.0 [24] and genetic correlations were estimated using bivariate analyses implemented in ASReml 3.0 [24] as well.

6.3.4 Genome-wide Association Study

The two-step 'GRAMMAR' approach was used to perform the GWAS using the GenABEL R Package [25]. The GWAS was performed in each population separately, and on the two populations combined. First, model (2) was applied to adjust the lice count data based on fixed (year group in the combined analysis) and polygenic effects (relationships between animals as measured by the genomic relationship matrix). Subsequently, the mmscore method [27] of GenABEL was applied to measure the association between individual SNPs and the residuals from model (2) (which are corrected for family relatedness). Significance thresholds were calculated using a Bonferroni correction to obtain genome-wide (0.05/number of all quality-control filtered SNPs, ~ 33 K) and chromosome-wide (0.05/number of SNPs on the corresponding chromosome) thresholds, respectively. For the SNPs that were closest to chromosome-wide significance (*i.e.* those with the lowest P values), allele substitution effects were estimated using model (2) in ASReml 3.0 [24] by including the fixed effects of SNP genotype and population. The additive effect (a) of the SNP was calculated as half the difference between the predicted phenotypic means of the two homozygotes, *i.e.* $(\mathrm{AA}-\mathrm{BB})/2$, and the dominance effect (d) was calculated as $\mathrm{AB}-[(\mathrm{AA}+\mathrm{BB})/2]$, where the AB represents the predicted phenotypic mean of the heterozygote. The additive genetic SNP variance (σ^2_{SNP}) was estimated using the following equation:

$$\sigma^2_{SNP} = 2\mathrm{pq}\big(a + d(\mathrm{q}-\mathrm{p})\big)^2, \qquad (3)$$

where p and q are the frequency of the major and minor alleles at the SNP, respectively. The proportion of variance explained by the SNP is then given by:

$\sigma^2_{SNP}/\sigma^2_a$ (4)

where σ^2_a is the total additive genetic variance of the trait when no SNP effects are included in the model.

6.3.5 Assessment of Genomic Prediction

The utility of genomic prediction for resistance to lice was assessed by cross-validation analyses under various scenarios (see below) in which (i) varying SNP densities (0.5 K, 1 K, 5 K, 10 K, 20 K (all chosen at random), and 33 K (full dataset)) and (ii) varying degrees of relationships between training and validation sets were applied.

6.3.5.1 Scenario (i): Random Selection

Within each population (which correspond to discrete 'year groups' of a commercial Atlantic salmon breeding programme), cross-validation analysis was performed by selecting five random non-overlapping training and validation sets as described previously [21]. At each SNP density (0.5 to 33 K SNPs), GBLUP was applied to predict the masked phenotypes of the validation sets and the resulting prediction accuracy was compared to that of pedigree-based BLUP (PBLUP), as described above. The average accuracy across the five cross-validation replicates for each SNP density was computed.

6.3.5.2 Scenario (ii): Sibling

Within each population, training and validation sets were established such that both sets contained representatives of each family. The same cross-validation analyses were performed as for Scenario (i).

6.3.5.3 Scenario (iii): Non-sibling

Within each population, training and validation sets were established such that full

siblings were not included in either set (*i.e*. different full-sibling families were used for training and validation sets). The resulting training and validation sets were more distantly related than for Scenarios (i) and (ii), although they did contain some half-sibs. The same cross-validation analyses were performed as for Scenarios (i) and (ii).

6.3.5.4 Scenario (iv): Across Populations

To assess prediction accuracy across populations per year group, population I was used as the training set and population II as the validation set, and *vice versa*. The same genomic prediction and cross-validation analyses were performed as for Scenarios (i) to (iii), but pedigree-based prediction was not possible since genetic links between the two populations were absent from the available pedigree.

6.3.6 Cross-Validation

The five-fold cross-validation analyses for each scenario described above were performed using the methods described in Tsai et al. [21]. Briefly, for the within-population analyses, populations I and II were each divided into a training (80 %) and validation (20 %) set. Phenotypes (*i.e*. lice counts) of the samples in the validation sets were then masked and GBLUP or pedigree-based BLUP (PBLUP) was applied to predict the phenotypes of the masked individuals using model (2) implemented in ASReml 3.0 [24]. The Pearson correlation coefficient of the estimated breeding values (EBV) [either genomic EBV (GEBV) or pedigree-based EBV (PEBV)] with the adjusted phenotype of the masked validation set. Accuracy was calculated as the correlation divided by the square root of heritability using all individuals, and then averaged across the five replicates (Figures. 6-2 and 6-3).

6.4 Results

6.4.1 General statistics and genetic parameters of resistance to lice and growth

Estimated heritability for lice count was moderate (~ 0.3) and relatively consistent when using pedigree relationship matrix (Table 6-1). Estimates of heritability for the growth-related traits (weight and length) were higher (~ 0.6), in line with previous estimates [21]. The two growth traits had a high positive phenotypic and genetic correlation with each other (~ 0.93 to 0.96), and correlations of the growth traits with lice count were either equal to zero or slightly negative (Table 6-2).

Table 6-1. General statistics and heritability estimates for lice count and growth traits.

	Population I		**Population II**	
	Mean (SD)	Heritability[1] (SE)	Mean (SD)	Heritability[1] (SE)
Lice[2]	25.8 (12.3)	0.33 (0.08) / 0.27 (0.08)	18.3 (9.1)	0.22 (0.06) / 0.27 (0.08)
Length	214.2 (16.1)[3]	0.61 (0.07) / 0.51 (0.11)[3]	206.2 (14.3)	0.51 (0.07) / 0.50 (0.10)
Weight	112.0 (21.0)[3]	0.61 (0.07) / 0.49 (0.10)[3]	89.9 (19.9)	0.50 (0.07) / 0.50 (0.10)

[1]Heritability was estimated based on the **G**-matrix / **A**-matrix

[2]The lice count data (number of lice per fish) used here was without data adjustment

[3]The results are from Tsai et al. [21]

SD is the standard deviation and SE is the standard error

Table 6-2 Estimates of genetic and phenotypic correlations between lice count and growth traits in populations I and II.

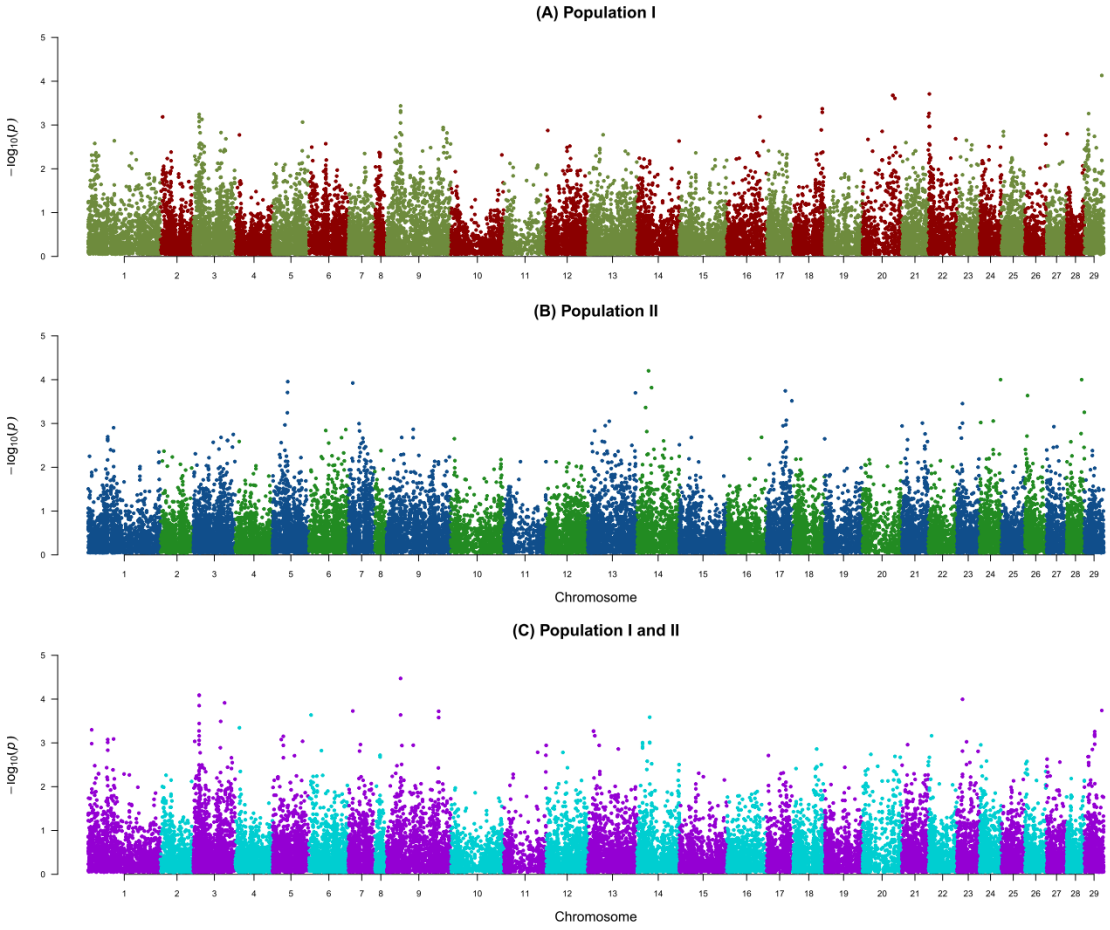
Phenotypic correlation			
	Lice	**Length**	**Weight**
Genetic correlation			
Population I			
Lice	-	-0.04	-0.06
Length	0.10	-	0.96
Weight	0.11	0.96	-
Population II			

Lice	-	-0.1	-0.1
Length	-0.3	-	0.93
Weight	-0.3	0.95	-

6.4.2 Genome-wide Association Study

The results of the GWAS suggest that lice resistance is a polygenic trait, with no SNPs surpassing the Bonferroni-corrected significance thresholds (Figure 6-1). Indeed, when each population was analysed separately, there was little overlap between regions that had the lowest P values (Figure 6-1a and 6-1b). When the two populations were combined (Figure 6-1c), SNPs with the lowest P values were located on chromosomes 1, 3, 9 and 23. The estimated proportion of additive genetic variance explained by these SNPs ranged from ~ 2 to 6 % each. The quantile-quantile (Q-Q) plots for each GWA analysis are in Figure S2 [See Additional file 2 Figure S2].

Figure 6-1. Manhattan plots of the genome-wide association study for populations I (A), II (B), and I and II combined (C). Top markers are close to chromosome-wide significance ($\alpha < 0.05$) but do not pass the threshold.

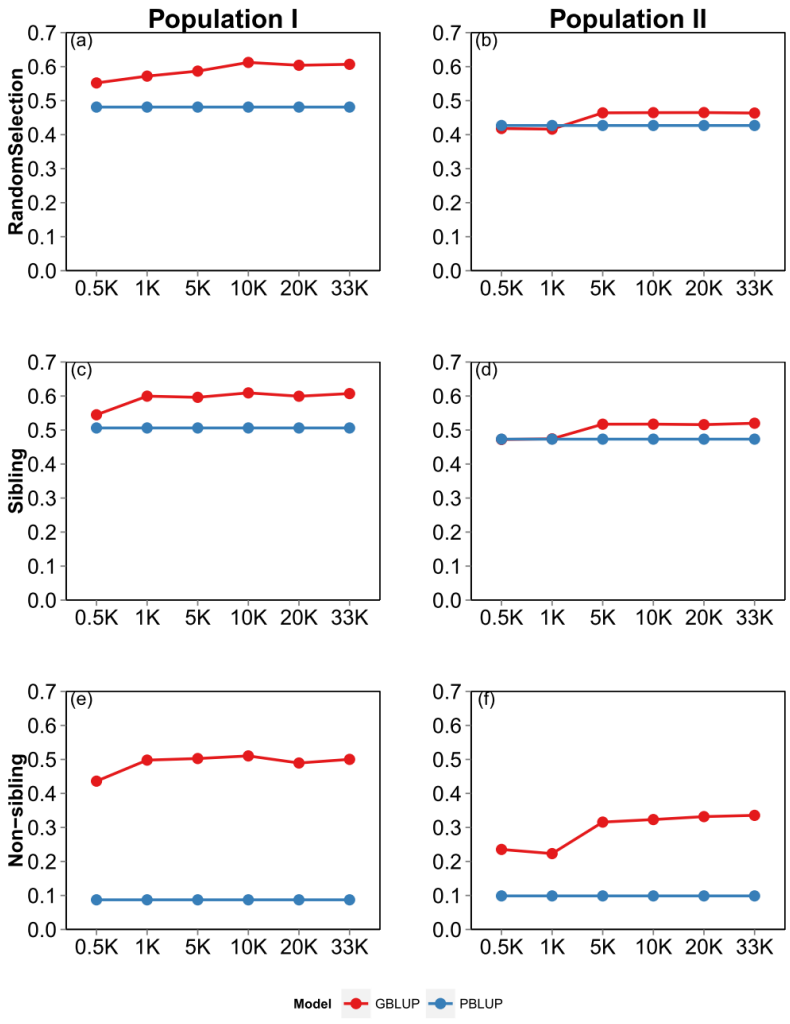


6.4.3 Accuracy of Prediction Breeding Values

The putative polygenic architecture of lice resistance in these populations means that genomic prediction may be a practical and effective method of predicting breeding values for lice resistance, which was tested using cross-validation analyses under different scenarios in which varying SNP densities and varying levels of relatedness between training and validation sets were applied (see 'Methods' for details). Accuracy of prediction using the genomic relationship matrix (GBLUP) was generally higher than that using the pedigree relationship matrix (PBLUP). Greater SNP density tended to improve prediction accuracy, but the asymptote was generally reached at ~ 5 K SNPs for both populations (Figure 6-2).

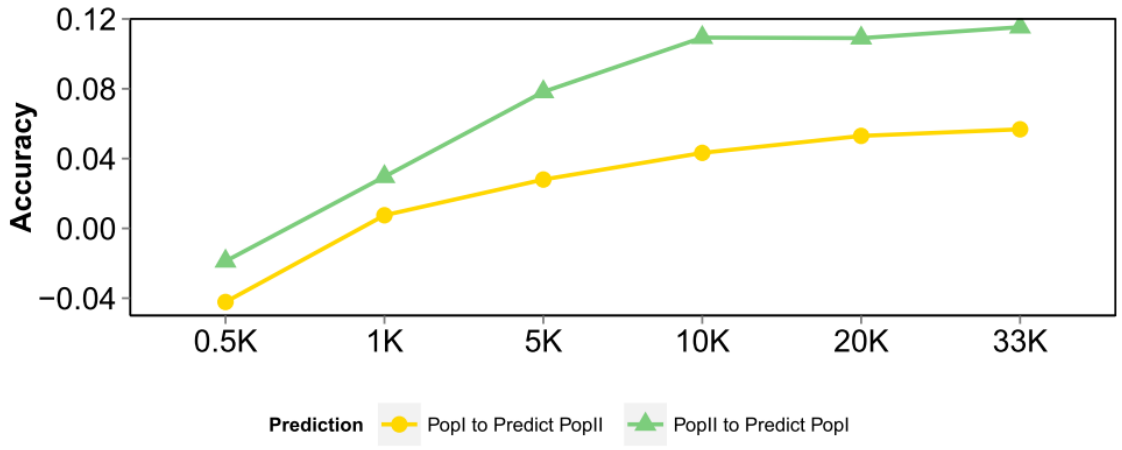
The results of genomic prediction under the "random selection" (where training and validation sets were chosen at random), and "sibling" (where full siblings from each family were deliberately included in both the training and validation sets) scenarios were very similar for both populations (Figures 6-2a to 6-2d). Therefore, including animals that share close relationships did not improve the accuracy of genomic predictions for these populations, which indicates that "random selection" will result in the presence of several closely-related fish across the training and validation data sets by chance. In both cases, GBLUP resulted in more accurate predictions of lice count in the validation data than PBLUP, with a relative advantage of approximately 27 % for population I and 10 % for population II (Figures 6-2a to 6-2d). Increasing marker density to more than ~ 5 K randomly chosen SNPs had little impact on prediction accuracy, which may be expected when the training and validation sets are closely related [9].

Figure 6-2. Accuracy of genomic and pedigree-based prediction within populations. Comparison of prediction accuracy (Y-axis) of two populations using increasing SNP densities from 0.5 to 33 K (X-axis) assessed by cross-validation analyses. "Random Selection" involved random assignment of individuals to training and validation sets (a) and (b); "sibling" involved assigning full siblings from each family to both the training and validation sets (c) and (d); and "non-sibling" involved avoidance of full-sibling animals in the training and validation sets (e) and (f). Panels (a), (c) and (e) represent results for population I and panels (b), (d), and (f) represent those for population II.



When the training and validation sets were less related, predictions of both pedigree- and genomic-based methods were less accurate, as expected. In the "non-sibling" scenario (where no full-siblings were included in both the training and validation sets), accuracies of prediction obtained with both GBLUP and PBLUP were substantially lower than those in the previous two scenarios. However, the benefit of genomic prediction was greatest under this scenario, with prediction accuracies increasing 5-fold (population I) and 2.5-fold (population II) relative to pedigree-based prediction accuracies. Perhaps surprisingly, there was little benefit from increasing SNP density above ~ 5 K SNPs under this scenario as well (Figures. 6-2e and 6-2f). When the accuracy of genomic prediction was assessed across the two populations (corresponding to two year groups of the Landcatch broodstock), accuracies were markedly lower (0.05 to 0.11) than with all of the within-population scenarios (0.34 to 0.61). Increasing SNP density did seem to yield incremental (albeit small) increases in prediction accuracies when predicting across populations (Figure 6-3), which suggested that this scenario was likely to benefit most from a high-density SNP array. However, these two populations were probably too small to achieve high prediction accuracy for these distantly-related animals, and a more thorough test of across-population prediction in salmon should use larger sample sizes.

Figure 6-3. Accuracy of genomic prediction across populations. Based on setting population I as the training set and population II as the validation set and *vice versa*. Accuracy of prediction (Y-axis) for the two populations was estimated using increasing SNP density from 0.5 to 33 K (X-axis).



6.5 Discussion

Genomic selection is an increasingly important component of modern aquaculture breeding schemes, with simulated and applied studies highlighting its benefits over pedigree-based selection [9, 28]. However, the substantial cost of genome-wide genotyping means that the traits targeted by GS are likely to be those of high economic value, particularly those that cannot be easily measured on the selection candidates themselves. Currently, sea lice present the largest threat to the sustainability of salmon farming, which relies heavily on expensive and potentially environmentally-damaging chemical treatments [10]. Host resistance to sea lice has consistently been shown to have a substantial genetic component [11]. Therefore, resistance to lice is an ideal candidate trait for the application of GS. In our study, lice count data and genome-wide SNP genotypes were collected for two pedigreed salmon populations from a commercial

breeding programme to assess the utility of genomic prediction of host resistance to sea lice under different scenarios, including a comparison to predictions based on pedigree records alone.

The heritability of resistance to lice was estimated at ~ 0.3 and 0.2 in populations I and II, respectively, which is similar to the findings of Gharbi et al. [11] (~ 0.3) and Gjerde et al. [29] (~ 0.2 to 0.3), and slightly higher than those of Ø degård et al. [9] (~ 0.13 to 0.14). However, it should be noted that the challenge experiments that are reported in Gharbi et al. [11], Gjerde et al. [13], and in our study, were all conducted in controlled tanks conditions, whereas the study of Ø degård et al. [9] was based on challenges in a sea-cage environment, which may display greater environmental variation. Furthermore, it should be noted that the higher heritability estimates for all traits in population I may be due in part to confounding between genetic and common environmental effects due to the family-specific rearing of the fry (compared to population II, for which individuals were mixed into a single tank as first feeding fry).

The GWAS indicated that host resistance to lice likely has a polygenic architecture, with no major QTL segregating in these populations (Figure 6-1). Therefore, it is likely that individual QTL for lice resistance explain only a small percentage of the genetic variance, and a proportion of the QTL may be population-specific. As such, GBLUP and similar methods of genomic prediction are likely to be suitable for predicting breeding values for host resistance to lice, particularly within populations.

The degree of the genetic relationships between training and validation sets is critical for the efficacy of genomic prediction. In our study, genomic prediction was found to be highly effective and showed a significant advantage in terms of accuracy over pedigree-based methods within populations (which correspond to year groups of a salmon breeding programme, Figure 6-2). The accuracy of prediction and the relative advantage of genomic prediction were lower for population II than for population I (Figure 6-2), which may reflect the lower estimated heritability in this population because a low heritability can contribute to low prediction accuracy [20, 30]. Also, the family structure

of population II was potentially less amenable to accurate prediction since it comprised a larger number of smaller families, which decreased the chance of having useful numbers of full siblings in the training and validation sets. Prediction accuracies were highest when training and validation sets were closely related, as was shown with the "Random selection" and "Sibling" scenarios. In addition, these results showed that deliberately including highly-related animals (*i.e.* full siblings) in the training and validation sets yielded little advantage over random assignment. This likely reflects the typical family structure of commercial salmon breeding populations, which consist of large full sibling families (thousands of fish per family) that result in close relationships between selection candidates and test individuals. However, the benefit of using genomic prediction over pedigree-based prediction was largest under the "Non-sibling" scenario, in which training and validation sets were established such that no full-siblings were included (*i.e.* the sets were less related than would be expected by chance, Figure 6-2). Prediction across populations or year groups (for which genetic relationships are more distant) was substantially less effective, with relatively low prediction accuracies (Figure 6-3). This may reflect, in part, inadequate sample size of the populations, or possibly differences in the experimental procedures between the two studies. However, our findings imply that either the GBLUP analyses did not efficiently capture short range LD between SNPs and QTL for resistance to sea lice, and/or that the QTL were population-specific. Therefore, in commercial salmon breeding schemes, regular phenotype data collection on animals that are closely-related to the selection candidates, combined with medium- or low-density (and cost) SNP panel genotyping appears to be the most effective means of using genomic prediction for resistance to lice. This strategy is supported by results from previous simulation studies (*e.g.* [28]).

Using data collected from a challenge trial performed in a sea cage environment, Ø degård et al. [9] also showed that genomic prediction of breeding values for lice resistance was more accurate compared to pedigree-based prediction. As in our study, the observed improvements depended partly on SNP density with ~ 32 (1 K SNPs) and 51 % (220 K SNPs) higher reliabilities than those obtained from predictions based on

pedigree records alone [9]. Interestingly, increasing SNP density above a threshold of around 5 K SNPs had little impact on accuracy of prediction in both studies (Figure 6-2, [9]). This may reflect the relatively close relationships between the training and validation sets, since higher SNP density did slightly improve the accuracy of cross-population predictions, as shown in our study, up to ~ 33 K SNPs (the highest density tested) (Figure 6-3). However, it seems unlikely that linkage alone is underpinning the predictions, since predictions with low SNP densities (< 1 K) and predictions based on an IBD (Identity-by-descent) genomic relationship matrix were less accurate [9]. Therefore, short or long range LD between SNP and QTL alleles may be an important component of prediction. Obviously, such LD can be captured by a relatively sparse SNP set, a finding that may be related to the relatively close relationships between training and validation sets, recent population admixture [9], or slower decay of LD due to the lack of male recombination in male salmon across much of the genome [31, 32].

A difference between simulation studies and those performed on experimental data is often observed in genomic prediction studies. Previous simulation studies indicated that the accuracy of breeding value prediction can reach values of 0.8 to 1.0 if the reference population size is sufficiently large (*e.g.* more than 100,000) [2, 33]. However, in practice, due to financial and practical limitations, research programs that use 'experimental' data usually involve the analysis of relatively small reference populations [9, 21, 34]. It is likely that if we had used larger population sizes, higher accuracies of prediction would have been obtained, particularly for predictions across the two distantly-related populations (subject to sufficient SNP density). As such, cost-effective means of generating high-density SNP data remain a relevant goal, and genotype imputation is likely to be increasingly important, particularly now that the majority of the Atlantic salmon reference genome has been assembled and ordered onto chromosomes (Genbank assembly accession GCA_000233375.4, [35]). Genotyping-by-sequencing may be crucial for reaching such high SNP density at moderate cost and its potential for genomic prediction in livestock has already been reported [36]. With a high SNP density across large sample sizes, one may expect to capture LD between SNPs and

QTL, and co-segregation of chromosome segments among related individuals, although the resolution of mapping causative variants may be hampered by the strong relationship structure in the population. Within populations/year groups, the requirement in terms of SNP density for accurate prediction is clearly lower and as few as 1 to 5 K informative SNPs are sufficient. However, while this points to the potential utility of cheaper and lower density genotyping platforms in aquaculture breeding, it is important to keep in mind that SNP informativeness can vary between populations.

6.6 Chapter Conclusions

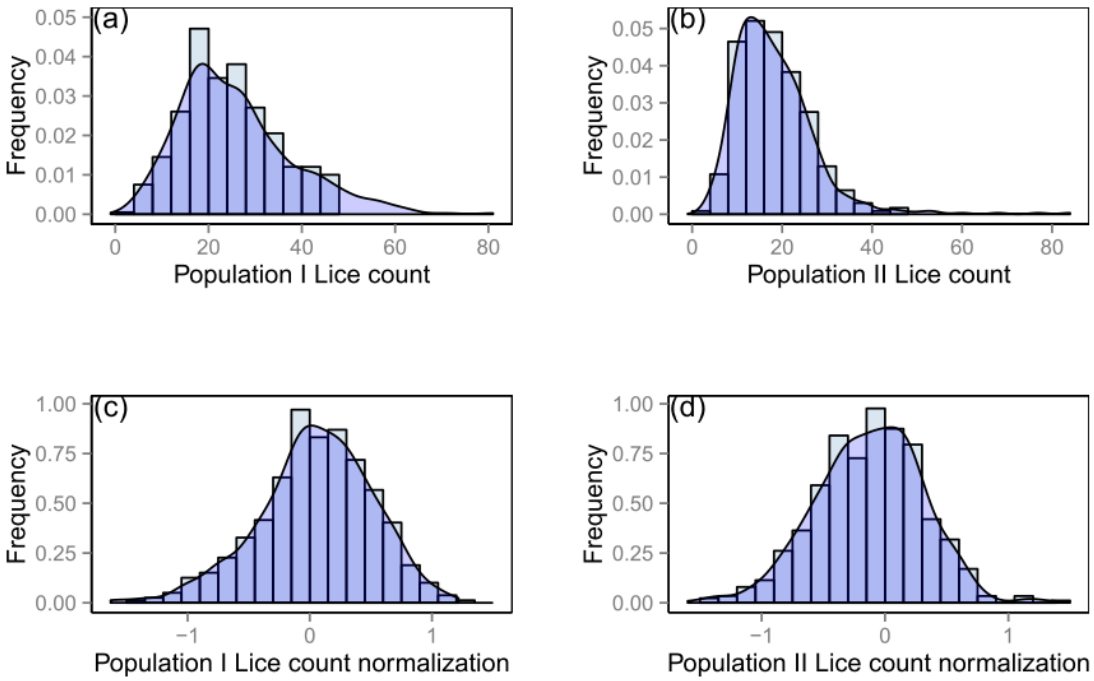
Genomic prediction is an effective method for predicting breeding values for host resistance to sea lice in Atlantic salmon populations from a commercial breeding programme. The GWAS results suggested that lice resistance is a polygenic trait. Cross-validation tests of genomic prediction highlighted the substantial improvements in prediction accuracy compared to that of pedigree-based prediction. The accuracy of GBLUP was highest when training and validation sets were closely related but the relative advantage over pedigree-based prediction within a population was largest when relationships were more distant. Relatively low SNP densities (from 1 to 5 K SNPs) were sufficient for accuracy to reach the asymptote under most of the scenarios tested. Prediction accuracy is generally much lower across distantly-related populations, although a trend was evident that increased marker density was advantageous in such situations. Therefore, larger population sample sizes and high-density SNP genotypes are probably necessary to improve across-population prediction. Given the economic importance of resistance to sea lice, and the efficacy of genomic prediction, it is likely that selective breeding for this trait using genomic data will become an important component of sea lice control.

6.7 Conclusions

Heritability of lice resistance traits was 0.2 to 0.3 in two separated populations, estimating by either genomic and pedigree relationship matrices. Genomic prediction showed that approximately 5 to 10 K SNPs can achieve the asymptote of accuracy in breeding value prediction in the within population containing closely related individuals. The accuracy was clearly improved by increased the markers in across two-population test, whilst the accuracy was apparently lower in comparison with within population test. Overall, the accuracy predicted by PBLUP was inferior to GBLUP in general. The Manhattan plots reflected the fact that lice resistance traits are polygenic, and may be population-specific. To conclude, genomic selection is likely to form an important component in sea lice control in salmon breeding.

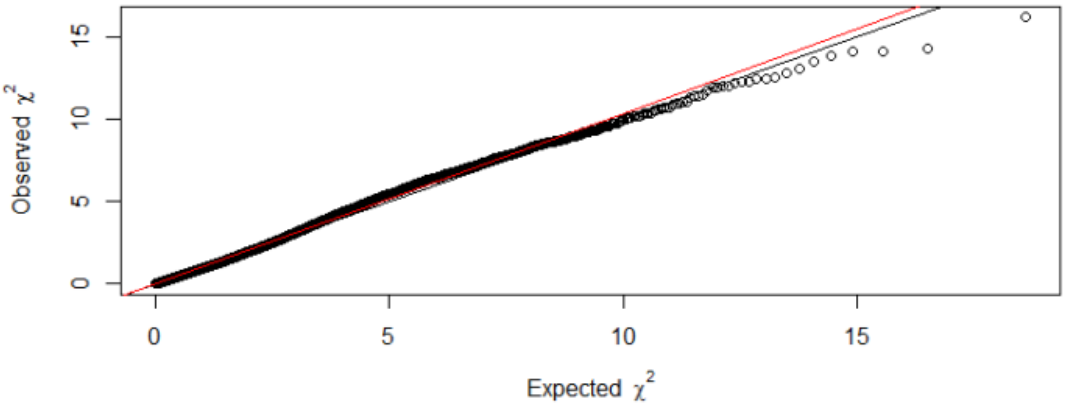
6.8 Additional Files

Additional file 1 Figure S1. Distributions of data for lice counts and after data normalization. Panels (a) and (c) represent results for population I, and panel (b) and (d) represent results for population II.

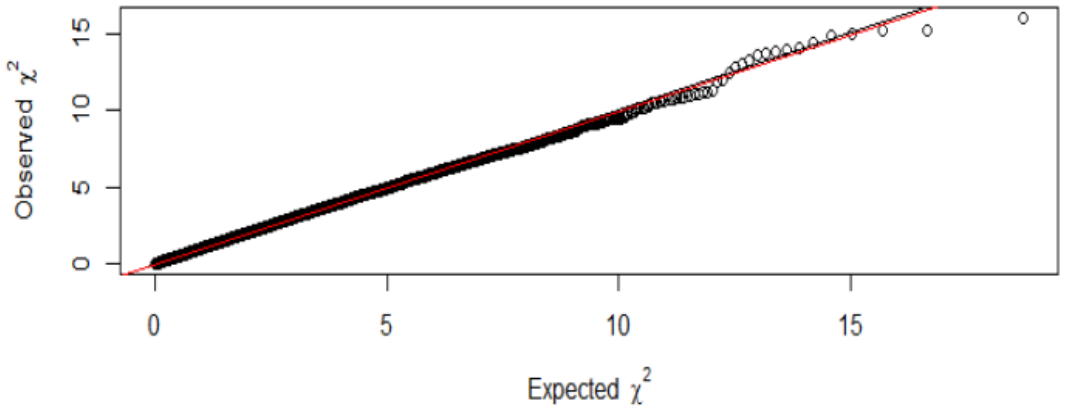


Additional file 2 Figure S2. Quantile-quantile (Q-Q) plot for the GWAS analysis. Three Q-Q plots are given in the file including population I (A), population II (B) and populations I and II combined (C).

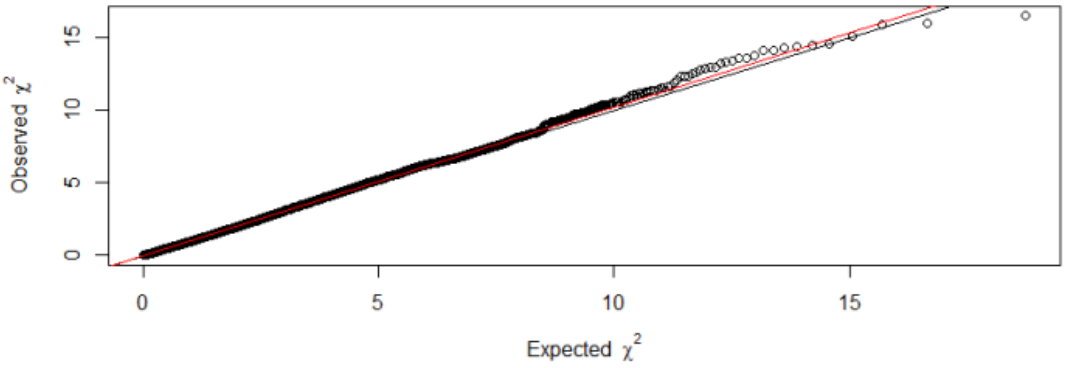
(A) Population I



(B) Population II



(C) Populations I and II combined



6.9 Acknowledgements

The authors gratefully thank the Edinburgh Genomics facility for SNP array genotyping, and Bill Roy and Matt Tinsley at the Marine Environmental Research Laboratory for assistance with trait data collection. This project received funding from the Technology Strategy Board (TP 5771-40299), Innovate UK (45266-329178), and Biotechnology and Biological Sciences Research Council (BBSRC) grants (BB/H022007/1, BB/J004235/1 and BB/J004324/1). JBT and JEB were partly supported by the MASTS pooling initiative (The Marine Alliance for Science and Technology for Scotland) for the completion of this study. MASTS is funded by the Scottish Funding Council (grant reference HR09011) and contributing institutions. HYT is supported by funding from the Ministry of Education, Taiwan.

6.10 References

1. Habier D, Fernando RL, Garrick DJ. Genomic BLUP decoded: A look into the black box of genomic prediction. Genetics. 2013;194:597–607.

2. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.

3. Wilkinson S, Wiener P, Archibald AL, Law A, Schnabel RD, McKay SD, et al. Evaluation of approaches for identifying population informative markers from high density SNP chips. BMC Genet. 2011;12:45.

4. Riggio V, Matika O, Pong-Wong R, Stear MJ, Bishop SC. Genome-wide association and regional heritability mapping to identify loci underlying variation in nematode resistance and body weight in Scottish Blackface lambs. Heredity (Edinb). 2013;110:420–9.

5. Bermingham ML, Bishop SC, Woolliams JA, Pong-Wong R, Allen AR, McBride SH, et al. Genome-wide association study identifies novel loci associated with resistance to bovine tuberculosis. Heredity (Edinb). 2014;112:543–51.

6. Gutierrez AP, Yáñez JM, Fukui S, Swift B, Davidson WS. Genome-wide association study (GWAS) for growth rate and age at sexual maturation in Atlantic salmon (*Salmo salar*). PLoS One. 2015;10:e0119730.

7. Yáñez JM, Newman S, Houston RD. Genomics in aquaculture to better understand species biology and accelerate genetic progress. Front Genet. 2015;6:128.

8. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010;185:1021–31.

9. Odegård J, Moen T, Santi N, Korsvoll SA, Kjøglum S, Meuwissen THE. Genomic

prediction in an admixed population of Atlantic salmon (*Salmo salar*). Front Genet. 2014;5:402.

10. Frazer LN, Morton A, Krkosek M. Critical thresholds in sea lice epidemics: evidence, sensitivity and subcritical estimation. Proc Biol Sci. 2012;279:1950–8.

11. Gharbi K, Matthews L, Bron J, Roberts R, Tinch A, Stear MJ. The control of sea lice in Atlantic salmon by selective breeding. J R Soc Interface. 2015;12:0574.

12. Kolstad K, Heuch PA, Gjerde B, Gjedrem T, Salte R. Genetic variation in resistance of Atlantic salmon (*Salmo salar*) to the salmon louse *Lepeophtheirus salmonis*. Aquaculture. 2005;247:145–51.

13. Gjerde B, Ø degård J, Thorland I. Estimates of genetic variation in the susceptibility of Atlantic salmon (*Salmo salar*) to the salmon louse *Lepeophtheirus salmonis*. Aquaculture. 2011;314:66–72.

14. Houston RD, Bishop SC, Guy DR, Tinch AE, Taggart JB, Bron JE, et al. Genome wide association analysis for resistance to sea lice in Atlantic salmon: application of a dense SNP array. In Proceedings of the 10th World Congress of Genetics Applied to Livestock Production: 17-22 August 2014; Vancouver. 2014.

15. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12:186.

16. Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G, Hickey JM. Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. Genetics. 2013;193:347–65.

17. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177:2389–97.

18. Wientjes YCJ, Veerkamp RF, Calus MPL. The effect of linkage disequilibrium and

family relationships on the reliability of genomic prediction. Genetics. 2013;193:621–31.

19. Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM, et al. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci. 2014;54:1476-88.

20. Vela-Avitúa S, Meuwissen THE, Luan T, Ø degård J. Accuracy of genomic selection for a sib-evaluated trait using identity-by-state and identity-by-descent relationships. Genet Sel Evol. 2015;47:9.

21. Tsai HY, Hamilton A, Tinch AE, Guy DR, Gharbi K, Stear MJ, et al. Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. BMC Genomics. 2015;16:969.

22. Houston RD, Taggart JB, Cézard T, Bekaert M, Lowe NR, Downing A, et al. Development and validation of a high density SNP genotyping array for Atlantic salmon (*Salmo salar*). BMC Genomics. 2014;15:90.

23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.

24. GilmourA.R, Gogel B J, Cullis BR, Thompson R. ASReml User Guide. 4th ed. Hemel Hempstead: VSN International Ltd; 2014.

25. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007;23:1294–6.

26. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.

27. Chen WM, Abecasis GR. Family-based association tests for genomewide association

scans. Am J Hum Genet. 2007;81:913–26.

28. Sonesson AK, Meuwissen THE. Testing strategies for genomic selection in aquaculture breeding programs. Genet Sel Evol. 2009;41:37.

29. Gjerde B, Pante MJR, Baeverfjord G. Genetic variation for a vertebral deformity in Atlantic salmon (S*almo salar*). Aquaculture. 2005;244:77–87.

30. Nirea KG, Sonesson AK, Woolliams JA, Meuwissen THE. Strategies for implementing genomic selection in family-based aquaculture breeding schemes: double haploid sib test populations. Genet Sel Evol. 2012;44:30.

31. Lien S, Gidskehaug L, Moen T, Hayes BJ, Berg PR, Davidson WS, et al. A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. BMC Genomics. 2011;12:615.

32. Gonen S, Lowe NR, Cezard T, Gharbi K, Bishop SC, Houston RD. Linkage maps of the Atlantic salmon (*Salmo salar*) genome derived from RAD sequencing. BMC Genomics. 2014;15:166.

33. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One. 2008;3:e3395.

34. Calus MPL. Genomic breeding value prediction: methods and procedures. Animal. 2010;4:157-64.

35. Davidson WS, Koop BF, Jones SJM, Iturra P, Vidal R, Maass A, et al. Sequencing the genome of the Atlantic salmon (*Salmo salar*). Genome Biol. 2010;11:403.

36. Gorjanc G, Cleveland MA, Houston RD, Hickey JM. Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet Sel Evol. 2015;47:12.

Chapter 7
General Discussion

7.1 Overview of Outcomes

In this thesis, I generated and applied genomic tools to study the genetic basis of economically important traits in farmed Atlantic salmon populations. The major outcomes in each chapter were specifically listed below.

In **chapter 2**, I constructed a high density linkage map using SNPs from a recently developed high density SNP array [1], and integrated the results with the most recent salmon reference genome assembly (Genbank Accession GCA_000233375.4, [2]). 96,396 SNPs were successfully assigned to 29 linkage groups corresponding to the karyotype of European Atlantic salmon. The number of SNPs mapped on 29 LGs is highly correlated with previous salmon linkage map reported by Lien et al. [3] (r = 0.94) and Gonen et al. [4] (r = 0.87). The physical length of LG showed high correlation with the number of assigned SNP in every chromosome (r = 0.95). Approximately 6.5 % of the unassigned reference genome contigs were mapped to existed chromosomes by referring to the result of linkage analysis. The female map was around 1.5-fold longer than the male map, covering 7,153 and 4,769 cM respectively. Comparison of male and female recombination rates reflected the significant difference between the two sexes, with male showing about 2 times higher than female in putative sub-telomeric regions.

In **chapter 3**, I performed a two-stage QTL mapping study to detect the chromosomes / QTLs associated with growth performance and quality traits in a large commercial adult salmon population (approximately 3 years post-hatching). The heritabilities of several fillet-related traits were significant and moderate to high (*e.g.* harvest and meat trait), ranging from 0.52 to 0.53. Chromosomes 13, 18, 19, and 20 were identified as harbouring QTLs affecting growth-related traits with genome-wide significance ($p < 0.05$) in sire-based analysis. Secondly, chromosomes 13, 18 and 20 were verified in the dam-based analysis, and explained about 6 % to 7 % of the within-family genetic variation in the traits of interest. These QTL may harbor genes affecting growth or quality traits, and may be useful for selective breeding applications.

In **chapter 4**, I performed a genome-wide association analysis to assess the population-wide association between individual SNPs on the SNP array, and juvenile growth traits in a commercial salmon population. 111,908 segregating SNPs passed through the quality control step were retained in the analysis. The heritability estimated by the model fitting the genomic relationship matrix was approximately 0.6 for body weight and length, while slightly lower using pedigree relationship matrix (~ 0.5). Results suggested that the genetic architecture associated with body weight and length of juvenile Atlantic salmon may be polygenic, with only one SNP surpassing the chromosome-wide significance in body length ($p < 0.05$). SNPs showing an arbitrary threshold of significance ($p < 0.005$, ~ top 0.5 % of markers) were selected for putative gene identification. Twenty candidate genes associated with growth traits were consequently identified. Genomic prediction was applied to predict the breeding values of traits in selection candidates, and the genomic BLUP approach was compared to the pedigree-based BLUP approach. The prediction results showed that the accuracy of GBLUP was notably higher than PBLUP, numbering 0.7 and 0.58 respectively. Different SNP marker densities were evaluated and an asymptote was seen at the density of 5 K in both traits, implying that 5,000 SNPs was sufficient to capture the genetic variation of phenotypes in this breeding program population.

In **chapter 5**, I selected sixteen candidate SNPs based on chapter 3 and 4, and to verify the association between the selected SNPs and performance traits in a separate population of farmed salmon. The population used was a subset of the QTL mapping population as described in chapter 3. The heritabilities of fillet-related traits were moderate to high, ranging from 0.52 to 0.53, which are identical with the results in chapter 3. Eight candidate SNPs were successfully genotyped with 1,152 fish across 198 families, and two of eight SNPs were significantly associated with several growth traits (*e.g.* fillet weight) in the two separate populations. SNP AX88270804 (*MEP1A* gene) was mapped to chromosome 16, explaining a small percentage of the overall variation in fillet traits (~ 1 %), waste traits (2 % to 3 %) and fat percentage (4 %). SNP

AX88141678 (*PCNT* gene) was mapped to chromosome 5, explaining nearly 1 % of the additive genetic variation in the weight-related traits. These two SNPs were likely to affect growth traits directly or may cause genetic variations due to population-wide linkage disequilibrium, reflecting the possibility of applying them as the candidates for marker-assisted selection in the breeding program.

In **chapter 6**, I used genomic prediction to estimate the breeding values of traits against sea lice in individuals in two separate populations. Estimation of genetic parameters suggested the trait of lice resistance was heritable (h^2 ~ 0.22 to 0.33). The prediction accuracy using genomic-based prediction was higher than pedigree-based method in four scenarios in both populations. The highest accuracy (accuracy ~ 0.6) was achieved when the training and validation sets contained closely related animals (*e.g.* sibling test), while the greatest advantage of genomic prediction over pedigree-based prediction was observed in more distantly animals (*e.g.* non-sibling test). The asymptote of accuracy was achieved at the marker density of 5 K to 10 K within population, which was similar to the results in chapter 4. Prediction accuracy across the two separate populations was lower, but improved with higher marker density. Finally, the GWAS showed that the lice resistance in the two populations were likely polygenic with no SNP surpassing the genome-wide significance threshold in a GWAS.

7.2 Discussion

7.2.1 Linkage Map

In the past 12 years, the number of genetic markers available for Atlantic salmon has increased rapidly (Table 7-1) (*e.g.* [1, 5]). The recently published reference genome assembly (GCA_000233375.4, [2]) enables us to look into the genetic architecture of Atlantic salmon with greater detail. Building a high density genetic map for species of interest has been challenging until recently. Until the past decade, the approaches and tools implemented to construct linkage maps typically used small number of markers

and limited sample size, mainly due to the limitation of marker availability and computation (reviewed by [6]). Fortunately, large computer clusters and efficient algorithms have been gradually developed [7], allowing us to map, order and calculate the genetic distance (cM) between markers relatively quickly and accurately. For Atlantic salmon, the relatively recent WGD event results in high levels of sequence similarity between paralogous regions, making analyses of sequence data more challenging. In addition, there is a large recombination rate difference between the two sexes of salmon [2, 8], which presents a challenge to constructing a high density linkage map for this species. As such, I removed all possible duplicated SNPs from our dataset (reducing the number from 132 K to 100 K), and integrated the genetic map with recent salmon reference genome assembly, resulting in a high density map comprising ~ 100 K SNPs. Benefitting from the genetic map, as 22 % of recent salmon reference genome assembly was not yet assigned to chromosome [2], our linkage map was used to map these unassigned contigs to possible chromosomes, which accounts for approximately 1 % of total reference genome assembly [9].

Table 7-1. Comparison of Atlantic salmon linkage map in previous and current studies.

Study	Mapped molecular markers	**Sex recombination rate difference (M:F)**	**Number of linkage group identified**
Gilbey et al. [10]	50 microsatellites	1:3.9	15
Moen et al. [11]	54 microsatellites 473 AFLP markers	1:8.3	25
Lien et al. [3]	5,650 SNPs	1:1.4	29
Gonen et al. [4]	6,458 SNPs	1:1.5	29
Current Study	96,396 SNPs	1:1.5	29

The unique recombination pattern along the chromosomes in salmonids and large difference in recombination rate between male and female salmon have been much discussed in the literature [3, 4, 12, 13]. To our knowledge, the mechanisms underpinning the recombination difference is still not yet clear, but may be related to the ancestral WGD. The recombination rate differences can affect the level of genetic variance, as higher recombination rate is likely able to increase the efficiency of breeding programs to turn genetic variation into genetic gain (response to selection) [14, 15]. To characterize the recombination rate patterns between two sexes across the entire genome, a high resolution genetic map is of both scientific and applied interest, such as the evolutionary study and breeding in salmonids. In chapter 2, I studied the recombination rate across the entire genome, and compared the difference between two sexes in corresponding chromosome regions. The results indicated that sub-telomeric regions of the 29 pairs of chromosomes showed on average ten times higher recombination rate than the rest regions of the genome. However, this drastic variation was observed for males, and recombination rate was relatively stable in females across the 29 pairs of chromosomes. In addition, earlier studies reported that the overall map length in female was longer than male (Table 7-1), and our result was comparable to Lien et al. [3] and Gonen et al. [4].

To conclude, in chapter 2, I built up a high resolution genetic map, this step facilitated to address the QTLs / loci associated with growth-related and disease resistance traits that I aimed to investigate in the later chapters (chapter 3, 4, 5 and 6).

7.2.2 Genetic architecture of growth traits in farmed salmon

In chapter 3, 4, and 5, I studied the molecular genetic basis of growth-related traits in Atlantic salmon. Abundant studies have been reported that the growth-related traits are moderate to high heritable in different Atlantic salmon populations (*e.g.* reviewed by [16] and *e.g.* [17–20]). Our results using both pedigree and genomic-based method also indicated that these traits are heritable with around 0.5 to 0.6 at heritability estimation at

juvenile and adult stage of fish (Table 7-2), but these traits are likely to have a polygenic genetic architecture.

Table 7-2. Comparison of heritability of growth traits in two current studies.

Study	Current study Chapter 3*	Current study Chapter 6
Body weight (Juvenile)	-	0.61 (2007 year group) 0.50 (2010 year group)
Body weight (Adult)	0.52 (1999 year group)	-
Body length (Juvenile)	-	0.61 (2007 year group) 0.51 (2010 year group)
Body length (Adult)	-	-

*The heritability was estimated using pedigree records, while the genomic-based method was used in chapter 6.

7.2.2.1 Genetic Fillet-related Traits

Chromosomes 13, 18, 19 and 20 were shown to harbor genome-wide significant QTLs affecting several growth-related traits in the sire-based analysis. Further, the chromosomes 13, 18 and 20 were confirmed in dam-based analysis in the 1999 year group. The reason to apply two-stage QTL mapping in salmon is due to the large recombination rate difference between two sexes of salmon [8], as it has been discussed in previous section. The advantage of initially using sire-based QTL analysis is that the characterization of lower recombination rate in male salmon can offer higher power to detect targeted QTLs with fewer SNPs per LG (*e.g.* ~ 2 markers per LG), then we are able to fine map the position of QTLs by dam-based QTL analysis with higher resolution of marker in potential LGs. This strategy enables us to reduce the cost of QTL mapping in Atlantic salmon, which has been applied in several relevant works (reviewed by [21]).

Compared the result with chapter 3 and 4, results in chapter 4 indicated that QTLs associated with growth traits were not exactly consistent with the separate population (2007 year group) and previous relevant studies [17-19], with only some of QTLs were commonly found in these studies (A review table was made in Table 3-5 in chapter 3). The GWA analysis showed those individual SNPs with the lowest P-value for growth traits were located on chromosomes 5, 16, 17, 20, 21, and 28, while no SNP surpassed the genome-wide significance threshold ($p < 0.05$). To verify the results from chapter 3 and 4, I consequently selected SNP markers located on significant QTLs (chapter 3) and SNPs with the lowest P-value (chapter 4) to examine the previous findings. However, there were only 2 of 8 SNPs successfully found to be significantly associated with growth-related traits in both 1999 and 2007 year group (chapter 5). Comparing the studies (chapter 3, 4 and 5) with previous literature, it is worthwhile to note that the age (juvenile or adult), family structure (half or full-sibling), and environmental factors (sea cage, indoor tank, temperature, and feeding) of fish reared in these investigations were highly variable [17-19, 22]. Therefore, the aforementioned factors may influence the growth performance of fish. Nevertheless, all the published evidence and the results from the current studies point to that the fact that growth traits are likely polygenic, and population specific. The identified QTLs are likely pleiotropic, as the QTLs had affects across several component weight-related traits even the harvest weight (body weight) was set as the covariate to distinguish the QTLs linked with overall body weight, rather than only specifically associated with other component weight traits (*e.g.* head and gut weight) in our populations [17, 23].

7.2.2.2 Flesh Colour and Fat Content

The heritabilities for non-visual traits such as fat content and fillet colour were relatively low in comparison with weight-related traits, which were about 0.14 to 0.18 (Table 7-3). A relevant study indicated that the heritabilities of astaxanthin and canthaxanthin were highly correlated with flesh colour [24]. However, the major components of fat content, including crude lipid, percentage n−3 long-chain polyunsaturated fatty acids (n−3LC-

PUFA) and absolute n−3LC-PUFA were significantly heritable (h^2 = 0.69, 0.77, and 0.34 respectively) [25]. Given that the major components of fat content are heritable, the heritability of fat content might have been expected to be higher than current result. The investigation in terms of the reason that causes the heritability difference between fat content and its components is of interest, and may be related to the methods of measuring fatness in salmon.

Table 7-3. Comparison of heritabilities (SE) of fat content, fillet colour and their components in current and previous studies.

	Quinton et al. [24]	Leaver et al. [25]	Current Study
Fat Content (%)	0.19 (0.08)	-	0.18 (0.03)
Crude Lipid (g/100 g)	-	0.69 (0.14)	-
Percentage n−3LC-PUFA	-	0.77 (0.14)	-
Absolute n−3LC-PUFA (mg/100 g)	-	0.34 (0.11)	-
Flesh Colour (20-34)	0.13 (0.07)	-	0.14 (0.03)
Astaxanthin (%)	0.09 (0.06)	-	-
Canthaxanthin (%)	0.11 (0.06)	-	-

There were no significant QTLs associated with fat content detected in chapter 3, but relevant studies reported that the putative QTLs were harbored with chromosomes 9 and 10 [22]. A high genetic correlation between weight-related traits and fat content was observed (r = 0.82 to 0.84) in chapter 3. Interestingly, in chapter 5, a SNP marker (aligned on putative gene *PCNT*) affecting fat content was identified, explaining about 4 % of the genetic variation. The role of *PCNT* is yet unclear in Atlantic salmon or even salmonids, but a research suggested that *PCNT* is involved in normal growth and lipid regulation in humans [26], raising the possibility that the sequence variants surrounding the gene *PCNT* may contribute to phenotypic variation in the growth and fat content-related traits.

Chromosomes 3 and 26 were previously detected to harbor QTLs associated with fillet colour trait [17]. We also identified that chromosome 17 showed chromosome-wide significance with fillet colour in farmed Atlantic salmon in chapter 3. Notably, a single locus *SCAR* marker was found to link with muscle flesh colour significantly ($p < 0.0001$) in Coho salmon [27].

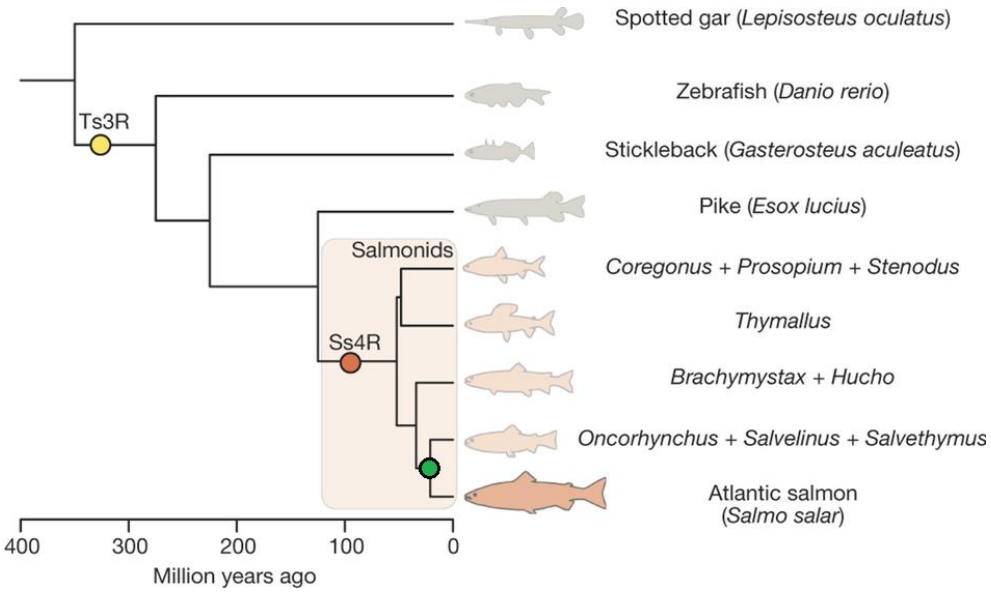


Figure 7-1. Phylogenetic relationship of Atlantic salmon (Salmo Salar) and Coho salmon (Oncorhynchus kisutch). The green point marks the divergent position of Atlantic salmon and Coho salmon. This illustration was adapted from Lien et al. [2].

Considering the phylogenetic relationship of Atlantic salmon and Coho salmon (Figure 7-1) [2], and also the low heritability of flesh colour, the QTL harboring the 'similar' *SCAR* marker in Atlantic salmon is of scientific and commercial interest.

7.2.2 Genomic Prediction of Growth Traits

Genomic prediction has been applied to predict breeding values associated with host resistance against disease in aquaculture species (*e.g.* [28–30]), but rarely for growth-related traits [31]. As the polygenic architecture of growth traits in salmon, the GWA

analysis reflected the genetic variations affecting major growth traits (*e.g.* body weight) were explained by many SNPs of small effect [31, 32]. As such, in order to improve the performance of fillet-related traits (*e.g.* muscle weight) in salmon, the application of genomic prediction may be an effective approach to achieve the goal. In chapter 4, the utility of SNP dataset for genomic prediction of the body weight and body length of juvenile fish was subsequently implemented to estimate the breeding values of traits in selection candidates in the family-based breeding program. Further, the appropriate marker density to use for genomic prediction was evaluated. The outcomes reflected that the genomic prediction is an applicable approach for improving growth-related traits, as few as 5,000 SNPs gives the highest prediction accuracy (~ 0.7) in weight and length traits of individual within population, implying that only relatively sparse marker platform can achieve the asymptote of accuracy prediction using genomic information with BLUP model (GBLUP). Odegard et al. [30] also reported that the prediction accuracy of fillet colour fitting genomic data to BLUP model was improved up to 4.7 % when comparing with pedigree-based method in within population test. As such, for those traits with polygenic architecture or relatively low heritability (*e.g.* fat content and flesh colour), genomic prediction is likely a useful tool for capturing the genetic variations of phenotypes based on well-developed SNP panel.

Nevertheless, the explanation in terms of why relatively lower marker density can achieve the equivalent prediction accuracy as high density SNP platform is still not that clear. But we have to note that a relatively small and closely relatedness population were utilized to predict the estimated breeding values of candidates in our study, thus higher marker density (> 5 K) may be essential in the future tests, especially for those investigations underlying less closely individuals, larger population size and targeted traits with low heritability. On the other hand, several studies also pointed out that increasing marker density beyond certain number (*e.g.* 10 K SNPs) did not obviously improve the accuracy in genomic prediction in simulated maize [33] and livestock population [34, 35], the similar outcomes were also observed in real salmon population [30].

7.2.3 Genetic architecture of host resistance to sea louse

The heritability of sea lice resistance (*Lepeophtheirus salmonis*) was moderate, which were around 0.2 to 0.3 using pedigree and genetic-based method respectively. Before calculating the heritability, I applied a formula (model (1) in chapter 6) to normalize the skewness observed in the data distribution of lice count in our population (corresponding result is given in additional file 1 in chapter 6). Gjerde et al. [36] estimated the heritability based on lice count and lice density respectively, pointing out the way to present phenotype to the animal model (**y** = **Xb** + **Zu** + **e**) to assess the heritability would result in different outcomes (Table 7-4). Given that the lice count showed skewness in the data distribution, the lice density is a better way to give accurate heritability, which was thus implemented in chapter 6. It is also worthwhile to note that the challenge trial in Odegard et al. [30] was conducted based on the sea-cage environment, which may potentially involve higher environmental effects while estimating the narrow sense heritability, as such, the heritability was relatively lower than the studies applying tank environment, such as current study (Table 7-4).

Table 7-4. Comparison of heritability of lice resistance using genomic (G-matrix) and pedigree relationship matrix (A-matrix) in the animal model in previous and current studies.

Study	Group	Heritability	Sample Size
Gjerde et al. [36]	Lice count	0.33 (A-matrix)	2206
	Lice density	0.26 (A-matrix)	
Gharbi et al. [37]		0.3 (G-matrix)	1479
Odegard et al. [30][$]	Test 1	0.13 to 0.14	1444
	Test 2	(G-matrix)	519
Correa et al. **[38]***		0.12	2628
		(G-matrix)	
	2007	0.33 (G-matrix)	621
Current Study	year group	0.27 (A-matrix)	
	2010	0.22 (G-matrix)	874
	year group	0.27 (A-matrix)	

[$]The lice challenge trial was conducted based on the sea-cage environment, and others were tank environment. [*]The sea lice species used for challenge trial was *Caligus rogercresseyi* which is a species identified in Chile, and others were *Lepeophtheirus salmonis*.

In chapter 6, the heritability of lice resistance in the 2007 year group was slightly higher than 2010 year group. The possible reason to explain the difference is that the 2007 year group was reared in family-specific tank at the fry stage, yet the 2010 year group was mixed together in single big tank, therefore, this disparity is likely due to the confounding of genetic variations with different common environment effects. Interestingly, Correa et al. [38] showed much lower heritability of *Caligus rogercresseyi* resistance in comparison with studies challenged with *Lepeophtheirus salmonis*, which implies that Atlantic salmon has different levels of resistance ability against two sea lice species, while further investigation may be required as different SNP platforms were implemented in these studies (Table 7-4).

The GWAS reflected that lice resistance (*Lepeophtheirus salmonis*) was a polygenic trait in both 2007 and 2010 year group, with no major QTL segregating. Similar results have been reported for host resistance against different sea louse species (*Caligus rogercresseyi*) in Chile, implying that resistance to both species is polygenic [28, 38]. Therefore, I employed genomic prediction as the tool to estimate the breeding values of lice resistance traits in selection candidates, and used different degrees of relatedness between training and validation sets to test the impact of genetic relationship on prediction accuracy. Overall, results showed that prediction accuracy was significantly altered by the genetic and pedigree relationship between training and validation sets. When the relationship between training and validation sets was close (*e.g.* full-sibling), the accuracy was apparently higher than those with less related populations. Nevertheless, the advantage of genomic prediction was clearly reflected in the tests with less related animals between training and validation sets (*e.g.* non-sibling). The accuracy measured by the prediction model fitting genetic relationship matrix was significantly higher than fitting pedigree relationship matrix in the within population test, highlighting the utility of genomic prediction is an applicable approach, in particular for the polygenic traits with several QTLs explaining only small proportion of genetic variations. A simulated study suggested that the family-based selection method is the most effective way to perform selective breeding in aquaculture breeding scheme [39]. As two experimental populations were involved in the analysis, the effect of family structure toward the prediction was also observed. Compared with 2007 year group, several small families existed in the 2010 year group, leading to some families with limited fish number were not able to provide sufficient information of relatedness to training and validation sets respectively (*e.g.* sibling test). As such, in order to maximize the prediction accuracy based on family-based selection method, the number of representative individual from each family is critical, when collecting the experimental samples from field.

In chapter 6, I tested three different scenarios, which were carried out depending on ‘within population’ situation, showing the highest accuracy was achieved at around 5 K

to 10 K SNP density. The accuracy had a little improvement above a certain marker density, which was observed in previous study using experimental population (rather than simulated population), which the similar result was also seen in current study (Figure 7-2). The possible reason that causes this kind of phenomenon has been described in "Genomic prediction of growth traits" section above.

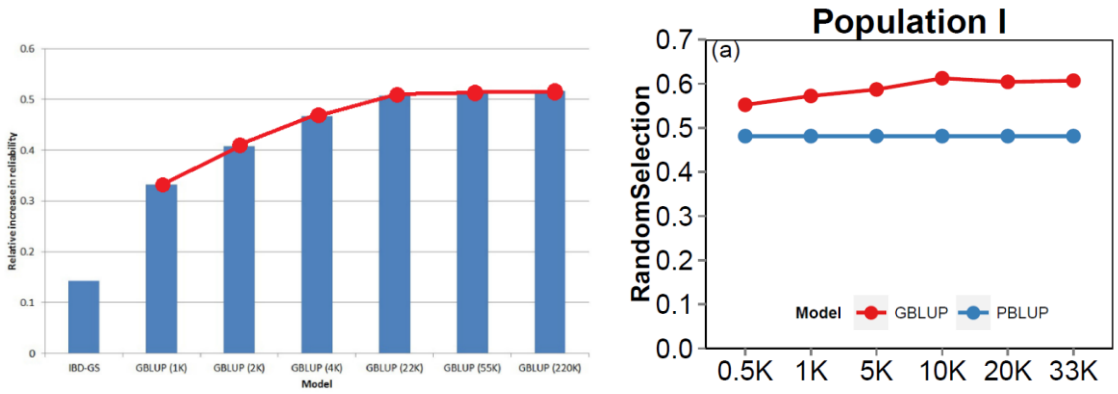


Figure 7-2. The tendency of accuracy from low to high marker density in Odegard et al. [30] (left) and current study (right). The left bar chart was adapted from Odegard et al. [30] and right line chart was adapted from chapter 6 in current study. The figure showed that there is only a little improvement in the accuracy when the marker density was over 22 K (left) and 10 K (right).

A similar tendency was also shown in the across populations test in the current study (Figure 6-3 in chapter 6), while the accuracy was generally low as the relatedness between two populations were more distant. This implies that the accuracy gained from medium marker density is comparable to high marker density, and the linkage disequilibrium (LD) is able to be efficiently captured by relatively sparse marker platform when the close related animals between training and validation sets were implemented in the genomic prediction.

7.2.4 Gene Annotation and Gene Associated with Growth Traits

In chapter 2, I successfully annotated 45.9 % of unique markers to putative genes, with 11 % of the genic SNPs were mapped to putative exonic regions (5,856 of 48,842 SNPs). In chapter 4 and 5, several markers associated with growth traits have been identified by means of GWAS. In order to address the potential function of these top markers and their predicted genes, I integrated the annotation result with the top markers that previously reported, which is summarized in Table 7-5.

The exonic SNP AX87963258 appeared to affect body weight and length, and was predicted to be a non-synonymous variant causing an amino acid substitution. Non-synonymous marker is a coding variant that may cause protein functional changes, and consequently result in phenotype alteration to individuals. Amino acid substitution within the same group (*e.g.* within hydrophobic group) are usually tolerable, while changing to other residues (*e.g.* hydrophobic group to polar or charged group) can be deleterious, leading to impact the protein function and structure.

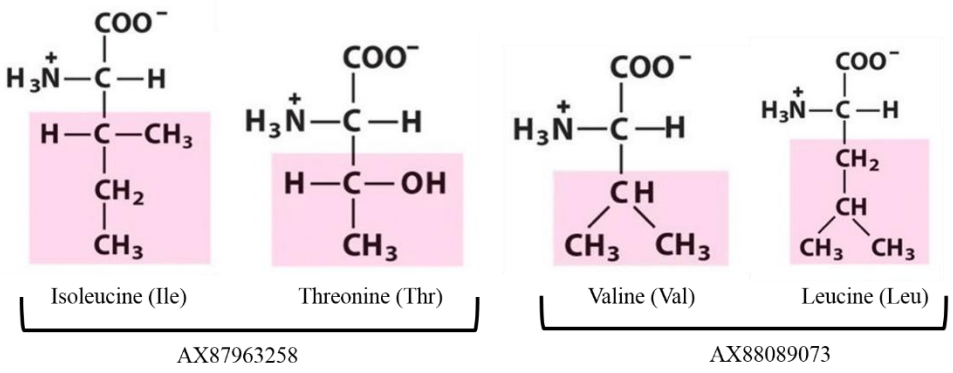


Figure 7-3. Structure of isoleucine, threonine, valine and leucine. The SNP AX87963258 (*RAI2*) caused amino acid alteration from hydrophobic residue (Ile) to polar uncharged residue (Thr). Other two amino acids alteration (Val and Leu) arisen by the SNP AX88089073 (*POMT1*) are belong to the same hydrophobic group.

The genotype A/G was characterized in the SNP AX87963258, therefore the predicted reading frame showed that the A/G genotype can have two different amino acids, including isoleucine (T**A**A, or represents as AUU in RNA form) and threonine (T**G**A, or represents as ACU in RNA form). Since it is known that the isoleucine is a hydrophobic amino acid while threonine is recognized as a polar uncharged amino acid, which reflects the alteration can result in protein function changes in corresponding gene *RAI2*. While the missense variant caused by another SNP AX88089073 were both in the same hydrophobic residue group, the substitution may be more tolerable than SNP AX87963258 (Figure 7-3).

The function of *RAI2* (AX87963258) and *POMT1* (AX88089073) have been described in the discussion section in chapter 4. Briefly, the *RAI2* is well known as the role in growth and differentiation in early life [40], and is involved in the regulation of bone formation and mineralization in Atlantic salmon [41]. *POMT1* can secret the *POMT* enzyme complex that is associated with muscular dystrophy in mammals [42]. Since both genes are linked with growth, the predicted missense variants identified from two top SNPs may be of interest for *in vitro* and even *in vivo* investigations whereby newly gene editing technology, such as CRISPR, could be applied to test causality of these variants [43].

Table 7-5. The gene annotation and effect prediction of top SNPs identified in chapters studied on growth-related traits (chapter 2, 3, 4 and 5).

Marker	P-value	Allele frequency		Additive effect (s.e.)	Dominance effect (s.e.)	PVE	Chromosome	Genomic Variant	Alternative Variant	Effect	Gene Annotation
		p	q								
Weight											
AX87888225	1.0E-04	0.94	0.06	7.00 (6.28)	23.83 (6.66)	0.06	28	C	G	Intron variant	*WAPL isoform x1*
AX88223695	1.2E-04	0.80	0.20	3.32 (2.76)	16.54 (3.02)	0.04	28	C	T	Intron variant	*Spata20*
AX87963258	1.4E-04	0.57	0.43	5.80 (1.47)	2.00 (2.04)	0.05	17	A	G	Missense variant: p.Ile218Thr	*RAI2*
Length											
AX87963258	1.7E-05	0.57	0.43	4.42 (0.99)	1.27 (1.37)	0.07	17	A	G	Missense variant: p.Ile218Thr	*RAI2*
AX87959512	9.1E-05	0.68	0.32	5.46 (1.48)	0.21 (1.55)	0.08	20	C	T	Intron variant	*POMT1*
AX88089073	1.6E-04	0.70	0.30	4.77 (1.62)	1.07 (1.65)	0.05	20	G	C	Missense variant: p.Val76Leu	*POMT1*

Bold: AX87963258 appears in both traits and surpasses the chromosome-wide significance level ($p < 0.05$).

WAPL isoform x1: Wings apart-like protein homolog isoform x1

Spata20: Spermatogenesis-associated protein 20

POMT1: Protein o-mannosyl-transferase 1 *RAI2*: Retinoic acid-induced protein

7.3 References

1. Houston RD, Taggart JB, Cézard T, Bekaert M, Lowe NR, Downing A, Talbot R, Bishop SC, Archibald AL, Bron JE, Penman DJ, Davassi A, Brew F, Tinch AE, Gharbi K, Hamilton A: Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). *BMC Genomics* 2014, 15:90.

2. Lien S, Koop BF, Sandve SR, Miller JR, Matthew P, Leong JS, Minkley DR, Zimin A, Grammes F, Grove H, Gjuvsland A, Walenz B, Hermansen RA, Schalburg K Von, Rondeau EB, Genova A Di, Samy JKA, Vik JO: The Atlantic salmon genome provides insights into rediploidization. *Nature* 2016, 533:200–205.

3. Lien S, Gidskehaug L, Moen T, Hayes BJ, Berg PR, Davidson WS, Omholt SW, Kent MP: A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genomics* 2011, 12.

4. Gonen S, Lowe NR, Cezard T, Gharbi K, Bishop SC, Houston RD: Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing. *BMC Genomics* 2014, 15:166.

5. Yáñez JM, Naswa S, López ME, Bassini L, Correa K, Gilbey J, Bernatchez L, Norris A, Neira R, Lhorente JP, Schnable PS, Newman S, Mileham A, Deeb N, Di Genova A, Maass A: Genome-wide single nucleotide polymorphism (SNP) discovery in Atlantic salmon (Salmo salar): validation in wild and farmed American and European populations. *Mol Ecol Resour* 2016, 16:1002–11.

6. Fierst JL: Using linkage maps to correct and scaffold de novo genome assemblies: Methods, challenges and computational tools. *Front Genet* 2015, 6.

7. Rastas P, Calboli FCF, Guo B, Shikano T, Merilä J: Construction of ultra-dense linkage maps with Lep-MAP2: stickleback F2 recombinant crosses as an example.

Genome Biol Evol 2016. In Press.

8. Hayes BJ, Gjuvsland A, Omholt S: Power of QTL mapping experiments in commercial Atlantic salmon populations, exploiting linkage and linkage disequilibrium and effect of limited recombination in males. *Heredity (Edinb)* 2006, 97:19–26.

9. Tsai HY, Robledo D, Lowe NR, Bekaert M, John B: Construction and annotation of a high density SNP linkage map of the Atlantic salmon (Salmo salar) genome. *Genes Genomes Genet* 2016, 6(July):2173–2179.

10. Gilbey J, Verspoor E, McLay A, Houlihan D: A microsatellite linkage map for Atlantic salmon (Salmo salar). *Anim Genet* 2004, 35:98–105.

11. Moen T, Hoyheim B, Munck H, Gomez-Raya L: A linkage map of Atlantic salmon (Salmo salar) reveals an uncommonly large difference in recombination rate between the sexes. *Anim Genet* 2004, 35:81–92.

12. Moen T, Hayes B, Baranski M, Berg PR, Kjøglum S, Koop BF, Davidson WS, Omholt SW, Lien S: A linkage map of the Atlantic salmon (Salmo salar) based on EST-derived SNP markers. *BMC Genomics* 2008, 9:223.

13. Volff J-N: Genome evolution and biodiversity in teleost fish. *Heredity (Edinb)* 2005, 94:280–294.

14. Mszros G, Gorjanc G, Jenko J, Woolliams JA, Hickey JM: Selection on Recombination Rate to Increase Genetic Gain. In *10th World Congr Genet Appl to Livostock Prod*. Vancouver, Canada; 2014:021.

15. Battagin M, Gorjanc G, Faux A-M, Johnston SE, Hickey JM: Effect of manipulating recombination rates on response to selection in livestock breeding programs. *Genet Sel Evol* 2016, 48:1–12.

16. Garcia de Leaniz C, Fleming I a, Einum S, Verspoor E, Jordan WC, Consuegra S, Aubin-Horth N, Lajus D, Letcher BH, Youngson a F, Webb JH, Vøllestad L a, Villanueva B, Ferguson a, Quinn TP: A critical review of adaptive genetic variation in Atlantic salmon: implications for conservation. *Biol Rev Camb Philos Soc* 2007, 82:173–211.

17. Baranski M, Moen T, Våge DI: Mapping of quantitative trait loci for flesh colour and growth traits in Atlantic salmon (Salmo salar). *Genet Sel Evol* 2010, 42:17.

18. Houston RD, Bishop SC, Hamilton A, Guy DR, Tinch AE, Taggart JB, Derayat A, McAndrew BJ, Haley CS: Detection of QTL affecting harvest traits in a commercial Atlantic salmon population. *Anim Genet* 2009, 40:753–5.

19. Gutierrez AP, Lubieniecki KP, Davidson EA, Lien S, Kent MP, Fukui S, Withler RE, Swift B, Davidson WS: Genetic mapping of quantitative trait loci (QTL) for body-weight in Atlantic salmon (Salmo salar) using a 6.5K SNP array. *Aquaculture* 2012, 358-359:61–70.

20. Powell J, White I, Guy D, Brotherstone S: Genetic parameters of production traits in Atlantic salmon (Salmo salar). *Aquaculture* 2008, 274:225–231.

21. Yáñez JM, Houston RD, Newman S: Genetics and genomics of disease resistance in salmonid species. *Front Genet* 2014, 5(November):415.

22. Sodeland M, Gaarder M, Moen T, Thomassen M, Kjøglum S, Kent M, Lien S: Genome-wide association testing reveals quantitative trait loci for fillet texture and fat content in Atlantic salmon. *Aquaculture* 2013, 408-409:169–174.

23. Tsai HY, Hamilton A, Guy DR, Tinch AE, Bishop SC, Houston RD: The genetic architecture of growth and fillet traits in farmed Atlantic salmon (Salmo salar). *BMC Genet* 2015, 16:51.

24. Quinton CD, McMillan I, Glebe BD: Development of an Atlantic salmon (Salmo salar) genetic improvement program: Genetic parameters of harvest body weight and carcass quality traits estimated with animal models. *Aquaculture* 2005, 247:211–217.

25. Leaver MJ, Taggart JB, Villeneuve L, Bron JE, Guy DR, Bishop SC, Houston RD, Matika O, Tocher DR: Heritability and mechanisms of n-3 long chain polyunsaturated fatty acid deposition in the flesh of Atlantic salmon. *Comp Biochem Physiol Part D Genomics Proteomics* 2011, 6:62–9.

26. Rauch A, Thiel CT, Schindler D, Wick U, Crow YJ, Ekici AB, van Essen AJ, Goecke TO, Al-Gazali L, Chrzanowska KH, Zweier C, Brunner HG, Becker K, Curry CJ, Dallapiccola B, Devriendt K, Dörfler A, Kinning E, Megarbane A, Meinecke P, Semple RK, Spranger S, Toutain A, Trembath RC, Voss E, Wilson L, Hennekam R, de Zegher F, Dörr H-G, Reis A: Mutations in the Pericentrin (PCNT) Gene Cause Primordial Dwarfism. *Sci* 2008, 319 (5864):816–819.

27. Araneda C, Neira R, Iturra P: Identification of a dominant SCAR marker associated with colour traits in Coho salmon (Oncorhynchus kisutch). *Aquaculture* 2005, 247:67–73.

28. Correa K, Lhorente JP, Bangera R, Yáñez JM: Evaluation of Genomic Selection for Sea Louse Resistance in Atlantic Salmon (Salmo salar L). In *Plant Anim Genome Conf*. San Diego, CA; 2016:P0461.

29. Fragomeni BO, Misztal I, Lourenço DAL, Vallejo RL, Palti Y: Weighted ssGBLUP Improves Genomic Selection Accuracy for Bacterial Cold Water Disease. In *Plant Anim Genome Conf*. San Diego, CA; 2016:P0472.

30. Odegård J, Moen T, Santi N, Korsvoll SA, Kjøglum S, Meuwissen THE: Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). *Front Genet* 2014, 5(November):402.

31. Tsai HY, Hamilton A, Tinch AE, Guy DR, Gharbi K, Stear MJ, Oswald M, Bishop SC, Houston RD: Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. *BMC Genomics* 2015, 16:969.

32. Gutierrez AP, Yáñez JM, Fukui S, Swift B, Davidson WS: Genome-Wide Association Study (GWAS) for Growth Rate and Age at Sexual Maturation in Atlantic Salmon (Salmo salar). *PLoS One* 2015, 10:e0119730.

33. Hickey JM, Crossa J, Babu R, de losCampos G: Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Sci* 2012, 52:654–663.

34. Vela-Avitúa S, Meuwissen THE, Luan T, Ø degård J: Accuracy of genomic selection for a sib-evaluated trait using identity-by-state and identity-by-descent relationships. *Genet Sel Evol* 2015, 47:9.

35. Gorjanc G, Cleveland M a, Houston RD, Hickey JM: Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genet Sel Evol* 2015, 47:12.

36. Gjerde B, Ø degård J, Thorland I: Estimates of genetic variation in the susceptibility of Atlantic salmon (Salmo salar) to the salmon louse Lepeophtheirus salmonis. *Aquaculture* 2011, 314:66–72.

37. Gharbi K, Matthews L, Bron J, Roberts R, Tinch A, Stear MJ: The control of sea lice in Atlantic salmon by selective breeding. *J R Soc Interface* 2015, 12:0574.

38. Correa K, Lhorente JP, Bassini L, López ME, Di Génova A, Maass A, Davidson WS, Yáñez JM: Genome wide association study for resistance to Caligus rogercresseyi in Atlantic salmon (Salmo salar L.) using a 50K SNP genotyping array.

Aquaculture 2016. In Press.

39. Sonesson AK, Meuwissen THE: Testing strategies for genomic selection in aquaculture breeding programs. *Genet Sel Evol* 2009, 41:37.

40. Niederreither K, Dollé P: Retinoic acid in development: towards an integrated view. *Nat Rev Genet* 2008, 9:541–553.

41. Ø rnsrud R, Lock EJ, Glover CN, Flik G: Retinoic acid cross-talk with calcitriol activity in Atlantic salmon (Salmo salar). *J Endocrinol* 2009, 202:473–482.

42. Manya H, Chiba A, Yoshida A, Wang X, Chiba Y, Jigami Y, Margolis RU, Endo T: Demonstration of mammalian protein O-mannosyltransferase activity: coexpression of POMT1 and POMT2 required for enzymatic activity. *Proc Natl Acad Sci U S A* 2004, 101:500–5.

43. Sander JD, Joung JK: CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol* 2014, 32:347–55.

Chapter 8
Conclusions

8.1 Objectives of Thesis

The aim of the thesis was to apply different genetic approaches to identify the genetic parameters associated with growth traits and host resistance to sea lice disease in farmed salmon populations. A high density genetic map was constructed using a previous developed 132 K SNP array. Ultimately, the practical outcomes of the studies could be in selective breeding to shorten the time to harvest, reduce the losses arisen by sea lice infection, improve the animal welfare, and allow breeders to establish sustainable genetic gains to the industries by enhancing selective breeding in salmon aquaculture.

8.2 Conclusions

8.2.1 Linkage Map

(i) I constructed a linkage map containing 96,396 SNP markers, and the number of mapped SNPs in every linkage map was highly correlated with the length of corresponding linkage map respectively (r = 0.95), indicating that the number of genetic variant is significantly associated with the length of corresponding chromosome.

(ii) Our linkage mapping result contributed approximately 6.5 % of the unassigned genome assembly to be anchored to known 29 chromosomes, of which was equal to around 1 % of entire reference genome assembly (Genbank Accession GCA_000233375.4).

(iii) The large recombination difference in male and female salmon was observed, with substantially higher recombination rate in sub telomeric regions in males. Overall, the ratio of male:female recombination rate was 1:1.5. However, the mechanism underlying these recombination rate differences is still not yet clear.

8.2.2 Genetic Parameters Associated with Growth and Lice Resistance

(i) Based on the results of QTL mapping, GWA analysis and previous literatures, the growth, fillet-related and host resistance to lice traits in Atlantic salmon are likely a polygenic genetic architecture and population-specific.

(ii) The identified QTLs tend to be pleiotropic as several weight-related traits were significantly associated with certain QTLs. However, there is no consistent QTL found to be associated with the traits according to current and previous studies.

(iii) Genomic prediction is an effective approach to improve the traits using family-based breeding schemes, as few as 5 to 10 K high-quality SNPs is able to reach the maximal prediction accuracy in growth and lice resistance traits when using GBLUP methods. The results show that BLUP with fitting genomic information is a better choice to estimate the breeding values of candidates than using traditional BLUP method with pedigrees.

8.3 Future Outlook

The outcomes of this thesis showed that growth traits and sea lice resistance in farmed salmon are both heritable. As such, genomic prediction can be a promising avenue for performing selective breeding. However, the remaining challenges are the cost of genotyping and the prediction accuracy of across population.

Alternative methods for reducing the cost of genotyping

The ultimate goal of the study is to enhance the growth rate and control the sea lice in farmed salmon. Since the cost of genotyping and sample collection are still the critical issues in performing selective breeding through genetic approach, the genotyping-by-sequencing (GBS) can obtain high density genotyping data at an affordable price in comparison with traditional methods. Additionally, genotype imputation based on LD and/or well-recorded pedigree information is also an alternative way to make contribution in genotype data collection. The GBS presents a relatively cost-effective method of discovering and genotyping numerous SNPs, and genotype imputation is able to impute uncharacterized genotypes with *in silico* manner. Both methods are developed to address the major limitation in genomic studies - the cost of genotyping. However, both the price of GBS and the correction rate of imputation require further optimization, which are the major goals of our future works.

Enhanced accuracy across populations

Even though the prediction accuracies of growth traits and lice resistance were robust and encouraging in the within population test, the accuracy of prediction across population test was still a challenge. Encouragingly, this accuracy clearly improved with increasing SNP density. The experimental results imply that genomic prediction can significantly improve the capture of genetic variations in those traits with polygenic architecture. The across-population prediction accuracy is likely to be higher with a larger reference population, and with incorporation of candidate causative SNPs rather than just anonymous markers.

Moving from association to causality

Several candidate SNPs have shown significant associations with traits in this and previous studies (*e.g.* growth traits in chapter 3 to chapter 5, and lice resistance trait in chapter 6). When SNP of interest is targeted, a series of experiments can be performed to study the potential function of the SNP of interest. Our past experience shows that relatively fewer SNPs are characterized on coding regions in comparison with noncoding regions, of which are categorized into two types, including synonymous and nonsynonymous SNPs. The former category of mutation does not affect the amino acid sequence, whilst the latter mutation can alter protein sequence, and it may induce structural changes to the corresponding protein. The sequence variant located in the promoter region may be associated with gene expression, and the variant in the signal peptide domain can influence the cellular localization of the protein. To uncover the function of targeted variants, we can carry out several trials to verify our discoveries, as described below.

For example, initially, we can use *in silico* approaches, such as protein structure analysis tools (*e.g.* http://www.ebi.ac.uk/Tools/structure/), to predict whether the mutation can induce any protein structural changes, and to map the variant with a possible domain to forecast its function changes toward the harboured salmon gene.
Next, whereby *in vitro* studies, we are able to establish the stable cell line or primary cell line (*e.g.* salmon muscle cell), and to quantify the RNA expression levels of the gene containing the targeted SNP (*e.g.* by real-time PCR) to know its impact on gene function. Additionally, recombinant protein expression can be analysed, and SDS-PAGE can be run to see if the SNP alters the post-transcriptional modification, and carry out semi-quantitative western blot to determine whether the variant may change the amounts of protein expression.

Fusing a gene carrying different alleles and a tag protein gene (*e.g.* GFP, green fluorescent protein) allows us to visualize the corresponding variant and changes in its cellular localization. Using immunofluorescence and coimmunoprecipitation can also reveal whether the SNP can make the differences of interaction between the targeted protein and its functional regulation network. Luciferase reporter gene assay

is able to provide information in terms of how the SNP influences promoter activity, and electrophoretic mobility shift assays are able to understand whether the SNPs in transcription factor binding sites can impact the activity of enhancer.

Lastly, the *in vivo* experiment, the generation of transgenic animals or cutting-edge gene editing technology can clarify the association between SNPs and phenotypic changes in real or model organisms. For instance, clustered regularly interspaced short palindromic repeats (CRISPR) technique can be used to edit the targeted allele and to examine its functional effect on phenotype in salmon (as well as in cell lines).

8.4 Summary

Conventional methods of selectively breeding salmon rely on pedigree records and outer appearance to select the parental candidates for mating and breeding, of which may be applicable for physical traits such as body weight traits. However, traits that are economically important but cannot be detected by visual inspection such as disease resistance and muscle traits, both cannot be precisely phenotyped from parental performance, therefore, advanced genetic approaches are capable of improving the drawbacks of traditional selective breeding, by considering the DNA-level information in the salmon breeding program. Chapter 2 of this thesis describes the high density linkage map used for high resolution mapping of SNPs underpinning the targeted traits we investigated in later chapters. Chapters 3-6 discuss how several established genetic methods, including QTL mapping, GWAS, and genomic prediction, were used to characterize the genetic architecture of growth, muscle-related (chapters 3-5), and sea lice resistance traits (chapter 6) in several separate adult and fry salmon populations. The impact of chapter 2 is that, the genetic map can help researcher to identify the position of locus associated with traits of interest, and to better the quality of assembly of reference sequences in Atlantic salmon. By using QTL mapping and GWAS, our findings facilitate industry to underpin the SNPs or QTLs linked with weight-related traits (chapters 3-5), or with sea lice resistance (chapter 6), respectively. Genomic prediction was also used for weight and lice resistance traits, which provides an effective method to predict the performance of traits in candidate from the genomic information and training population. Using the proposed method, the genomic prediction can potentially avoid years of field data collection and reduce the cost of performing a salmon breeding program for the industry. Additionally, although Atlantic salmon has benefitted from modern selective breeding in comparison with most aquaculture species worldwide, only approximately 10 % of total aquaculture production is harvested from selective breeding via genetic-based approaches. The results of this thesis also provide an example of selective breeding for other freshwater and marine farming species in the foreseeable future.

Ultimately, based on our current results, and the *in silico*, *in vitro*, and *in vivo* approaches discussed above, all of which can be potentially utilized when the SNP / QTL of interest is identified via association analysis. The outcomes of this thesis improve understanding of the biological basis of key production traits and can aid genetic enhancement of Atlantic salmon for aquaculture breeding and production.