
THE UNIVERSITY
*of* EDINBURGH

# Mapping and Functional Characterisation of the Atlantic salmon Genome and its Regulation of Pathogen Response

**Serap Gonen**


THE UNIVERSITY
*of* EDINBURGH

Doctor of Philosophy
2011 – 2015

*One night*

*I dreamt of a legend.*

*In the morning*

*when I woke up*

*they told me the legend was gone.*

*But not really.*

*Because legends never really die, do they?*

*This thesis is for the only real legend I ever knew.*

*Professor Stephen Bishop*

*I am grateful beyond words.*

*So I give you my science instead…*

*This thesis is also dedicated to my mother, Fatma Gonen…*

# Declaration

I hereby declare that I am the author of this thesis and that I did all of the work described herein, unless otherwise specified. In chapters 2, 3, 4 and 5, where data was generated prior to the project start or was obtained from collaborative/online resources, these have been clearly specified, and my contributions to the work have been clearly noted. This thesis describes the work carried out by me whilst studying for the degree of Doctor of Philosophy at the University of Edinburgh, between the years 2011 – 2015.

Serap Gonen

# List of Publications

## PEER-REVIEWED PUBLICATIONS

**Gonen** et al. 2014. Linkage maps of the Atlantic salmon (*Salmo salar*) genome derived from RAD sequencing. *BMC Genomics*, **15**:166.

**Gonen** et al. 2015. Mapping and validation of a major QTL affecting the resistance of Atlantic salmon (*Salmo salar*) to pancreas disease. *Heredity*.

**Gonen** et al. Under review. Exploring the utility of cross-laboratory RAD-Sequencing datasets for phylogenetic analysis. *BMC Research Notes.*

## CONFERENCE PRESENTATIONS

The International Symposium on Genetics in Aquaculture XII (ISGA), 2052. Oral presentation: Exploring the Genetic Basis of Resistance to Pancreas Disease in Atlantic salmon (Salmo salar). **Gonen S**, Baranski M, Thorland I, Norris A, Grove H, Arnesen P, Bakke H, Lien S, Bishop SC and Houston RD.

World Congress in Genetics Applied to Livestock Production (WCGALP), 2014. Oral presentation: Genetic Variation in Resistance to Pancreas Disease in Atlantic Salmon. **Gonen S**, Norris A, Arnesen P, Bishop SC, Houston RD.

Roslin Institute Student day, 2014. Poster presentation: Mapping Quantitative Trait Loci (QTL) Linked to Resistance Against Pancreas Disease in Atlantic Salmon Fry. **Gonen S**, Norris A, Arnesen P, Bishop SC, Houston RD.

BBSRC Institutes Conference, 2013. Poster presentation: Heritability of Resistance to Pancreas Disease (Salmonid Alphavirus) in Atlantic salmon Fry. **Gonen S**, Norris A, Arnesen P, Bishop SC, Houston RD.

The International Symposium on Genetics in Aquaculture XI (ISGA), 2012. Oral presentation: Comparative genomics of a disease resistance QTL region in Atlantic salmon. **Gonen S**, Bekaert M, Taggart JB, Bron JE, Davey JW, Bridgett S, Gharbi K, Bishop SC, Houston RD.

4th International Conference Quantitative Genetics, Edinburgh (ICQG), 2012. Poster presentation: Heritability of Resistance to Pancreas Disease in Atlantic salmon Fry. **Gonen S**, Norris A, Arnesen P, Bishop SC, Houston RD.

Advanced breeding programmes for sustainable crop and livestock production International Workshop, 2012. Poster and oral presentation: Mining for disease resistance candidate genes using comparative genomics in Atlantic salmon. **Gonen S**, Bekaert M, Taggart JB, Bron JE, Davey JW, Bridgett S, Gharbi K, Bishop SC, Houston RD.

Roslin Institute Student day, 2012. Poster presentation: Comparative genomics of a disease resistance QTL region in Atlantic salmon. **Gonen S**, Bekaert M, Taggart JB, Bron JE, Davey JW, Bridgett S, Gharbi K, Bishop SC, Houston RD.

# TABLE OF CONTENTS

# Abstract

Atlantic salmon is a species of both scientific and economic importance, and Atlantic salmon farming is a highly profitable industry worldwide. One of the biggest challenges being faced by farms, which affects production efficiency and results in severe economic loss, is disease. In livestock production, one of the approaches taken to limit the impact of disease outbreaks is to selectively breed for improved resistance within farmed populations. Although traditional family-based resistance breeding programs have shown improvements in resistance to a variety of bacterial, viral and parasitic diseases on Atlantic salmon farms, response to selection can be slow. One way of increasing selection efficiency is through the incorporation of genetic markers into breeding programs, for marker-assisted or genomic selection. However, genomic resources for cultured aquatic species are sparse, and the generation of new and denser resources for use in selective breeding programs would be advantageous. The main focus of this thesis is the development of genomic resources in Atlantic salmon and the application of those resources to gain a better understanding of the salmon genome, particularly in the genetic basis of host resistance to infectious diseases.

The first aim of this thesis was to develop improved genomic resources for Atlantic salmon, and to characterise the Atlantic salmon genome via construction and analysis of a SNP linkage map derived from RAD-Sequencing (RAD-Seq). Approximately 6,500 SNPs were assigned to 29 linkage groups, and ~1,800 male-segregating, and ~1,400 female-segregating SNPs were ordered and positioned. Overall map lengths and recombination ratios were relatively consistent between the sexes and across the linkage groups (~1:1.5, male:female). However, a substantial difference in the degree of marker clustering was seen between males and females, which is reflective of the difference in the positions of chiasmata between the two sexes. Using this map, ~4,000 Atlantic salmon reference genome contigs were assigned to a linkage group, and 112 contigs were assigned to multiple linkage groups, highlighting regions of homeology (large sections of duplicated chromosomal regions) within the salmon

genome. Alignment of SNP-flanking sequences to the stickleback and rainbow trout genomes identified putative gene-associated SNPs and cross-species chromosomal orthologies, and provided evidence in support of the salmonid-specific genome duplication.

In addition, based on this and other publically available RAD-Seq datasets, the utility of RAD-Seq-derived data from different species and laboratories for population genetics analyses was tested. Short RAD-Seq contigs in Atlantic salmon and nine other teleost fish were used to identify cross-species orthologous genomic relationships. Several thousands of orthologous RAD loci were identified across the species, with the number of RAD loci decreasing with evolutionary distance, as expected. Previously published broad-level relationships between orthologous chromosomes were confirmed. The identified cross-species orthologous RAD loci were used to estimate evolutionary relationships between the ten teleost fish species. Previously published relationships were recovered, suggesting that RAD-Seq data derived from different laboratories is useful for this purpose.

The second aim was to characterise the genetic architecture of resistance to two viral diseases affecting Atlantic salmon production on farms: pancreas disease (PD), and infectious pancreatic necrosis (IPN). Using data and samples collected from a large population of salmon fry challenged with PD, a high heritability for resistance was estimated ($h^2$ ~0.5), and four QTL were identified, on chromosomes 3, 4, 7 and 23. The QTL explaining the highest within-family variation for resistance was located on chromosome 3. This QTL has been confirmed in a population of post-smolts by an independent research group, highlighting the potential for its incorporation into breeding programs to improve PD resistance. For IPN, the major resistance QTL had previously been mapped to linkage group 21. However, the mutation(s) underlying this QTL effect and the consequences of these mutation(s) on the affected genes and relevant biological resistance mechanisms are unknown. To generate a list of candidate genes within the vicinity of the IPN QTL, QTL-linked DNA sequences were aligned to four model fish genomes. This identified two QTL-orthologous regions in each of the species, and gene order within these regions was highly

conserved across species. Analysis of gene expression patterns between IPN resistant and susceptible salmon in a viral challenge experiment revealed that the five most significantly differentially-expressed genes mapped to the QTL-orthologous region on linkage group II of stickleback. Pathway enrichment analysis across all differentially-expressed genes suggests that biological pathways influencing viral infection stress response/entry/replication, cellular energy production and apoptosis may be involved in resistance during the initial stages of IPN virus (IPNV) infection. These results have provided the basis for further study of the putative involvement of these candidate genes and pathways in genetic resistance to IPNV.

In summary, the results and resources presented in this thesis extend our current understanding of the salmon genome and the genetic basis of resistance to two viral diseases, and provide resources with the potential to be used in Atlantic salmon selective breeding programs to tackle disease outbreaks.

# Lay Summary

Atlantic salmon are a widely farmed aquaculture species, with high global annual production and economic profit. As such, factors such as diseases which affect production from farms are highly detrimental, economically, but also in terms of fish welfare. Although measures for disease control are in place, these are not always fully effective when implemented alone. One way of tackling diseases on farms is through the use of selective breeding programs, to improve resistance amongst fish stock. This involves the identification of the most resistant individuals within the farmed population, which are then used as parents for the next generation. However, the identification of the most resistant individuals requires the exposure individuals to pathogens. As well as causing fish welfare issues, pathogen-exposed individuals cannot be used for breeding, since they may still be carriers of the disease. An alternative way of identifying the most resistant individuals without exposure to disease causing pathogens is through the use of genomic technologies. Genotyping individuals at previously defined regions of the genome which are known to play a role in the biological resistance pathways (resistance loci) can be used to predict whether an individual is likely to be resistant or susceptible to a disease causing pathogen. However, these resistance loci must first be identified, and this requires the generation of a larger repertoire of genomic resources, followed by screening of these for association with resistance. This thesis describes the generation of new genomic resources for Atlantic salmon, and further, describes the implementation of these and other available genomic resources for the identification and characterisation of loci underlying resistance to two important viral diseases, infectious pancreatic necrosis and pancreas disease.

# Chapter 1

# Introduction

## 1.1 Sustainable farming: Aquaculture

With increasing global human population sizes, demands for increased and sustainable production from plant, aquaculture and livestock farming have risen. In response, major advances in farm production efficiency have been documented. For example in cattle, annual milk productions per cow have increased by 16% within a ten year time interval (2003–2013), resulting in reductions in herd size (and requirement for farm land) despite no significant losses in total global milk yields (DEFRA, 2014). In addition, alternate methods of farming to increase production efficiency, as well as the potential for domestication of as yet uncultured species, are being explored (Fjalestad *et al*, 2014; USDA, 2014a). For example, the domestication of new aquatic species (such as muscles), and production efficiency from established aquaculture farms, have both experienced a dramatic increase over the past few decades, with a reported annual overall growth rate of 7%.

Currently, more than 50% of the total global fish production is from aquaculture farms (FAO, 2014a). As such, aquaculture species are of huge economic importance, promising a sustainable resource of high-quality protein and long-chain fatty acids (AquaGen, 2013; FAO, 2013; FAO, 2014b; Gjerde *et al*, 2014). For some species, such as the salmonids, farming practices are well established (CSIRO, 2012; Gjerde *et al*, 2014; IcelandicFisheries, 2014). For others, this is an ongoing process, and best practices for efficient culturing of a large diversity of fish, crustacean and mollusc species are still being investigated (CanadianAquaculture, 2012; Fjalestad *et al*, 2014; USDA, 2014b). The farmed species of choice within specific countries or continents varies, and may be affected by climatic conditions as well as the availability of natural populations to source founders for the purposes of breeding. For example, currently, the highest shrimp producer is China (FAO, 2014c). Norway,

Chile and North America are highly ranked amongst the main producers of Atlantic salmon (FAO, 2014b; MarineHarvest, 2014).

## 1.2 Atlantic salmon: Farming and breeding

Although still relatively new, farming in salmonid species has experienced a substantial increase in the last few decades. In particular, Atlantic salmon (*Salmo salar*) is one of the main contributors to aquaculture farming. Recent reports of Atlantic salmon global annual production values exceed 2 million tonnes of production, corresponding to a value of just over $10 billion (FAO, 2012). As well as their significance at an industrial level, Atlantic salmon remain a traditional food source and livelihood for smaller communities in close locality to natural habitats (for example in North America) (Guy *et al*, 2009; Davidson *et al*, 2010; FDA, 2013), and are an important recreational sport fishing species.

As with livestock and plant species farming, diseases on Atlantic salmon farms present a major challenge for efficient production. Depending on the geographic location, lifecycle stage and time of year, farmed populations of Atlantic salmon are exposed to a variety of parasitic, bacterial and viral diseases (Toranzo *et al*, 2005; Lhorente *et al*, 2012; Saravanan *et al*, 2013; Madhun *et al*, 2014). These diseases do not originate from a farmed environment, and are known to affect wild populations (Waknitz *et al*, 2002; Waknitz *et al*, 2003; Thorstad *et al*, 2008; Whelan, 2010; Murray *et al*, 2011; Ruane and Jones, 2013). However, the likelihood of disease epidemics after an initial outbreak are higher within a farm setting, due to the higher density of fish within a given area, the increased levels of stress experienced by fish, and the greater possibility of a sustainable pathogen vector on sites (Taksdal *et al*, 1998; Heuch and Mo, 2001; Bjorn and Finstad, 2002; Skilbrei and Wennevik, 2006; Krkosek *et al*, 2007). Therefore, farm management techniques, such as improving site hygiene, vaccination, and in extreme cases, culling of infected stock, are being implemented in an attempt to control and limit the impact of disease outbreaks.

In common with terrestrial livestock species, breeding programs to select for improvements in traits of economic value on Atlantic salmon farms are in place.

Selected traits include faster growth rates, flesh and fillet quality (including pigmentation, texture and low fat content), harvest weight, feed conversion efficiency, survival, reproduction, late sexual maturity and disease resistance (Leon, 1975; Aas *et al*, 2006; Baranski *et al*, 2010; CSIRO, 2012; Gjerde *et al*, 2014; IcelandicFisheries, 2014). For many traits, including disease resistance, related individuals show more similar trait phenotypes, on average, compared to unrelated individuals (i.e. the trait is heritable). Therefore, identifying the best performing individuals for a given trait, and then using these as breeding parents in selective breeding programs over a number of generations, should result in improvements in the selected trait.

In general, selectively breeding for a trait of economic value such as growth involves the recording of the trait phenotype of a potential breeding parent (i.e. growth rate/time taken to reach harvest size), identifying the individuals with the best score (i.e. those that grow the quickest), and then using these as the breeding parents for the next generation. Using this method of selective breeding, improvements in selected traits of economic value have been observed. For example, farmed salmon have been reported to be, on average, 2.5 fold heavier than wild salmon when reared under hatchery conditions (Glover *et al*, 2009). In Atlantic salmon breeding, the large number of eggs obtained from a single female, combined with the ability to conduct external fertilisation, has enabled the design of highly-controlled mating systems within populations of farmed fish and the production of large full- and half-sibling families (Sonesson, 2005). This has further increased the selection efficiency within selective breeding programs.

Selectively breeding for improved resistance to diseases amongst farmed populations, as a component of a disease management strategy, has been implemented in many domesticated species for a variety of diseases. For example, reports of up to a 5% increase in resistance to mastitis in cattle have been recorded between lines selected for resistance and lines selected solely for high milk production (i.e. unselected controls) (Steine, 1998; Heringstad *et al*, 2000).

For disease resistance as a trait of economic importance, the identification of the best performing individuals (i.e. those most resistant to pathogen infections) requires the exposure of potential breeding candidates to the disease causing agent within disease challenge experiments. However, survivors of disease challenge experiments might not always be available for the purposes of breeding, for the following reasons.

First, although individuals survive a challenge, the infection may have resulted in a reduction in their overall performance. Therefore, they are no longer considered as the top candidates for selective breeding for improvements in other traits. For example, mean family weight and mean family resistance to taura syndrome virus in the Pacific shrimp were reported to be negatively correlated (genetic correlation: -0.46 ± 0.18) (Argue *et al*, 2002). As such, selecting for improved resistance could result in a reduction in harvest weight in subsequent generations. Second, for some diseases, survivors of challenge experiments may not have been able to completely clear the pathogen from their systems, and they remain as pathogen carriers/vectors. This may increase the likelihood of a second epidemic within the population if pathogen naïve and susceptible individuals remain, or may result in vertical transmission of the pathogen from the carrier parent to their offspring.

Instead of directly selecting upon survivors of infection, breeding candidates may be selected based on the performance of their offspring, full-siblings and/or other relatives in a pathogen challenge experiment (Gjedrem, 1985; Gjøen and Bentsen, 1997). For example, in rainbow trout, the breeding potential of parents for obtaining improvements in resistance to diseases such as viral haemorrhagic septicaemia has been tested based on the performance of their offspring in experimental challenges [e.g. Henryon *et al* (2005)]. When using information from full-siblings, families are split into two groups of challenged and naïve individuals. The disease response (e.g. survival, viral load, etc.) of the challenged group determines whether their naïve full-siblings are used as breeding parents (Falconer and Mackay, 1996). This family-based method of selection has and is been applied for many diseases, with improvements in resistance documented through subsequent generations. For example, family-based selection for Marek's disease resistance in chickens has been

reported to result in up to a 78% difference in mortality between lines selected for increased resistance and those selected for increased susceptibility (Cole, 1964; Cole, 1969). Family-based selection for disease resistance has been an ongoing part of several Atlantic salmon breeding programs, with promising improvements in resistance seen for several diseases (Fjalestad *et al*, 1993; Storset *et al*, 2007).

While effective, selection for resistance based on family member phenotypes ignores within-family variation, hence, will be less effective in comparison to approaches which utilise both within- and between-family variation (Falconer and Mackay, 1996; Meuwissen and Goddard, 1996). Another potential way of improving resistance involves the identification and application of selection based on parameters which minimise the between- and/or within-family variation. However, this would result in a reduction in genetic diversity and an increase in inbreeding, resulting in inbreeding depression and a potential reduction in fitness and performance in other traits. Therefore, alternative methods able to exploit the within-family variation for resistance whilst maintaining diversity are desirable (Falconer and Mackay, 1996; Stead, 2002; Oldenbroek (ed), 2007; Hill, 2013). One such method involves the incorporation of genetic markers into breeding programs.

## 1.3 Genomic technologies in breeding

The ability to exploit both the between- and within-family variation and select based at the individual rather than family level has meant that the incorporation of genetic markers [traditionally microsatellite markers, and more recently, single nucleotide polymorphisms (SNPs)] into livestock and plant breeding programs has become popular in recent decades. This has been particularly useful when selecting for carcass traits such as fillet quality and for improvements in disease resistance, where selection based on individual performance is not possible (Meuwissen and Goddard, 1996; Haley and Visscher, 1998; Sonesson, 2005). This is because an individual's response to infection with a pathogen can be inferred based on marker genotype, without exposing the individual to the pathogen. Furthermore, genetic information is available at birth, or at least as soon as DNA can be collected, and selection can potentially take place earlier in the lifecycle of an individual. With the identification

of fully informative markers (i.e. markers able to predict resistance at the population level and able to capture the within-family variation), the need for family-based challenge tests may decrease and may even no longer be required. This will reduce stress levels experienced by fish due to disease exposure, and decrease costs associated with the breeding and maintenance of individuals used in challenge studies, which will have no final economic value (Meuwissen and Goddard, 1996; Houston *et al*, 2008; Goddard *et al*, 2010).

The identification of genetic markers significantly associated with the trait of interest involves the collection of trait phenotypes for a large number of individuals, genotyping these individuals using a set of genetic markers, and then using both datasets in association or linkage-based studies to discover markers linked to quantitative trait loci (QTL). Depending on the genetic architecture of the trait, i.e. the number of significant QTL and the size of their effects, markers tightly linked to these QTL can be used for marker-assisted or genomic selection in breeding programs. For traits which appear to segregate in a Mendelian fashion (i.e. one or a few major QTL identified), marker-assisted selection is a useful breeding strategy, whereby selection is performed based on genotypes at one or a few significant markers. Alternatively, if the trait is influenced by many QTL with small effect, genomic selection based on a larger number of QTL-marker associations across the genome can be implemented. QTL mapping, and the use of QTL-linked markers for selection, has been successfully incorporated into many livestock breeding programs [e.g. selection for improvements in forelimb-girdle muscular anomaly in Japanese black cattle (Masoudi *et al*, 2007)].

The identification of QTL-trait associations is the first step in determining the underlying causal variants which are directly influencing disease resistance. Further fine-mapping and the identification of causal gene(s) and/or mutation(s) would increase selection efficiency, reduce the need for continuous sib-challenge experiments, and improve the understanding of the underlying biological mechanisms involved in resistance (Li *et al*, 2011; Houston *et al*, 2012). Furthermore, knowledge of the genes and pathways involved in resistance may help

in the long term development of treatments and more effective vaccination strategies (Biering *et al*, 2005). Direct selection based on the causative variant(s) can be incorporated into breeding programs and can be applied across populations, since unlike marker-based selection, it does not rely on the conservation of linkage between marker and QTL across populations with differing genetic backgrounds (Misztal, 2006).

## 1.4 QTL mapping in farmed fish populations

QTL mapping experiments in farmed fish species have some advantages compared to other farmed animals. Firstly, the large number of offspring in full- and half-sibling families results in an increased power to detect QTL segregating within the population of interest (Darvasi and Soller, 1992; Hayes *et al*, 2009). Secondly, since external fertilisation is possible in many farmed fish, large numbers of families with controlled, experiment-specific family structures and mating designs can be generated. For example, full-sibling families can be nested within a half-sibling family structure, and, unlike large livestock species where the sire is the common parent, half-sibling families with the dam as the common parent can be produced if required (Sonesson, 2005; Gjedrem and Baranski, 2009).

In particular for Atlantic salmon, characteristics of the genome such as the large difference in recombination rates between males and females (see section 1.6 for further description), mean that QTL mapping experiments can be undertaken using a cost-effective, two-step, family linkage-based approach. The much lower recombination rates reported in males means that markers located on the same chromosome will appear tightly linked in sire-based linkage analyses of marker segregation patterns. As such, genotyping a few markers per linkage group and analysing the segregation of these markers from sire to offspring will identify chromosomes potentially harbouring trait-associated QTL (step one, sire-based linkage analysis). More accurate positioning of the identified QTL can be obtained through genotyping a denser set of markers only on these significant chromosomes, and analysing their segregation patterns from dam to offspring (step two, dam-based linkage analysis) (Darvasi and Soller, 1992; Hayes *et al*, 2006; Houston *et al*, 2008;

Hayes *et al*, 2009). This approach further takes advantage of the large full- and half-sibling family structures available on farms, by enabling the identification of significant chromosomes using large half-sibling families where the sire is the common parent.

QTL mapping, and the use of marker-QTL associations for breeding, is starting to become more feasible in aquaculture breeding programs (Devlin *et al*, 1991; Nirea *et al*, 2012; AquaGen, 2013; Palaiokostas *et al*, 2013). In Atlantic salmon, marker-assisted selection is currently being implemented to improve host resistance to infectious pancreatic necrosis (IPN) (Houston *et al*, 2008; Moen *et al*, 2009; Houston *et al*, 2010; Houston *et al*, 2012), and the utility of genomic selection for other traits of economic value is being investigated (Sonesson and Meuwissen, 2009; Nirea *et al*, 2012).

## 1.5 Diseases in Atlantic salmon farming

At the time of writing, four of the top five major diseases on Atlantic salmon farms are caused by a viral pathogen, all of which are listed as having 'no treatment' [Table 1.1; FAO (2014b)]. Although management practices are in place, additional and alternative control measures, particularly in the prevention of initial disease outbreaks, are essential.

**Table 1.1: Top five major disease problems on Atlantic salmon farms, taken from FAO**

| DISEASE | AGENT | TYPE | SYNDROME | MEASURES |
|---|---|---|---|---|
| ISA (Infectious salmon anaemia) | Orthomyxovirus | Virus | Lethargy; appetite loss; gasping at water surface; pale gills & heart; fluid in body cavity; dark liver; haemorrhages in internal organs | No treatment; statutory controls; biosecurity; bloodwater treatment |
| VHS (Viral Haemorrhagic Septicaemia) | Rhabdovirus | Virus | Bulging eyes and, in some cases, bleeding eyes; pale gills; swollen abdomen; lethargy | No treatment; statutory controls; vaccines being developed |
| IPN (Infectious Pancreatic necrosis) | Birnavirus | Virus | Erratic swimming, eventually to bottom of tank where death occurs | No treatment; statutory controls; biosecurity; broodstock screening; vaccines being developed |
| SPDV (Salmon Pancreas Disease virus) | Togavirus | Virus | Weight loss; emaciation; mortalities | No treatment; withholding feed; vaccination |
| Furunculosis | Aeromonas salmonicida | Bacterium | Inflammation of intestine; reddening of fins; boils on body; pectoral fins infected; tissues die back | Antibiotics; vaccination |

As described above, improvements in resistance to many diseases are being achieved using family-based selective breeding programs in livestock, plant and aquaculture species (Soller and Andersson, 1998; Nicholas, 2005). However, the utility of family-based selection for improved resistance is limited, since the within-family variation cannot be exploited and the determination of the most resistant families still requires pathogen challenge experiments (Haley and Visscher, 1998; Sonesson, 2005). As such, the use of genetic markers to infer the resistance status of an individual without pathogen exposure is becoming common practice in livestock and plant species (Soller and Andersson, 1998; Gibson and Bishop, 2005; Nicholas, 2005; Yang and Francis, 2005; Misztal, 2006; Miedaner and Korzun, 2012; Ortega and Lopez-Vizcon, 2012; Recknagel *et al*, 2013a). This requires the characterisation of the underlying host genetic basis to resistance.

In this thesis, analyses into furthering the understanding of the host genetic basis to two of the top four major viral diseases in Atlantic salmon are described. These are infectious pancreatic necrosis (IPN) and pancreas disease (PD).

### 1.5.1 Infectious pancreatic necrosis (IPN)

The causative agent of IPN is the aquabirnavirus infectious pancreatic necrosis virus (IPNV). The genome of the virus is composed of two double-stranded RNA molecules: segment A and segment B (Figure 1.1). Segment A encodes four proteins: VP2, which is an external capsid protein and contains the viral antigenic sites recognised by the host immune system; VP3, which is an internal capsid protein; VP5, which may be involved in preventing apoptosis of infected host cells and promoting viral survival; and a non-structural protein involved in cleaving the polyprotein product of segment A. Segment B encodes a single protein VP1, which is an RNA dependent RNA polymerase involved in viral genome replication (Duncan *et al*, 1991; Blake *et al*, 2001; Chiu *et al*, 2010). Many strains of IPNV have been isolated in different geographic localities, including Scotland, Norway, Chile, and North and South America, and these show varying levels of virulence (Smail *et al*, 1992; Smail *et al*, 1995; Marjara *et al*, 2011; Skjesol *et al*, 2011).

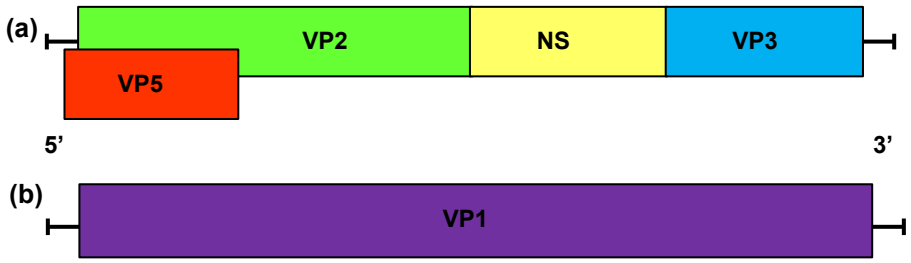



**Figure 1.1: The genome of IPNV**
The IPNV genome is composed of two double stranded RNA molecules: (a) Segment A, and (b) Segment B. Segment A encodes three structural and one non-structural (NS) protein. Segment B encodes for a single structural protein.

The clinical signs upon infection with IPNV have been well characterised, and include necrosis of pancreatic acinar and liver cells, a swollen abdomen and lethargy (Wolf and Quimby, 1971; MERCK, 2014). Infected fish lie on their sides and hyperventilate, appear darker in colour and show a reduced appetite and growth rate compared to naïve uninfected fish (Roberts and Pearson, 2005). In addition, infection with IPNV has been suggested to increase host susceptibility to other pathogens. For example, a higher sea lice burden has been reported in infected vs. naïve fish (Roberts and Pearson, 2005). Clinical signs of infection are typically seen 5–6 weeks after transfer to seawater in post-smolts, and the peak in mortalities is observed 8 weeks post-transfer (Guy *et al*, 2009).

A number of disease management techniques are currently being implemented on farms in an attempt to limit the effect of IPN outbreaks. These include: site hygiene, biosecurity measures and vaccination methods (Guy *et al*, 2006; Storset *et al*, 2007; Kjøglum *et al*, 2008; Guy *et al*, 2009). In particular, vaccination strategies have been reported to show improvements in survival rates compared to unvaccinated controls in the same tank, with improvements in survival rates reported even within highly-resistant groups. However, trials have shown varying levels of vaccine efficacy and are not always reproducible, thus the use of this strategy as a control measure has not been fully implemented (Frost and Ness, 1997; Mikalsen *et al*, 2004; Ramstad and Midtlyng, 2008; Kumari *et al*, 2013). Despite the extensive control measures, at the peak of epidemics, disease outbreaks have been reported to result in mortality levels as high as 80% in Atlantic salmon fry and 30% in post-smolts (Guy *et al*, 2009;

Houston *et al*, 2010). As such, additional control measures, such as selectively breeding for improved resistance, were required.

Several studies using data from natural outbreaks or challenge experiments have provided evidence to suggest that the level of innate host resistance has a significant effect on mortality levels, suggesting that response to infection is heavily influenced by host genetics (Taksdal *et al*, 1998; Storset *et al*, 2007; Kjøglum *et al*, 2008; Ramstad and Midtlyng, 2008; Guy *et al*, 2009). This, together with the high heritability estimates obtained in natural outbreaks and tank challenge trials ($h^2$ range: 0.31–0.69) has indicated that selecting for improved resistance would be plausible (Guy *et al*, 2006; Kjøglum *et al*, 2008; Guy *et al*, 2009; Moen *et al*, 2009).

Traditional family-based IPN resistance selective breeding programs have been in place since 1997, and have been implemented as part of larger breeding schemes (Guy *et al*, 2006; Storset *et al*, 2007; Kjøglum *et al*, 2008; Guy *et al*, 2009). Although improvements in resistance have been recorded (Storset *et al*, 2007), IPN epidemics were still resulting in high levels of mortality. In addition, the detection of viral particles in asymptomatic salmon and on salmon eggs suggests that the virus carrier status is a possibility amongst fish stock, and that vertical transmission of the virus is possible (Wolf *et al*, 1963; Bootland *et al*, 1991; Ruane *et al*, 2007; Orpetveit *et al*, 2008). Therefore, there was a clear need for alternative methods of disease control, based on individual- rather than family-level resistance. The reported moderate-to-high heritabilities for resistance suggested that the characterisation of the underlying genetic variation in host response to IPN through the identification of marker-QTL associations, and the incorporation of these into resistance breeding programs, could be one way of exploiting individual-level variation for resistance.

Using the two-step QTL mapping approach described in section 1.3 above, the major QTL explaining 21–32% of the within-family phenotypic variation and 98% of the additive genetic variation for resistance to IPN was identified and mapped to linkage group 21 (Houston *et al*, 2008; Moen *et al*, 2009; Houston *et al*, 2010). Although initially mapped to a 10cM region (using a microsatellite linkage map), subsequent

studies have reduced the confidence interval of the QTL to 2cM (Houston *et al*, 2012). The same QTL has been shown to influence response to infection at both the fry and post-smolt stages of the salmon lifecycle, suggesting a common mechanism of virus clearance across both life stages.

Although this QTL is clearly of large effect, the predicted nature of this effect varies across different studies. Some studies estimate a non-additive, dominant effect to the QTL, with groups of individuals with at least one resistant QTL allele showing no/negligible levels of mortality (Houston *et al*, 2010). Other studies report no mortality amongst QTL homozygote resistant individuals and a small percentage of mortalities amongst heterozygotes, suggesting some additive effect to the QTL (Moen *et al*, 2009; Houston *et al*, 2012). The likely reason for this discrepancy may be due to the number of mortalities observed (i.e. disease prevalence) within a given study, which is influenced by the force of infection (i.e. exposure of the fish to the virus, and whether the fish becomes infected once exposed) (Bishop and Woolliams, 2010a; Bishop and Woolliams, 2014).

Currently, SNP and microsatellite markers are being used for marker-assisted selection for improved resistance to IPN in commercial breeding schemes (Houston *et al*, 2008; Moen *et al*, 2009; Houston *et al*, 2012; AquaGen, 2013), and major improvements in resistance, as well as a substantial decrease in disease outbreaks and levels of mortality, have been reported (AquaGen, 2013). As described previously, a more efficient selection method would involve the identification of, and selection based on, the causative variant(s) underlying the QTL. This is because marker-QTL linkage associations can vary across populations and generations, and must be frequently re-tested (Misztal, 2006). However, the 2cM QTL confidence interval is still relatively large, and may contain hundreds of putative candidate genes. This, together with the lack of a fully-assembled and annotated Atlantic salmon reference genome [ASalBase, http://www.asalbase.org/sal-bin/index; Davidson *et al* (2010)], has hindered the identification of the underlying causal variant(s). Recent reports suggest that the underlying single causative mutation has been identified (Moen and

Ødegård, 2014). However, the identity of this causative mutation has not yet been reported in the public domain.

### 1.5.2 Pancreas Disease (PD)

PD was first described in Scotland in 1976, and the causative viral agent was subsequently identified as a salmonid alphavirus (SAV) (Munro *et al*, 1984). Six subtypes of SAV have been isolated in different parts of the world, including in Scotland, Norway and Chile (Munro *et al*, 1984; Rowley *et al*, 1998; Hodneland *et al*, 2005; Rodger and Mitchell, 2007; Fringuelli *et al*, 2008; Graham *et al*, 2012; Hjortaas *et al*, 2013; Graham *et al*, 2014). Subtypes are geographically specific, and farms within the same locality typically show infection with the same subtypes (Kristoffersen *et al*, 2009; Graham *et al*, 2012). For example, the two SAV subtypes in Norway (SAV2 and SAV3) have been shown to affect distinct sites (SAV2 in the north and SAV3 mainly in the south of Norway), with no overlap or co-infection within sites (Hjortaas *et al*, 2013; Jansen *et al*, 2014).

The SAV genome is comprised of one single-stranded RNA molecule (Figure 1.2). The 5' region of the genome encodes four non-structural proteins which are thought to be involved in viral replication. The 3' region encodes five structural proteins: the capsid; the antigenic proteins E1 and E2 which contain the epitopes recognised by the immune system; E3, which is involved in the correct folding of E2; and the 6K protein, which is involved in mediating viral entry (Vogel *et al*, 1986; Fuller, 1987; Gaedigknitschko and Schlesinger, 1990; Lusa *et al*, 1991; Cheng *et al*, 1995; Loewy *et al*, 1995; Sanz *et al*, 2003; McLoughlin and Graham, 2007).

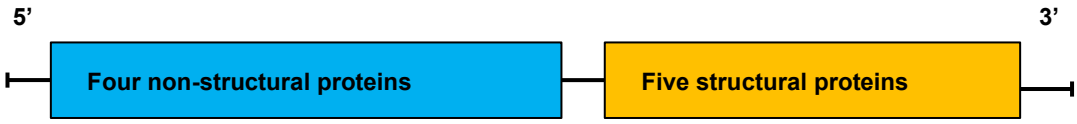



**Figure 1.2: The genome of SAV**
The SAV genome is comprised of a single-stranded RNA molecule, encoding four non-structural proteins in the 5' region, and five structural proteins (capsid, E3, E2, 6K and E1 in that order) in the 3' region.

Clinical signs of SAV infections include shrinkage of the nucleus of pancreatic acinar cells in the first stage, with later stages characterised by changes to skeletal

muscle, necrosis of cardiomyocytes, and a loss of pancreatic acinar cells (McLoughlin *et al*, 2002; Taksdal *et al*, 2007; Norris *et al*, 2008; Taksdal *et al*, 2014). Clinical signs appear gradually and are similar to clinical signs associated with IPNV infection, making the diagnosis of PD based on histology alone difficult (McLoughlin *et al*, 2002; McLoughlin and Graham, 2007; Taksdal *et al*, 2007; Collet *et al*, 2013). Additional diagnostic methods to complement histological evidence of infection include viral particle or antibody detection (Taksdal *et al*, 2007), PCR techniques (Jansen *et al*, 2010a) and tissue culture methods (Collet *et al*, 2013).

Natural infections with SAV have only been reported in Atlantic salmon post-smolts. Although epidemics may result in high levels of mortality (ranging from 10–50%), the peak in mortality can take many months to be observed, and may not be seen at all (Weston *et al*, 1999; McLoughlin *et al*, 2002; McLoughlin *et al*, 2006; McLoughlin and Graham, 2007; Rodger and Mitchell, 2007; Taksdal *et al*, 2007; Kristoffersen *et al*, 2009; Jansen *et al*, 2010b; Jansen *et al*, 2010a; Jensen *et al*, 2012). Further, long term sub-clinical infections are common, with a high level of morbidity, caused by a reduced appetite and poor growth rate amongst survivors, being the main factor responsible for economic loses in PD outbreaks (Norris *et al*, 2008; Cano *et al*, 2014; Jansen *et al*, 2014). At present, the best measure for morbidity caused by PD is not clear (e.g. it could be growth rate, viral load, etc.), thus the analysis of morbidity as a potential trait for selection is limited.

In an attempt to control and limit the effect of PD outbreaks on farms, management techniques similar to those described for IPN are being implemented. As for IPN, vaccination strategies have shown reductions in the levels of mortality, although repeatability of vaccine efficacy has not been achieved, and vaccination procedures are still under development (Rodger and Mitchell, 2007; Jansen *et al*, 2010b). In addition, the feeding rate of fish has been suggested to impact levels of mortality, where fish with higher feeding rates appear to be more susceptible to infection with SAV than fish fed at reduced rates. The biological mechanisms influencing this variation caused by feeding rate differences are unclear (Rodger and Mitchell, 2007).

Studies of PD disease dynamics have reported a complex, multifactorial basis to resistance, which is influenced by the time of year of infection outbreak (early summer/autumn typically show peak mortality levels) (Kristoffersen *et al*, 2009), temperature (virus growth and spread is temperature dependent) (Graham *et al*, 2008; Jansen *et al*, 2010b; Jansen *et al*, 2010a; Stene *et al*, 2013) and viral strain (McLoughlin *et al*, 2006; Graham *et al*, 2014; Taksdal *et al*, 2014). In addition, host-related factors, such as age and stress conditions, have been reported to influence resistance (Kristoffersen *et al*, 2009; Gadan *et al*, 2013; Grove *et al*, 2013; Herath *et al*, 2013). Results from natural and challenge experiments suggest that there is a significant host genetic component to resistance (Ruane *et al*, 2005; Norris *et al*, 2008; Xu *et al*, 2012). At present, only a single study investigating genetic variation in resistance to PD has been published, and a moderate heritability of 0.21 (±0.005) was estimated (Norris *et al*, 2008).

Although selective breeding programs to improve resistance to PD are in place, these are based on selection at the family-level, using full-sibling challenge experiments (Taksdal *et al*, 2007). As described previously, methods, such as marker-assisted or genomic selection able to exploit the within-family variation for resistance may increase the efficiency of response to selection. This is further necessary for PD, since like IPNV, the possibility of vertical transmission of SAV has not been ruled out (Jansen *et al*, 2010b). Given the estimated moderate heritability, the observed variation in PD susceptibility across individuals, and reports of a genetic component to resistance across multiple studies, individual selection via the incorporation of genetic markers into breeding schemes would be plausible. As yet, no published QTL mapping studies exist, and the underlying genetic architecture of resistance to this disease remains unclear.

## 1.6 Atlantic salmon: Genome structure and genomic resources

One reason for the relatively low number of examples of the implementation of genetic technologies into aquaculture selective breeding programs is the general lack of genomic resources [such as high-density linkage and physical maps, SNP and

gene expression arrays and expressed sequence tag (EST) libraries] for farmed aquatic species. The generation of high-density genomic resources for non-model organisms without a fully-assembled and annotated reference genome can be more challenging than in a model species (Moen *et al*, 2008; Andreassen *et al*, 2010; Lien *et al*, 2011; Houston *et al*, 2014a; Palti *et al*, 2014). For Atlantic salmon and salmonid species, the generation of genomic resources is an even greater challenge, due to specific evolutionary events in the salmonid lineage affecting characteristics of the genome.

Salmonidae originate from a common ancestor whose genome underwent a duplication event approximately 80–100 million years ago (MYA) (Allendorf and Danzmann, 1997; Volff, 2005; Shedko *et al*, 2013; Berthelot *et al*, 2014; Macqueen and Johnston, 2014). A recent estimate obtained from analysis of nuclear and mitochondrial sequences suggests a date of 60–100 MYA (Crête-Lafrenière *et al*, 2012), however, this study was not exclusive to paralogous sequence data only. Extant salmonid species have not yet fully recovered the diploid state. Areas of the genome still show evidence of tetraploid segregation and retention of duplicate gene copies, with high levels of similarity (>85%) reported between sequences of paralogous genes (McKay *et al*, 2004; Berthelot *et al*, 2014).

This 'pseudo-tetraploid' state of the genome makes salmonid species suitable models for the investigation of mechanisms involved in genome re-diploidisation, including gene silencing, gene divergence, inversions and chromosomal rearrangements (McKay *et al*, 2004; Ng *et al*, 2005; Volff, 2005; Lubieniecki *et al*, 2010; Lien *et al*, 2011; Leitch *et al*, 2013; Berthelot *et al*, 2014). These mechanisms of genome re-diploidisation are thought to cause a gradual reduction in the ability of homeologous chromosomes to form quadrivalent structures during meiosis, and to potentially be the cause of species' divergence (McKay *et al*, 2004; Moen *et al*, 2004a; Volff, 2005).

In Atlantic salmon, the formation of quadrivalent structures during meiosis appears to be specific to males, and loci appear to segregate in a diploid fashion in females

(Lubieniecki *et al*, 2010). Homeologous pairing within quadrivalent structures in males is thought to occur only at the telomeres, since it takes place after homologous chromosome pairing is complete. This secondary quadrivalent structure has been postulated to be responsible for the distinct lack of recombination observed in male salmon (Allendorf and Danzmann, 1997; Gilbey *et al*, 2004; Danzmann *et al*, 2008; Lien *et al*, 2011). While in most species the heterogametic sex often shows reduced recombination rates compared to the homogametic sex (Barthes *et al*, 2011), the ratio in Atlantic salmon is one of the highest reported amongst the vertebrates [up to 1:17 (male:female), Danzmann *et al* (2005)]. The formation of these quadrivalent structures is also thought to cause two phenomena in males: pseudolinkage and double reduction.

Pseudolinkage is the apparent linkage and parallel segregation of loci located on different chromosomes, which appear to segregate independently in females (Gilbey *et al*, 2004). The frequency of pseudolinkage is debated in the literature, and may be influenced by the parental strains used in locus detection. Pure strain fish are reported to show reduced levels of pseudolinkage compared to inter-strain hybrids, which is possibly due to the greater genomic compatibility between chromosome pairs in pure strains compared to inter-strain hybrids (Sakamoto *et al*, 2000; Danzmann *et al*, 2008; Moen *et al*, 2008). The occurrence of pseudolinkage in the Atlantic salmon genome is thought to be less frequent compared to other salmonid species (such as brown trout) (Danzmann *et al*, 2008; Moen *et al*, 2008; Lubieniecki *et al*, 2010). For example, inter-strain trout species appear to show pseudolinkage even in females (Davisson *et al*, 1973).

Double reduction is a frequently reported phenomenon amongst tetraploid species, and occurs as a result of crossovers between the locus of interest and the centromere of chromosomes within a quadrivalent (Levings and Alexander, 1966). This results in the collateral segregation of loci originally on sister chromatids to the same gamete. For example, when considering a cross between parents of genotype Aaaa and aaaa, offspring with the AAaa genotype can only be produced as a result of double reduction gametes (Allendorf and Danzmann, 1997). The likelihood of

observing a double reduction event at a particular locus is dependent on its position relative to the centromere, since the likelihood of a recombination event occurring between the centromere and locus of interest is increased with distance. Therefore, loci at the distal ends of chromosomes are more likely to show a double reduction event. The hypothesised higher recombination rates at telomeric regions of chromosomes in Atlantic salmon males may increase the likelihood of observing an offspring which derives from a male gamete produced as a result of a double reduction event (Levings and Alexander, 1966; Allendorf and Danzmann, 1997; Luo *et al*, 2004).

Existing Atlantic salmon linkage maps also highlight a marked difference in the distribution of putative crossover events between the sexes. Generally, equal marker dispersion is observed along female chromosomes, in contrast to telomere-specific recombination and little or no recombination at centromeric regions of male chromosomes (Gilbey *et al*, 2004; Moen *et al*, 2004a; Lien *et al*, 2011). Male maps have been reported to show 80–90% lower recombination frequencies between adjacent markers compared to females (Gilbey *et al*, 2004), with the majority (>80%) of male recombination events taking place at the distal ends of chromosomes (Lubieniecki *et al*, 2010).

In contrast to other studies, recent high-density SNP linkage maps with increased marker density at the telomeric regions of chromosomes have reported much lower estimates of recombination ratios between the sexes [male:female 1.38:1, Lien *et al* (2011)]. In congruence with previously published studies, these denser maps suggest that the distribution of recombination events between the sexes is markedly different, with male recombination events being more concentrated at the telomeres.

The duplicated and highly repetitive nature of the Atlantic salmon genome, combined with the differences in recombination rate between the sexes, can complicate the generation of genomic resources. The presence of long range (~1,500) repetitive elements, together with the short read lengths obtained from current sequencing technologies which are not yet long enough to span repetitive regions,

has hampered the assembly and annotation of a high-quality salmon reference genome [ASalBase, http://www.asalbase.org/sal-bin/index; Davidson *et al* (2010)]. In addition, the recent salmonid-specific genome duplication presents a problem for the identification of genetic markers for use in marker-assisted breeding programs. For example, the differentiation of true genomic polymorphisms (such as SNPs) from those appearing as polymorphic due to originating from paralogous regions of the genome is difficult, and may only be inferred based on excess genotype heterozygosity (Guryev *et al*, 2006; Hohenlohe *et al*, 2011). Furthermore, the construction of linkage maps based on male linkage analyses is difficult due to the potential incorrect linkage group assignment of pseudo-linked markers, and the reduced recombination rate and confidence of marker ordering within linkage groups. As a result, marker orders and positions are more reliably estimated in female-specific linkage maps [e.g. Moen *et al* (2004a), Moen *et al* (2008)].

Despite these challenges, the genomic resources for Atlantic salmon are amongst the most extensive of all aquaculture species, and include several genetic maps, a physical map, an extensive EST database of approximately 500,000 tags, several microarrays, and a recent dense (~130 K) SNP array (Moen *et al*, 2004a; Rise *et al*, 2004; Ng *et al*, 2005; Davidson *et al*, 2010; Leong *et al*, 2010; Gidskehaug *et al*, 2011; Brenna-Hansen *et al*, 2012; NCBI, 2013a; NCBI, 2013b; Houston *et al*, 2014a). The Atlantic salmon genome is also in the process of being sequenced and assembled (first draft assembly: NCBI Assembly GCA_000233375.1; www.ncbi.nlm.nih.gov/Traces/wgs/?val=AGKD01; latest assembly: http://www.ncbi.nlm.nih.gov/assembly/GCA_000233375.3).

Linkage maps currently available for Atlantic salmon include those based on amplified fragment length polymorphisms (AFLPs), microsatellites, and more recently, SNPs (Gilbey *et al*, 2004; Moen *et al*, 2004a; Moen *et al*, 2008; Lien *et al*, 2011). Microsatellites are generally most informative for linkage mapping due to their variability and multi-allelic nature, and tend to be the preferred markers for first-pass linkage map construction in non-model species [e.g. Zhu *et al* (2014)]. However, bi-allelic SNPs are increasingly being utilised due to their abundance in

the genome and amenity to accurate high-throughput scoring [e.g. Barchi *et al* (2011), Helyar *et al* (2011), Hohenlohe *et al* (2011), Kakioka *et al* (2013), Ogden *et al* (2013)]. The most recent salmon linkage map is comprised of 5,650 SNP markers in 29 linkage groups (Lien *et al*, 2011). However, the high-density and high-quality genomic resources required for more effective QTL mapping studies for use in selective breeding programs are still lacking and need to be generated.

## 1.7 Generation of genomic resources in non-model organisms

Prior to the advancements in next-generation sequencing technologies, the generation of genomic resources (such as high-density linkage maps) for non-model organisms without a reference genome and for which the characteristics of the genome (such as size, GC content, repetitive content, patterns of locus segregation, etc.) are unknown was a highly labour-intensive and costly process (Helyar *et al*, 2011). For example, specifically for linkage map construction, depending on the type of genetic marker being identified (AFLP, RAPD, microsatellite, etc.), laboratory based testing of random DNA primers, PCR amplification, bacterial cloning procedures, testing of markers for the elimination of false-positives, and the genotyping of validated markers across a larger number of individuals for the construction of genetic linkage maps, may have been necessary (Griffiths and Tiwari, 1993; Naqvi *et al*, 1995; Korpelainen *et al*, 2007; Visendi *et al*, 2013). As well as the extensive amount of research effort and time required, the high error rates associated with the amplification/cloning procedures made the discovery and genotyping of markers in uncharacterised genomes difficult, and often generated a sparse marker set [e.g. Gilbey *et al* (2004), Moen *et al* (2004a), Guyomard *et al* (2012)].

The recent advances in high-throughput sequencing have greatly facilitated the study of genetics and genomics across all species (Willing *et al*, 2011; Deschamps *et al*, 2012; Poland and Rife, 2012; Recknagel *et al*, 2013a; Cruaud *et al*, 2014). In part, this has been achieved by the coupling of existing methods of genome interrogation (i.e. restriction enzymes) with the new sequencing platforms into what is now known as genotyping-by-sequencing (GBS). GBS first requires the generation of reduced

representation libraries of the genome under study, using restriction enzyme digests of genomic DNA. DNA fragments from restriction enzyme digests are then sequenced at high depth across a pooled sample of individuals. Alignment of reads within and across individuals enables the concurrent identification and genotyping of markers across all individuals within the study (Davey *et al*, 2011).

GBS technologies have many advantages over traditional sequencing studies, including: the ease of sample library preparation for sequencing, absence of cloning procedures, improvements in error rates of sequencing machines, potential of pooling barcoded individual samples, reduction in experimental costs, and the possibility of obtaining large volumes of data from single sequencing experiments (Deschamps *et al*, 2012; Poland and Rife, 2012; Hipp *et al*, 2014). GBS technologies have proven useful for studies in model organisms for which a dense genomic resource already exists (Aslam *et al*, 2010; Bruneaux *et al*, 2013; Pavlopoulos *et al*, 2013; Jessri and Farah, 2014). In non-model organisms, the concurrent identification and genotyping of markers across individuals, coupled with the fact that no prior knowledge of the genome is required, has encouraged the widespread use of GBS technologies (Willing *et al*, 2011; Deschamps *et al*, 2012; Poland and Rife, 2012; Recknagel *et al*, 2013a; Cruaud *et al*, 2014).

Several methods of GBS exist, and include complexity reduction of polymorphic sequences (CRoPS) and low coverage genotyping (Davey *et al*, 2011). One of the most popular of these, which is being applied in both model and non-model aquatic species, is Restriction site-Associated DNA sequencing (RAD-Seq). The RAD-Seq methodology was first developed by Baird *et al* (2008), and has since evolved to accommodate different experimental designs, restriction enzyme types (depending on the frequency of the restriction enzyme recognition site in the genome and length of DNA fragments required), library preparation methods, next-generation sequencing platforms, single- or paired-end sequencing of fragments, and downstream bioinformatic pipelines (Miller *et al*, 2007; Catchen *et al*, 2011; Davey *et al*, 2011; Etter *et al*, 2011; Willing *et al*, 2011; Fraser and Davey, 2012; Peterson *et al*, 2012; Catchen *et al*, 2013).

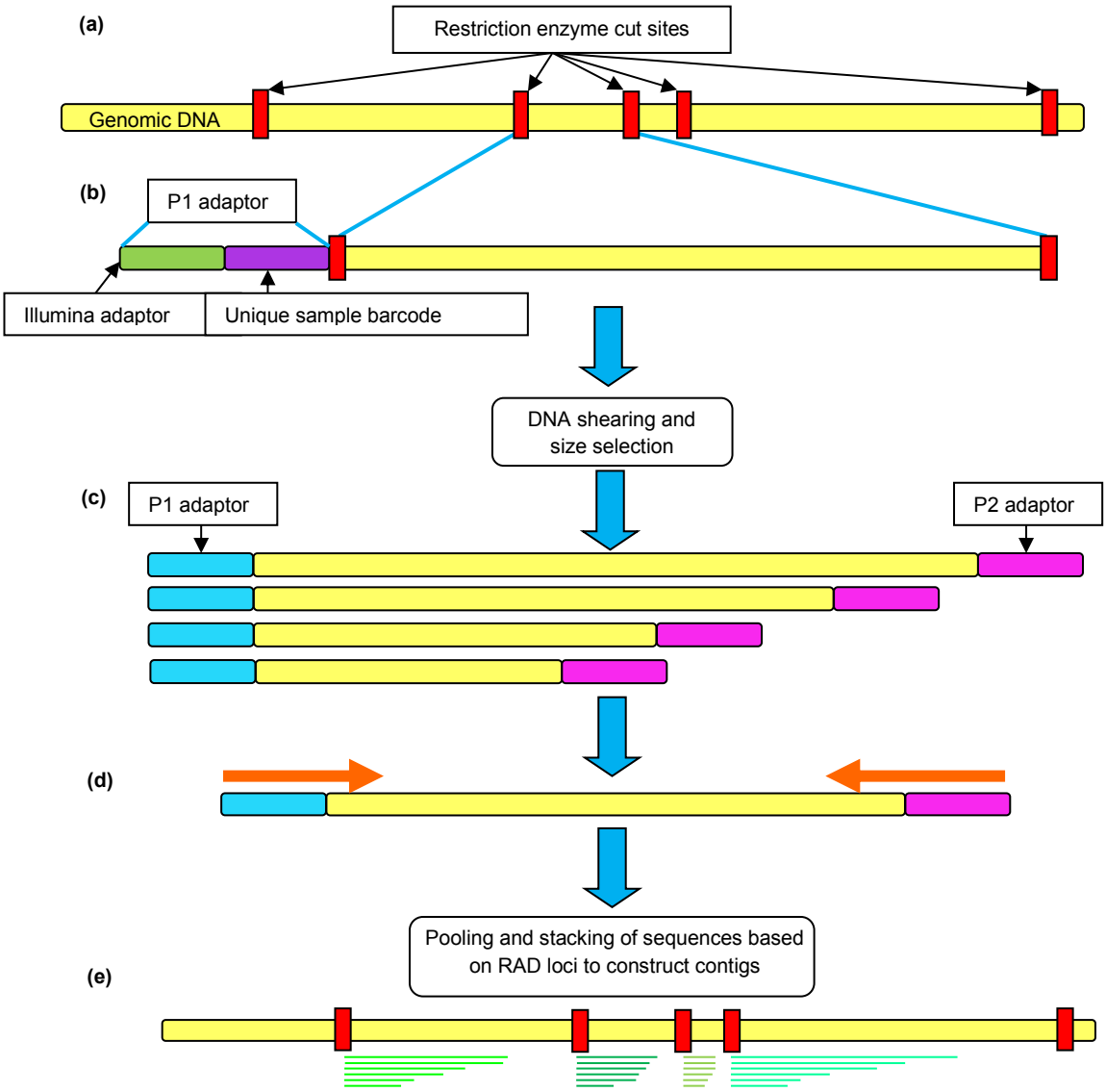



**Figure 1.3: Generation and assembly of paired-end sequence contigs from Restriction site-Associated DNA (RAD) Sequencing**

(a) DNA is extracted and subjected to digestion by the restriction enzyme of choice; (b) A P1 adaptor, composed of an Illumina adaptor and a unique sample barcode, is ligated on to the end of the fragments produced by restriction enzyme digest. P1-ligated fragments are pooled across samples then randomly sheared and size selected (typically 200–500bp) using gel electrophoresis; (c) Fragments with a P1 adaptor are identified, and a P2 adaptor is attached on to the sheared end of the fragment; (d) Fragments are sequenced from both adaptors on the Illumina sequencing platform; (e) Sequence reads are demultiplexed by barcode and are grouped within individual samples based on the RAD locus from which they originate. Stacking then collapsing of reads belonging to the same RAD locus within then across samples generates 95bp mini contigs around the original restriction enzyme cut site. Multiple reads from the paired-end sequence in the same region allows the creation of high-quality paired-end contigs of up to 500bp in length. Reads can be positioned by alignment to a reference genome (if available). Assembled contigs can be utilised in the identification of SNPs across all sequenced samples (Baxter *et al*, 2011; Rowe *et al*, 2011; Willing *et al*, 2011).

The paired-end RAD-Seq methodology using an Illumina sequencer is implemented as follows [also see Figure 1.3, Chapter two Materials and Methods, and Houston *et al* (2012)]. Briefly, individual genomic DNA samples are digested using the restriction enzyme of choice. Resulting DNA fragments are ligated to a P1 adaptor, which is comprised of an Illumina sequencing primer and a unique sample barcode. Fragments are then pooled across individuals, sheared and size selected [typically to between 200–500 base pairs (bp)], and a P2 adaptor is ligated on to the sheared end of the fragment. High coverage sequencing of fragments is performed using the Illumina sequencing platforms. Raw reads from the sequencing experiment are demultiplexed into individual samples based on barcode, and reads are grouped into RAD loci. Assuming no polymorphisms at a restriction cut site, the use of restriction enzyme digests enable the sequencing of homologous RAD loci across individuals. Stacking of reads across individuals for each RAD locus generates a small consensus sequence contig (95bp if single-end, ~500bp if paired-end) and enables the identification of thousands of polymorphisms (typically SNPs) segregating within the population of interest (Catchen *et al*, 2011; Davey *et al*, 2011; Etter *et al*, 2011; Willing *et al*, 2011; Catchen *et al*, 2013).

RAD-Seq-derived SNP and associated flanking sequences have been utilised in non-model aquatic organisms for a variety of purposes, including QTL mapping, QTL fine-mapping, candidate gene discovery [e.g. Houston *et al* (2012), Gagnaire *et al* (2013), Hohenlohe *et al* (2010)], linkage mapping [e.g. Everett *et al* (2012)], genome assembly and characterisation (Barchi *et al*, 2011; Baxter *et al*, 2011; Brieuc *et al*, 2014; Penaloza *et al*, 2014), evolutionary and conservation studies (Seeb *et al*, 2014), and for the inference of evolutionary relationships [e.g. Rubin *et al* (2012), Eaton and Ree (2013), Hipp *et al* (2014)].

## 1.8 Aims of the thesis and chapter descriptions

Overall, there is a clear need to improve our understanding of the Atlantic salmon genome, and how genetic variation between individuals influences the response to selection for traits of economic value. This can be achieved through the generation of high-density Atlantic salmon genomic resources (such as linkage maps) and analyses

towards the identification of QTL underlying variation in traits of economic value. With the implementation of next-generation sequencing technologies and methodologies, the generation of genomic resources for non-model species such as Atlantic salmon is becoming more feasible. In the absence of a reference genome, alternative methods of analysing the Atlantic salmon genome (e.g. based on orthology to reference genomes of closely related species), need to be explored.

This thesis presents the application of next-generation sequencing technologies (in particular RAD-Seq) in: (i) the characterisation of the Atlantic salmon genome; (ii) the exploration of the genetic architecture underlying variation in traits of economic value in farmed populations; (iii) the generation of genomic resources; and (iv) the investigation of genome evolution through comparative orthology analyses.

First, this thesis attempts to address the problem of genomic resource generation in Atlantic salmon. In chapter two, the construction of the first high-density RAD-Seq-derived SNP linkage map for Atlantic salmon is described. To gain a better understanding of the specific genomic characteristics of Atlantic salmon (such as the salmonid-specific genome duplication and the disparity in recombination rates between the sexes), analyses of mapped SNP marker positions (i) between Atlantic salmon males and females, and (ii) across other fish species for the inference of orthologous relationships, were conducted. To enable further use of the constructed map in QTL fine-mapping and the characterisation of QTL through the identification of candidate/causative variant(s), analyses towards the identification of gene-associated mapped markers are presented.

Second, this thesis presents the exploration of the host genetic basis of resistance to two of the top five most problematic diseases on Atlantic salmon farms: IPN and PD. At present, the understanding of the genetic control of host response to these viral infections is at different stages. For PD, a moderate heritability for resistance has been estimated. However, very little else is known about the underlying genetics of host resistance. Chapter three of this thesis presents the quantification and characterisation of the genetic variation in resistance to PD in Atlantic salmon, using

survival data from a fry population challenged with SAV. To determine whether genetic variation for PD resistance exists within the population, the heritability of resistance was estimated. To further characterise this variation and identify QTL-marker associations for use in selective breeding programs, QTL mapping and association studies were undertaken. For IPN, although a major QTL involved in resistance has been identified and repeatedly verified in independent populations, the mechanisms behind host response which are underlying this QTL [i.e. causative variant(s) and pathways involved in host response] are unknown. Chapter four outlines the approaches taken in an attempt to further characterise the major IPN resistance QTL, through the identification of putative candidate genes and resistance-influencing pathways.

RAD-Seq is an exciting new technology, and has opened up many opportunities towards the understanding of non-model genomes. Although RAD-Seq data has been utilised across different areas of biological research, best practices for its use are still being explored. For example, the reproducibility of homologous RAD loci across populations and across species, and the performance of these sequences in evolutionary relationship estimation, is still much debated in the literature. To assess this, RAD-Seq datasets across populations and laboratories for a given species, and across ten distantly related teleost fish species was obtained from published studies. Using this data, in chapter five, the reproducibility of RAD loci across populations of the same species was investigated, and the ability to identify cross-species orthologous RAD loci was explored. To evaluate the potential use of these RAD loci in inferring evolutionary relationships, evolutionary relationships amongst the ten species were reconstructed and compared to published phylogenies.

Chapter six is a summary and exploration of the implications of the results presented within in a broader context. Potential future investigations to complement, advance and develop upon the results and ideas presented are discussed.

Appendices, acknowledgements and the bibliography are given at the end of the thesis.

# Chapter 2

# Linkage maps of the Atlantic salmon genome derived from RAD-Sequencing

## 2.1 Abstract

Genetic linkage maps are useful tools for mapping quantitative trait loci (QTL) influencing variation in traits of interest in a population. Genotyping-by-sequencing approaches, such as Restriction-site Associated DNA sequencing (RAD-Seq), now enable the rapid discovery and genotyping of genome-wide SNP markers suitable for the development of dense SNP linkage maps, including in non-model organisms such as Atlantic salmon. In this study, *Sbf*I RAD-Seq-derived SNP markers identified in two Atlantic salmon reference families were used to construct high-density sex-specific SNP linkage maps of the salmon genome. Approximately 6,500 SNPs were assigned to 29 linkage groups, utilising markers from known genomic locations as anchors. Resulting map lengths were comparable between the sexes. However, the distribution of the SNPs showed sex-specific patterns, with a greater degree of clustering of sire-segregating SNPs to single regions within linkage groups. Linkage maps were integrated with the Atlantic salmon draft reference genome contigs, allowing the unique assignment of ~4,000 contigs to a linkage group. 112 genome contigs mapped to two or more linkage groups, highlighting regions of putative homeology within the salmon genome. Sequence orthology-based analyses with the stickleback reference genome identified putative genes closely linked to approximately half of the ordered SNPs, and blocks of orthology between the Atlantic salmon, stickleback and rainbow trout genomes were identified. This high-density Atlantic salmon RAD-Seq linkage map and additional resources generated herein are valuable for salmonid genomics research, as RAD-Seq becomes increasingly common.

## 2.2 Introduction

Linkage maps are important resources for investigating a variety of different biological questions, in both model organisms with high-quality reference genomes (such as humans) and non-model organisms for which reference genomes are still under development, such as Atlantic salmon. Linkage maps currently available for Atlantic salmon include those based on amplified fragment length polymorphisms (AFLPs), microsatellites, and more recently, single nucleotide polymorphisms (SNPs) (Gilbey *et al*, 2004; Moen *et al*, 2004a; Moen *et al*, 2008; Lien *et al*, 2011). These maps have been utilised in a variety of studies, including in the identification of cross-species chromosome orthologies (Li *et al*, 2011) and in QTL mapping studies (Moen *et al*, 2004b; Moen *et al*, 2007; Houston *et al*, 2008). In addition, these maps have allowed partial characterisation of features of the Atlantic salmon genome, for example in the identification of homeologous regions of the genome originating from the recent salmonid-specific genome duplication (Danzmann *et al*, 2005; Danzmann *et al*, 2008; Phillips *et al*, 2009), and in investigations of the large difference in recombination rates and distribution of recombination events between males and females (Moen *et al*, 2004a; Lien *et al*, 2011).

In the absence of an Atlantic salmon reference genome, a more detailed characterisation of the Atlantic salmon genome and further fine-mapping of QTL identified in population genetics analyses may be achieved using high-density linkage maps. In addition, high-density linkage maps can assist in the mapping and assembly of genome contigs to chromosomes for non-model organisms such as Atlantic salmon, where a genome sequencing project is underway (Davidson *et al*, 2010). At present, the majority of high-density linkage maps are constructed using SNP markers [e.g. Ryynanen and Primmer (2006), Aslam *et al* (2010), Everett *et al* (2011)], due to their higher frequency in the genome compared to other marker types. However, in the absence of reference genomes and other genomic resources, the initial identification of SNP markers can be difficult (Moen *et al*, 2008; Lien *et al*, 2011; Houston *et al*, 2014a).

Recently, genotyping-by-sequencing (GBS) technologies are offering an alternate, efficient and cost-effective platform for the generation of dense SNP panels in non-model organisms. GBS technologies, such as Restriction-site Associated DNA sequencing (RAD-Seq), enable the concurrent identification and genotyping of SNP markers across large populations of individuals within a single sequencing experiment [e.g. Baird *et al* (2008), Baxter *et al* (2011), Davey *et al* (2011), Etter *et al* (2011), Helyar *et al* (2011), Willing *et al* (2011), Peterson *et al* (2012)]. These methods generate denser marker sets (in comparison to microsatellite maps) for use in studies such as population genetics [e.g. QTL mapping (Houston *et al*, 2012; Gagnaire *et al*, 2013), GWAS (Slavov *et al*, 2014)] and comparative orthology analyses (Kakioka *et al*, 2013). RAD-Seq has the added advantage of providing short SNP-flanking sequence contigs [~95 base pairs (bp) in single-end, and up to ~500bp in paired-end RAD-Seq]. These ~500bp mini-contigs can be utilised in genome assembly and assignment of large genome contigs to chromosomes (Dasmahapatra *et al*, 2012), and in the identification of candidate genes in close proximity to significant SNPs from population genetic studies (Hegarty *et al*, 2013).

As yet, no RAD-Seq SNP linkage map exists for Atlantic salmon. Therefore, the main aim of this chapter was to construct a high-density SNP linkage map of the Atlantic salmon genome, using SNP markers derived from *Sbf*I RAD-Seq applied in two Atlantic salmon reference families. Additional aims were to utilise this linkage map to: (i) investigate the differences in recombination rate and distribution between males and females; (ii) integrate the new RAD-Seq linkage map with existing linkage/physical maps and the draft Atlantic salmon reference genome; (iii) identify putative genes proximal to the SNPs in the linkage map using comparative genomics; and (iv) investigate the salmonid genome duplication by comparative orthology analysis to the rainbow trout and stickleback genomes.

## 2.3 Methods

SNP identification and genotyping was conducted prior to the start of this project, using a RAD-Seq approach in two reference SalMap families (Br5 and Br6) (Danzmann *et al*, 2005). This is described briefly below, and a summary of the

general RAD-Seq protocol and bioinformatic pipeline is given in Figure 1.3. All subsequent SNP filtering, linkage map construction and genome characterisation analyses were conducted by me.

### 2.3.1 RAD library preparation and sequencing

The two SalMap families (Br5 and Br6) (Danzmann *et al*, 2005) used in this study are from a salmonid genetics resource population, and studies using these samples have been previously published (Danzmann *et al*, 2008; Phillips *et al*, 2009; Andreassen *et al*, 2010; Lubieniecki *et al*, 2010; Lukacs *et al*, 2010; Quinn *et al*, 2010; Lien *et al*, 2011). Therefore, no new biological experiments or sampling was carried out for this study. Genomic DNA samples for the fish in these two families were obtained (two parents and 46 offspring per family, total n=96) and quality checked [quantification using spectrophotometry (Nanodrop); agarose gel electrophoresis to confirm genomic integrity]. RAD libraries for each individual were prepared according to the methodology described in Etter *et al* (2011) with modifications as described in Houston *et al* (2012).

Briefly, each sample (1.5µg DNA per sample for parent libraries/0.25µg DNA per sample for offspring libraries) was digested with *Sbf*I-HF (NEB) (recognition cut site 5‘ CCTGCA/GG 3’ and 3’ GG/ACGTCC 5’). Within each library, a P1 adaptor containing an individual-specific nucleotide barcode was ligated to the digested DNA fragments of each sample. Details of the library composition and nucleotide barcodes are given in Appendix A, Table A1. Samples within each library were pooled into eight RAD libraries, with two parent libraries (n=2 per library) and six offspring libraries (each n=14–16). Since each library was subsequently sequenced on an individual lane of the Illumina HiSeq 2000, this design ensured higher sequence coverage of the parents compared to the offspring (Appendix A, Table A1). Fragments were sheared (Covaris S2 sonicator; Covaris Inc., Woburn, USA), size selected (agarose gel electrophoresis; size range: 250–500bp) and ligated to a P2 adapter. All eight libraries were amplified (18 cycles of PCR amplification), and a final size selection of fragments was conducted (Agilent Bioanalyser electrophoresis; size range: 300–500bp) prior to sequencing at the GenePool Genomics Facility,

University of Edinburgh (http://genepool.bio.ed.ac.uk/; now part of Edinburgh Genomics).

### 2.3.2 RAD-Seq bioinformatic pipeline and SNP identification

SNP genotype data for all individuals was generated as follows. Raw Illumina reads were 'demultiplexed' and assigned to individual samples according to their nucleotide barcode [RADpools v1.2.1 (Fraser and Davey, 2012)], resulting in an individual FASTQ file per animal. Raw reads originating from the same *Sbf*I cleavage site ('RAD locus') within each individual were grouped (allowing only a single base mismatch), and the consensus sequence at each side of the *Sbf*I cleavage sites was generated using ustacks and cstacks v0.992 (Catchen *et al*, 2011). Paired-end sequences at each *Sbf*I flanking site were assembled using clc assembly cell v3.22 and aligned back to the assembly using stampy 1.0.13 (Lunter and Goodson, 2011) ('PE contigs'). PCR duplicates were detected with Picard MarkDuplicates v1.55 (http://picard.sourceforge.net/) and excluded.

Overall, 482,547 consensus RAD contigs were generated, of which 366,219 were from the RAD loci and 116,328 were from the PE contig. 2% of RAD loci had more than one associated PE contig. Of the 366,219 RAD loci, 76,034 were identified in at least 50 individuals (of 96), and these were retained for further analyses. 85,725 PE contigs were associated with these RAD loci. SNPs within these RAD loci and PE contigs were called using samtools v0.1.18 (Li *et al*, 2009) and filtered using 'vcfutils' to ensure a minimum overall read count at the locus of 500, a maximum of 2,000 (to exclude potential repeat regions) and an overall SNP quality score of 60. Individual SNP genotypes with a quality score of at least 20, a read depth of at least 6, and with genotypes in both parents of a family were retained. SNP genotypes which failed the quality check were assigned no call.

### 2.3.3 SNP genotype quality control and filtering

The RAD-Seq bioinformatic pipeline described above resulted in a set of 28,415 putative SNPs, originating from both single- and paired-end consensus sequences. Due to variation in sequence coverage between individuals, there was a large number

of missing genotypes in the dataset (Figure 2.1). In addition, paralogous sequence variants (PSVs) within duplicate regions of the genome with a very high sequence similarity will be retained by the pipeline above.

Therefore, the data were filtered to remove (i) individuals and SNPs with excess missing data (RQTL, http://www.rqtl.org/), (ii) putative PSVs (inferred by excess heterozygosity; PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/) and (iii) apparent Mendelian errors (VIPER, http://bioinformatics.roslin.ac.uk/viper/). Given parental genotypes, a Mendelian error was defined as a highly improbable offspring genotype at a given SNP. The thresholds set for each of the filtering criteria were as follows. Individuals with <25% of SNPs genotyped and/or >200 Mendelian errors were removed. SNP markers with <50% of individuals genotyped, ≥2 Mendelian errors, and/or showing >70% heterozygosity across both mapping families, were removed from further analyses. SNPs with a single Mendelian error were set to missing only for the genotype in question (Table 2.1).

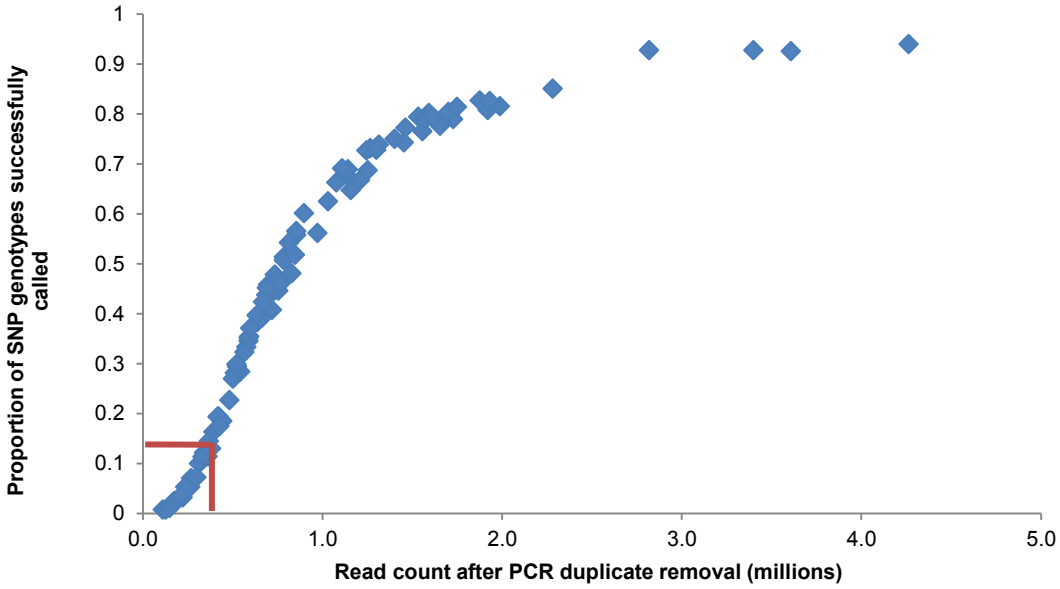



**Figure 2.1: Relationship between read depth and call rate**
The number of reads per individual following exclusion of PCR duplicates (x-axis) plotted against the proportion of SNP genotypes successfully called for all putative SNPs (y-axis). The red lines on the graph indicate the thresholds below which individuals were removed due to an excess of missing genotypes.

**Table 2.1: SNP and individual filtering procedure**

| Filtering step | Number of SNPs eliminated | Number of SNPs remaining | Number of individuals eliminated | Number of individuals remaining |
|---|---|---|---|---|
| Raw RAD-Seq processing | 0 | 28,415 | 0 | 96 |
| Missing genotypes–SNPs (>50%) | 14,778 | 13,637 | 0 | 96 |
| Missing genotypes–Individuals (>25%) | 0 | 13,637 | 15 | 81 |
| Excess heterozygosity–SNPs (PSVs; >70%) | 4,895 | 8,742 | 0 | 81 |
| Mendelian errors–SNPs (≥2) | 485 | 8,257 | 0 | 81 |
| Mendelian errors–Individuals (>200) | 0 | 8,257 | 4 | 77 |

Stringent quality control filtering was applied to the initial set of 28,415 putative SNPs generated from the processing of raw RAD-Seq reads. Filtering parameters for SNPs included removing excess missing data (>50%), excess Mendelian errors (≥2 individuals) and excess heterozygosity (putative PSVs; >70%). The final number of SNPs left for linkage map construction was 8,257. Individuals were removed if they showed excess missing genotypes (>25%) and/or excess Mendelian errors (>200 SNPs). 77 individuals remained for linkage map construction post-filtering.

### 2.3.4 Linkage map construction

Linkage maps using the filtered set of SNP markers were constructed separately for the four parents in the two mapping families, using a two-stage process. In stage one, SNPs were clustered into putative linkage groups based on linkage relationships with 116 anchor markers selected from existing studies (microsatellites, minisatellites, allozymes and SNPs; Appendix A, Table A2). These markers had previously been genotyped and used in linkage map construction for the four parents in the two reference SalMap families (Danzmann *et al*, 2005). Markers were chosen so that at least one informative marker was present within each linkage group. To enable ease of comparison between this map and that of Danzmann *et al* (2005), the nomenclature and numbering of linkage groups was retained as expressed in Danzmann *et al* (2005), where the 29 Atlantic salmon linkage groups are labelled 1 to 32, with linkage groups 26, 27 and 29 absent (due to later resolution of joining of linkage groups previously identified as two separate groups). The two-point linkage between anchor markers and the RAD-Seq SNPs was calculated using the 'twopoint' option of the CRI-MAP software package [Green *et al* (1990); version 2.4 as modified by Xuelu Liu (Monsanto)]. Based on these two-point linkage LOD scores, SNP markers were assigned to linkage groups using the 'autogroup' option, starting at a LOD of 40 and applying a stepwise decrease in LOD score threshold to a minimum of 4.

In stage two, the segregation type (aaXaa; aaXbb; aaXab; abXaa; abXab) of the linkage group-assigned SNPs within each family was determined, and SNPs showing informative segregation patterns for linkage map construction within family (aaXab–female parent segregating marker; abXaa–male parent segregating marker) were identified. The best estimated order of informative SNPs (without the inclusion of anchor markers) on each linkage group was calculated using the 'order.seq' algorithm of the OneMap software package [CRAN package OneMap; Margarido *et al* (2007); modified for parallelised computing by Marcelo Mollinari (Department of Genetics, University of São Paulo)], with the following parameters: n.init=5, THRES=4, draw.try=FALSE. SNP marker order was confirmed as the best order using the 'ripple.seq' function of OneMap, with a word size of 7 and applying a LOD

threshold of 4. The map position [in centiMorgans (cM)] for each SNP was calculated according to the Haldane mapping function (assumes independence and no interference between adjacent recombination events). The full OneMap R script written for linkage map construction in this study is given in Appendix B.

Despite the stringent filtering parameters applied, marker genotypes may still contain errors. These can appear as putative recombination events and may result in erroneous positioning of SNPs some distance removed from other markers at the distal ends of linkage groups. Therefore, maps for each parent and each linkage group were investigated manually, and SNPs with a gap of greater than 30cM in male maps and 20cM in female maps from the neighbouring SNP at the ends of the linkage groups were removed. Resulting maps were visualised using the MapChart software package (Voorrips, 2002).

### 2.3.5 Recombination ratios and comparison of marker distribution between the sexes

Linkage maps were constructed for the four mapping parents individually, based on sex-segregating markers within family. Therefore, male and female maps within families did not contain the same markers, and direct estimation of recombination ratios based on relative marker positions was not possible. Instead, recombination ratios were estimated by comparing relative map lengths.

To compare the distribution of markers between the sexes, maps for each linkage group were split into intervals of equal size, and the number of markers within intervals was compared. To define the intervals, the shorter of the parental maps was split into 5cM intervals. If map length was not long enough to produce at least 5 x 5cM intervals, a smaller cM interval was chosen. The longer parental map was then split into an equal number of intervals. For example for family Br5, linkage group 1, the male map was shorter than the female map (85cM and 229cM respectively). Therefore, the male map was split into 18 intervals of 5cM. The female map was then split into 18 marker intervals of approx. 13cM each. The percentage of SNPs mapping to each interval for both sexes was calculated, and the five intervals with

the highest percentage of markers were identified. The averages of the percentages in the top five intervals for each sex across all linkage groups and both families was calculated and compared.

### 2.3.6 Assignment of Atlantic salmon reference genome contigs to linkage groups using mapped RAD SNPs

To assign Atlantic salmon reference genome contigs to linkage groups, both the sequences associated with mapped SNPs ("mapped RAD contigs"), and the Atlantic salmon genome contigs (NCBI Assembly GCA_000233375.1; www.ncbi.nlm.nih.gov/Traces/wgs/?val=AGKD01), were repeat-masked using the RepeatMasker software package [RepeatMasker package; Smit *et al* (1996-2010)] and the "Salmon Raw Repeat DB v1.6" database (http://web.uvic.ca/grasp/) and aligned [BLASTN, BLAST+ version 2.2.25+; Zhang *et al* (2000)]. Alignment significance was taken at threshold E-values of $1e^{-30}$ for RAD loci and $1e^{-80}$ for PE contigs. The difference in chosen thresholds was due to the sensitivity of the reported BLAST E-value to variations in sequence length (Rognes, 2001; Agostino, 2012). RAD loci or PE contigs aligning to multiple (>2) reference genome contigs were excluded as potential uncharacterised repeat regions.

### 2.3.7 Identification of SNP-associated putative genes

To identify SNPs within or close to putative genes, a two-stage strategy based on sequence orthology to all known three-spined stickleback (*Gasterosterus aculeatus*) gene nucleotide sequences was implemented [stickleback gene sequences downloaded from Ensembl BioMart http://www.ensembl.org/biomart/martview/; Database=Ensembl Genes 72, Dataset=*Gasterosterus aculeatus* genes (BROADS1)]. In the first stage, the mapped RAD contigs were directly screened for sequence similarity to stickleback genes using TBLASTX (E-value<$1e^{-5}$). TBLASTX was chosen as it is more sensitive to protein homologies between distantly related species using DNA sequence data, since it translates sequences in all three frames before alignment, thus overcoming problems of detecting open reading frames in the context of frame shifts [e.g. see Parra *et al* (2003), Weng *et al* (2005)].

In the second stage, to detect mapped RAD SNPs in close proximity to, but not within, a stickleback gene ortholog, the linkage group-assigned and repeat-masked Atlantic salmon reference genome contigs were aligned to the stickleback gene sequences (TBLASTX, E-value for significance<$1e^{-5}$). Only the two most significant stickleback gene alignments were retained, in an attempt to avoid spurious alignment to multiple genes from different stickleback linkage groups (for example, due to sequence similarity of genes from the same gene family).

### 2.3.8 Identification of chromosomal orthologous relationships between Atlantic salmon and stickleback

To identify orthologous relationships between Atlantic salmon and stickleback linkage groups, mapped RAD contigs originating from the same RAD locus and showing significant alignment to a stickleback gene were grouped into a single RAD locus. This was done to ensure that each RAD locus was counted once only, so as to prevent bias in the inference of stickleback–Atlantic salmon orthologies.

For each Atlantic salmon linkage group, the number of RAD loci showing significant alignment to a gene on a particular stickleback linkage group was counted. One-to-one orthologous relationships between linkage groups of the two species was assigned only if the number of significant alignments was twice (or more) than the number of significant alignments to genes on any other stickleback linkage group, and if this relationship was seen across all four mapping parents. The only exceptions to this were in cases where RAD loci on a single Atlantic salmon linkage group showed an equal number of alignments to two stickleback linkage groups. In these cases, both stickleback linkage groups were assigned to that salmon linkage group.

## 2.4 Results

### 2.4.1 RAD-Sequencing

A total of 96 individuals belonging to the two SalMap reference families (denoted Br5 and Br6; 46 offspring and two parents per family) were sequenced and genotyped using paired-end RAD-sequencing. To maximise the chances of detecting segregating SNPs in the parents, parent libraries were sequenced at a substantially

higher depth than the offspring (Appendix A, Table A1). The average sequencing depth was 11 million reads per parent and 2 million reads per offspring, which was reduced to 3.5 million and 0.8 million following the removal of putative PCR duplicates [see Davey *et al* (2011); Appendix A, Table A1].

Following the merging of reads into RAD loci across individuals, 76,034 distinct RAD loci were detected, which is indicative of 38,017 *Sbf*I cleavage sites in the Atlantic salmon genome. This number is comparable to a previous RAD-Seq study in families of farmed fish (Houston *et al*, 2012), and 99% of the RAD loci were common to both studies (see chapter 5). This demonstrates that *Sbf*I RAD-Seq is sampling the same sites in the Atlantic salmon genome across wild and farmed populations, with positive implications for its reproducibility as a genotyping technique.

### 2.4.2 SNP discovery, filtering and genotyping

In total, 28,415 putative SNPs were discovered across the 76,034 RAD loci, with an overall genotyping rate of 50%. Of these, 11,103 were detected in the RAD loci and 17,312 were detected in the PE contigs. SNP genotypes were quality filtered (Table 2.1) to remove individuals and SNPs with a high level of missing data, Mendelian errors, or excess heterozygosity [suggestive of paralogous sequence variants (PSVs)].

Approximately half of the SNPs were eliminated due to missing genotypes in >50% of the individuals in the study, leaving 13,637 SNPs in total. 15 individuals genotyped at <25% of these remaining SNPs were removed from further analyses. The proportion of missing genotypes of an individual was inversely related to the sequence coverage for that individual (Figure 2.1). This was due to the removal of genotypes at individual SNP loci where sequence depth was below the threshold chosen to ensure high confidence in the genotype call (indicated by red lines in Figure 2.1). It is clear that a read depth of *ca.* 1 million reads (following removal of PCR duplicates) is required to ensure high levels of high confidence genotype calls for a given Atlantic salmon individual.

A total of 4,895 SNPs showing excess (>70%) heterozygosity across both families were removed from the dataset as putative PSVs [see Gonen *et al* (2014), Additional file 2]. These are a useful resource for excluding PSVs in future *Sbf*I RAD-Seq of Atlantic salmon and other salmonids. Following this, four individuals with >200 genotypes defined as Mendelian errors were removed from the analysis, and 485 SNPs with two or more Mendelian errors were discarded. 603 SNPs with one Mendelian error were set to missing for the genotype in question. The final filtered dataset consisted of 77 individuals (36 offspring and 2 parents in Br5; 37 offspring and 2 parents in Br6) and 8,257 SNP markers, with an overall genotyping rate of 76% (Table 2.1).

### 2.4.3 Linkage map construction

Following the SNP filtering process, the linkage arrangements between the remaining 8,257 putative SNPs were assessed. A total of 6,458 SNPs were assigned to 29 Atlantic salmon linkage groups (equal to the number of chromosomes; average 220 SNPs per linkage group) based on two-point linkage scores (LOD≥4) between the SNPs and a set of 116 anchor markers (details of anchor markers in Appendix A, Table A2), using the CRI-MAP software package [Green *et al* (1990); version 2.4 as modified by Xuelu Liu (Monsanto)] (Table 2.2, column 3). 5,787 of the linkage group-assigned SNPs were from the RAD loci and a further 671 were from the PE contigs. The lower number of PE SNPs is likely to be due to the lower sequence coverage of PE contigs compared to the RAD loci, as expected from the RAD-Seq protocol. Of the 6,458 linkage group assigned SNP markers, 3,640 and 3,699 in families Br5 and Br6 respectively showed informative segregation patterns for sex-specific linkage map construction (i.e. were either: heterozygous in the mother and fixed in the father, or heterozygous in the father and fixed in the mother; Table 2.3).

Using these sex-specific segregating markers, linkage maps were constructed individually for each of the four parents, using the OneMap software package [CRAN package OneMap; Margarido *et al* (2007), modified by Marcelo Mollinari (Department of Genetics, University of São Paulo)]. Approximately 1,400 SNP markers were ordered and positioned in each of the female parents, and the order and

position of approximately 1,800 SNP markers was estimated in each of the male parents [Table 2.2, columns 4–7; see Gonen *et al* (2014), Additional file 4; Figure 2.2]. No linkage map was constructed for linkage group 19 for the female parent of family Br6, due to very few female-segregating markers being assigned to this linkage group.

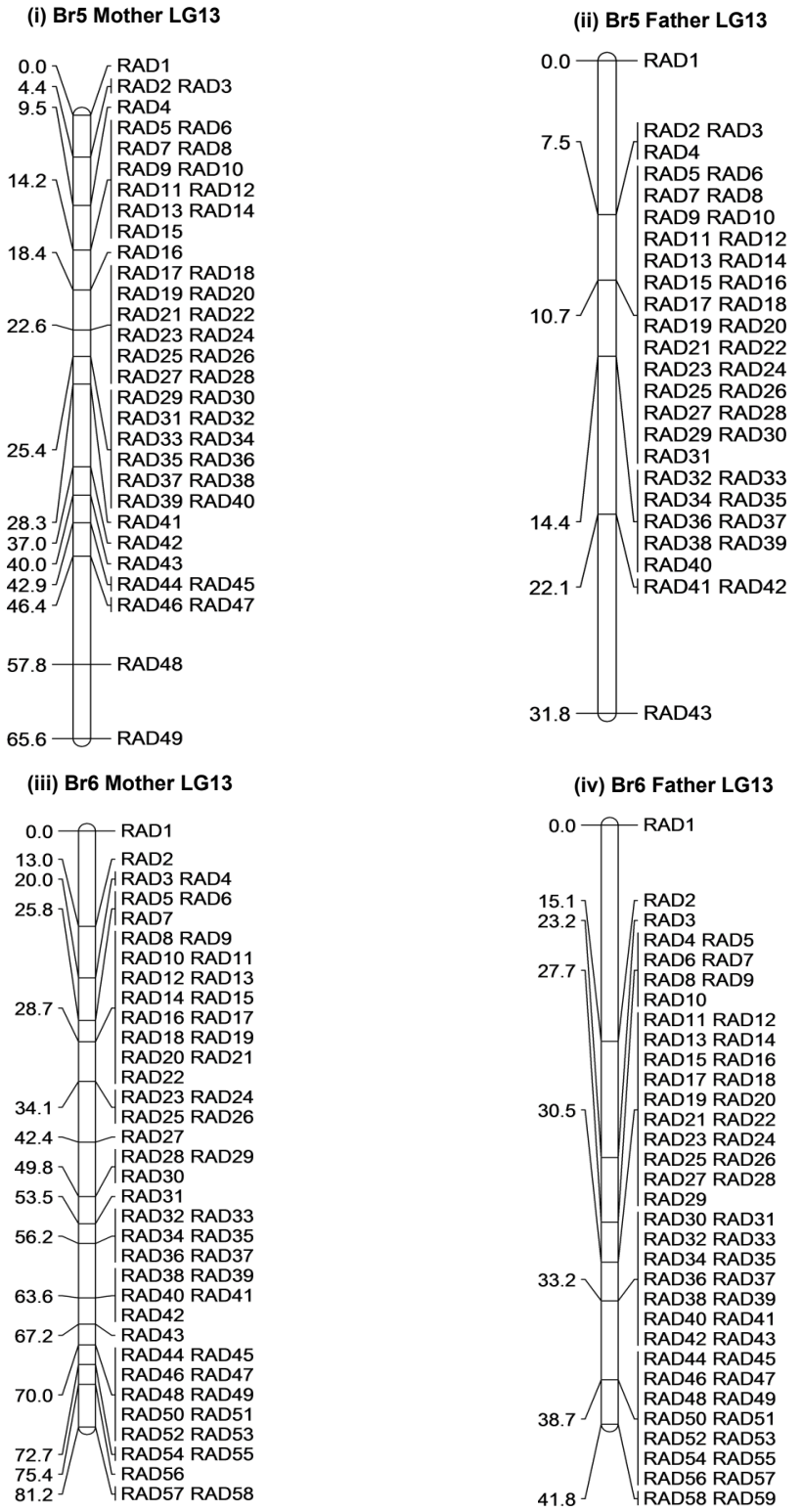



**Figure 2.2: Example linkage map**

Maps for linkage group 13 for: (i) Br5 mother; (ii) Br5 father, (iii) Br6 mother; (iv) Br6 father. Map lengths were shorter in males and markers were more widely spaced in the female maps. Marker names are coded as RAD1-RADX depending on the ordered position of the marker on the linkage group, and do not represent the same markers across individuals.

**Table 2.2: Number of SNPs assigned to linkage groups**

| Atlantic salmon linkage group | Atlantic salmon chromosome | Number of SNPs on linkage group (CRI-MAP) | Final number of SNPs ordered on each linkage group (OneMap) | | | |
|---|---|---|---|---|---|---|
| | | | Br5 Mother | Br6 Mother | Br5 Father | Br6 Father |
| 1 | 2 | 244 | 59 | 23 | 73 | 68 |
| 2 | 10 | 350 | 102 | 86 | 88 | 79 |
| 3 | 14 | 197 | 31 | 35 | 68 | 76 |
| 4 | 6 | 283 | 47 | 43 | 84 | 79 |
| 5 | 13 | 306 | 67 | 84 | 84 | 85 |
| 6 | 12 | 257 | 72 | 46 | 81 | 71 |
| 7 | 24 | 138 | 27 | 47 | 51 | 39 |
| 8 | 15 | 520 | 78 | 47 | 69 | 84 |
| 9 | 11 | 226 | 50 | 64 | 67 | 49 |
| 10 | 9 | 394 | 95 | 94 | 113 | 103 |
| 11 | 3 | 336 | 48 | 92 | 98 | 103 |
| 12 | 5 | 224 | 24 | 29 | 74 | 68 |
| 13 | 19 | 197 | 49 | 58 | 43 | 59 |
| 14 | 21 | 152 | 38 | 40 | 47 | 33 |
| 15 | 27 | 132 | 31 | 39 | 40 | 43 |
| 16 | 18 | 209 | 33 | 58 | 65 | 67 |
| 17 | 1 | 442 | 70 | 78 | 130 | 129 |
| 18 | 23 | 155 | 36 | 33 | 55 | 51 |
| 19 | 8 | 42 | 8 | 0 | 13 | 15 |
| 20 | 25 | 115 | 20 | 33 | 26 | 23 |
| 21 | 26 | 113 | 25 | 25 | 30 | 35 |
| 22 | 17 | 158 | 19 | 44 | 33 | 64 |
| 23 | 16 | 215 | 58 | 49 | 66 | 55 |
| 24 | 7 | 169 | 22 | 17 | 57 | 56 |
| 25 | 20 | 237 | 65 | 77 | 62 | 73 |
| 28 | 4 | 220 | 19 | 32 | 80 | 83 |
| 30 | 29 | 116 | 43 | 23 | 36 | 37 |
| 31 | 28 | 132 | 34 | 37 | 42 | 39 |
| 32 | 22 | 179 | 41 | 66 | 58 | 51 |
| **TOTAL** | - | 6,458 | 1,311 | 1,402 | 1,833 | 1,817 |

SNPs were assigned to a linkage group using previously mapped anchor markers in CRI-MAP. SNPs were then ordered on each linkage group using OneMap, for each mapping parent separately.

**Table 2.3: Number of linkage group-assigned SNPs (out of 6,458 SNPs) showing sex-specific segregation patterns and the total map length for each mapping parent**

| Mapping parent | No. of segregating SNPs | Total map length (cM) |
|---|---|---|
| Br5 Mother | 1,688 | 2,807 |
| Br5 Father | 1,952 | 2,170 |
| Br6 Mother | 1,804 | 2,358 |
| Br6 Father | 1,895 | 1,426 |

### 2.4.4 Recombination ratios and distribution of recombination events across the genome between males and females

One of the striking features of the Atlantic salmon genome is the large difference in recombination rate and distribution of recombination events observed between the sexes (Gilbey *et al*, 2004; Moen *et al*, 2004a; Danzmann *et al*, 2005; Lubieniecki *et al*, 2010; Lien *et al*, 2011). To investigate this phenomenon using the RAD-Seq linkage map, and to estimate the male-to-female recombination ratio, map lengths for each linkage group were compared for each parent within a family (Appendix A, Table A3).

For family Br5, resulting map lengths were 2,807cM for the female map, and 2,169cM for the male map, giving a recombination ratio of 1:1.3. This similarity of map length was generally consistent across most linkage groups, although for linkage groups 2, 21 and 31, the female map was longer (ratio>1:3; Appendix A, Table A3). For family Br6, the female map length was 2,358cM and the male map length was 1,426cM, resulting in a larger ratio of 1:1.7 compared to that estimated in family Br5. The larger ratio and smaller male map observed in family Br6 is likely related to two features of the Br6 male parent map. Firstly, the markers on linkage group 31 all clustered at 0cM (no ratio could be calculated for this linkage group). This clustering may be due to linkage group 31 being the smallest Atlantic salmon linkage group, and, therefore, the fewer expected recombination events. Secondly, linkage group 9 in family Br6 showed an extreme male:female map distance ratio of 1:10, which again, resulted in the overall reduction of the Br6 male parent map length.

In addition to the overall heterochiasmy in salmonids, previous studies have presented evidence for major differences in the distribution of putative recombination events between males and females, with a higher frequency of male recombination events occurring at the telomeres of chromosomes (Moen *et al*, 2004a; Lien *et al*, 2011). To investigate the distribution of putative recombination events between male and female maps in the current study, for each linkage group in each family, the shortest parental map was split into n x 5cM intervals. The longer map for the same linkage group (derived from the parent of the opposite sex) was then split into an equal number of evenly sized intervals (n x m cM intervals, where m is determined by the map length). For each map, the five intervals with the highest proportion of SNPs were identified, and an overall average of the percentage of markers in the top five most populated intervals was calculated and compared for the two sexes across both families Br5 and Br6 (Figure 2.3; see section 2.3.5).

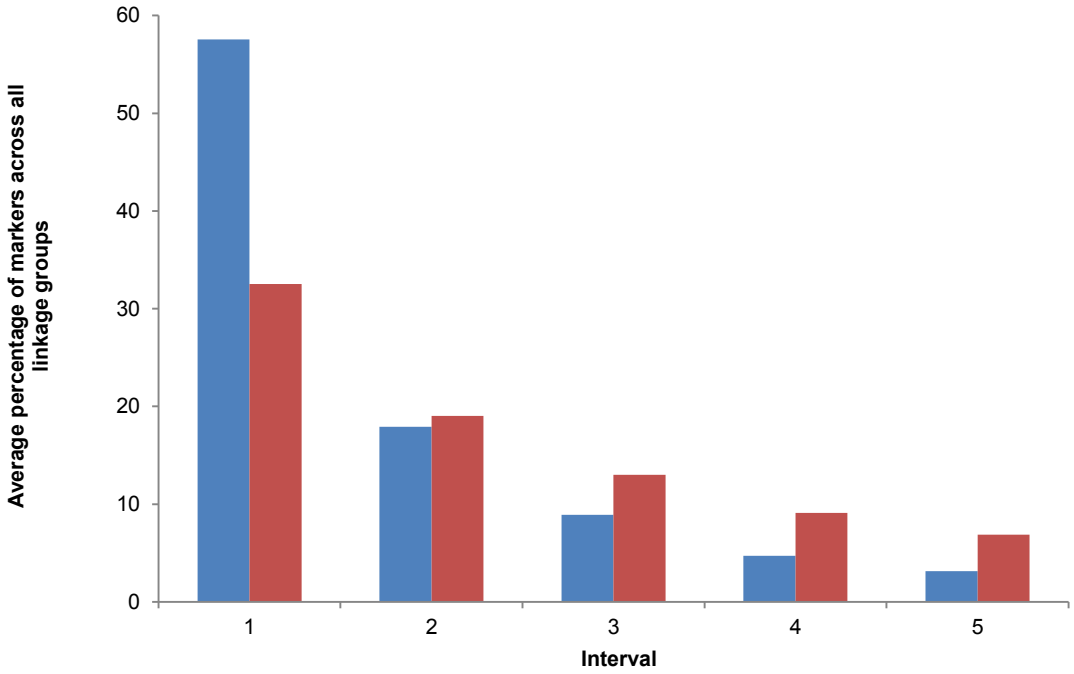



**Figure 2.3: Comparison of marker clustering between male and female linkage maps**
For each linkage group for each parent, the five intervals with the highest percentage of markers were identified. For each of these intervals, an average percentage of markers was calculated across all linkage groups and both families Br5 and Br6. Blue bars=Male average percentages; Red bars=Female average percentages. A greater clustering of markers to a single interval is apparent in the male maps.

Overall, in the male maps, markers formed one or two clear clusters of high marker density, corresponding to putative recombination deserts. These are postulated to be at the centromeric regions of chromosomes. Conversely, fewer markers were found in intervals closer to the extremes (putative telomeres) of male linkage groups. For example for family Br5, the average percentage of markers located at the extremes of the linkage groups was 8% in males compared to 19% in females. This is suggestive of more frequent recombination events in putative telomeric regions in males, which is in line with previous salmonid linkage mapping studies (Gilbey *et al*, 2004; Moen *et al*, 2004a; Moen *et al*, 2008; Lien *et al*, 2011).

### 2.4.5 Integration of the RAD-Seq maps with the Atlantic salmon reference genome, and inference of homeologous linkage group relationships

Alignment of the mapped RAD contigs (RAD loci and PE contigs) to the Atlantic salmon draft reference genome contigs allowed the assignment of 4,367 Atlantic salmon reference genome contigs (corresponding to 57Mb of sequence) to at least one linkage group [Table 2.4; see Gonen *et al* (2014), Additional file 5]. Of these, 110 genome sequence contigs showed significant sequence similarity to two different linkage groups, and two contigs to three linkage groups, which is indicative of homeology resulting from the recent salmonid-specific genome duplication [Table 2.5; see Gonen *et al* (2014), Additional file 6]. For example, 25 contigs aligned to both Atlantic salmon linkage groups 4 and 11, and homeology between these two linkage groups has previously been inferred (Danzmann *et al*, 2005; Danzmann *et al*, 2008; Phillips *et al*, 2009; Lien *et al*, 2011). Overall, 31 homeologous relationships across the Atlantic salmon genome were identified (Table 2.5), 22 of which have previously been inferred [based on sharing regions derived from the same proto-Actinopterygian ancestral chromosomes, as defined by Danzmann *et al* (2008)].

### 2.4.6 Identification of SNP-associated putative genes

To identify genes associated with the mapped and ordered RAD-Seq SNPs, a two-stage strategy based on sequence orthology to known stickleback genes was employed. The stickleback was chosen since it is the most closely related species to

Atlantic salmon for which there is a near-complete and annotated reference genome sequence available. In stage one of the gene identification procedure, the repeat-masked flanking sequences of mapped SNPs (including both the RAD locus and the PE contig; hereafter referred to as 'mapped RAD contigs') were screened for sequence similarity to all known stickleback gene sequences (Database=Ensembl Genes 72). Significant sequence similarity to a stickleback gene was observed for approximately 17% of the mapped RAD contigs (Table 2.6).

However, these contigs are relatively short (95bp for RAD loci; 450–600bp for PE contigs). As such, genes close to, but not within, the mapped RAD contigs may be undetected. Hence, in stage two, the 4,367 linkage group assigned Atlantic salmon reference genome contigs were repeat-masked and aligned with the stickleback gene sequences. Significant sequence similarity to stickleback genes was observed for 2,840 contigs (65%), 80 of which aligned to two Atlantic salmon linkage groups. In total, ~50% of the mapped SNPs were associated with a putative stickleback gene ortholog [Table 2.6; see Gonen *et al* (2014), Additional file 4]. Across all individuals and linkage groups, Atlantic salmon orthologs for 2,030 stickleback genes were identified and mapped. On average, 70% of the genes identified in stage one for each linkage group were also identified in stage two; the discrepancy likely being due to not all mapped RAD contigs being associated with a genome contig. These data will increase the utility of this linkage map for QTL fine-mapping and candidate gene identification.

### 2.4.7 Identification of orthologous relationships between Atlantic salmon and stickleback linkage groups

To investigate regions of conserved orthology between the Atlantic salmon and stickleback genomes, the stickleback linkage group positions of the genes associated with the mapped RAD contigs were recorded. For each of the salmon linkage groups, the stickleback linkage groups to which the mapped RAD contigs most frequently aligned to was determined. In total, conserved orthologous relationships for 26 of the 29 Atlantic salmon linkage groups with a stickleback linkage group were identified. No clear pattern of orthology with a stickleback linkage group was observed for

Atlantic salmon linkage groups 5, 19 and 22 (chromosomes 13, 8 and 17 respectively; Table 2.4, column 5).

### 2.4.8 Investigating the salmonid-specific genome duplication

To confirm the 31 homeologous relationships identified within the Atlantic salmon genome based on sharing of reference genome contigs, the Atlantic salmon linkage groups with orthologous relationships to the same stickleback linkage group were identified (from Table 2.4; summarised in Appendix A, Table A4). This confirmed 10 of the 31 homeologous relationships, and identified a further two putative homeologous relationships (Table 2.5).

To further test the theory of a salmonid-specific genome duplication, orthologous relationships between Atlantic salmon, rainbow trout and stickleback linkage groups were analysed using previously published data (Danzmann *et al*, 2005; Danzmann *et al*, 2008; Phillips *et al*, 2009). For 10 of the 12 Atlantic salmon homeologies identified due to sharing a single stickleback linkage group (Appendix A, Table A4), two rainbow trout linkage groups were identified, providing some support for a salmonid-specific duplication event. However, a 1:1 correspondence between single Atlantic salmon linkage groups and rainbow trout linkage groups was not observed (Table 2.4).

**Table 2.4: Reference genome contigs assigned to Atlantic salmon linkage groups**

| Atlantic salmon linkage group | Atlantic salmon chromosome | No. of genome contigs* | Amount of sequence data (Kb)** | Stickleback linkage group | Rainbow trout linkage group |
|---|---|---|---|---|---|
| 1 | 2 | 177 | 2,030 | 20 | 2/27/29/31 |
| 2 | 10 | 262 | 3,410 | 19 | 6/8/27 |
| 3 | 14 | 156 | 2,170 | 3/10 | 3/23/29 |
| 4 | 6 | 230 | 2,880 | 11 | 2/9/24 |
| 5 | 13 | 209 | 2,530 | NA | 9/22 |
| 6 | 12 | 211 | 3,090 | 17/9 | 2/29 |
| 7 | 24 | 94 | 1,160 | 13 | 10 |
| 8 | 15 | 175 | 2,220 | 18 | 21/23 |
| 9 | 11 | 180 | 2,160 | 2 | 1/10/18 |
| 10 | 9 | 298 | 3,850 | 1 | 3/4/25/26 |
| 11 | 3 | 232 | 3,480 | 11/3 | 2/9/13 |
| 12 | 5 | 173 | 2,110 | 20 | 3/27/31 |
| 13 | 19 | 110 | 1,500 | 21/5 | 17/19/22 |
| 14 | 21 | 111 | 1,570 | 16 | 5/31 |
| 15 | 27 | 118 | 1,330 | 10/20 | 16 |
| 16 | 18 | 159 | 1,900 | 6 | 6/21 |
| 17 | 1 | 254 | 3,780 | 14/6 | 8/14/30 |
| 18 | 23 | 103 | 1,300 | 8 | 24 |
| 19 | 8 | 33 | 454 | NA | 14/20 |

| | | | | | |
|---|---|---|---|---|---|
| 20 | 25 | 60 | 809 | 16 | 31 |
| 21 | 26 | 97 | 1,330 | 2 | 10/18 |
| 22 | 17 | 125 | 1,530 | NA | 7/12/15 |
| 23 | 16 | 164 | 2,210 | 19 | 6/16/27 |
| 24 | 7 | 109 | 1,370 | 4 | 7/15 |
| 25 | 20 | 196 | 3,100 | 13 | 11/19 |
| 28 | 4 | 165 | 2,200 | 7 | 14/20 |
| 30 | 29 | 76 | 992 | 21 | 7/17 |
| 31 | 28 | 84 | 1,130 | 5 | 17/22 |
| 32 | 22 | 120 | 1,520 | 17 | 12 |
| **TOTAL** | - | 4,367 | 57,402 | - | - |

Atlantic salmon reference genome contigs were assigned to linkage groups by BLASTN alignment to mapped RAD contigs. Column 5 shows the stickleback linkage groups orthologous to the Atlantic salmon linkage groups identified by this study. Column 6 shows the Atlantic salmon-rainbow trout orthologous linkage groups as defined by Phillips et al. (2009) (red) and Danzmann et al. (2008) (blue) individually, and those identified in both studies (green). Half of the stickleback-rainbow trout relationships suggested in this table have previously been identified (Guyomard et al. 2012).

* Total includes genome contigs assigned to more than one linkage group only once, thus is less than the sum of genome contigs per linkage group.

** Total includes sequence data of genome contigs assigned to more than one linkage group only once, thus is less than the sum of sequence (Kb) assigned per linkage group.

**Table 2.5: Homeologous relationships between Atlantic salmon linkage groups**

| Atlantic salmon linkage groups | Atlantic salmon chromosomes | No. of shared contigs | Shared ancestral linkage group |
|---|---|---|---|
| 4/11§* | 6/3 | 25 | E |
| 10/25 | 9/20 | 11 | G/H |
| 1/12§* | 2/5 | 9 | B |
| 9/21* | 11/26 | 9 | J |
| 3/11 | 14/3 | 8 | M |
| 22/24§* | 17/7 | 8 | K |
| 1/6§* | 2/12 | 5 | D |
| 3/15 | 14/27 | 4 | B |
| 2/23 | 10/16 | 3 | M,J/K |
| 2/18 | 10/23 | 3 | M |
| 7/25 | 24/20 | 3 | I |
| 19/28§* | 8/4 | 3 | - |
| 5/17 | 13/1 | 2 | I |
| 9/28 | 11/4 | 2 | G/H |
| 3/14 | 14/21 | 1 | - |
| 5/28 | 13/4 | 1 | G/H |
| 6/32§ | 12/22 | 1 | L |
| 16/17§ | 18/1 | 1 | D |
| 17/31 | 1/28 | 1 | D |
| 22/23§* | 17/16 | 1 | K |
| 22/30 | 17/29 | 1 | K |
| 17/32 | 1/22 | 1 | - |
| 12/17 | 5/18 | 1 | - |
| 10/32 | 9/22 | 1 | - |
| 9/25 | 11/20 | 1 | G/H,I |
| 4/6 | 6/12 | 1 | - |
| 2/25 | 10/20 | 1 | - |
| 2/6 | 10/12 | 1 | - |
| 9/14 | 11/21 | 1 | - |
| 3/13/17 | 14/19/1 | 1 | 3&13=M |
| 1/6/32 | 2/12/22 | 1 | 1&6=D |
| **TOTAL** | - | 112 | - |

112 Atlantic salmon reference genome contigs showed alignment to two or more linkage groups. Column 3 gives the total number of shared contigs between two or more Atlantic salmon linkage groups. Column 4 shows the proto-Actinopterygian ancestral linkage group shared between Atlantic salmon linkage groups, as defined by Danzmann et al. (2008).
§ Also identified in Danzmann et al. (2008)
* Also identified in Phillips et al. (2009)

**Table 2.6: RAD SNPs located proximal to a putative gene**

| Parent | No. of ordered and positioned SNPs | No. of gene-associated SNPs (stage 1) | Percentage of gene-associated SNPs (stage 1) | No. of gene-associated SNPs (stage 2) | Percentage of gene-associated SNPs (stage 2) |
|---|---|---|---|---|---|
| Br5 Mother | 1,311 | 212 | 16.2 | 541 | 41.3 |
| Br6 Mother | 1,399 | 227 | 16.2 | 621 | 44.4 |
| Br5 Father | 1,833 | 327 | 17.8 | 843 | 46.0 |
| Br6 Father | 1,817 | 298 | 16.4 | 815 | 44.9 |

The number of RAD SNPs located within or close to genes based on direct alignment of mapped RAD contigs (Stage 1; columns 3 and 4) or mapped Atlantic salmon reference genome contigs (Stage 2; columns 5 and 6) to known stickleback nucleotide gene sequences. Column 2 gives the total number of sex-specific segregating SNPs ordered and positioned within a linkage group using the OneMap software package.

## 2.5 Discussion

This study describes the construction and characterisation of the first high-density RAD-Seq-derived SNP linkage map in Atlantic salmon. As RAD-Seq becomes increasingly utilised as a cost- and time-efficient method of SNP discovery and genotyping in salmonid genomic studies, this map, and the additional resources generated herein, will provide a framework for orientation of the marker genotypes with the Atlantic salmon reference genome, and identification of putative candidate genes in population genomic studies.

### 2.5.1 Linkage map construction

For each of the four parents in the two SalMap families used in this study, sex-specific microsatellite linkage maps are available [ASalBase, http://www.asalbase.org/sal-bin/index; Danzmann *et al* (2005)]. Using a selection of these previously mapped markers as anchors in the current study enabled the assignment of ~6,500 RAD-Seq SNPs to salmon linkage groups, thus allowing a partial integration of the existing linkage maps with a dense SNP linkage map. However, final maps constructed in the current study were comprised only of RAD-Seq-derived SNP markers, since reliable ordering and positioning of anchor markers and SNPs in a combined linkage map was not possible. This was likely due to the different properties associated with the inheritance of the different marker types, as well as constraints in the number of informative meioses due to the small sample size in this study.

The large number of markers that can be discovered and scored in a single sequencing experiment is an advantage of the RAD-Seq approach. However, stringent filtering must be applied to avoid false-positive SNPs, particularly in the recently duplicated salmonid genomes. In this study, more than 28,000 putative SNPs were discovered, however, over half of these were removed due to an excess of missing genotypes. The large proportion of missing genotypes in the dataset was partially due to a degree of irregularity in the sequence coverage across individuals, but also due to the inevitably lower sequence coverage for PE contigs compared to the RAD loci, which reduces the confidence of SNP genotype calls. A large

proportion of the SNPs removed at this stage were from PE contigs. Overall, a strong relationship between sequence coverage and proportion of successful SNP genotype calls across individuals was observed (Figure 2.1), despite ensuring near equal quantities of offspring genomic DNA in each library. Therefore, to avoid high proportions of missing genotypes in future experiments using RAD-Seq, it is important to (i) strive for identical quantity and quality of input genomic DNA per individual and (ii) to account for the uneven read distribution across individuals and scale up the projected read coverage per individual accordingly.

Post-filtering, the average genotyping rate in the dataset across 77 individuals and ~8,500 SNPs was 76%. This is a substantial increase in the average genotyping rate in the unfiltered data (*ca.* 50%). However, given the relatively small sample size in this study (and, therefore, the low number of informative meioses), any missing data will reduce the resolution of the constructed maps.

To confirm the unbiased nature of RAD-Seq in sampling the whole genome, the number of SNPs initially assigned to each linkage group using the CRI-MAP software package and anchor marker information in this study was compared to the only other published high-density SNP linkage map (Lien *et al*, 2011). Despite the use of different SNP discovery, sequencing and genotyping technologies and map construction methods (*de novo* vs. anchor marker oriented), a strong positive correlation between the number of SNPs assigned to each linkage group in the two studies was observed (Figure 2.4). This suggests that *Sbf*I RAD-Seq is yielding an unbiased sample of the salmon genome, and that the number of SNPs per linkage group in both studies is likely to be related to chromosome size.

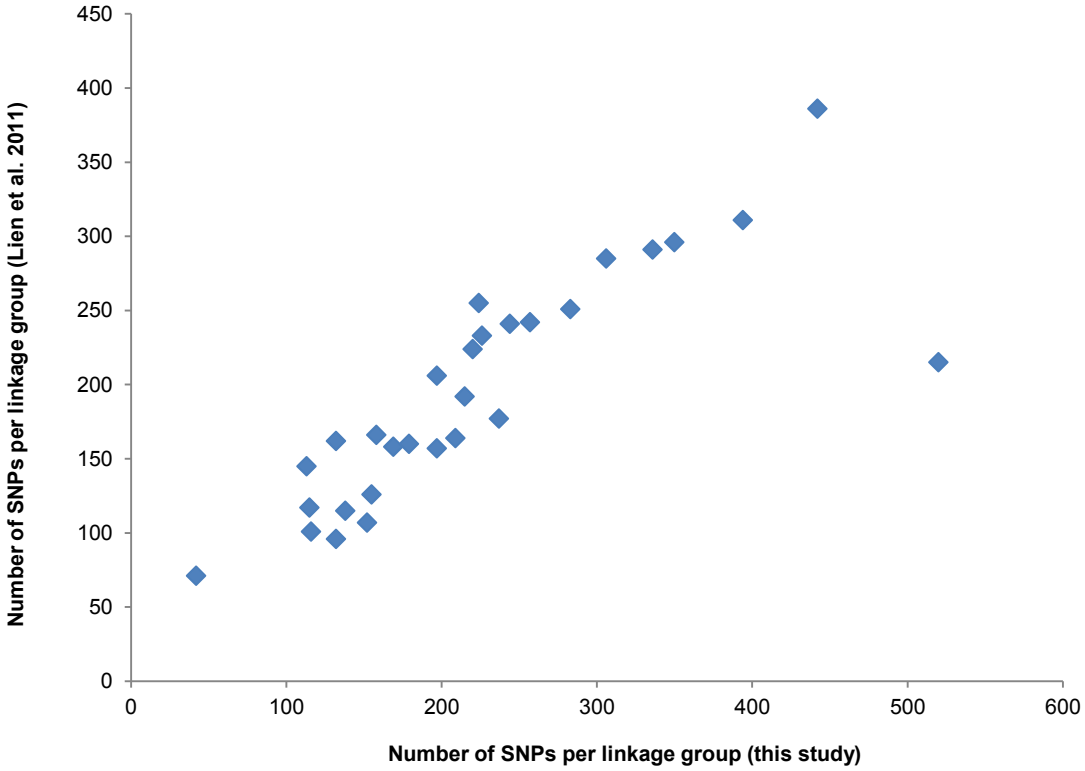



**Figure 2.4: Comparison of the number of SNPs per linkage group with a previously published map**
The number of SNPs per linkage group (assigned using the CRI-MAP software package) in the current study (x-axis) and in Lien et al. (2011) (y-axis). The number of linkage group-assigned SNPs in the two studies was highly correlated ($r^2$=0.83).

The length of genomic DNA sequenced at each RAD locus, including the RAD locus itself and the PE contig, is approximately 500bp. Therefore, multiple SNPs originating from a single locus may be observed. Recombination between these SNPs is unlikely, therefore, they are expected to map to the same position. To test this, the positions of SNPs from RAD loci and PE contigs originating from the same restriction cut site were analysed. A total of 26 restriction cut sites with mapped SNPs from both the RAD locus and PE contig were identified. In approximately 60% of these cases, SNPs in both the RAD locus and the associated PE contig mapped to identical map positions. Where this did not occur, PE SNPs were found to be positioned at the terminal ends of linkage groups. Given the lower read coverage for the PE contig due to the nature of the RAD-Seq protocol, SNPs derived from PE contigs may have a higher genotyping error rate than those from the RAD locus. A common feature of linkage mapping software packages is the positioning of markers

with higher error rates at the ends of linkage groups, and this may explain the instances where RAD loci and PE contig derived SNPs did not co-localise on linkage maps.

### 2.5.2 Map lengths and recombination ratios

The large difference in recombination rate between Atlantic salmon males and females, and the distribution of the recombination events along the chromosome, have been a subject of much discussion in the literature (Gilbey *et al*, 2004; Moen *et al*, 2004a; Moen *et al*, 2008; Lien *et al*, 2011). In the current study, only a relatively small difference in overall map lengths between the two sexes was seen (~1:1.5), which is comparable to that reported in the other Atlantic salmon high-density SNP linkage maps [1:1.38; Lien *et al* (2011)].

Furthermore, in accordance with previously published studies, an increased clustering of male-segregating markers compared to female-segregating markers was identified, supporting the hypothesis that the major difference between the sexes is in the positions, and not the overall frequency, of recombination events. However, it should be noted that since sex-specific markers were used, there were no common SNPs between the male and female maps within a family. Therefore, interpretations of the map distance differences and recombination events are based on overall patterns of linkage group length and marker clustering, rather than direct comparisons between marker positions.

### 2.5.3 Cross-species orthology and investigations of the salmonid-specific genome duplication

To increase the utility of the constructed linkage map in population genomic studies, SNPs associated with putative genes were identified, based on alignment of mapped SNP flanking sequences (generated from *Sbf*I RAD-Seq in the current study, or Atlantic salmon genome contigs containing a mapped SNP) with known annotated three-spined stickleback genes. The high proportion (~50%) of gene-associated mapped SNPs identified in this study provides additional evidence to support previous hypotheses of a bias of *Sbf*I RAD-Seq to genic regions of the genome. This

is likely due to the GC bias in the *Sbf*I recognition sequence [e.g. Everett *et al* (2012)]. This highlights the potential for *Sbf*I RAD-Seq in novel gene discovery, QTL fine-mapping and candidate gene identification, and cross-species sequence homology analyses [due to the expected higher conservation of gene sequences relative to other regions of the genome (Cooper and Brown, 2008; Bergmiller *et al*, 2012); see chapter five].

To investigate genome orthology between Atlantic salmon and stickleback, the linkage group positions of the gene-associated mapped SNPs in Atlantic salmon were compared to the linkage group positions of the gene orthologs in the stickleback genome. Due to the large evolutionary distance between stickleback and salmon, extensive chromosomal rearrangements are likely to have occurred in both species. Therefore, a direct conservation of gene order is not expected. Furthermore, the stickleback genome is comprised of 21 chromosomes (annotated as linkage groups in Ensembl), which is fewer than Atlantic salmon (*viz.* 29, equal to the number of linkage groups). As such, the same stickleback linkage group could show orthology to more than one Atlantic salmon linkage group. Most salmon linkage groups were assigned to at least one stickleback group, with three salmon linkage groups (5, 19 and 22) remaining unassigned. This was possibly due to the lower number of gene-associated markers on these linkage groups, which were used to infer chromosomal orthologies in this study (Table 2.4, column 5; summarised in Appendix A, Table A4). Overall, identified orthologous relationships were consistent with published literature (Danzmann *et al*, 2008; Phillips *et al*, 2009; Lien *et al*, 2011).

To further investigate the signatures of the salmonid-specific genome duplication, linkage group orthologies between Atlantic salmon, stickleback and rainbow trout were analysed, using published data (Danzmann *et al*, 2008; Phillips *et al*, 2009; Guyomard *et al*, 2012). Overall, a 1:1 correspondence between Atlantic salmon and rainbow trout linkage groups was not observed (Table 2.4), and this may be explained by the genomic rearrangements that have occurred in the two different genomes post-diploidisation. However, in most cases where two Atlantic salmon linkage groups shared a single stickleback linkage group (Atlantic salmon

homeologous linkage groups), two orthologous rainbow trout linkage groups were identified. This 1:2:2 correspondences between stickleback, Atlantic salmon and rainbow trout linkage groups respectively provides some support for the salmonid-specific ancestral genome duplication. With the recently published rainbow trout reference genome (Berthelot *et al*, 2014), the Atlantic salmon–rainbow trout chromosomal orthologies presented herein may prove useful for investigating regions of interest within the Atlantic salmon genome.

Homeologous linkage group relationships within the Atlantic salmon genome have previously been identified, based on shared, duplicated microsatellite markers, and by shared homology to the same ancestral proto-Actinopterygian linkage groups (Danzmann *et al*, 2008; Phillips *et al*, 2009). In the current study, 31 Atlantic salmon homeologies were identified, based on shared Atlantic salmon reference genome contigs (Table 2.5). It should be noted that the Atlantic salmon genome assembly used in this study was the first published draft, thus may contain assembly errors and chimeric contigs with sequence from multiple linkage groups. Furthermore, despite repeat-masking of all sequences used in this study, it is possible that the presence of repetitive elements within contigs could create spurious homeologies between linkage groups.

Therefore, these homeologies were further confirmed; first, based on sharing regions derived from the same ancestral karyotype [as characterised by Danzmann *et al* (2008)] (Table 2.5), and second, using comparative genomics, by looking for shared stickleback linkage group orthologies (Table 2.4; summarised in Appendix A, Table A4). Of the 31 Atlantic salmon homeologous relationships identified, 12 were confirmed based on common orthology to a single stickleback linkage group, both in this study and other published studies [e.g. Lien *et al* (2011)]. Furthermore, 8 of the 12 homeologous relationships were previously described in Danzmann *et al* (2008) based on sharing of regions derived from the ancestral karyotype, and 7 have been identified in Phillips *et al* (2009), providing strong support for these homeologous relationships (highlighted in Table 2.5). These Atlantic salmon homeologous relationships are valuable for studies such as that presented in this chapter, for

example in resolving false linkage relationships between pseudo-linked male markers.

## 2.6 Conclusion

This study describes the construction and characterisation of a high-density SNP linkage map of the Atlantic salmon genome in an outbred population, using SNPs derived from paired-end RAD-Seq. Analysis of the pattern of recombination events in male and female mapping parents revealed a difference in the distribution of putative recombination events across the linkage groups, in line with previously published literature. Comparative sequence orthology analyses with the stickleback genome allowed the identification of genes proximal to (or containing) the mapped RAD-Seq SNPs. Homeologous regions within the Atlantic salmon genome, and the putative orthologs of the salmon linkage groups in the stickleback and rainbow trout genomes were identified and confirmed, providing support for a salmonid-specific genome duplication. RAD-Seq is an increasingly popular tool for QTL mapping and population genomics, and this new map, and the additional resources generated herein, will provide a useful framework for future genomics studies.

# Chapter 3

# Estimating genetic parameters and mapping of QTL affecting Pancreas Disease resistance in Atlantic salmon

## 3.1 Abstract

Pancreas disease (PD), caused by a salmonid alphavirus (SAV), has a large negative economic and animal welfare impact on Atlantic salmon aquaculture. Evidence for genetic variation in host resistance to this disease has been reported ($h^2$=0.21±0.005), suggesting that genetic selection for improved resistance can form an important component of disease control. The aim of this study was to explore the genetic architecture of resistance to PD, using survival data collected from a freshwater fry SAV challenge experiment. Analyses of these binary survival data revealed a high heritability for PD resistance of ~0.5, and this estimate was consistent across the different models applied. QTL mapping analyses based on sire- then dam-linkage information of SNP marker segregation patterns detected four putative QTL influencing resistance to PD, on chromosomes 3, 4, 7 and 23. The QTL on chromosome 23 reached genome-wide significance in the sire-based QTL mapping analysis. The QTL on chromosome 3 reached chromosome-wide significance in both sire- and dam-based QTL mapping analyses, and explained the largest proportion of the within-family variation for resistance. SNP markers showing significant association with PD resistance on this chromosome have been identified for potential use in marker-assisted selection for resistance on Atlantic salmon farms. Importantly, this QTL has recently been independently replicated in a post-smolt SAV challenge experiment. The independent mapping of the QTL on chromosome 3 in two populations validates this QTL, and suggests a common mechanism for PD resistance across both Atlantic salmon life stages. These results are of economic importance to breeding companies, and suggests that fry PD challenges can be used as models to study disease dynamics.

## 3.2 Introduction

Pancreas disease (PD), an alphaviral disease, is currently one of the most problematic infectious diseases on Atlantic salmon farms, resulting in high levels of mortality and morbidity (FAO, 2014b). Six subtypes of the PD-causing salmonid alphavirus (SAV) have been isolated in different parts of the world, including Scotland, Norway and Chile (Fringuelli *et al*, 2008). Subtypes are geographically specific, and farms within the same locality typically show infection with the same subtypes (Kristoffersen *et al*, 2009; Graham *et al*, 2012). For example, the two SAV subtypes in Norway (SAV2 and SAV3) have been shown to affect distinct sites (SAV2 in the north and SAV3 mainly in the south of Norway), with no overlap or co-infection within sites (Hjortaas *et al*, 2013; Jansen *et al*, 2014).

Natural infections with SAV have only been documented in the post-smolt stage of the salmon lifecycle, shortly after transfer from freshwater to sea water. Infection with SAV has been shown to result in histological changes in the heart, skeletal muscle and the pancreas of post-smolt salmon, as well as causing signs of morbidity such as a loss of appetite and lethargy (McLoughlin *et al*, 2002; Rodger and Mitchell, 2007; Taksdal *et al*, 2007). Long term sub-clinical infections are common, and the peak in mortalities associated with natural outbreaks is often seen many months after infection (Karlsen *et al*, 2012). Survivors of infection can show chronic long term illness and a reduced growth rate, leaving them vulnerable to infection with other pathogens and thus with drastically reduced economic value (Fringuelli *et al*, 2008; Cano *et al*, 2014; Taksdal *et al*, 2014). Response to infection may be influenced by many factors, such as feeding rate, season, temperature, stocking density, co-infection with other pathogens, and host genetics (McLoughlin *et al*, 2002; Rodger and Mitchell, 2007; Graham *et al*, 2008; Norris *et al*, 2008; Kristoffersen *et al*, 2009; Jansen *et al*, 2010a; Stene *et al*, 2013).

Management techniques such as site fallowing, hygiene, and vaccination strategies currently employed in an effort to prevent the spread of the virus and limit its impact on affected farms have not been fully effective (Rodger and Mitchell, 2007; Jansen *et*

*al*, 2010a; Karlsen *et al*, 2012; Graham *et al*, 2014; Jansen *et al*, 2014). As such, there is a need for additional methods to complement or enhance current control measures, such as breeding salmon that are more resistant to PD.

Resistance to PD in farmed Atlantic salmon post-smolts in natural PD outbreaks has been shown to be moderately heritable [$h^2$=0.21; Norris *et al* (2008)]. Therefore, family-based selection for enhanced PD resistance would be possible, based on the performance of relatives in viral challenge experiments. As described previously, the disadvantage of this family-based method of selection is that it does not utilise the within-family variation for resistance, and, therefore, response to selection is likely to be slower compared to that observed when direct individual phenotypes are used. This is because the use of survivors of SAV challenge experiments as breeding parents is not possible, due to the possibility of vertical viral transfer from parents to offspring (Jansen *et al*, 2010b). Furthermore, survivors of a SAV infection often show a reduced performance (i.e. appear lethargic and morbid, and show a much reduced overall growth rate) in comparison to naïve fish, and are unlikely to be selected as breeding candidates (Norris *et al*, 2008; Cano *et al*, 2014; Jansen *et al*, 2014). Instead, unchallenged naïve breeding parents are selected based on the performance of their relatives (generally full-siblings) in viral challenge experiments. Therefore, alternative strategies able to select based on individual performance, such as the use of genetic markers to infer individual resistance without exposure of the individual to the virus, are required. This requires the characterisation of the genetic architecture of PD resistance and the identification of marker-QTL associations influencing variation in resistance for potential use in marker-assisted or genomic selection.

As yet, no published PD QTL mapping study exists, and the underlying mechanisms for the observed variation in host response to infection with SAV are unclear. This is partially due to the dynamics of the disease. Natural infections with SAV have only been reported at the post-smolt stage of the salmon lifecycle and mortalities from natural or challenge outbreaks are generally observed many months post infection (Weston *et al*, 1999; McLoughlin *et al*, 2002; McLoughlin *et al*, 2006; Rodger and

Mitchell, 2007; Taksdal *et al*, 2007; Kristoffersen *et al*, 2009; Jansen *et al*, 2010b; Jansen *et al*, 2010a; Jensen *et al*, 2012; Jansen *et al*, 2014), making it difficult to conduct challenge experiments required to estimate genetic parameters. As such, alternative challenge models, such as infecting salmon at the juvenile freshwater fry stage of the lifecycle, are being explored (Cano *et al*, 2014). However, although fry can be infected with SAV, the differences in physiology between the fry and post-smolt stages means that it is not yet clear whether genetic parameters and QTL identified using fry challenge models can be used to select for resistance at the post-smolt stage.

The overall aim of this study was to explore the genetic architecture of PD resistance in a large population of Atlantic salmon fry, challenged with SAV. The specific aims of this study were to: (i) estimate the heritability of resistance to PD; (ii) detect and position QTL influencing resistance to PD; and (iii) identify markers in population-level association with the resistance QTL, for potential use in MAS.

## 3.3 Materials and Methods

### 3.3.1 Experimental population

A total of 218 full-sibling (83 paternal half-sibling) families were selected from the 2010 year class of Marine Harvest (MH) for inclusion in the PD challenge experiment. The Marine Harvest fish stock originate from the Mowi strain from the River Bolstad and the River Aaroy in Norway, and breeding programs were established in the 1960s (Glover *et al*, 2009). The family structure of these 218 families was as follows. Of the 83 sires, 10 were mated to 4 dams each, 45 to 3 dams each, 15 to 2 dams each and 13 to 1 dam each. From each full-sibling family, 30 eggs (total=6,540 eggs) were selected for the challenge experiment. 605 eggs failed to hatch, and a further 122 mortalities post-hatch and before challenge start were observed, leaving 5,813 fish to be included the challenge experiment. Fish were reared to fry stage (51 days post-hatch; average weight: 0.5 g) prior to challenge.

### 3.3.2 Salmonid alphavirus challenge

A total of 100 Atlantic salmon parr (i.e. young maturing salmon between the fry and smolt stage of the salmon lifecycle; average weight ~38 g) were intraperitoneally-injected with the SAV3 strain of the salmonid alphavirus, which is the most abundant strain in Norway (Hodneland *et al*, 2005). The viral challenge dose used was $3.3e^5$ TCID50 per shedder. Parr were allowed to shed virus into a tank of 216 L in volume for one week, after which effluent water from this tank was passed into the fry challenge tank. The 5,813 fry to be included in the challenge experiment were starved for 24 hours and then transferred into the challenge tank. The fry challenge was conducted within a single tank of 10 L in volume, thereby avoiding potential environmental confounding factors associated with tank effects (Kjøglum *et al*, 2008; Guy *et al*, 2009). Water temperature in this fry challenge tank was maintained at 12℃ and water flow was >1 L/Kg/min., in order to maintain an $O_2$ saturation level of >75%. The challenge was allowed to continue until mortalities were negligible (start date: 15/06/2010; end date: 11/08/2010; challenge profile is given in Figure 3.1). Ten fry from the main mortality period (i.e. after >10% mortality was observed) were sampled and measured for viral load to confirm infection.

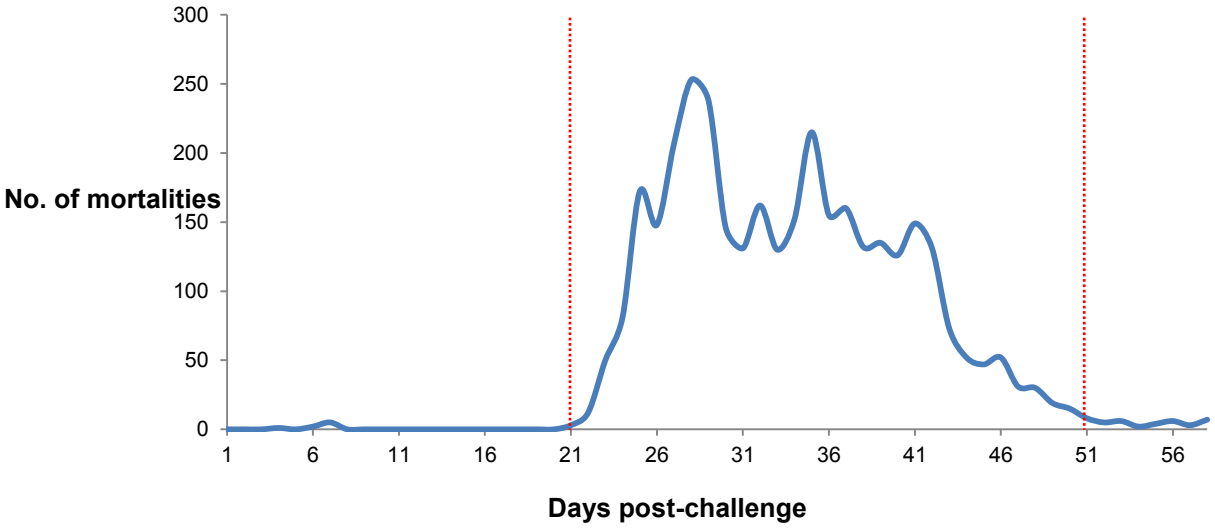



**Figure 3.1: Mortality profile across the challenge duration**
Number of mortalities observed per day over the course of the challenge. The peak in mortalities was observed 28 days post-challenge. Mortalities occurring between 21 and 51 days post-challenge (indicated by the vertical dotted red lines) were assumed to be due to PD, and were used in all further analyses.

Mortalities occurring during the course of the challenge experiment were collected daily. Tissue samples of mortalities (posterior half of the body; Figure 3.2) collected on the same day were stored in 70% ethanol in the same tube, and kept at -80°C. Surviving fry were collected at the challenge termination date and were sampled in the same manner as the mortalities. DNA from the survivors was extracted at MH from a piece of the posterior body tissue sample, and a family assignment analysis was performed using a panel of nine microsatellite loci (GenoMar, Co/Glastad Invest AS, Fridtjof Nansens Plass 3, Oslo, Norway). In total, 2,102 of the 2,328 surviving fry (90%) were successfully assigned back to family. Following this, the remainder of the posterior body tissue samples of these 2,102 family-assigned survivors were placed in individual wells in 96 well plates containing 70% ethanol, and labelled with individual IDs according to the well address, challenge year ID (V2526) and plate number, and sent to The Roslin Institute (RI) for genotyping (see below).

### 3.3.3 Samples received from Marine Harvest post-challenge

Three sets of samples were received from MH: survivors from the PD challenge experiment (total=2,102), mortalities from the PD challenge experiment (total=3,455), and parents of the challenged individuals (282 in total; samples were available from 202 of the 218 dams, and 80 of the 83 sires in the study). These sample sets arrived in different formats, and, therefore, required different levels of processing before genotyping. For the 282 parental samples, DNA had previously been extracted by MH. Therefore, these were sent directly for genotyping (see section 3.3.5). For surviving and mortality offspring, tissue samples of the posterior half of the body (Figure 3.2) were sent to RI for DNA extraction (see section 3.3.4 for DNA extraction protocol).

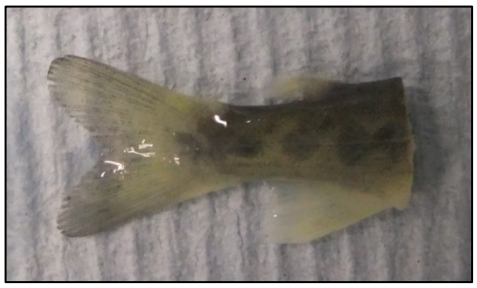


**Figure 3.2: Posterior body tissue samples from surviving and mortality fry received from Marine Harvest post-challenge**

For the purposes of this study, mortalities between 21 (05/07/10) and 51 (04/08/10) days post-challenge were included in the genotyping and subsequent analyses, in order to distinguish mortalities due to PD from the low level of baseline mortalities (Figure 3.1). This interval contained 3,415 of the 3,455 total mortalities, with negligible mortalities outwith these dates. Since challenged fish were too small to be PIT tagged, they had not been assigned an individual ID prior to challenge, and family assignment for mortalities had not yet been conducted. For mortality fish, individual IDs were assigned during sample processing, comprising the tube number and the order of the fish taken out of the tube. The processing of mortality samples prior to DNA extraction was as follows. Fish were individually removed from tubes, and any excess ethanol was dried off. Tail fins of each fish were removed and placed in individual wells in 96 well plates. The ID of the fish was recorded on a plate map, and 96 well plates of dry tissue samples were frozen at -20°C prior to DNA extraction.

The 2,102 survivor samples with assigned IDs and known family structures were shipped to RI in 96 well plate format. Tissue samples were processed as described above for mortalities, and stored at -20°C. For the purposes of this study, sample processing and subsequent DNA extraction was conducted for only 640 of the 2,102 survivors, since these were the individuals from the 20 half-sibling families selected for QTL mapping (see section 3.3.8). Survivor IDs previously assigned by MH were retained.

### 3.3.4 DNA extraction

DNA extraction of all samples (3,415 mortalities and 640 survivors) in 96 well plates (36 plates for mortalities, 7 plates for survivors) was conducted using the Qiagen DNeasy 96 protocol Blood & Tissue kit (96 well plate format), with some modifications (Appendix C). DNA quality and quantity post-extraction was estimated for a random selection of four or five samples per plate, using Nanodrop (Thermo Scientific, Delaware, USA). 100 µl of DNA was taken from the eluted DNA sample and placed in new 96 well plates, which were heat-sealed and stored at -20°C prior to shipping for genotyping.

### 3.3.5 Genotyping and parentage assignment

DNA samples of parents, mortalities, and survivors were shipped to LGC Genomics Ltd. (Herts, EN11 0WZ, UK) for genotyping. Genotyping was conducted using the Kompetitive Allele Specific PCR (KASP) technology, which involves DNA amplification through two rounds of PCR using allele specific forward primers, followed by the addition of a fluorescent complementary primer and a final DNA elongation step (www.lgcgroup.com/products/kasp-genotyping-chemistry/).

All mortalities and parents were initially genotyped using a sparse panel of 69 SNPs [taken from Moen *et al* (2008); Appendix A, Table A5], chosen so that each of the 29 Atlantic salmon chromosomes contained between 1 and 3 informative SNPs. Parentage assignment of mortalities was carried out at RI, using these genotypes and three software packages: SNPPIT (maximum likelihood algorithm) (Anderson, 2010), FAP (genotype exclusion algorithm) (Taggart, 2007), and Vitassign (genotype exclusion algorithm) (Vandeputte *et al*, 2006). Successful parentage assignment of mortalities was taken only if agreement was seen between outputs from at least two of the software packages. A total of 2,455 of the 3,415 mortalities were successfully assigned to family.

Surviving offspring had previously been assigned to family by MH, using a panel of nine microsatellite loci (see section 3.3.2). The 640 survivors in the 20 half-sibling families selected for QTL mapping (see section 3.3.8) were identified and genotyped using the same sparse SNP panel described above. Assignment for the 640 survivors in the 20 half-sibling families selected for QTL mapping was confirmed at RI, using the same three software packages used for the assignment of mortalities and sparse SNP panel genotypes. Eight survivors with poor quality genotyping information (i.e. genotyped at fewer than 35 of the 69 SNPs) were removed. Following this, 26 offspring with a different sire and dam assignment and two offspring with the correct sire but incorrect dam assignment compared to MH assignments were excluded. Of the remaining 604 offspring, 141 were assigned to multiple families by all three

software packages and were removed, leaving 463 family-assigned survivors for subsequent QTL mapping analyses (see section 3.3.8).

In addition to family assignment, the sparse SNP panel genotyped in parents, mortalities and the 463 survivors was used for the sire-linkage based stage one of the two-stage QTL mapping approach implemented in this study (see section 3.3.8).

To better position the chromosome- or genome-wide significant QTL identified on the four chromosomes in stage one, a denser SNP panel of 36 SNPs [taken from Lien *et al* (2011) and the linkage map in chapter two of this thesis, and including the sparse panel of SNPs used in stage one for genotyping; Appendix A, Table A6] was genotyped across the parents, mortality and surviving offspring in the 20 families used for QTL mapping. Based on this denser SNP panel, linkage maps for these four chromosomes containing significant QTL were constructed, using the Lep-MAP software package (Rastas *et al*, 2013) (Appendix A, Table A6). This software package implements a Bayesian algorithm to linkage map construction, and was chosen since it accommodates differences in recombination patterns between the sexes. The upper and lower LOD score thresholds to use in the assignment of SNP markers to chromosomes were estimated using the 'EstimateLODLimit' function. These thresholds are used as the maximum and minimum restrictions to estimate the empirical distribution of LOD thresholds used to infer marker linkage. SNP markers were assigned to chromosomes using the 'SeparateChromosomes' option, with lower and upper LOD score limits of 0.6 and 1.4 respectively. The low LOD thresholds estimated by the software packages are likely due to there being only 36 markers to assign. The order of SNP markers within chromosomes was estimated using the 'OrderMarkers' option. These linkage maps and genotypes were used in the dam-linkage based stage of the QTL mapping approach (stage two).

### 3.3.6 Data filtering, quality control, and processing prior to analysis

To generate a dataset of sufficient power (i.e. large number of individuals) with complete challenge survival information for all included offspring, genotype data and family assignments were combined and then filtered as follows. Of the 282 parental

samples genotyped using the sparse SNP panel (69 SNPs, see section 3.3.5), 33 parents with poor quality genotyping output (i.e. excess missing genotypes) were removed, leaving 185 full-sibling (76 half-sibling) families. Following parentage assignment of mortality offspring to family, 10 full-sibling families with no assigned mortality (based on RI assignment) or surviving offspring (based on MH assignment) were removed, since they would not contribute to the subsequent quantitative genetic analyses. Of the remaining 177 full-sibling (76 half-sibling) families, 157 (71) had mortalities assigned, 173 (75) had survivors assigned, and 153 (76) families had both mortalities and survivors assigned.

To increase the power for conducting heritability and QTL mapping studies, the 177 full-sibling families were further filtered to retain those with a minimum of 15 offspring (sum of mortalities and survivors). This resulted in a final dataset comprised of 3,949 offspring in 150 full-sibling (72 half-sibling) families, with an average mortality of 61% (59%) and a range in family size of 16–42 (18–106) offspring per family.

While a maximum of 30 offspring per family were initially expected, in some cases, a higher number based on genotype assignment was observed. This is likely to be due to errors during egg transfer when setting up the challenge experiment, or due to the initial incorporation of a greater number of eggs per family, to account for the possibility of high hatch failure rates and/or post-hatch mortalities within full-sibling families. Alternatively, false parent-offspring assignments may be partially responsible for the discrepancy in family size, which could be due to relatedness between the parents, since parents used in the study were from the same population and broodstock. However, this is unlikely to be the case, since assignments were confirmed using three software packages implementing different assignment algorithms (genotype exclusion vs. maximum likelihood). As well as this, parentage assignment of survivors conducted by MH was verified at RI using the same sparse SNP genotyping panel and the same three software packages used for mortality assignment.

### 3.3.7 Quantitative genetic parameter estimation

The 150 full-sibling (72 paternal half-sibling) families (parents not closely related) described above were used for the estimation of genetic parameters, i.e. the additive genetic variation and heritability for PD resistance. Variance components were estimated by fitting the following linear mixed model in the ASReml software package (Gilmour *et al*, 2009):

$$Y_{ijk} = \mu + Sire_i + Dam_{ij} + e_{ijk}$$

where $Y_{ijk}$ is the observed SAV challenge outcome for individual $k$ with sire $i$ and dam $j$; $\mu$ is the population mean; $Sire_i$ and $Dam_{ij}$ are the random additive genetic effects of the $i$th sire and $j$th dam; and $e_{ijk}$ is the residual variance. Sire and Dam were fitted as random effects and assumed to be normally distributed, with variances $\sigma^2_{SIRE}$ and $\sigma^2_{DAM}$, respectively. The total additive genetic variance was estimated as $2(\sigma^2_{SIRE}+\sigma^2_{DAM})$. Since challenge outcome was scored as a binary variable (1=mortality, 0=survived), the heritability of PD resistance was estimated on the observed binary scale, and by using a logit-link or probit-link function to account for the binary data. Assuming a continuous underlying liability for the binary challenge outcome, the observed binary scale heritability ($h^2_{01}$) was converted to the underlying liability scale ($h^2$) using the formula given in Falconer and Mackay (1996):

$$h^2 = h^2_{01}(1-p)/i^2p$$

where $p$ is the disease incidence, i.e. proportion of mortalities in the 150 full-sibling families (0.61); and $i$ is the estimated underlying mean liability of affected individuals in the population.

The heritability of days to death was also estimated on the observed scale (days to death analysed as a continuous trait) using the same linear mixed model described above, and data from mortalities only.

### 3.3.8 QTL mapping

To increase the power of detecting QTL segregating within family, a total of 20 paternal half-sibling families with intermediate levels of mortality were selected as follows. First, half-sibling families were ranked with respect to percentage mortality. Families with intermediate levels of mortality (40–70%) were identified and ranked

based on the total number of offspring and number of full-sibling families. The 20 paternal half-sibling (55 full-sibling) families with intermediate levels of mortality (half-sibling family mortality final range: 45–71%; average mortality 56%), >40 offspring (half-sibling family size range: 41–90 offspring) and at least 2 full-sibling families were selected for QTL mapping (Figure 3.3). These families comprised 463 survivors and 810 mortalities, to total 1,273 offspring. Assuming QTL of intermediate frequency, the use of paternal half-sibling families with intermediate mortality levels increases the power for the detection and mapping of QTL segregating within family.

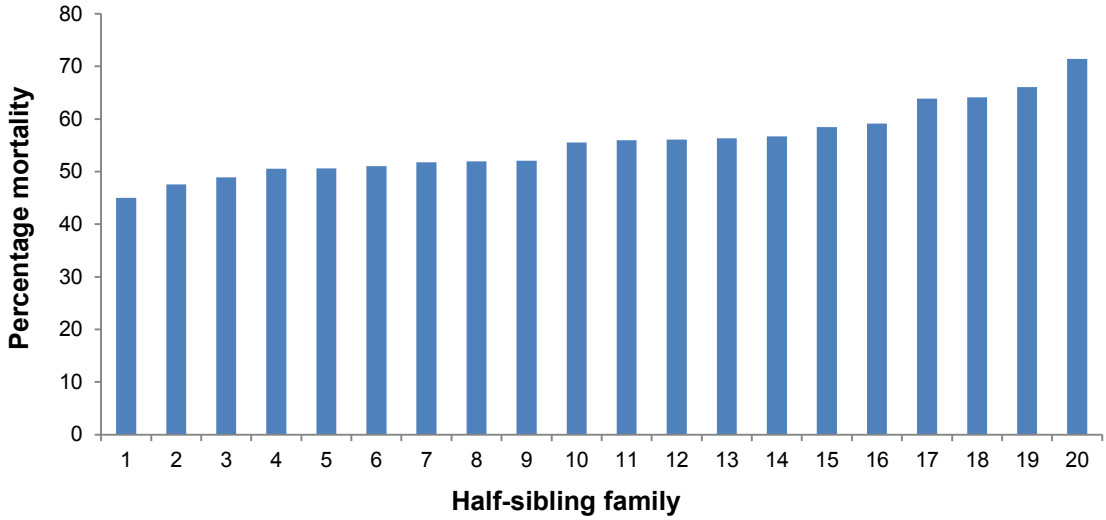



**Figure 3.3: Percentage mortality across the 20 paternal half-sibling families selected for QTL mapping**

Taking advantage of these large paternal half-sibling families, a half-sibling (HS) QTL mapping analysis was conducted using the GridQTL software package (Allen *et al*, 2012), which uses a linear regression-based interval mapping approach to QTL identification [for details, see Knott *et al* (1996)]. Briefly, using a multiple marker based approach, the probability of inheriting a particular allele at a given marker location is calculated and used to estimate the information content of each marker. At each centiMorgan (cM) interval, the phenotypes (challenge outcome) are regressed on to the probabilities of inheriting particular alleles. The strength of evidence for a QTL is calculated and expressed as an F ratio test, with the number of numerator degrees of freedom equal to the number of parents informative at a given marker (Knott *et al*, 1996).

The chromosome-wide critical F ratio threshold was determined using 10,000 permutations in the GridQTL software package. A genome-wide critical F ratio threshold was calculated by obtaining a Bonferroni corrected P-value at the 5% significance level (given the 29 pairs of chromosomes of Atlantic salmon, adjusted P-value=0.05/29) and then obtaining the genome-wide critical F ratio threshold at this adjusted P-value, using 10,000 permutations (Churchill and Doerge, 1994). The QTL F ratio was compared to the chromosome- and genome-wide critical F ratio, and if the QTL F ratio was larger than either one of these, then this QTL was determined as significantly affecting PD resistance in this population. For each genome-wide significant QTL, confidence intervals for the location parameter were estimated using the bootstraps with resampling method and 10,000 iterations (Visscher *et al*, 1996).

This HS QTL mapping analysis was implemented using a two-stage approach (Hayes *et al*, 2006; Houston *et al*, 2008), which takes advantage of the large disparity in recombination rates between Atlantic salmon males and females. The low recombination rate in males means that the inheritance of large sections of the genome from sire to offspring can be tracked by genotyping individuals using a sparse marker set (Hayes *et al*, 2006). This allows the identification of chromosomes harbouring QTL significantly influencing the trait of interest (stage one). In this first sire-linkage based stage, the sparse SNP marker panel (69 SNPs across the 29 Atlantic salmon chromosomes) was utilised. A dam-linkage analysis using the same approach and sparse SNP panel was also conducted in order to identify QTL segregating in dams but not sires, recognising that with sparse markers true QTL may be missed.

In the second stage of QTL mapping, a denser set of SNP markers (36 in total; Appendix A, Table A6) was genotyped across the four chromosomes identified as containing chromosome- or genome-wide significant QTL in stage one. In order to confirm QTL and estimate their position on these significant chromosomes, the inheritance patterns of these markers from dam to offspring with respect to challenge

outcome were analysed, using a dam-linkage analysis in the GridQTL software package.

To exploit the full-sibling family structure, QTL mapping was also conducted using a sib-pair (SP) approach, in the GridQTL software package. This approach is based on the principle that siblings who inherit more QTL alleles identical-by-descent (IBD) tend to be more similar in phenotype, i.e. the difference between their phenotypes tends to be smaller the more QTL alleles they share IBD (Haseman and Elston, 1972). IBD probabilities are calculated at 1cM intervals, and the squared difference of the phenotypes (i.e. residuals estimated from the IBD probability analysis) is regressed onto the IBD probabilities (Haseman-Elston approach) (Haseman and Elston, 1972; Knott and Haley, 1998). QTL significance and confidence intervals for QTL position were determined using permutation testing, as described for the HS analysis above. The SP analysis was conducted using the sparse set of SNP markers, and was repeated for the four chromosomes harbouring significant QTL using the denser set of SNPs.

The proportion of within-family variance explained by each significant QTL (PVE) was estimated using the HS analyses results, and the following formula. For the sire-based linkage analysis: $h^2_{QTL}=4[1-(MSE_{full}/MSE_{red})]$; for the sire- and dam-based linkage analyses combined: $h^2_{QTL}=2[[1-(MSE_{full}/MSE_{red})_{SIRE}] + [1-(MSE_{full}/MSE_{red})_{DAM}]]$; where $MSE_{full}$ and $MSE_{red}$ are the mean square errors for the models with and without the QTL, respectively.

### 3.3.9 Association analysis

To identify individual markers associated with mortality at the population level, SNPs from the denser SNP marker panel distributed across the four chromosomes harbouring significant QTL were independently analysed for population-wide association with PD resistance, using the ASReml software package (Gilmour *et al*, 2009). Since marker density was at most 12 SNPs per chromosome, markers were assumed to be unlinked and segregating independently. Therefore, SNP association was conducted by fitting the same mixed model used to estimate the heritability, with

the added step of fitting each SNP individually as a fixed effect. Assuming SNP alleles A and G, the additive effect, dominance effect, average allelic substitution effect, and percentage of additive genetic variance explained by significant SNPs were calculated as follows: Additive effect, a=(AA-GG)/2; Dominance effect, d=AG-[(AA+GG)/2]; Average allelic substitution effect, $\alpha=a+d(p-q)$; Percentage additive genetic variance explained by SNP, $\%Va=100[2pq(a+d(q-p))^2]/V_A$, where p=frequency of allele A, q=frequency of allele G, $V_A=2(\sigma^2_{SIRE}+\sigma^2_{DAM})$.

## 3.4 Results

### 3.4.1 Challenge outcomes and parentage assignment

A total of 3,456 mortalities were observed over the course of the 59 day challenge. For all analyses conducted, 3,415 mortalities occurring between 21 and 51 days post-challenge (05/07/10 and 04/08/10) were assumed to be due to PD. Of these, 2,455 were successfully assigned to 157 full-sibling (71 half-sibling) families. 2,328 survivors remained at the challenge termination date (11/08/10), 2,102 of which were successfully assigned (MH) to 203 full-sibling (81 half-sibling) families.

### 3.4.2 Estimated heritabilities

The heritability of PD resistance was estimated using 150 full-sibling (72 half-sibling) families with at least 15 offspring (3,949 offspring in total, 2,367 mortalities, 1,582 survivors). Mortality levels in these families ranged from 0–100%, with an average mortality of 61% (Figure 3.4). The heritability of PD resistance across these families was estimated as 0.34 (± 0.05) on the observed binary scale, which equated to approximately ~0.5 on the underlying liability scale (Table 3.1). This estimate was relatively consistent across the different models applied (underlying liability, logit-link and probit-link models). Non-genetic effects associated with full-sibling family were not significant. The heritability of time to death was estimated as 0.06 (± 0.02), which is substantially lower than that estimated for the binary survival trait.

**Table 3.1: Estimated heritabilities for resistance to PD in Atlantic salmon fry**

| Method | Heritability (±SE) |
|---|---|
| Observed binary scale | 0.34 (±0.05) |
| Underlying liability scale | 0.55 |
| Probit-link scale | 0.54 (±0.07) |
| Logit-link scale | 0.46 (±0.06) |

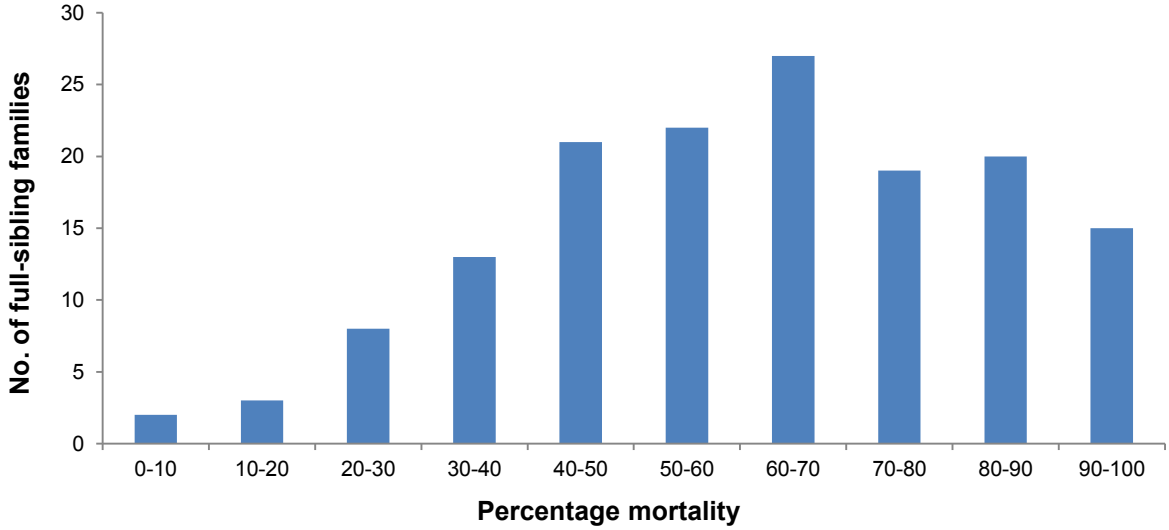



**Figure 3.4: Percentage mortality across the 150 full-sibling families used to obtain heritability estimates**
The number of full-sibling families across the 150 full-sibling families with a given percentage mortality. Family mortalities ranged from 0–100%, with an average mortality of 61%.

### 3.4.3 QTL mapping

To identify QTL significantly influencing PD resistance in Atlantic salmon fry, a two-step QTL mapping approach using a HS analysis was applied, utilising first a sparse (stage one), then a denser SNP panel (stage two), and further taking advantage of the unequal recombination rates between males and females. The initial sire-linkage analysis using a sparse SNP panel identified three putative QTL affecting PD resistance, on chromosomes 3, 7 and 23 (Table 3.2). The QTL on chromosome 23 was significant at the genome-wide level, whereas the QTL on chromosomes 3 and 7 were significant at the chromosome-wide level. The QTL on chromosome 3 was confirmed in the dam-linkage analysis using the sparse SNP panel, and an additional QTL on chromosome 4 was identified (both were significant at the chromosome-wide level). The SP analysis using the same sparse SNP panel also identified the

QTL on chromosomes 3 and 4, with the QTL on chromosome 4 reaching genome-wide significance (Table 3.2).

To estimate the chromosomal position of significant QTL, a further 28 SNPs (to total 36, including SNPs from the sparse panel) were genotyped across the four chromosomes for which QTL were identified at the chromosome- or genome-wide significance level in stage one (i.e. on chromosomes 3, 4, 7 and 23). These were used in a dam-linkage analysis (stage two). This confirmed and positioned the QTL on chromosomes 3 and 4 towards the ends of the chromosomes, at map positions 135cM and 74cM respectively (Table 3.2; Chromosome 3 QTL position shown in Figure 3.5). The SP analysis using the denser set of markers confirmed the QTL on chromosomes 3 and 4 (both reached chromosome-wide significance), and estimated QTL locations overlapped with those obtained from the HS analysis (chromosome 3 at 129cM and chromosome 4 at 75cM; Table 3.2). The confidence interval for the QTL on chromosome 4 was narrowed to 13cM in the SP analysis, using bootstrapping.

Overall, the four QTL were individually estimated to explain between 4 and 9% of the within-family variance (PVE) for PD resistance (Table 3.2). The PVE values estimated using the sire-based linkage analysis only were comparable between the three sire-segregating QTL. The QTL on chromosome 3 was the only QTL identified in both the sire- and dam-based linkage analysis, and explained the highest proportion of within-family variance for resistance when both sire- and dam-based analyses were considered together (Table 3.2).

### 3.4.4 Association analysis

To detect SNPs in linkage disequilibrium with PD resistance QTL, SNPs from the dense marker panel across the four chromosomes harbouring significant QTL were individually analysed for significant association with PD mortality. SNP associations were tested using a mixed model approach, across all 20 genotyped half-sibling families. Overall, two SNPs on chromosome 3 were identified as significantly associated with resistance to PD (SNP1: consensus46559_56; SNP2:

consensus110127_55; Table 3.3). SNP1 was estimated to explain ~2% of the additive genetic variance, whilst SNP2 was estimated to explain ~30% of the additive genetic variance for PD resistance (Table 3.3; all SNP allele frequencies and PVEs in the dense maps for the four significant chromosomes are given in Appendix A, Table A6). However, it should be noted that since estimates were obtained using allele frequencies from the 20 half-sibling families with intermediate levels of mortality, estimates may be biased if the allele frequencies in this subset are not representative of the population as a whole.

**Table 3.2: Summary of genome-wide and chromosome-wide significant QTL across all linkage analyses**

| Analysis | | Chromosome | No. of SNPs | Map length (cM) | QTL position (cM) | F ratio | Chromosome-wide F ratio threshold # | No. of segregating parents | PVE (%) |
|---|---|---|---|---|---|---|---|---|---|
| HS | Sparse Sire | 3 | 1 | NA | NA | 2.3* | 1.7 / 2.1 | 3 | 7.6 / 9.2 ‡ |
| | | 7 | 3 | NA | NA | 2.2* | 1.6 / 2.0 | 3 | 7.4 / 5.5 ‡ |
| | | 23 | 2 | NA | NA | 2.3** | 1.7 / 2.0 | 2 | 8.3 / 4.8 ‡ |
| | Sparse Dam | 3 | 1 | NA | NA | 1.6* | 1.5 / 1.8 | 5 | 7.6 / 9.2 ‡ |
| | | 4 | 2 | NA | NA | 1.8* | 1.5 / 1.7 | 6 | NS / 6.3 ‡ |
| | Dense Dam | 3 | 12 | 135 | 135 | 1.6* | 1.5 / 1.7 | 5 | 10.1 |
| | | 4 | 6 | 77 | 74 | 1.5* | 1.4 / 1.7 | 6 | 9.6 |
| | | 7 | 7 | 65 | NA | 1.1 NS | 1.5 / 1.7 | 4 | NS |
| | | 23 | 11 | 71 | NA | 0.8 NS | 1.5 / 1.7 | 3 | NS |
| SP | Sparse | 3 | 1 | NA | NA | 9.7* | 3.2 / 6.6 | NA | NA |
| | | 4 | 2 | NA | NA | 20.9** | 4.9 / 9.4 | NA | NA |
| | Dense | 3 | 12 | 135 | 129 | 12.5* | 6.7 / 11.5 | NA | NA |
| | | 4 | 6 | 77 | 75 | 10.1* | 6.3 / 11.5 | NA | NA |

PVE: Proportion of within-family variance explained by QTL.
HS: Half-sibling analysis in GridQTL
SP: Sib-pair analysis in GridQTL
NS: Not significant
* Chromosome-wide significant
** Genome-wide significant
# Chromosome-wide thresholds at P<0.05/P<0.01
‡ PVE estimate obtained from sire only/sire+dam sparse SNP analyses

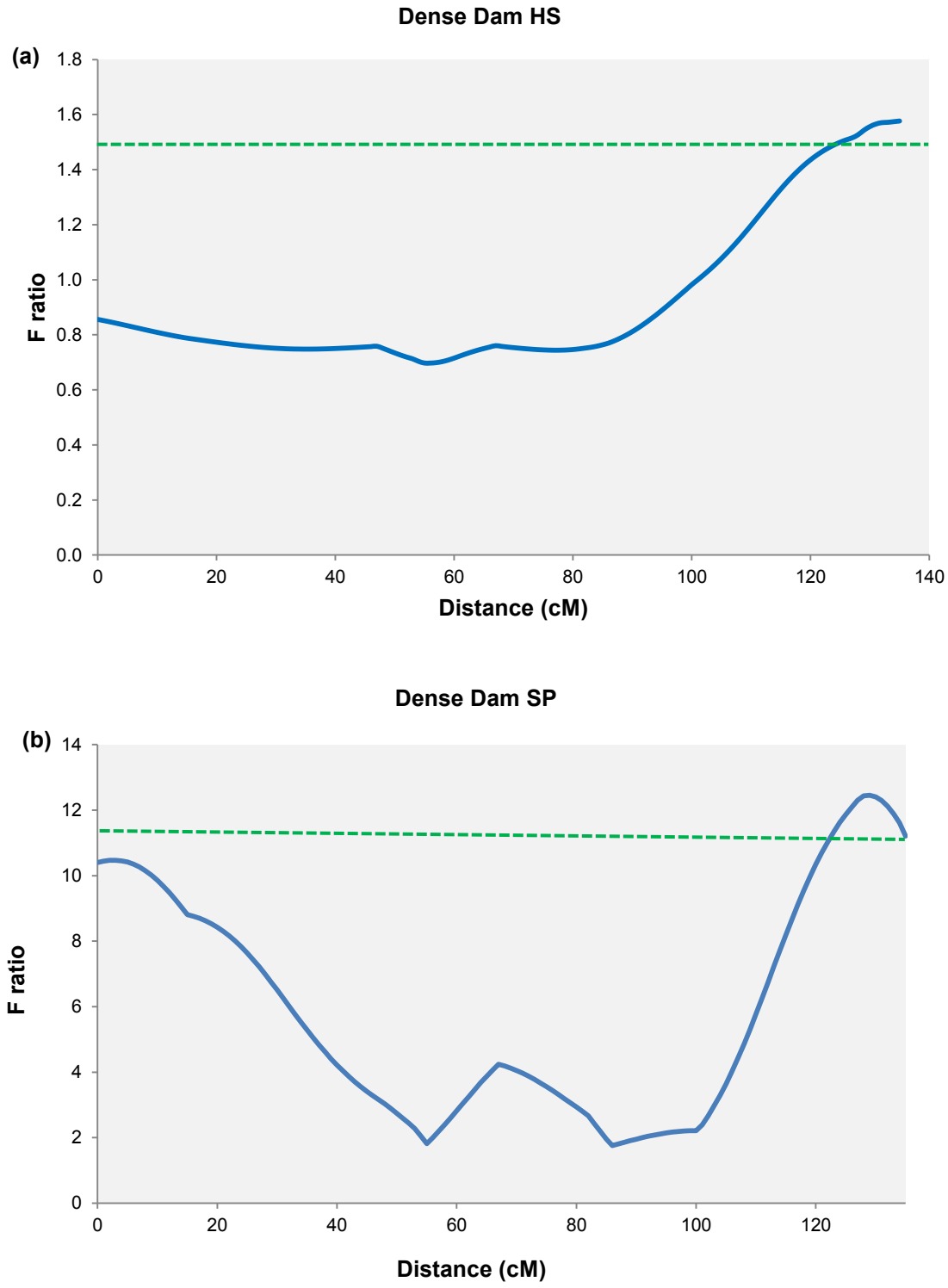



**Figure 3.5: Chromosome 3 QTL location**
The estimated position of the QTL on chromosome 3 identified using the dense SNP marker panel in the (a) half-sibling (HS) and (b) sib-pair (SP) analyses. The horizontal dotted green lines represent chromosome-wide F ratio thresholds.

**Table 3.3: The additive effect, dominance effect, allelic substitution effect and percentage of genetic variation explained by the two significant SNPs on chromosome 3**

| SNP ID | Position on chromosome 3 (cM) | Allele frequencies | Additive effect (±SE) | Dominance effect (±SE) | Average allelic substitution effect (±SE) | Percentage genetic variation explained by SNP |
|---|---|---|---|---|---|---|
| SNP1 consensus46559_56 | 67 | p=0.57;q=0.43 | 0.048 (±0.022) | 0.036 (±0.029) | 0.053 (±0.022) | 1.7 |
| SNP2 consensus110127_55 | 82 | p=0.09;q=0.91 | 0.268 (±0.123) | 0.148 (±0.127) | 0.389 (±0.161) | 29.8 |

Both SNPs were significant at P<0.001.

Average allelic substitution effect: Average change in genotypic value that results when one SNP allele is substituted for the other.

## 3.5 Discussion

To quantify the genetic variation in PD resistance and to explore its genetic architecture, a large population of farmed Atlantic salmon fry was challenged using SAV3, the most prevalent strain of salmonid alphavirus in Norway. Using the binary trait of challenge outcome (died or survived), a high heritability for resistance to PD was obtained (~0.5). Following this, a QTL mapping study identified four putative QTL involved in resistance to PD, on chromosomes 3, 4, 7 and 23. The most convincing and robust evidence for a QTL was detected near the distal end of chromosome 3. This QTL explained the largest proportion of within-family variance for PD resistance, and SNP markers showing population-level association with PD resistance located in close proximity to the QTL peak have been identified.

Recently, genetic parameter estimation and QTL mapping for PD resistance has been conducted in an independent population of Atlantic salmon post-smolts, using broodstock from the 2009 year class belonging to the breeding company SalmoBreed AS. Briefly, 4946 post-smolts (body weight ~85g and 333 days post-hatch) belonging to 284 full-sibling (120 paternal half-sibling) families were intraperitoneally injected with the same strain of the PD-causing salmonid alphavirus (SAV3) as the fry in this study. The challenge lasted for 16 days, with a total mortality of 3,058 fish (62%) at challenge termination. All survivors and mortalities were genotyped using the 6 K Atlantic salmon Illumina iSelect SNP array developed by the Centre for Integrative Genetics (www.cigene.no). The underlying liability for resistance in this population of post-smolts was estimated at ~0.4, and three QTL, on chromosomes 2, 3, and 14, were identified. Analyses within this population were conducted as described for the fry population in this chapter. Results obtained in this study of post-smolts and the current study in fry have been submitted as a joint publication (Gonen *et al*, 2015). Results obtained in the current fry study are discussed in light of results obtained in both populations.

The high heritability obtained in this study is similar to some of the larger estimates reported for disease traits in Atlantic salmon [e.g. 0.55 for infectious pancreatic necrosis (IPN) and 0.51–0.62 for furunculosis (Kjøglum *et al*, 2008; Drangsholt *et al*,

2011)]. Notably, this estimate is almost double that obtained by previous studies when analysing data from natural PD outbreaks in farmed Atlantic salmon smolts [$h^2$=0.21±0.005 on the underlying liability scale; Norris *et al* (2008)]. The underlying liability heritability estimated for PD resistance in the post-smolt population was similar in magnitude to that obtained in the current study for fry ($h^2$ ~0.4) (Gonen *et al*, 2015). The differences in heritability estimates obtained for the fry population in the current study and the other two studies [post-smolts in Gonen *et al* (2015) and Norris *et al* (2008)] may be explained by the difference in life stage (fry vs. smolts), challenge model (tank challenge vs. natural sea challenge) and possibly due to the different and independent origins (and therefore genetic background) of the populations across the three studies. Despite the differences, there is consistent evidence for high heritable variation for PD resistance. This, combined with the large variation in mortality seen across all three independent populations, strongly suggests that selection for PD resistance is plausible.

In addition to heritability for resistance, the heritability for time to death was estimated in this study, and this estimate was much lower than that estimated for resistance. It may be that this trait truly exhibits a low heritability. However, it is important to note that the power of this estimate is lower than for the binary trait of resistance, due to the following reasons. Firstly, the challenge protocol implemented may result in a low viral challenge dose per fish. In addition, the exposure of fish to the virus and fish infection times is expected to be stochastic. Both of these may result in increased noise in the data, and, therefore, lower heritability estimates (Bishop and Woolliams, 2010b; Woo *et al*, 2011) (see section 1.5 of this thesis for a more detailed explanation). Secondly, the time to death estimate was obtained using mortality offspring only. Therefore, fewer individuals were used in this estimate, which again, would result in a lower power to detect heritability.

For the purposes of QTL mapping in this study, subsets of families showing intermediate levels of mortality were deliberately chosen from a larger set of families. This increases the power of QTL detection by linkage analysis by increasing the likelihood of having QTL-segregating parents in the dataset (Darvasi

and Soller, 1992; Hayes *et al*, 2009). Using these families, QTL mapping was conducted using a cost-effective, two-step approach, which combines genotype inheritance patterns with the large differences in recombination rates between Atlantic salmon males and females. The low recombination rate across much of the genome in male Atlantic salmon results in tight linkage between markers, so that large chromosome segments are inherited with minimal or no recombination. This characteristic poses difficulties in some aspects of Atlantic salmon genomics (e.g. linkage map construction; see chapter two of this thesis), but has previously been reported to be advantageous for QTL mapping in Atlantic salmon.

First, using a sparse SNP map, three chromosome-wide and one genome-wide significant QTL were identified in this study. The genome-wide significant QTL on chromosome 23 was segregating in only two sires, albeit with large effect (estimates of effect on mortality proportion: 0.71±0.16 and 0.48±0.17), which is consistent with a relatively rare variant with a large effect on PD resistance segregating within this commercial population. To support this, only two (of 55) dams reached significance in the dam-based QTL analyses, and both were estimated to have a large within-family effect. However, the overall dam-based analysis for chromosome 23 did not reach significance. The use of a denser SNP panel in stage two enabled the positioning of the QTL on chromosomes 3 and 4 towards the distal ends of the respective chromosomes.

Both the sire- and dam-based HS analysis and the SP analysis identified the QTL on chromosome 3, and this QTL also explained a larger proportion of the within-family variance for PD resistance than the other three QTL. Importantly, this QTL was independently identified in the QTL mapping study conducted in the population of post-smolts, where it reached genome-wide significance (Gonen *et al*, 2015). To confirm that this QTL maps to the same region of chromosome 3 in both populations, common markers between the linkage maps used in QTL mapping in the two populations were identified, and used as anchors to orientate the maps relative to each other. Likelihood profile maps were plotted to demonstrate that the QTL mapped to the same region of chromosome 3 (Figure 3.6). This suggests a common

QTL influencing resistance in both populations, and across both the fry and post-smolt stages of the Atlantic salmon lifecycle.

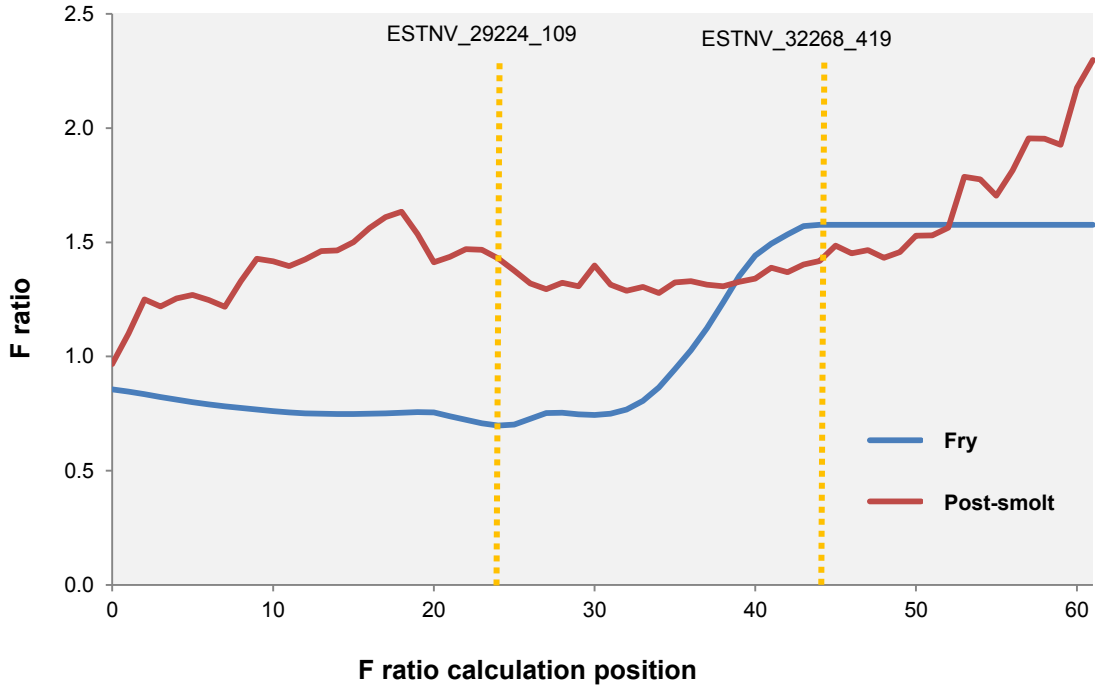



**Figure 3.6: Confirmation of QTL location identified in the fry challenge (this study) and the post-smolt challenge (Gonen *et al*, 2015)**

F ratio calculation positions (x-axis) are given according to the map used for QTL mapping in the post-smolt population. ESTNV_29224_109 and ESTNV_32268_419 are markers common to the two studies, and were used to orientate the two maps relative to each other in order to confirm the overlap in QTL location.

If the causal factors underlying the QTL on chromosome 3 (or any of the four QTL identified in this study) are related to an immune response, they are likely to be part of a general innate immune response against pathogenic infections, since the adaptive immune response is undeveloped at the juvenile fry stage. Interestingly, chromosomes 3, 4 and 23 have also been found to contain QTL influencing resistance to the parasite *Gyrodactylus salaris* (Gilbey *et al*, 2006), which raises the possibility of common resistance QTL for the two diseases. A significant induction of innate genes such as IFN and Mx genes has previously been implicated in the response of salmon to SAV challenge (Grove *et al*, 2013; Herath *et al*, 2013). The

innate defence pathways in which these genes are involved may be partially regulated by the causal factor(s) underlying the QTL identified in this study. Alternatively, the causative variant(s) underlying the resistance QTL may be involved in blocking the progression of the viral life cycle, by preventing viral entry, cell internalisation or replication.

If the mechanisms underlying the QTL on chromosome 3 are innate immune related, then this may have influenced the observed difference in the results obtained from the fry challenge in this study and in the post-smolt challenge (Gonen *et al*, 2015) as follows. Although both studies were able to detect the QTL on chromosome 3, a difference in the QTL significance level and effect was observed (i.e. chromosome-wide only in fry vs. genome-wide in post-smolts). This may be explained by the differences in challenge protocol and the resulting challenge mortality profiles. In the intraperitoneally-injected post-smolt challenge, a single sharp peak in mortalities was seen, and the challenge duration was short (18 days). In comparison, the exposure of fry to SAV was conducted using free viral particles in tank water, the challenge duration was much longer (56 days), and at least two peaks in mortality were observed (days 28 and 35). Given the longer challenge duration in fry, the first peak may represent mortalities due to differences in innate immune mechanisms (i.e. chromosome 3 QTL effect). After this stage, it could be expected that the adaptive immune response is activated, and the second peak in mortalities may be due to differences in adaptive immune response at other QTL. Combined analysis of all mortalities as implemented in this study may dilute the effect of the QTL on chromosome 3 and reduce the power of detecting the initial chromosome 3 QTL effect influencing the innate immune response to viral infection.

Natural outbreaks of PD are recorded almost exclusively at the post-smolt stage of the Atlantic salmon lifecycle. Although viral isolates have been detected in freshwater, no outbreaks of PD in the fry stage have been recorded (Jansen *et al*, 2010b). The practicality of challenge experiments within a single tank at the fry stage has meant that disease challenge experiments are often carried out at the fry stage of the salmon lifecycle, as a model for post-smolt PD outbreaks [e.g. see Cano *et al*

(2014)]. However, differences in behaviour, environment and physiology between the freshwater and marine stages mean that the genetic architecture of PD resistance could differ between the two life stages. As such, PD resistance QTL identified under the fry challenge model would need to be evaluated in post-smolts in a marine environment. The mapping of the QTL on chromosome 3 in independent Atlantic salmon fry and post-smolt populations, and the similarities in the heritability estimated obtained across the two studies, suggest that a PD fry challenge may be a suitable model for the estimation of genetic parameters and informing selection decisions for increased PD resistance at the post-smolt stage.

Marker-assisted selection (MAS) has been successfully applied in aquaculture breeding, including for female monosex Chinook production (Devlin *et al*, 1991), for resistance to lymphocystis disease in Japanese flounder (Fuji *et al*, 2007), and for resistance to IPN in Atlantic salmon (Houston *et al*, 2008; Moen *et al*, 2009; Houston *et al*, 2010). In the current study, two SNPs showing population-level association with the PD resistance QTL on chromosome 3 have been discovered, which may be used in MAS to select for improved PD resistance in commercial Atlantic salmon.

However, it is possible that these markers are still a considerable distance from the causative variant(s). In addition, although both SNPs were significantly associated ($P<0.001$), a large difference between their relative estimated effects was observed, where SNP2 (estimated to be closest in distance to the QTL) was estimated to explain ~30% of the additive genetic variance, and the other, SNP1 (more distantly positioned), was estimated to explain ~2%. The reason for this could simply be that SNP2 is more closely located to the QTL, and, therefore, the linkage relationship between this SNP and the QTL is more frequently observed compared to SNP1 and the QTL, due to the lower likelihood of a recombination event within smaller genetic distances. Therefore, SNP2 would be a better predictor of QTL effect. Alternatively, the higher estimate at SNP2 could be due to a bias in allele frequency at that SNP in the selected 20 half-sibling families, compared to the population average. The strategy of family selection used in this study to detect QTL (i.e. families showing intermediate levels of mortality) was chosen to increase the chance of detecting a

segregating QTL, by increasing the minor allele frequency (MAF), and, therefore, maximising the additive genetic variance explained. As such, the estimated proportions of genetic variance explained for both SNPs are likely to be overestimates, since the population level MAF is likely to be lower than in the selected sample. However, the MAF of SNP2 in the selected sample is low (0.09). As a result, only a small proportion of individuals in the sample were homozygous for the minor allele, resulting in large standard errors for all estimates. Since the MAF of this SNP in the overall population is unknown, the confidence in the effect estimated for SNP2 (given the large standard error) using the sampled families is reduced.

Further refinement of the QTL position and eventual identification of candidate genes would be advantageous for both applied MAS (more accurate marker predictors of QTL genotype) and our understanding of the biological basis of genetic resistance to PD. This could be achieved by genotyping and testing a larger number of SNPs in the region of the QTL, using, for example, a high-density SNP array (Houston *et al*, 2014a), or by re-sequencing of alternate homozygotes at the QTL on chromosome 3. Additionally, positional and functional candidate genes for the QTL may be generated by taking a comparative genomics approach, as demonstrated for Infectious Salmon Anaemia resistance (Li *et al*, 2011). This will be greatly assisted by improvements in the assembly and annotation of the Atlantic salmon genome (Davidson *et al*, 2010).

## 3.6 Conclusion

Using a population of Atlantic salmon fry challenged with SAV3, a high heritability for resistance to PD was obtained ($h^2$ ~0.5), demonstrating the feasibility of family selection for PD resistance. A QTL mapping analysis conducted within this population identified four chromosomes (3, 4, 7 and 23) harbouring putative PD resistance QTL. The QTL on chromosome 3 was replicated in both a sire- and dam-based linkage analysis, and further, explained the largest within-family variation for resistance. Importantly, this QTL has been independently identified in a population of post-smolts, also challenged with SAV3. Concordance in the estimated position of

the QTL, on the distal end of chromosome 3, was obtained across both populations, suggesting a common mechanism for PD resistance across both life stages. SNPs on this chromosome showing population-wide association with PD resistance were identified, and these could be implemented in MAS for improved PD resistance. Higher density SNP resources, coupled with the availability of the Atlantic salmon genome reference, will facilitate further fine-mapping and characterisation of candidate genes underlying this QTL, leading to more effective selection for resistance to this important disease.

## 3.7 Acknowledgements

I would like to thank Dr John Taggart (Institute of Aquaculture, Stirling) for providing me with the required files to use the FAP parentage assignment software package and for his assistance in its usage, Marine Harvest for the use of their data, and Matthew Baranski (Nofima, As, Norway) for

# Chapter 4

# Identification of candidate genes and biological pathways affecting host resistance to Infectious Pancreatic Necrosis virus using comparative genomics

## 4.1 Abstract

Infectious pancreatic necrosis (IPN) is one of the most problematic viral diseases affecting productivity on Atlantic salmon farms. The estimated high heritability for resistance to this disease obtained from both field and challenge experiments ($h^2$ range: 0.31–0.61) has meant that family-based selection for IPN resistance has been incorporated into breeding programs. With the mapping of a major resistance QTL to linkage group 21 and the identification of genetic markers tightly linked to this QTL, improvements in selection efficiency through marker-assisted selection and in resistance to IPN amongst fish stock have been achieved. Further improvements in selection efficiency may be gained through the identification of causative gene(s)/variant(s) underlying this major QTL. The current Atlantic salmon reference genome assembly is still fragmented, and, as such, the identification of positional candidate genes is challenging. The aim of this study was to generate a list of putative candidate genes within the vicinity of the resistance QTL. Mapping of QTL-linked sequences to four published model teleost genomes identified two QTL-orthologous regions in each fish species. Conservation of gene order across species within these regions was observed, highlighting the possible utility of comparative genomics for identifying positional candidate genes. Microarray analysis of gene expression in IPN resistant and susceptible salmon before and after viral challenge identified lists of functional candidate genes which showed significant differential expression. Subsequent fine-mapping of the IPN-QTL region and mapping of the

differentially-expressed functional candidate genes to the stickleback genome highlighted a single QTL-orthologous region on stickleback linkage group II. QTL region and genome-wide pathway enrichment analysis of differentially-expressed genes suggested that viral entry/replication, cell energy production and apoptotic pathways may potentially be involved in resistance. The genes and pathways identified provide candidates for further investigation. Such studies could look for causative factors for potential cis-regulatory effects, which might be responsible for the marked phenotypic difference observed between IPN resistant and susceptible fish.

## 4.2 Introduction

Infectious pancreatic necrosis (IPN), caused by the IPNV aquabirnavirus, has been amongst the most problematic viral diseases affecting aquaculture farms (FAO, 2014b). General management practices implemented to limit the effect of the disease have not been fully effective, and the disease has been reported to cause mortalities as high as 80% in Atlantic salmon fry and 30% in post-smolts (Guy *et al*, 2009). Several studies analysing data from natural field and experimental challenge populations have suggested a strong host genetic component to IPN resistance (heritabilities range from 0.31–0.61), and family-based selection for improved resistance has been applied since 1997 (Guy *et al*, 2006; Kjøglum *et al*, 2008; Guy *et al*, 2009). Although improvements in resistance amongst fish stock using family-based selection has been recorded, the four-year generation interval of Atlantic salmon has meant that response to selection has been limited, and the disease remained a significant problem for the industry (Storset *et al*, 2007; Houston *et al*, 2008; Moen *et al*, 2009). As such, methods, such as the incorporation of genetic markers in to breeding programs, which are able to select based on individual rather than family-level resistance and potentially offer more rapid progress, became desirable.

The consistent estimates of a moderate-to-high heritability for resistance to IPN across populations (Guy *et al*, 2006; Kjøglum *et al*, 2008; Guy *et al*, 2009) and the subsequent identification of the major IPN resistance QTL on linkage group (LG) 21

(Houston *et al*, 2008; Moen *et al*, 2009) has enabled the use of marker-assisted selection (MAS) in breeding programs. This QTL was estimated to explain 21–32% of the within-family phenotypic variation, and 83–98% of the additive genetic variation for resistance (Houston *et al*, 2008; Moen *et al*, 2009; Houston *et al*, 2010). Single nucleotide polymorphism (SNP) and microsatellite markers tightly linked to this QTL have been identified, and MAS using these markers has resulted in improved selection efficiency and a decline in the number of IPN mortalities [(AquaGen, 2013), non-peer-reviewed publication].

One limitation of MAS is that, even with confirmed marker-trait associations, the linkage disequilibrium between genetic markers and QTL might break down or vary across different populations. Therefore, marker-QTL associations must be confirmed in populations of interest prior to being implemented in breeding programs. A more robust selective breeding strategy would be to select directly for the resistance causative variant(s) (Haley and Visscher, 1998; Sonesson, 2005). As well as improving selection efficiency, knowledge of the genes and pathways involved in resistance can help in the long term development of treatments and vaccines, as has been applied for viral pathogens causing respiratory diseases in cattle (Glass *et al*, 2012). Currently, the gene(s) underlying this QTL, and the biological pathways involved in resistance, are not publically known. Recent unpublished reports have suggested that a single (undisclosed) gene in the QTL region contains the causative variant(s) underlying resistance (Moen and Ødegård, 2014).

The overall aim of this study was to generate a list of candidate genes potentially underlying the IPN resistance QTL in Atlantic salmon. In the absence of a well assembled and annotated Atlantic salmon reference genome, a comparative mapping approach was adopted. To identify QTL-orthologous regions within published genome sequences of model teleost fish species, QTL-linked sequences derived from Bacterial Artificial Chromosome (BAC)-end sequencing and Restriction-site Associated DNA sequencing (RAD-Seq) were aligned to the three-spined stickleback, zebrafish, medaka and green spotted puffer fish genomes. To identify putative positional candidate genes, genes within the QTL-orthologous regions in all

four model fish species were identified, and conservation of gene order across species was investigated.

To infer putative functional candidate genes within the orthologous regions, gene expression data from an IPNV challenge experiment in resistant and susceptible Atlantic salmon fry were utilised. The microarray probes corresponding to the genes in the orthologous regions were identified and investigated for evidence of differential expression between resistant and susceptible individuals, prior to and after IPNV challenge. Analysis of gene differential expression can highlight candidate genes potentially harbouring mutations within regulatory regions (such as promoters), which may be influencing their expression.

Alternatively, pathway enrichment analyses of all differentially-expressed genes could highlight candidate genes involved in the regulation of biological pathways. These genes may not be differentially expressed themselves, but may be harbouring the underlying QTL causative mutation(s) that causes downstream effects on gene expression within the same pathway(s). To identify biological pathways likely to be involved in resistance to IPN, pathways with enrichment for differentially-expressed genes were identified. The concordance between the comparative positional candidate genes and the functional candidate genes identified from the gene expression and pathway analysis was investigated. Overall, this study has identified a list of potential candidate genes and pathways which may be influencing resistance. These results will be useful in directing future investigations into the resistance conferring causative variant(s).

## 4.3 Materials and Methods

### 4.3.1 Identification of IPN QTL-orthologous regions in teleost fish genomes

All datasets used in this study, i.e. viral challenge experiments, QTL-linked markers and sequences and microarray gene differential expression analyses, were obtained prior to the start of this project. All downstream bioinformatic and comparative analyses were conducted by me, using these available resources. A brief description

of the procedures used to generate these resources is provided herein, followed by the more detailed analyses procedures conducted by me.

#### 4.3.1.1 Generation of QTL-linked sequences

To generate a list of putative candidate genes within the initially identified 10 cM IPN QTL region (Houston *et al*, 2008), QTL-linked sequences were generated using two different sequencing methodologies. First, short length sequences (range: 95–457bp; Table 4.1) within the vicinity of the QTL were generated, using Restriction-site Associated DNA sequencing (RAD-Seq). RAD-Seq was conducted within two families [labelled B and C in Houston *et al* (2010)] originating from the Landcatch Natural Selection Broodstock (Cooperage Way, Alloa, FK10 3LP UK). These families were chosen since they appeared to be segregating at the IPN QTL (i.e. both parents were QTL heterozygotes) (Houston *et al*, 2010). From each family, seven QTL homozygote resistant (RR) and seven QTL homozygote susceptible (SS) offspring were identified [QTL genotypes for parents and offspring were determined using a microsatellite genotyping panel, as described by Houston *et al* (2010)].

DNA from fin tissue of all parents and offspring (4 parents, 14 RR offspring, 14 SS offspring) was digested with the *Sbf*I restriction enzyme and RAD libraries for each individual were prepared. Libraries were sequenced on the Illumina sequencer to generate paired-end RAD contigs [full protocol in Houston *et al* (2012); the paired-end RAD-Seq approach is detailed in Figure 1.3 and in chapter two, Materials and Methods]. Briefly, fragments produced by the restriction enzyme digest were ligated to a P1 adaptor, which is comprised of a unique sample barcode and an Illumina sequencing primer. Enrichment for P1 adaptor-ligated fragments, followed by shearing and size selection (300–700bp), generated a second set of fragments. Sheared ends of fragments were ligated to a P2 adaptor, and fragments were sequenced on the Illumina sequencing platform from both adaptors (paired-end). Pooling of resulting reads based on sample barcode and the stacking of reads into RAD loci within and then across individuals created a mini-contig for each identified RAD locus.

In total, 71,404 RAD loci in family B and 70,938 RAD loci in family C were identified, of which 69,286 were common to both families (Houston *et al*, 2012). RAD contigs were aligned to QTL-linked SNP-flanking sequences, and those showing significant alignment were identified. Overall, 44 and 78 IPN QTL-linked RAD contigs, corresponding to 6 and 15 RAD loci for families B and C respectively (more than one contig is generated per RAD locus), were identified.

Second, BAC clones from in and around the IPN-QTL region were generated. Briefly, alignment of QTL-linked marker flanking sequences (microsatellite and SNP) to the Atlantic salmon BAC physical library [(Ng *et al*, 2005); data available in the cGRASP database, 2009] identified Contig fsp378 as the contig closest to the IPN-QTL region. A minimum tiling path of BAC clones across this contig were sequenced using a 454-sequencing platform by The GenePool (now part of Edinburgh Genomics). *De novo* assembly of raw sequences using the Newbler software package (GS Data Analysis package; http://www.454.com/products/analysis-software/, version 3) generated 28 BAC contigs, ranging in length from 122–40,983 base pairs (bp) (total sequence available: 242,005bp) (Table 4.1).

To summarise, the data available for analysis prior to further quality control of QTL-linked sequences were: 28 BAC contigs (length range: 122–40,983bp) from the BAC clone assembly; 44 RAD contigs from family B (length range: 95–410bp); and 78 RAD contigs from family C (length range: 95–457bp) (Table 4.1).

#### 4.3.1.2 Filtering of QTL-linked sequences: Elimination of repetitive elements and contigs of bacterial origin

To identify and remove sequences of bacterial origin (contamination due to the cloning procedure), BAC contigs were aligned to the NCBI non-redundant database [BLASTN, BLAST+ package version 2.2.25+; Zhang *et al* (2000)]. In total, 15 BAC contigs comprised uniquely of bacterial sequences were eliminated, leaving 13 BAC contigs for further analysis.

Previously published studies suggest that the genome of Atlantic salmon is highly repetitive, which may partially be due to the recent and ancient whole genome duplication events (McKay *et al*, 2004; Danzmann *et al*, 2005; Danzmann *et al*, 2008; Koop and Davidson, 2008; Guyomard *et al*, 2012; Berthelot *et al*, 2014). In comparative mapping studies, sequences originating from repetitive regions of the genome may align to multiple locations within the compared genome, making the identification of cross-species chromosomal orthologous relationships more challenging (Li *et al*, 2011). In an attempt to minimise this, both the BAC and RAD contigs were repeat-masked, using the online Atlantic salmon repeat masking software (http://lucy.ceh.uvic.ca/repeatmasker/cbr_repeatmasker.py). Many of the 13 BAC contigs were heavily masked (average masking of contigs: 42%). After repeat-masking, 32 of the 44 (73%) QTL-linked RAD contigs in family B, and 73 of the 78 (94%) QTL-linked RAD contigs in family C remained for further analysis.

To summarise, 13 BAC contigs (total sequence ~160Kb), 32 RAD contigs from family B (total sequence ~5Kb), and 73 RAD contigs from family C (total sequence ~14Kb) remained post-filtering (Table 4.1).

**Table 4.1: IPN QTL-linked sequence descriptions**

| Contig type | Originating from | Pre-filter | | | Post-filter | | | |
|---|---|---|---|---|---|---|---|---|
| | | Total | Length (bp) | Total sequence (bp) | Total | Length (bp) | Total sequence (bp) | Number of contigs (%) with significant alignment to stickleback genes |
| BAC | Newbler assembly | 28 | 122–40,983 | 242,005 | 13 | 126–40,983 | 158,991 | 3 (10.7) |
| RAD | Family B | 44 | 95–430 | 5,961 | 32 | 95–410 | 4,731 | 5 (11.4) |
| | Family C | 78 | 95–457 | 14,581 | 73 | 95–457 | 13,782 | 20 (27.4) |

#### 4.3.1.3 Comparative genomic exploration using published teleost genomes

Previous comparative genome mapping studies suggest that, despite the ancient origin and millions of years of divergence of teleost species (Near *et al*, 2012; Rosindell and Harmon, 2012; Berthelot *et al*, 2014), large regions of conserved orthologous relationships can be identified across species, although gene order may not necessarily be conserved (Danzmann *et al*, 2008; Phillips *et al*, 2009; Davidson *et al*, 2010). Based on this assumption, to identify candidate genes within the IPN QTL region in Atlantic salmon, a comparative mapping approach was adopted in this study, using QTL-linked sequences.

At the time of the analysis, five teleost reference genomes were available in Ensembl (Ensembl 69, http://oct2012.archive.ensembl.org/index.html): *Danio rerio* (zebrafish, Ensembl Dataset=*Danio rerio* genes, Zv9), *Gasterosterus aculeatus* [three-spined stickleback, Ensembl Dataset=*Gasterosterus aculeatus* genes (BROADS1)], *Oryzias latipes* [medaka, Ensembl Dataset=*Oryzias latipes* genes (HdrR)], *Takifugu rubripes* [fugu, Ensembl Dataset=*Takifugu rubripes* genes (FUGU4.0)] and *Tetraodon nigroviridis* [green spotted puffer fish, Ensembl Dataset=*Tetraodon nigroviridis* genes (TETRAODON8.0)]. The fugu genome assembly was still in a fragmented state, with 90% of the genome assigned to 1,118 scaffolds, most of which were <1Mb in length (Fugu Genome Project, http://www.fugu-sg.org/; Ensembl 69, http://oct2012.archive.ensembl.org/Takifugu_rubripes/Info/Annotation/#genebuild). Therefore, further analysis using fugu was not performed.

To identify putative IPN QTL-orthologous regions within sequenced and annotated teleost genomes for the purposes of resistance candidate gene identification, filtered BAC and RAD contigs were aligned [BLASTX; BLAST+ package version 2.2.25+; Zhang *et al* (2000)] to the four teleost reference genomes in the Ensembl database. BLASTX was chosen as the alignment algorithm since it identifies protein coding regions only, which are more likely to be conserved across species compared to non-genic regions of the genome (Hardison *et al*, 1997; Brenner *et al*, 2002; Santini *et al*, 2003). For each significant alignment identified (E-value<$1e^{-5}$), the position (bp) of

the alignment on the reference teleost chromosome, E-value, alignment score and percentage identity was recorded.

To infer IPN QTL-orthologous chromosomes in the published teleost fish genomes, the number of times the salmon QTL-linked sequences showed significant alignment to genes on a particular chromosome in the other fish genome was counted. For some RAD loci, multiple paired-end contigs may be obtained, which may be caused by read assembly errors or due to the incorrect grouping of paralogous regions of the Atlantic salmon genome to a single locus. Since these are expected to align to the same location in the compared genome, they may cause a bias in the number of significant alignments identified and used to infer chromosomal orthologies. To address this and to prevent bias in count, RAD contigs derived from the same RAD locus were grouped together.

For each of the teleost chromosomes identified as orthologous to the IPN QTL, a region of orthology (in bp) was defined, based on the chromosomal locations of significant alignments. These regions were used as chromosome coordinates, and sequences of genes residing within these regions were obtained from BioMart (http://www.ensembl.org/biomart/martview/). To identify conserved orthologous relationships and to explore gene conservation across the four teleost species, genes common to all of the QTL-orthologous regions were identified and counted.

### 4.3.2 Narrowing the QTL-orthologous region in the three-spined stickleback

Although analysis of the RAD and BAC contigs enabled the identification of QTL-orthologous regions in teleost genomes, these regions were often large (see Results). To reduce the size of the QTL orthologous regions, fine-mapping of the QTL, followed by comparative mapping analyses of sequences more tightly linked to the QTL region was conducted, as described below.

#### 4.3.2.1 IPN QTL fine-mapping

Using a panel of microsatellite markers, the first mapping study positioned the IPN QTL to within a 10 centiMorgan (cM) (~10Mb) confidence interval on LG 21 (Houston *et al*, 2008). Subsequent studies utilising SNP markers were able to reduce this confidence interval to ~2cM (~2Mb) (Houston *et al*, 2012). To facilitate the integration of the map utilised in Houston *et al* (2012) with existing dense SNP linkage maps, eleven evenly spaced SNP markers previously mapped to LG 21 in the dense SNP map described in Lien *et al* (2011) were genotyped in families B and C. These eleven SNPs were integrated with the IPN QTL SNP linkage map in Houston *et al* (2012) using the CRI-MAP software package [Green *et al* (1990); version 2.4, as modified by Xuelu Liu (Monsanto)]. Re-estimation of the location and confidence interval for the QTL using this integrated SNP map and families B and C identified two SNPs within the 2cM confidence interval of the QTL (SSA0019 and RAD010201).

#### 4.3.2.2 Comparative genomic analyses of the 2cM QTL region

To generate a list of candidate genes within this narrower QTL region, the flanking sequence of these SNPs were aligned (TBLASTX, E-value<$1e^{-5}$) to the stickleback genome only. Of all four teleost species available, the stickleback reference genome was chosen for further refinement and characterisation of this QTL region since it is a high quality assembled and annotated model reference genome, and has previously been shown to be useful for the identification of conserved orthologous relationships with salmonid species [e.g. Li *et al* (2011), Guyomard *et al* (2012)].

The sequence associated with SSA0019 (798bp) aligned to stickleback LG II at 8.33Mb, to the gene poly(A)-binding protein, nuclear 1-like (cytoplasmic) (ENSGACG00000015428). The sequence associated with RAD010201 (95bp) did not show significant alignment to the stickleback genome. To obtain a longer sequence associated with RAD010201 for further investigation, the RAD010201 sequence was aligned [BLASTN; BLAST+ package version 2.2.25+; Zhang *et al* (2000)] against the Atlantic salmon draft genome assembly (first draft assembly was utilised; NCBI Assembly GCA_000233375.1;

www.ncbi.nlm.nih.gov/Traces/wgs/?val=AGKD01). The identified contig (reference genome contig ID: gi|354363830|gb|AGKD01095206.1| Salmo salar Contig_095218) was analysed for the presence of putative genes (TBLASTX against known stickleback reference gene sequences, E-value<$1e^{-5}$). The gene closest to RAD010201 was NEDD8 activating enzyme E1 subunit 1 (ENSGACG00000015403), which mapped to 8.25Mb on LG II in stickleback [~80.4Kb from poly(A)-binding protein, nuclear 1-like (cytoplasmic)]. Since the QTL position in salmon was estimated to within a 2cM (~2Mb) confidence interval, the narrower QTL-orthologous region on stickleback LG II was inferred as 7.3–9.3Mb (1Mb on either side of 8.3Mb, the approximate corresponding position of these two QTL-linked SNP markers).

To identify putative positional IPN resistance candidate genes, sequences of annotated stickleback genes from within this 2Mb region were obtained from BioMart (gene set 1). In addition, sequences for all LG II genes (gene set 2), as well as all known and annotated gene sequences from the stickleback genome (gene set 3), were extracted from BioMart [Ensembl 69; *Gasterosterus aculeatus*, Ensembl Dataset=*Gasterosterus aculeatus* genes (BROADS1)].

### 4.3.3 Inferring functional roles for positional putative candidate genes

Up- or down-regulation of the expression of genes within biologically relevant pathways is often observed in response to external stimuli (such as viral infections). The differential regulation of a gene as a response to external stimuli can highlight it as a potential candidate for further analysis, and may enable the inference of biological pathways involved in the response. In addition, previous studies suggest that the clustering of genes to the same genomic location which show co-ordinated expression and which act within the same biological pathway is a widespread phenomenon, and many examples of non-random conservation of gene order across eukaryotes has been described [see Hurst *et al* (2004) for a summary]. This could occur if, for example, genes are under the control of a cis-acting QTL, due to a mutation within a promoter or enhancer for several genes, or due to a mutation in a gene encoding a shared transcription factor.

To investigate whether the putative positional candidates within the 2Mb QTL-orthologous region in stickleback play a functional role in resistance, positional candidates were tested for differential expression between IPNV resistant and susceptible individuals after viral challenge. This was done using available microarray gene expression data generated prior to the start of the project. A brief description of this data is given below, followed by descriptions of the functional analyses conducted by me using this information.

#### 4.3.3.1 Microarray differential expression analysis

To identify genes which appear to be differentially regulated between IPN resistant and susceptible individuals upon exposure to the virus, challenge experiments for analysis of gene expression patterns were set up as follows [described in Houston *et al* (2010)]. 20 families of Atlantic salmon fry were challenged with IPNV, with two replicate tanks of fry challenged for each family. For each family, the level of mortality was averaged across the two replicate tanks, and mortalities across these families ranged from 0-34% upon challenge termination (see Appendix D, Figure D1). Based on the levels of mortality, families J and N were designated susceptible, families Q and T appeared resistant and families I, P, B, O, D, S, C and L were designated as intermediate. To ascertain the QTL genotype of parents of challenged offspring within these families, a fin sample from each parent was removed and genotyped at the IPN QTL-linked microsatellite markers given in Houston et al. (2010). Families B and C were identified as 'double heterozygote' families where both parents were putative heterozygotes for the QTL, and, therefore, subsequent gene expression data was considered for these two families only.

Gene expression patterns between resistant and susceptible offspring within families B and C was analysed as follows. Each family was represented by three tanks each containing 100 fry, one of which was terminated and sampled at 1 day post-challenge ('time point 1'), one at 7 days post-challenge ('time point 2') and one at 20 days post-challenge ('time point 3'). In addition, a sample of 100 fry from all families was taken prior to challenge ('time point 0'). To ascertain QTL genotype of sampled

individuals at each time point, a fin sample from each offspring was removed and genotyped at the IPN QTL-linked microsatellite markers given in Houston et al. (2010). At each time point, RNA was extracted from six fish of each QTL genotype (i.e. homozygote resistant at the IPN QTL: RR; or homozygote susceptible at the IPN QTL: SS) and hybridised to the Agilent 44K (Atlantic salmon) Oligo Array (Martin *et al*, 2007). This microarray is comprised of 43,661 probes (partial gene sequences), representing ~90% of the known Atlantic salmon expressed sequence tags (ESTs) [Rise *et al* (2004); http://web.uvic.ca/grasp/microarray/array.html].

Preliminary studies conducted by our group suggested a significant up-regulation of innate immune genes in susceptible individuals by day 7 (time point 2), which was not seen in resistant individuals (see Appendix D, Figure D2 for volcano plots showing up- and down-regulation of transcripts in susceptible and resistant families at time points 1, 2 and 3). This suggests that by this point, an immune response to viral infection has already been activated in susceptible individuals, but no large-scale activation of the immune system was observed within resistant individuals. Therefore, the causative mechanisms associated with initial genetic resistance are likely to be active prior to, or just after, initial viral infection (i.e. time point 0 or 1).

To incorporate this information into the current study, probe microarray signals (proxy for expression levels) for all included resistant and susceptible individuals were combined across the pre-challenge and post-challenge time points (0 and 1). Significant differential expression of probes was determined by comparing the average microarray signal across both time points, using a 3-way ANOVA [factors = QTL genotype (resistant vs. susceptible), family (B or C), and time point (0 or 1)]. To avoid exclusion of genes of potential biological relevance, a nominal threshold of $P<0.05$ for significance was chosen (i.e. P-values were not corrected for multiple testing). A total of 1,924 differentially-expressed probe sequences were identified, and the top 100 differentially-expressed genes are given in Appendix A, Table A7.

#### 4.3.3.2 Functional roles of positional IPN resistance candidate genes

To assess whether expression of putative positional candidate was altered between resistant and susceptible genotypes before and after IPNV challenge, sequences of genes within the stickleback QTL-orthologous region were aligned to the salmon microarray probe sequences. First, to determine the number of stickleback genes with representative probes on the microarray, all salmon microarray probe sequences (regardless of expression patterns) showing significant alignment (BLASTN, E-value<$1e^{-5}$) to genes in stickleback gene sets 1–3 were identified. To investigate the enrichment of differentially-expressed genes to the 2Mb QTL-orthologous region on stickleback LG II relative to the whole of LG II and to the whole stickleback genome, the numbers and proportions of differentially-expressed genes in each gene set were identified based on alignment to differentially-expressed probes and compared.

#### 4.3.3.3 Pathway enrichment

Candidate gene identification using microarray differential expression between resistant and susceptible individuals is a way of predicting the involvement of a gene in the underlying disease resistance pathways. However, the causative gene(s) may be involved in the regulation of downstream genes in biological pathways involved in disease progression, and differential expression of the QTL causative gene(s) itself may not necessarily be observed.

To identify biological pathways which may be differentially regulated between resistant and susceptible individuals (and, therefore, may be involved in conferring resistance), pathway analysis of all differentially-expressed microarray probes (corresponding to gene sequences) was conducted. First, differentially expressed probe sequences were assigned *Homo sapiens* Entrez IDs based on gene orthology, using the KOBAS software package (KOBAS package; http://kobas.cbi.pku.edu.cn/home.do). Of the 1,924 differentially-expressed probes, 1,602 (83%) had identifiable corresponding *Homo sapiens* Entrez IDs. *Homo sapiens* Entrez IDs were used in order to include as many probes in the analysis as possible, since fish Entrez IDs are sparse and relatively poorly annotated with pathway

information. Pathway analysis based on these Entrez IDs was conducted using the Ingenuity Pathway Analysis (IPA) software package (INGENUITY package; http://www.ingenuity.com/) (see Appendix E for IPA parameters). This analysis produced a list of pathways enriched for putative genes which were differentially-expressed between QTL resistant and susceptible individuals. One or many of these pathways may be harbouring a gene containing the mutation underlying resistance.

To identify and select genes located in the stickleback 2Mb QTL-orthologous region which map to one or more of these differentially-expressed pathways (regardless of whether the gene itself was differentially-expressed), all gene ontology (GO) terms (i.e. biological pathways) for genes within this region were identified (GO terms were available with Ensembl BioMart gene sequence downloads). Genes within the 2Mb region which mapped to one or more of the differentially-expressed pathways were extracted.

## 4.4 Results

### 4.4.1 Identification of IPN QTL-orthologous regions in sequenced teleost genomes

To determine the location of IPN QTL-orthologous regions in four model teleost species (three-spined stickleback, medaka, green spotted puffer fish and zebrafish) for the purposes of candidate gene identification, a set of IPN QTL-linked Atlantic salmon RAD and BAC sequence contigs were aligned to the reference genomes of the four fish species. Overall, 11% of the BAC contigs, and 11% and 26% of the QTL-linked RAD contigs in families B and C respectively, showed significant alignment to at least one of the four teleost genomes in the Ensembl database (Table 4.1). Analysis of the positions of significant alignments identified two putative QTL-orthologous regions located on two different chromosomes within each species. These were: zebrafish chromosomes 7 and 25; medaka chromosomes 3 and 6, stickleback linkage groups II and XIX; and puffer fish chromosomes 5 and 13 (Figure 4.1 and Table 4.2).

**Table 4.2: Regions of orthology to the IPN QTL in sequenced fish genomes**

| Fish | Chromosome / linkage group | Start (Mb) | End (Mb) | Length (Mb) | Number of genes (BioMart) |
|---|---|---|---|---|---|
| Zebrafish (Danio rerio) | Chr 7 | 14.5 | 69.1 | 54.6 | 825 |
| | Chr 25 † | 0.5 | 36.7 | 36.2 | 617 |
| Stickleback (Gasterosterus aculeatus) | LG II † | 3.1 | 19.4 | 16.3 | 507 |
| | LG XIX | 7.2 | 12.2 | 5.0 | 238 |
| Medaka (Oryzias latipes) | Chr 3 † | 9.6 | 33.0 | 23.4 | 483 |
| | Chr 6 † | 2.1 | 12.5 | 10.4 | 267 |
| Green spotted puffer fish (Takifugu rubripes) | Chr 5 | 2.6 | 13.3 | 10.7 | 491 |
| | Chr 13 | 7.3 | 12.3 | 5.0 | 263 |

†Chromosomes/linkage groups with large gaps in the positions of significant alignments

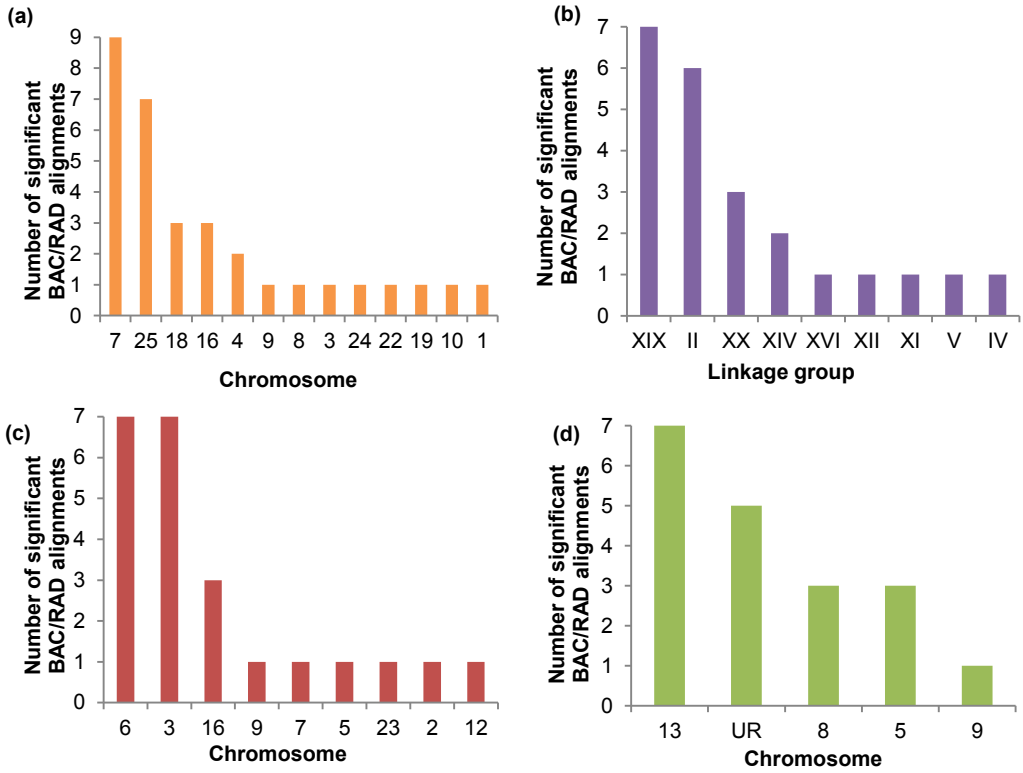



**Figure 4.1: Distribution of BAC/RAD contig alignments in the sequenced fish genomes**
Significant alignment of the IPN QTL-linked BAC/RAD contigs occurred most frequently on: (a) Zebrafish chromosomes 7 and 25; (b) Stickleback linkage groups II and XIX; (c) Medaka chromosomes 3 and 6; and (d) Green spotted puffer fish chromosomes 13 and Un_Random (UR). Since Un_Random is a group of sequences yet to be assigned to a chromosome, chromosome 5 was chosen (chromosome 8 corresponded to a single gene alignment of multiple Atlantic salmon sequences).

In general, the identified QTL-orthologous regions were large, and in some cases, almost represented the full length of the chromosome (Table 4.2). These large regions may partially be explained by the genomic rearrangements expected to have occurred in the ~275 MY since the last most recent common ancestor of these fish and Atlantic salmon (Near *et al*, 2012; Berthelot *et al*, 2014). The possibility of chromosomal gene rearrangements is also supported by the large gaps identified between significant alignments on some of the QTL-orthologous chromosomes (Table 4.2). For example, the region identified on LG II in stickleback may be split into two smaller regions of enrichment for significant alignments (3.08–3.12Mb and 15.10–19.40Mb). This was also the case for medaka chromosome 3, whereby significant alignment of a single RAD locus resulted in the extension of the QTL-orthologous region by ~17Mb (the narrower region would have been at 25.4–33.0Mb).

Genes within these QTL-orthologous regions were extracted (BioMart). Comparisons of the presence of the same combinations of genes across the QTL-orthologous regions in the four fish identified two groups of orthologous relationships: orthologous group (OG) A, which contained zebrafish chromosome 7, stickleback linkage group II, medaka chromosome 3 and green spotted puffer fish chromosome 5; and OG B, which contained zebrafish chromosome 25, stickleback linkage group XIX, medaka chromosome 6 and green spotted puffer fish chromosome 13 (Figure 4.2). In general, fewer genes were common across the four fish in OG B compared to OG A (Appendix D, Figures D3 and D4).

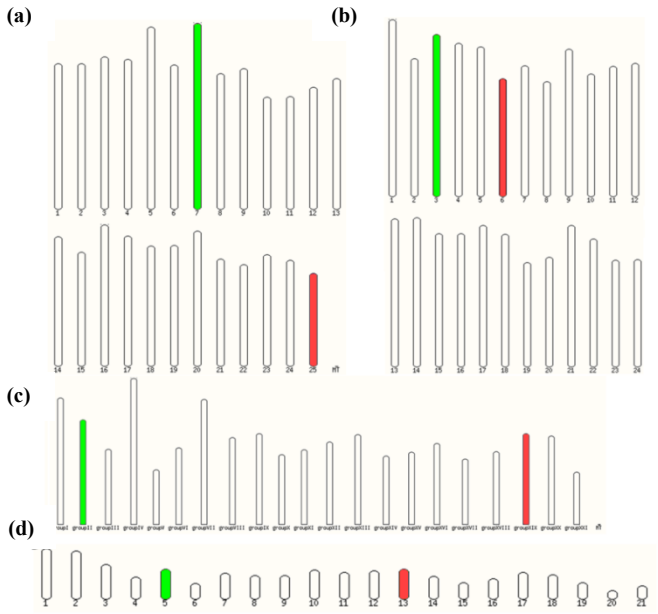



**Figure 4.2: IPN QTL-orthologous teleost chromosomes**
(a) Zebrafish (ZF); (b) Medaka (MD); (c) Stickleback (SB); (d) Green spotted puffer fish (GP). Comparisons of the number of common genes between IPN QTL-orthologous regions identified two groups of conserved orthology. Orthologous group A is labelled in green, and is comprised of ZF-7, MD-3, SB-II and GP-5. Orthologous group B is labelled in red, and is comprised of ZF-25, MD-6, SB-XIX, GP-13.

Current estimates of evolutionary relationships suggest that of the four teleost species, stickleback, medaka and puffer fish are most closely related (last most recent common ancestor ~110 MYA) (Near *et al*, 2012; Rosindell and Harmon, 2012; Berthelot *et al*, 2014), with evolutionary relationships amongst these three species differing depending on the type of data analysed (i.e. single locus vs multi-locus, nuclear or mitochondrial DNA) (Volff, 2005; Near *et al*, 2012; Whittington and Moerland, 2012; Broughton *et al*, 2013; Berthelot *et al*, 2014; Joerger *et al*, 2014). Zebrafish is the most distantly related to all species in this study (last most recent common ancestor ~275 MYA) (Near *et al*, 2012).

Overall, the number of common genes between pairs of species within OG A and OG B individually was concordant with evolutionary distance (Figures 4.3 and 4.4). For example, a higher proportion of genes in the QTL-orthologous regions were common between stickleback and medaka (>90% in OG A and ~50% in OG B), whereas 95% and 87% of the genes in zebrafish in OG A and OG B respectively were unique to zebrafish (Figure 4.3, Appendix D, Figures D3 and D4). Alternatively, the higher

proportion of unique zebrafish genes identified could be due to the identification of an overall larger QTL-orthologous region in zebrafish compared to the other fish. This would result in the zebrafish region potentially containing the genes common to stickleback and medaka, as well as many additional ones. For green spotted puffer fish, only 29% and 10% of genes in OG A and OG B respectively were unique when compared to stickleback and medaka, with some suggestion that medaka and green spotted puffer fish share a greater proportion of their genes in OG B relative to stickleback and medaka (Appendix D, Figure D4).

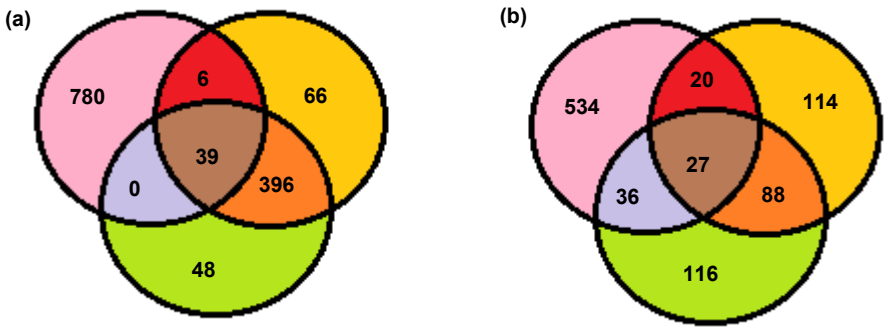



**Figure 4.3: Number of genes in common between the IPN QTL-orthologous regions of zebrafish (pink), stickleback (yellow) and medaka (green) in (a) orthologous group A and (b) orthologous group B**

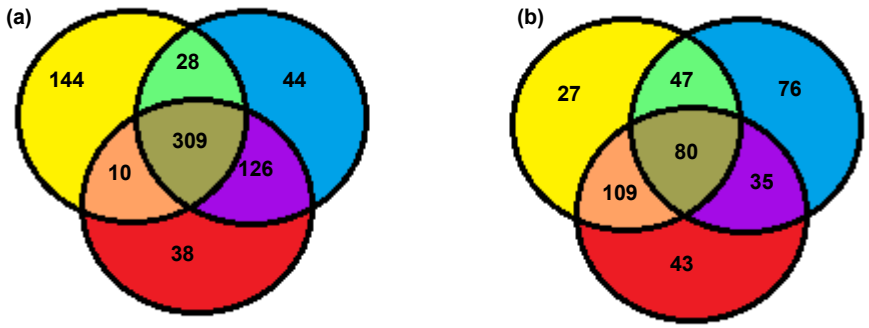



**Figure 4.4: Number of genes in common between the IPN QTL-orthologous regions of green spotted puffer fish (yellow), stickleback (blue), and medaka (red) in (a) orthologous group A and (b) orthologous group B**

For both OG A and OG B, comparisons of common gene locations by synteny mapping suggested that a greater number of gene rearrangements may have occurred within the zebrafish orthologous regions relative to the other three fish. As above, this may possibly be a reflection of the closer evolutionary relationship between

these three fish (Appendix D, Figures D3 and D4). Within OG A, mapping of common gene locations across species suggested a potential inversion or genome mis-assembly on stickleback LG II, where gene order appeared more highly conserved if part of the QTL-orthologous region on LG II (~6–18Mb) was inverted (Appendix D, Figures D3 and D5 show gene order conservation before and after inversion of this stickleback region, respectively). This putative inversion in stickleback did not present a problem for subsequent analysis of the 2Mb QTL-orthologous region of LG II, since this narrower region was mapped to within the apparent inversion event (7.3–9.3Mb).

### 4.4.2 Narrowing down the QTL-orthologous region in stickleback

The IPN QTL-linked Atlantic salmon sequences utilised in this study originated from within the 10cM confidence interval on LG 21, to which the major IPN QTL was initially mapped (Houston *et al*, 2008). Although these sequences were useful in identifying QTL-orthologous regions in published teleost genomes, these regions were large and contained a substantial number of genes (Table 4.2). Subsequently published studies using SNP panels have fine-mapped the QTL to within a 2cM confidence interval on LG 21 [Houston *et al* (2012) plus the integrated map described in the Materials and Methods]. To reduce the interval of the QTL-orthologous region in stickleback, flanking sequences of two SNPs within the 2cM confidence interval of the QTL (and, therefore, closely linked to the QTL) were aligned to the stickleback genome. This enabled the identification of a smaller QTL-orthologous region on stickleback LG II, between 7.3–9.3Mb. The apparent inversion seen in the larger stickleback region (described above) does not affect gene order within this 2Mb region, and gene order remains highly conserved in this region when compared to the other species (Appendix D, Figure D6).

### 4.4.3 Potential functional roles of positional candidate genes

Within the 2Mb QTL-orthologous region on stickleback LG II, 92 genes have previously been sequenced and annotated. To explore the potential biological involvement of these 92 positional candidates in resistance to IPN based on differential expression after challenge with IPNV, sequences of all known

stickleback genes within this 2Mb region were downloaded from BioMart (gene set 1; Table 4.3). To test whether the 2Mb region was enriched for differentially-expressed genes in comparison to the rest of the stickleback genome, sequences of all genes on stickleback LG II and all known and annotated stickleback genes in the genome were extracted from BioMart (gene sets 2 and 3 respectively; Table 4.3).

**Table 4.3: Number of stickleback genes (downloaded from BioMart) and the proportion which showed significant alignment to Atlantic salmon microarray probes (whole microarray vs differentially-expressed probes)**

| Region in Stickleback | Number of genes | Number (%) with sig. align. to probes | Number (%) with sig. align. to diff. expr. probes * |
|---|---|---|---|
| LG II, 7.3–9.3Mb (gene set 1) | 92 | 68 (73.9) | 18 (26.5) |
| LG II (gene set 2) | 861 | 685 (79.6) | 164 (23.9) |
| All known stickleback genes (gene set 3) | 22,456 | 17,461 (77.8) | 3,430 (19.6) |

* Percentage is relative to those that aligned to any probe on the microarray (regardless of probe expression level). E.g. gene set 1, (18/68)*100=26.5%

#### 4.4.3.1 Enrichment for differential gene expression within the QTL-orthologous region

Of the 43,661 probes on the Atlantic salmon microarray, 1,924 (5%) showed significantly different levels of hybridisation between resistant and susceptible individuals (nominal $P<0.05$). The measure of differential expression was a comparison of the average array probe signal between resistant and susceptible individuals just before and one day after challenge with IPNV (time points 0 and 1). While the genes differentially-expressed at individual time points are likely to differ, this combined measure across the two earliest time points was used as a pragmatic overall indication of the possible different patterns of gene expression between resistant and susceptible fish (as opposed to running separate analyses for each time point).

To assess whether the 2Mb QTL-orthologous region in stickleback was enriched for differentially-expressed genes, two nucleotide BLAST databases were created from

repeat-masked probe sequences: one containing sequences of all the probes on the microarray, and one containing the 1,924 differentially-expressed probe sequences only. To determine the number of genes within each of the three gene sets with representative probes on the salmon microarray, each set was individually aligned to the whole microarray nucleotide database. The majority of genes across all three gene sets (>70%) were represented by probes on the Atlantic salmon microarray (Table 4.3). However, this may be an overestimate, resulting from the alignment of a single probe to multiple genes. This could occur if, for example, the short length probe sequences (range: 122–7835bp) are derived from conserved gene family regions.

To determine the proportion of differentially-expressed genes in each set, the three gene sets were aligned to the 1,924 differentially-expressed probe sequences. Of the 17,461 genes in gene set 3 (all known stickleback gene sequences) that were represented on the microarray, 3,430 (19.6%) showed significant differential expression between resistant and susceptible individuals (Table 4.3; the distribution of alignments across the stickleback genome is given in Figure 4.5). As described above, the alignment of gene sequences to multiple microarray probes may be the reason why 3,430 genes were identified as significantly differentially-expressed, despite only 1,924 probes being differentially-expressed on the whole microarray. Overall, an enrichment of differentially-expressed genes on stickleback LG II (OG A) can be seen. No enrichment of differentially-expressed genes was seen on LG XIX (OG B) of stickleback, other than that seen as background differential expression across the whole genome (Figure 4.5).

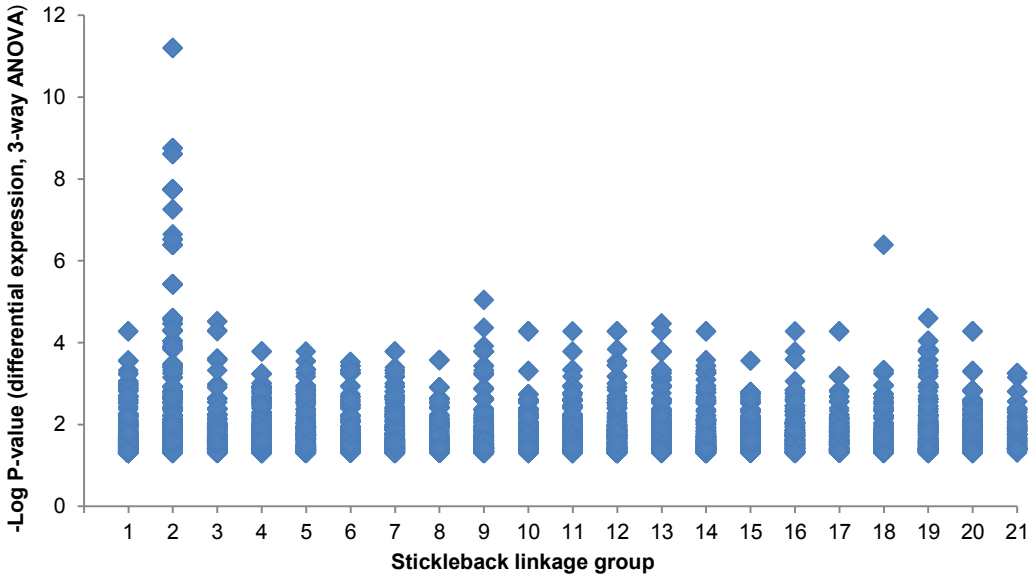



**Figure 4.5: Mapping of differentially-expressed Atlantic salmon microarray probes to stickleback linkage groups**

Each point represents a probe (partial gene sequence) which was differentially-expressed between IPN resistant and susceptible Atlantic salmon after IPNV challenge across two families (B and C) and time points (0 and 1). Probe sequences showing differential expression corresponded to genes located mainly on stickleback LG II.

Of the 861 genes in gene set 2 (whole of LG II), 685 aligned to a salmon microarray probe, 164 (23.9%) of which were significantly differentially-expressed (Table 4.3; Figure 4.6). To focus specifically on the most significantly differentially-expressed probes mapping to LG II, a threshold of –log P-value > 6 on probe differential expression was applied, and the genes to which these probes aligned were identified (Figure 4.6). Three probes (Ssa#S35541768, Ssa#STIR18537, Ssa#35575706) were eliminated due to poor alignment with the stickleback gene sequence, alignment to potential repetitive regions or conserved protein family domains, or because there was no stickleback gene name or gene description associated with the sequence aligning to the probe. The top five most differentially-expressed probes corresponded to: Sorbitol dehydrogenase, Cancer susceptibility candidate 4, Eukaryotic translation initiation factor 3 subunit M, NEDD8 activating enzyme E1 subunit 1 and DWT domain containing 1 (Figure 4.6).

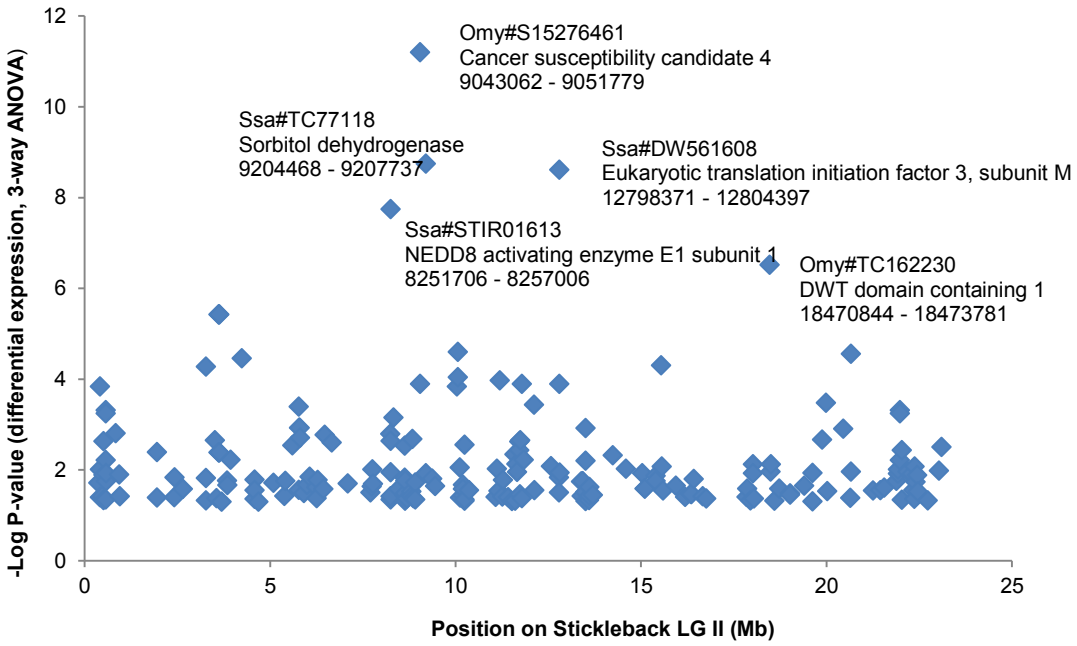



**Figure 4.6: The top five most differentially-expressed probes and the stickleback genes to which they align**
Probe IDs and the stickleback gene name and position (bp) to which they align are given.

Of the 68 genes in gene set 1 (2Mb region on LG II) with representative probes on the Atlantic salmon microarray, 18 (27%) showed significant alignment to differentially-expressed probes (Figure 4.7; Table 4.3). The slightly higher proportion of differentially-expressed genes within this QTL orthologous region compared to the whole of stickleback LG II (24%) and to the whole of the stickleback genome (20%) may suggest an enrichment of differentially-expressed genes in this 2Mb region. Interestingly, three of the top five most differentially-expressed probes aligned to three genes within this 2Mb region in stickleback: Cancer susceptibility candidate 4, Sorbitol dehydrogenase and NEDD8 activating enzyme E1 subunit 1 (Figure 4.7). These genes are both putative positional and differentially-expressed candidates for future investigations to identify variants or mutations influencing resistance to IPN in Atlantic salmon.

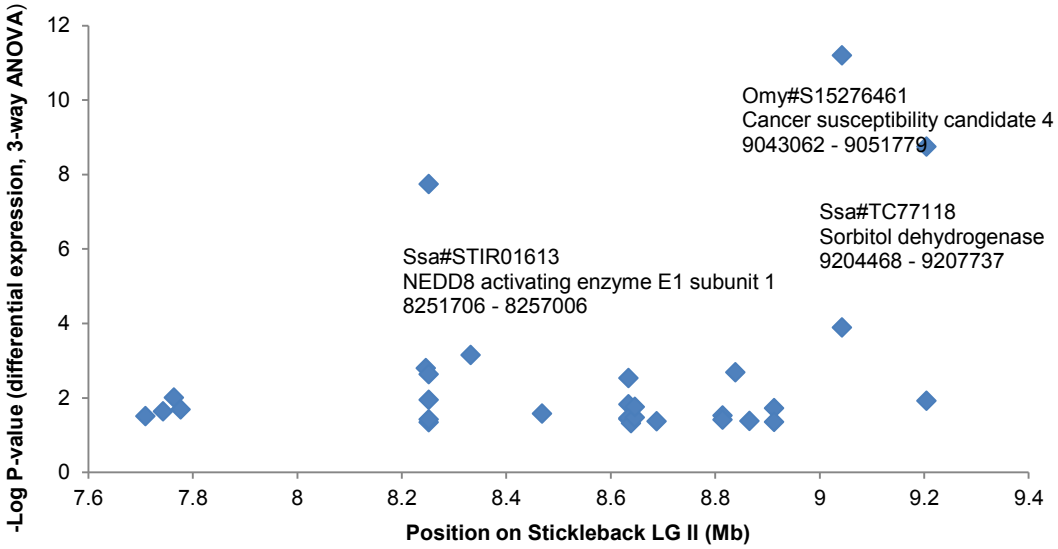



**Figure 4.7: Differential expression within the 2Mb IPN QTL-orthologous region on stickleback LG II**
Atlantic salmon differentially-expressed probes aligned to three genes within this region.

### 4.4.3.2 Pathway enrichment analysis

Analysis of gene differential expression patterns between resistant and susceptible animals can be applied to identify a list of candidate genes potentially involved in response to pathogen infection. However, gene differential expression may be a downstream effect of the causal factor(s) underlying disease resistance QTL. Identification of the biological pathways to which the differentially-expressed genes map could give greater insight into the host response to disease. Analysis of all genes within these pathways could highlight other potential resistance candidates which were not detected based on analysis of gene expression. To address this, pathway analysis of the 1,924 differentially-expressed probes was conducted, using the Ingenuity Pathway Analysis (IPA) software package. Overall, the 1,924 probes mapped to 255 biological pathways (the top 20 pathways with enrichment for differentially-expressed genes are given in Table 4.4). 12 of the 92 genes in the 2Mb QTL-orthologous region in stickleback have previously been mapped to 20 of the 255 pathways (Table 4.5). The four genes involved in the greatest number of pathways were brain-derived neurotrophic factor (5 pathways), eukaryotic translation initiation factor 3, subunit J (3 pathways), EPH receptor A6 (3 pathways), and cadherin 15, type 1, M-cadherin (myotubule) (3 pathways), and none of these genes had been determined as being significantly differentially-expressed.

**Table 4.4: The top 20 of the 255 pathways most enriched for differentially-expressed genes between IPN resistant and susceptible individuals after viral challenge**

| Differentially-expressed pathway | -Log P-value * | Number of genes in pathway ** | Number (%) of differentially-expressed probes *** |
|---|---|---|---|
| RhoA Signaling | 3.33 | 53 | 19 (35.8) |
| Rac Signaling | 2.86 | 58 | 17 (29.3) |
| Signaling by Rho Family GTPases | 2.48 | 74 | 29 (39.2) |
| Fatty Acid Metabolism | 2.48 | 70 | 18 (25.7) |
| Protein Ubiquitination Pathway | 2.37 | 68 | 31 (45.6) |
| Granzyme B Signaling | 2.32 | 19 | 5 (26.3) |
| Pyruvate Metabolism | 2.31 | 87 | 12 (13.8) |
| Alanine and Aspartate Metabolism | 2.17 | 65 | 8 (12.3) |
| Tumoricidal Function of Hepatic Natural Killer Cells | 2.17 | 28 | 6 (21.4) |
| Purine Metabolism | 2.15 | 261 | 32 (12.3) |
| Endoplasmic Reticulum Stress Pathway | 2.08 | 25 | 5 (20.0) |
| Regulation of Actin-based Motility by Rho | 2.08 | 37 | 13 (35.1) |
| Mitotic Roles of Polo-Like Kinase | 2.07 | 36 | 11 (30.6) |
| Aminoacyl-tRNA Biosynthesis | 2.04 | 64 | 7 (10.9) |
| Starch and Sucrose Metabolism | 2.02 | 99 | 11 (11.1) |
| Assembly of RNA Polymerase II Complex | 1.99 | 14 | 9 (64.3) |
| Acute Myeloid Leukemia Signaling | 1.98 | 39 | 12 (30.8) |
| N-Glycan Degradation | 1.91 | 43 | 6 (14.0) |
| Apoptosis Signaling | 1.87 | 62 | 13 (21.0) |
| PPARα/RXRα Activation | 1.86 | 108 | 21 (19.4) |

* -Log P-value as estimated by IPA. This is a measure of the likelihood that the genes within the input gene set are mapped to a specific biological pathway by chance. A pathway is enriched/over-represented if more of the input genes are associated with it than expected by chance, taking into account the number of known molecules within that pathway (for more details, see: http://www.ingenuity.com/wp-content/themes/ingenuity-qiagen/pdf/ipa/functions-pathways-pval-whitepaper.pdf)
** As given by https://targetexplorer.ingenuity.com where available, or http://www.genome.jp/kegg-bin/get_htext?br08901.keg
*** Number of differentially-expressed probes on whole microarray which map to this pathway

**Table 4.5: List of genes in the 2Mb region of stickleback LG II which map to at least one of the 255 pathways showing enrichment for gene differential expression between resistant and susceptible individuals after IPNV challenge**

| Gene | Pathway |
|---|---|
| Brain-derived neurotrophic factor (ENSGACG00000015502) | Axonal Guidance Signalling<br>CDK5 Signalling<br>Huntington's Disease Signalling<br>Neuropathic Pain Signalling In Dorsal Horn Neurons<br>Thyroid Cancer Signalling |
| Cadherin 15, type 1, M-cadherin (myotubule) (ENSGACG00000015445) | Gα12/13 Signalling<br>RhoGDI Signalling<br>Signalling by Rho Family GTPases |
| Dihydroorotate dehydrogenase (quinone) (ENSGACG00000015466) | Mitochondrial Dysfunction |
| Dipeptidase 1 (renal) (ENSGACG00000015454) | Eicosanoid Signalling |
| EPH receptor A6 (ENSGACG00000015384) | Axonal Guidance Signalling<br>Ephrin A Signalling<br>Ephrin Receptor Signalling |
| Ephrin-B2 (ENSGACG00000015366) | Axonal Guidance Signalling<br>Ephrin Receptor Signalling |
| Eukaryotic translation initiation factor 3, subunit J (ENSGACG00000015551) | eIF2 Signalling<br>mTOR Signalling<br>Regulation of eIF4 and p70S6K Signalling |
| G protein-coupled receptor kinase 1 (ENSGACG00000015322) | Phototransduction Pathway |
| Protein S (alpha) (ENSGACG00000015305) | Coagulation System |
| Ribosomal protein, large, P2 (ENSGACG00000015499) | eIF2 Signalling |
| Spleen focus forming virus (SFFV) proviral integration oncogene spi1 (ENSGACG00000015525) | Acute Myeloid Leukemia Signalling<br>Production of Nitric Oxide and Reactive Oxygen Species in Macrophages |
| Tight junction protein 1 (ENSGACG00000015573) | Tight Junction Signalling |

## 4.5 Discussion

Resistance to IPNV infection in Atlantic salmon has been shown to have a strong genetic component, and a major QTL explaining almost all of the genetic variance associated with response to infection has been mapped to LG 21 (Houston *et al*, 2008; Moen *et al*, 2009; Houston *et al*, 2010). As yet, no published study has been able to identify the causative variant(s) underlying this QTL, although reports of the identification of the (undisclosed) causative gene exist (Moen and Ødegård, 2014). In this study, a list of positional candidate genes potentially involved in resistance were generated, by alignment of QTL-linked Atlantic salmon sequences to published teleost genomes. To infer functional involvement of these positional candidates in resistance mechanisms, positional candidates were analysed for differential expression between resistant and susceptible individuals. To identify biological pathways which may be important during the initial stages of infection, pathway enrichment analysis for differentially-expressed genes was conducted.

Alignment of QTL-linked sequences to four published teleost fish genomes identified two IPN QTL-orthologous regions in each fish genome. The identification of two regions of conserved orthology in each sequenced fish genome was unexpected, and suggests that the origins of the IPN QTL region may lie prior to the divergence of salmonids from the other teleost fish species, and after the teleost specific genome duplication, approx. 300–400 MYA (Volff, 2005; Berthelot *et al*, 2014).

Comparisons of gene presence across QTL-orthologous regions suggested that orthologous regions share a significant number of genes in common, with high gene order conservation across species. This suggests that both regions have remained highly conserved across the teleost species after the teleost whole genome duplication, despite the ~300 MY since the divergence of teleost species. Indeed, reconstruction of the ancestral proto-Actinopterygian linkage groups prior to the teleost duplication (13 LGs labelled A–M) revealed that paralogous (duplicated) regions originating from the teleost duplication can still be identified, and that these remain conserved across many fish genomes (Danzmann *et al*, 2008). For example,

LG 21 and part of LG 9 in Atlantic salmon, chromosomes 7, 10, 18 and 25 in Zebrafish, and chromosomes 3, 6 and 20 in medaka were all found to derive from ancestral linkage group J. Atlantic salmon LGs 9 and 21 were also identified as sharing regions of homeology in chapter two of this thesis. Also in chapter two, these two Atlantic salmon linkage groups were suggested to contain regions of orthology to stickleback LG II. These results obtained across different analyses correspond well with the results obtained in the current study, and supports the possible ancestral origin of the QTL region.

Although the QTL-linked sequences utilised in this study originated from within a 10cM (~10Mb; cM estimate from female linkage map) confidence interval on LG 21, large QTL-orthologous regions of up to 55Mb (Zebrafish chromosome 7: 825 genes) were identified. This was mainly due to large gaps in these regions where salmon QTL-linked sequences did not align. This discontinuity in regions identified by comparative sequence mapping across species has previously been reported. For example, based on alignment of BAC sequences, the Infectious Salmon Anaemia (ISA) resistance QTL in Atlantic salmon was mapped to LG 24 of medaka to 2.5–3.5Mb and 15–25Mb (Li *et al*, 2011). One explanation for this discontinuity could be the potential chromosomal rearrangements or gene shuffling events which may have occurred since the last most recent common ancestor of Atlantic salmon and the other four fish. Alternatively, these large gaps may be caused by undetected repetitive elements within the QTL-linked sequences, which may result in the incorrect alignment of these sequences to other regions of the genome.

These large orthologous regions contain several hundreds of genes. Therefore, direct inference of positional candidate genes based on these broad genomic regions is limited. To improve upon this approach, two further analyses were conducted: (i) a smaller (2Mb) QTL-orthologous region in the stickleback genome was identified; and (ii) the potential involvement of genes within this smaller region was investigated, based on gene expression data. The five genes with the strongest statistical evidence (lowest P-value) for differential expression were located on LG II, and importantly, three of these were located within this 2Mb QTL-orthologous

region: NEDD8 activating enzyme E1 subunit 1, cancer susceptibility candidate 4 and sorbitol dehydrogenase.

The close proximity of these differentially-expressed genes suggests that there may be a functional reason for the clustering of these genes to the same genomic location. In eukaryotes, it has been hypothesised that gene order is non-random, due to the requirement for gene co-expression, sharing of a common cis-acting promoter, sharing of a common enhancer element, or due to sharing of a transcription factor or suppressor binding site (Hurst *et al*, 2004; Dewey, 2011). Many examples of clustering of genes which act in the same pathway and which show similar or coordinated expression patterns are known. For example, in yeast, genes which act in the same phase of the cell cycle map to the same location in the genome (Cho *et al*, 1998). One of the most studied gene clusters is the *Hox* gene cluster, which is highly conserved across vertebrates (Santini *et al*, 2003). Similarly, causative variant(s) within regulatory elements which result in differential expression of co-located genes could underlie response to IPNV infection.

Alternatively, the observed joint differential expression of these adjacent genes could be due to experimental procedures implemented in this study. Since resistant and susceptible individuals used in the IPNV challenge and microarray hybridisation experiment were from the same family, individuals are likely to be, on average, more similar to each other across the genome compared to unrelated individuals, but consistently different at the IPN QTL region. Therefore, differential expression of genes within the QTL region may be due to hitchhiking of genes linked to the causative mutation within the QTL region, caused by a high degree of relatedness amongst individuals in the study, and not due to an expression of resistance.

Nonetheless, this study has provided a list of putative candidate genes to direct future studies. The potential roles of these genes in IPN resistance was investigated using published literature, and this is discussed briefly below. In addition, results from the pathway analysis based on enrichment of differentially-expressed genes and their potential involvement in resistance to IPN based on existing literature is discussed

below. Finally, genes within the QTL-orthologous region in stickleback which were not differentially-expressed but which mapped to at least one differentially-expressed pathway are discussed. Clearly, further laboratory experiments and *in silico* analyses would be required to determine involvement of the identified genes and pathways in IPN resistance *in vivo*.

### 4.5.1 Differentially-expressed genes

*NEDD8 activating enzyme E1 subunit1*

The NEDD8 activating enzyme (NAE) has a role in the activation of the ubiquitin-like protein NEDD8, and is involved in the binding and activation of many proteins involved in regulation of protein activity and degradation, protein-protein interactions, stress response and subcellular localisation of proteins (Swords *et al*, 2010; Milhollen *et al*, 2011). Inhibition of NAE (and therefore no activation of NEDD8) has been reported to result in apoptosis in a wide range of cells, since downstream targets of NEDD8 involved in cell survival remain inactive. For example, two cullin proteins SCF and CRL4 which are known to be downstream targets of NEDD8 have been implicated in DNA damage and cell cycle check point activation, leading to cell cycle arrest in the S phase and apoptosis (Chen *et al*, 2003; Lin *et al*, 2010; Milhollen *et al*, 2011). Cullin inactivation has also been suggested to result in apoptosis of virus infected cells (Nascimento *et al*, 2012). This mechanism may play a role in the host response to IPNV, since apoptosis of IPNV-infected cells has been implicated as an important host-defence mechanism to infection (O'Brien, 1998; Imajoh *et al*, 2005). In addition, the VP5 gene protein product encoded by the IPNV genome is thought to have a role in preventing host cell apoptosis in the early stages of infection, which could be an important way of establishing infection before the host is able to remove infected cells (Hong *et al*, 2002; Imajoh *et al*, 2005). NAE was one of the two genes in close proximity to the QTL-linked SNPs, and is therefore an interesting candidate worth further investigation.

*Cancer susceptibility candidate 4*

Much of the published literature on cancer susceptibility candidate 4 (CASC4) is regarding its overexpression in association with HER-2/neu proto-oncogene

overexpression and human breast/ovarian cancers (Oh *et al*, 1999; NCBI, 2012). The closest homologue of CASC4, Golgi phosphoprotein 2 (GOLPH2), has been implicated in a variety of cellular activities, including protein trafficking and transport, mediating protein-protein interactions, cell survival, and in response to viral infection (Oh *et al*, 1999; Riener *et al*, 2009; Zhou *et al*, 2011). As well as hijacking of host cell survival mechanisms [e.g. the VP5 protein of IPNV in apoptosis (Hong *et al*, 2002; Imajoh *et al*, 2005)], efficient trafficking of assembled viruses and protein-protein interactions for mediating viral entry and exit into host cells are important parts of the viral lifecycle. Given their roles in normal cell survival and protein trafficking, controlling the expression of CASC4 and/or GOLPH2 may influence viral lifecycle progression, and play an important role in host defence to viral infections (Pous *et al*, 2005). In addition, pathway enrichment results presented in this thesis suggest that cell survival pathways (such as those involved in cancer progression) may play an important role in response to IPNV, providing further support for the importance of host cell survival in viral lifecycle completion.

*Sorbitol dehydrogenase*

The main role of sorbitol dehydrogenase is in the polyol (or sorbitol) pathway, which is activated during hyperglycaemia. It is mainly responsible for the conversion of sorbitol to fructose, but has also been shown to bind to other sugar alcohols and could have a wide range of cellular activities (Gabbay, 1975; Carr and Markham, 1995). Inactivation of this enzyme has been shown to result in eventual cell death through the accumulation of sorbitol and osmotic stress (Carr and Markham, 1995). Elevated serum levels of sorbitol dehydrogenase in woodchucks chronically infected with the hepatitis B virus have been reported to result in host cell necrosis (Zhou *et al*, 2000). Necrosis of pancreatic cells is the main clinical sign of IPNV infection, and it could be that an elevation in sorbitol dehydrogenase is one of the triggers for the activation of cell necrosis pathways.

### 4.5.2 Differentially-expressed pathways

Differential expression between resistant and susceptible fish is a clear way of identifying genes involved in viral replication, and has been used in many analyses (Sadasiv, 1996; Marjara *et al*, 2011). However, differential expression of a gene may not be indicative of it harbouring the causative variant(s) underlying the QTL. Rather, the causative variant(s) may be within upstream regulatory elements of differentially-expressed genes within the same pathway. For example, if differentially-expressed genes on LG II are controlled by the same transcription factor, individual analysis of differentially-expressed genes may not lead to the causative mutation. The identification of pathways involved in conferring resistance will improve the understanding of the underlying biology resulting in the expression of resistance, and may highlight novel pathways to target for vaccine development. Further, this may highlight putative candidate genes which would not be considered within gene expression studies.

Given the nature of the disease and clinical signs observed, innate immunological and apoptosis-controlling mechanisms are likely to be involved in disease progression. Genes involved in preventing virus entry or replication may also be involved, although IPNV can be isolated from asymptomatic fish and survivors of challenge experiments (Ruane *et al*, 2007), suggesting that viral entry may not be entirely blocked in resistant individuals. In addition, published studies have shown that uptake of other aquabirnaviruses by resistant cell lines does occur, albeit at a slower rate than in susceptible cell lines (Imajoh *et al*, 2003). This suggests that mechanisms involved in limiting (but not fully preventing) viral entry may be involved in conferring resistance.

To identify pathways which may be involved in resistance, mapping of all 1,924 differentially-expressed probes to their respective pathways was conducted. Pathway enrichment analysis identified 255 pathways which were potentially differentially regulated between IPN resistant and susceptible individuals. The top four pathways, and their possible involvement in IPN resistance, are discussed below.

*RhoA Signalling*

The RhoA signalling pathway is known to be involved in many cellular processes, including in the regulation of gene transcription, wound repair, and in the organisation of the actin cytoskeleton, for cell cytokinesis and endocytosis of external material in to the cell (Aspenström *et al*, 2004; Jaffe and Hall, 2005; Zhou and Zheng, 2013). Importantly, this signalling pathway has been shown to be involved in virus internalisation. For example, the internalisation and virulence of influenza A virus has been shown to be dependent on the calcium-dependent endocytosis pathway, which in turn, activates the RhoA signalling pathway. Inactivation of this pathway (e.g. as in RhoA double mutant cells) has been shown to result in the inhibition of influenza A virus internalisation and infection relative to the wild-type (Fujioka *et al*, 2013). For other viruses, such as the Vaccinia virus, the expression of the RhoA protein during the early stages of infection (2 hours post-infection) has been suggested to mediate virus cell entry. During the latter stages of infection, RhoA is inhibited by the F11 protein encoded by the Vaccinia virus genome, and the inability of the F11 protein to bind to and inactivate RhoA has been shown to result in a reduction of the spread of the virus across cells (Arakawa *et al*, 2007; Cordeiro *et al*, 2009; Handa *et al*, 2013). This pathway was the top pathway identified, and is also known to regulate, and is regulated by, the second most important pathway identified in this study, i.e. Rac signalling.

*Rac Signalling*

In this study, the Rac signalling pathway was identified as the second most important pathway in regulating the initial stages of IPNV infection. Activation of this pathway is thought to enhance cell-cell adhesions and to promote cell migration. Importantly, this pathway is known to be differentially regulated with the RhoA signalling pathway, i.e. activation of Rac signalling results in the inactivation of RhoA signalling, and vice versa (Caron, 2003; Nimnual *et al*, 2003). In Marek's disease virus infections, inhibition of the Rac signalling pathway resulted in an increase in viral plaque sizes, whereas inhibition of the Rho signalling pathway had the opposite effect. In addition, the inactivation of the Rac signalling pathway resulted in a reduction in the number of cell-cell contact regions, suggesting that this pathway is

important in viral spread (Richerioux *et al*, 2012). The human HIV virus has been shown to activate the Rac signalling pathway in order to enhance cell-cell adhesion and enable virus entry into neighbouring cells (Harmon and Ratner, 2008). The identification of both the Rac and RhoA signalling pathways in the current study, as well as the involvement of both pathways in a number of other viral infections, strongly suggests that these pathways may play a major role in the initial stages of IPNV infection and in contributing to genetic resistance/susceptibility. Further studies aimed at understanding this involvement and the interactions between these two pathways in IPNV infection are required.

*Signalling by Rho family GTPases*

GTPases are small effector molecules known to play a role in many cellular functions, including regulation of transcription, controlling the actin cytoskeleton, vesicle trafficking, apoptosis and activation of immune cells. The three main GTPases are Rac, Cdc42 and Rho. This pathway results in the activation of these GTPases, and, therefore, activates the Rac and RhoA signalling pathways (Karnoub and Der, 2000; Liang *et al*, 2004; Schwartz, 2004; Spiering and Hodgson, 2011). As described above, these pathways are clear candidates for further research.

*Fatty Acid Metabolism*

Fatty acid metabolism is the main energy production pathway through the degradation of fatty acids in the Krebs cycle. Importantly, this pathway is thought to play a major role in enhancing viral replication. Inhibition of upstream regulators of enzymes within this pathway resulted in a reduction in the levels of Rift Valley fever virus and West Nile Virus genomic RNA and mRNA relative to control. This is indicative of a restriction of viral replication (Yamaguchi *et al*, 2005; Martín-Acebes *et al*, 2011; Moser *et al*, 2012; Greseth and Traktman, 2014). RNAi and inhibitor targeted knockdown of components of the fatty acid synthesis pathway such as fatty acid synthase resulted in a decrease in Dengue virus replication in a variety of cell lines, using a Dengue virus-luciferase replicon to measure viral replication. In addition, this virus was shown to cause targeted relocation of the fatty acid synthase enzyme to the cell nucleus, where viral replication takes place (Heaton *et al*, 2010).

Interestingly, studies characterising the ability of Atlantic salmon to retain high levels of total lipid and high n-3 long-chain polyunsaturated fatty acid flesh contents when fed a vegetable oil diet suggest that families with high lipid phenotypes also showed improved survival in IPN challenge experiments (Morais *et al*, 2012). These observations, together with the abundant literature on the requirement of fatty acid biosynthesis/metabolism for replication of many other viruses, strongly suggests that this pathway is likely to be involved in resistance to IPNV.

### 4.5.3 Genes within the QTL-orthologous region which map to differentially-expressed pathways

Differential expression of a gene in response to viral infections is a useful way of detecting genes which may be involved in enhancing or limiting the viral lifecycle. However, gene differential expression does not directly imply a causative role for a gene in the differential regulation of virus response, i.e. whether an individual is resistant or susceptible to a viral infection. To investigate the involvement of genes which were not differentially-expressed between IPN-resistant and susceptible individuals upon viral challenge, the genes within the QTL-orthologous region of stickleback which mapped to pathways differentially regulated between resistant and susceptible individuals were identified. The four genes involved in the greatest number of pathways were brain-derived neurotrophic factor (5 pathways), eukaryotic translation initiation factor 3, subunit J (3 pathways), EPH receptor A6 (3 pathways), and cadherin 15, type 1, M-cadherin (myotubule) (3 pathways), and their potential involvement in resistance to IPNV is discussed briefly below. The mapping of these genes to multiple pathways potentially involved in resistance suggests that these genes could be influencing resistance at multiple levels, by causing differential expression of genes within the same pathways.

*Brain-derived neutrotrophic factor*

Of all the genes within the 2Mb QTL-orthologous region in stickleback, brain-derived neutrotrophic factor mapped to the highest number of pathways differentially-expressed between IPNV resistant and susceptible individuals. The

main role of this gene is in promoting and enhancing the development, maturation and survival of neurons within the central nervous system. As such, this gene has been suggested to be a good candidate in gene therapy treatments of neurological diseases, and much of the literature is focussed on the best vectors for efficient and effective delivery of this gene (Di Polo *et al*, 1998; Benraiss *et al*, 2001; Jia *et al*, 2002). However, recent reports have suggested a role for this protein in adult T-cell leukemia, which is caused by infection with the retrovirus human T-cell leukemia virus type 1 (HTLV-1). Specifically, this virus has been shown to result in a significant up-regulation of brain-derived neutrotrophic factor gene transcription, thus enhancing the survival of infected cells compared to uninfected control cells (Polakowski *et al*, 2014).

*Cadherin 15, type 1, M-cadherin (myotubule) CDH15*

The main role of CDH15 is in promoting the fusion of muscle cells to form multinucleate myotubule structures. Upon fusion, these muscle cells lose the ability to differentiate and DNA replication is inactivated. In addition, it is thought that immobilisation of membranous material occurs, limiting endocytosis and exocytosis within the cell (Holtzer *et al*, 1975; Salvatori *et al*, 1995; Bergstrom *et al*, 2002). As such, these cells present a challenge for viruses, since the genomes of DNA viruses cannot be replicated and exit of assembled viruses may be limited. However, many viruses have evolved mechanisms which can overcome this effect through reprogramming of cell fate and degeneration of multinucleate mytotubules into mononucleated cells (Holtzer *et al*, 1975). Amongst many other genes and pathways, the degeneration of myotubules requires the inactivation of muscle cell fusion enhancing proteins, including CDH15 (Sunadome *et al*, 2011). This phenomenon has been observed in cells infected with Moloney murine sarcoma virus (Birnbaum *et al*, 1993) and Rous sarcoma virus (Holtzer *et al*, 1975).

*EPH receptor A6*

Eph receptors are the largest known group of receptor tyrosine kinases (RTKs). Eph receptors are located on the cell surface membrane and are fundamental in the cell response to environmental cues, in cell-cell interactions, cell migrations and in cell

growth and survival. These receptors are known to be involved in immune response to pathogens, and are up-regulated in response to inflammatory cytokines released from neighbouring cells. Ephrin signalling has also been implicated in the activation of apoptotic pathways, particularly in response to viral infections, through interactions of Eph receptors with the p53 family of proteins (Kullander et al, 2001; Surawska et al, 2004; Egea and Klein, 2007; Coulthard et al, 2012).

*Eukaryotic translation initiation factor 3, subunit J*

Eukaryotic translation initiation factor (eIF) 3 is a large protein complex comprised of 13 subunits (labelled A-M). The main function of this complex is in the initiation of cap-dependent protein mRNA translation through interactions with the ribosome and the G subunit of eIF4, and inhibition of the interaction between eIF3 and eIF4 results in a decline in cap-dependent cell protein synthesis (Morris-Desbois *et al*, 2001). As such, this mechanism a prime target for viruses such as poliovirus, which are able to utilise the cap-independent protein synthesis pathway. Cleavage of the G subunit of eIF4 by the poliovirus protease 2A protein inactivates cap-dependent translation of host cell mRNAs, thus effectively high-jacking the host protein synthesis machinery without affecting translation of its own RNA (Wyckoff *et al*, 1990; Wyckoff *et al*, 1992; Gradi *et al*, 1998). Influenza A virus has also been reported to cause lysosomal degradation of the B subunit of eIF4. This targeted degradation of eIF4 results in the inhibition of interferon synthesis and dampens the immune response to the influenza A virus, thus enabling its replication within host cells (Wang *et al*, 2014). In yellow fever virus infections, the L subunit of eIF3 has been shown to interact with the NS5 protein of the virus, and overexpression of eIF3L resulted in a decrease in yellow fever virus replication (Morais *et al*, 2013). Interestingly, the chicken infectious bursal disease virus (IBDV), another birnavirus known to be closely related to IPNV, has been demonstrated to high-jack host cell protein synthesis pathways much like the poliovirus, through targeted cleavage of eIF4 (Tacken *et al*, 2004; Busnadiego *et al*, 2012). In this study, although the J subunit of eIF3 was not differentially-expressed, another subunit of this gene (subunit M) also found on LG II of stickleback but just outside of the 2Mb QTL-orthologous region was highly differentially-expressed between resistant and

susceptible individuals. This provides strong support for the involvement of the eIF3 protein complex in IPNV infection.

## 4.6 Conclusion

Using IPN QTL-linked sequences in a series of comparative, gene expression and pathway analyses, this study aimed to identify potential candidate genes involved in resistance to IPN in Atlantic salmon. Alignment of QTL-linked sequences to four reference teleost fish genomes identified two chromosomes in each fish species potentially harbouring QTL-orthologous regions. Gene order within QTL-orthologous regions appears to be conserved across species which suggests that identification of comparative positional candidate genes is possible. Using flanking sequences of SNPs closely linked to the QTL, a 2Mb QTL-orthologous region on stickleback LG II was identified. Analysis of gene expression patterns after an IPNV challenge between resistant and susceptible individuals revealed an enrichment of differentially-expressed genes within this region relative to the whole genome, and highlighted putative candidate genes requiring further investigation. Pathway enrichment analyses of differentially-expressed genes suggested that pathways involved in preventing viral entry/replication, apoptosis, and cell energy production, may be involved in response to infection with IPNV. Overall, this study presents results towards improving the understanding of the biology for IPN resistance in Atlantic salmon, and highlights putative candidate genes and pathways requiring further investigation.

# Chapter 5

# Exploring the utility of cross-laboratory RAD-Sequencing datasets for phylogenetic analysis

## 5.1 Abstract

Restriction site-Associated DNA sequencing (RAD-Seq) is a next-generation sequencing technique which can be used to generate genome-wide sequence and genetic marker datasets. RAD-Seq has been applied in many population genetic studies, including for the construction of phylogenetic trees. However, the consistency of RAD-Seq data generated in different laboratories, the use of these data in inferring cross-species orthologous loci for relationship estimation, and the filtering parameters to apply to remove false orthologies whilst maintaining sufficient sequence data with adequate phylogenetic signal, have not been widely investigated. This study is an assessment of the use of RAD locus consensus sequences derived from different populations, laboratories, and bioinformatic pipelines for the estimation of evolutionary relationships amongst ten finfish species with previously established phylogeny. As expected, the number of cross-species orthologous RAD loci identified decreased with increasing evolutionary distance, ranging from ~3,000 salmonid-species specific loci to ~450 loci between more distantly related species. Interspecific single nucleotide variants at each orthologous RAD locus were identified, and estimated relationships using concatenated sequences of variants were congruent with previously published phylogenies. The inclusion of RAD loci which were absent in varying proportions of the analysed species did not affect estimated relationships, and improvements in node support over complete datasets were observed. Overall, this study has demonstrated the reproducibility and utility of cross-laboratory RAD-Seq data, both across populations and across species, for the inference of orthologous RAD loci for use in the estimation of evolutionary relationships.

## 5.2 Introduction

The recent advancements in next-generation sequencing (NGS) technologies have meant that genotyping-by-sequencing technologies (such as RAD-Seq) are being widely applied across both model and non-model organisms in many different investigations across the biological sciences. This includes in the estimation of evolutionary relationships.

The estimation of evolutionary relationships is traditionally conducted based on morphological comparisons (i.e. cladistics approach). With improvements in sequencing technologies, single locus sequences have been utilised in the estimation of relationships [e.g. Claiborne Stephens and Nei (1985), Olsen *et al* (1985), Lebarbenchon *et al* (2010), Joerger *et al* (2014)]. Although useful, gene trees obtained in single locus analyses do not always agree with those obtained from morphological studies, due to incomplete lineage sorting at the chosen locus (Lynch, 1999; Gontcharov *et al*, 2004; Maddison and Knowles, 2006; Castresana, 2007). One way of minimising this is by constructing species trees based on a combined analysis of multiple loci (Maddison and Knowles, 2006). However, obtaining sequence information across multiple loci and species is difficult, and until recently, studies were typically restricted to 10–30 loci [but see Zeng *et al* (2014)].

One of the main advantages of RAD-Seq for the estimation of relationships is the ability to sample loci from multiple regions distributed throughout the genome. In addition, the use of restriction enzymes for the digestion of genomic DNA means that theoretically, assuming no polymorphisms in the restriction enzyme cleavage site, the same genomic regions (i.e. homologous RAD loci) are sampled and sequenced across all individuals, making the inference of cross-species orthologous loci more feasible. As such, data generated from RAD-Seq experiments have become popular for use in inferring evolutionary relationships across species [e.g. *Drosophila* (Rubin *et al*, 2012; Arnold *et al*, 2013; Cariou *et al*, 2013), bamboo (Wang *et al*, 2013), broomrape family of flowering plants (Eaton and Ree, 2013), American oak (Hipp *et al*, 2014) and ground-beetles (Cruaud *et al*, 2014)]. Overall, relationships

estimated using RAD data have been congruent with those seen in previously published literature, suggesting that RAD data would prove useful in non-model taxa for which the evolutionary relationships are unknown.

Although RAD-Seq has been useful across many phylogenetic studies, there remain areas of the data filtering and processing steps which are yet to be standardised. First, *in silico* studies suggest that phylogenetic inference using RAD data may be restricted to species with less than 100 million years (MY) since the last most recent common ancestor (Rubin *et al*, 2012). This is due to the expected reduction in the number of cross-species orthologous RAD loci identified with increasing evolutionary distance. To my knowledge, the performance of experimentally-derived RAD loci in relationship estimation for more distantly related species has not been investigated.

Second, the inference of cross-species orthologous RAD loci in published phylogenetic studies involves the collection of samples and locus identification uniquely within each study. Currently, RAD-Seq data is already available for a variety of species generated across different laboratories. Therefore, a large number of RAD-Seq datasets already exist, which could potentially be combined into large datasets for phylogenetic or other population genetic studies. Furthermore, although *in silico* studies have been conducted to investigate the effect of incorporating RAD loci which are not observed in all included species on the estimation of evolutionary relationships [e.g. Huang and Knowles (2014)], the reduction in the number of inferred orthologous RAD loci due to the combined effects of increased evolutionary distances and differences in the identified loci across differing RAD-Seq protocols has not been quantified.

Currently, phylogenetics studies utilise concatenated sequences across all RAD loci for each included individual for the estimation of evolutionary relationships. However, analysis of the potentially many thousands of putative orthologous RAD loci identified using the available phylogenetic software packages may require an extensive amount of time and computing power, which may not be available for

small scale projects. Analysis time will increase further with the number of species and number of individuals included per species. Instead, within-species consensus RAD loci or interspecific single nucleotide variants could be used to infer evolutionary relationships, but this is yet to be tested. Therefore, the overall aim of this study was to investigate the reproducibility and utility of published cross-laboratory RAD-Seq data in closely and distantly related finfish species with previously published phylogenies (Kitano *et al*, 1997; Broughton, 2010; Crête-Lafrenière *et al*, 2012; Near *et al*, 2012; Shedko *et al*, 2013). The specific aims of the study were to: (i) investigate the reproducibility of RAD data by aligning RAD sequences within species and across laboratories; (ii) investigate the performance of RAD data in the inference of cross-species orthologous loci and evolutionary relationships; and (iii) investigate the effect on relationship inference of inclusion or omission of RAD loci which appear absent in some species (i.e. minimum taxon coverage per locus).

## 5.3 Materials and Methods

### 5.3.1 Sequence data

In a typical RAD-Seq bioinformatic pipeline, sequence reads derived from the flanking regions of the restriction enzyme are collapsed into a 'RAD locus' (Baird *et al*, 2008). For each locus, sequence reads are aligned within and across individuals, and a single population level 'consensus sequence' is generated (see Figure 1.3 and chapter two for details on the RAD-Seq protocol). Single- or paired-end RAD consensus sequences were obtained for Atlantic salmon (*Salmo salar*), rainbow trout (*Onchorhynchus mykiss*), three-spined stickleback (*Gasterosterus aculeatus*), gudgeon (*Gnathopogon sp.*), Chinook salmon (*Onchorhynchus tshawytscha*), sockeye salmon (*Onchorhynchus nerka*), spotted gar (*Lepisosteus oculatus*), lake whitefish (*Coregonus clupeaformis*), Baltic sea herring (*Clupea harengus*), and Atlantic halibut (*Hippoglossus hippoglossus*) (Table 5.1). To facilitate the identification of orthologous RAD loci across species, only single-end sequences from studies utilising the *Sbf*I restriction enzyme were considered in the analysis. *Sbf*I RAD-Seq studies were chosen since this is the most commonly used protocol within aquatic species, therefore, had the most publically available data.

For rainbow trout and Atlantic salmon, data from four and two different studies respectively were obtained. For stickleback, consensus RAD sequences were generated within individuals (n=46) and aligned to the reference genome, and population-level consensus sequences were unavailable (Table 5.1). For each of these three fish species, a single file of common RAD loci was produced using BLASTN alignments of all sequences, where common RAD loci were defined if sequence for that locus was observed in more than a certain threshold number of populations/individuals (see Appendix F for full details and scripts).

### 5.3.2 Data filtering, processing and characterisation

All ten consensus sequence files were processed as follows. To avoid bias in alignment parameters due to differences in sequence lengths (Rognes, 2001; Agostino, 2012), all sequences were trimmed to 60 base pairs (bp) (the shortest read length amongst the studies). To limit the misleading alignment of sequences to multiple regions due to genomic repeat elements, known repeats were masked using RepeatMasker (Smit *et al*, 1996-2010) and the Atlantic salmon repetitive element database (download: http://web.uvic.ca/grasp/salmon_v1.6).

Previous studies suggest that, given the GC-rich recognition site of the *Sbf*I restriction enzyme, RAD loci obtained from *Sbf*I RAD-Seq analyses may be biased towards gene-rich regions of the genome [e.g. Amores *et al* (2011); Everett *et al* (2012); Bruneaux *et al* (2013)]. To investigate the proportion of RAD loci found in coding regions for each of the species, trimmed and repeat-masked sequences were individually aligned [TBLASTX; BLAST+ package version 2.2.25+; Zhang *et al* (2000)] to a custom-made database of nucleotide gene sequences. This database comprised gene sequences originating from Atlantic cod (*Gadus morhua*), pufferfish (*Takifugu rubripes*), medaka (*Oryzias latipes*), platyfish (*Xiphophorus maculatus*), spotted gar (*Lepisosteus oculatus*), three-spined stickleback (*Gasterosterus aculeatus*), Tetraodon (*Tetraodon nigroviridis*), tilapia (*Oreochromis niloticus*) and zebrafish (*Danio rerio*) [Ensembl 78 http://www.ensembl.org/index.html; (Flicek *et al*, 2014)].

**Table 5.1: Descriptions of the RAD sequences obtained from the different studies**

| Species | Reference | Consensus seq availability § | Initial number of seqs | Seq length (bp) | Post-processed number of seqs | Number of mapped RAD loci | Protocol and pipeline details | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | RAD-Seq library preparation protocol | Fragment size selection window (bp) | Sequencing platform | Sequence analysis pipeline | Minimum depth coverage per locus |
| Chinook salmon (Oncorhynchus tshawytscha) | Brieuc et al., 2014. G3, 4(3) | Online (SE) | 62,249 | 75 | 62,249 | 7,304 | Baird et al,. 2008 | 200-500 | Illumina GAII/HiSeq | STACKS | Locus sequenced in 135 (85 %) individuals |
| Sockeye salmon (Oncorhynchus nerka) | Everett et al., 2012. BMC Genomics, 13(521) | Provided by authors (SE) | 64,613 | 60 | 64,613 | 2,175 | Baird et al,. 2008 Etter et al., 2011 | 400-800 | Illumina GAII/HiSeq | Custom-written Perl scripts, Bowtie, Novoalign | 10 reads per allele per locus per individual |
| Rainbow trout (Oncorhynchus mykiss) | Hecht et al., 2012. G3, 2(9) | Provided by authors (SE) | 12,073 | 67 | 32,027 | 587 | Miller et al., 2007 Baird et al., 2008 | 200-500 | Illumina GAII/HiSeq 2000 | Perl scripts from Miller et al., 2012, Novoalign | 5 reads per locus per individual |
| | Hale et al., 2013. G3, 3(8) | Provided by authors (SE) | 277,469 | 89 | | - | Miller et al., 2012 | 300-600 | Illumina HiSeq | Perl scripts from Miller et al., 2012, Novocraft | 5 reads per locus per individual |
| | Hohenlohe et al., 2013. Molecular Ecology, 22(11) | Online (PE) | 77,141 | 147-552* | | - | Etter et al., 2011 | 330-400 | Illumina HiSeq | STACKS | Locus sequenced in 1/60 (2 %) individuals after pooling across individuals |
| | Miller et al., 2012. Molecular Ecology, 21(2) | Online (SE) | 40,649 | 68 | | - | Baird et al., 2008 Hohenlohe et al., 2010 | 200-500 | Illumina HiSeq | Custom-written Perl scripts, Novoalign | Locus sequenced in 3 individuals |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Atlantic salmon (Salmo salar) | Gonen et al., 2014. BMC Genomics, 15(166) | Provided by authors (PE) | 366,219 | 95 | 65,758 | 5,004 | Etter et al., 2011 with modifications from Houston et al., 2012 | 250-500 | Illumina HiSeq 2000 | RADtools, STACKS | 500 reads per locus across 96 individuals |
| | Houston et al., 2012. BMC Genomics, 13(244) | Provided by authors (PE) | 66,073** | 95 | | - | Baird et al,. 2008 Etter et al., 2011 | 250-500 | Illumina GAIIx/HiSeq 2000 | RADtools | 5 reads per allele per locus per individual |
| Lake whitefish (Coregonus clupeaformis) | Gagnaire et al., 2013. Evolution, 67(9) | Provided by authors (SE) | 193,258 | 69 | 193,258 | 3,438 | Baird et al,. 2008 | 200-500 | Illumina HiSeq 2000 | STACKS | Locus is present in at least one mapping parent |
| Three-spined stickleback (Gasterosterus aculeatus) | Roesti et al., 2013. Molecular Ecology, 21(12) | Provided by authors (SE) | 31,118≠ | 64 or 138‡ | 31,118 | 29,041# | Baird et al,. 2008 | 200-500 | Illumina HiSeq 2000 | Novoalign, SAMtools | 12 reads per locus across 284 individuals |
| Atlantic halibut (Hippoglossus hippoglossus) | Palaiokostas et al., 2013. BMC Genomics, 14(566) | Provided by authors (SE) | 83,678 | 96 | 83,678 | 5,711 | Baird et al,. 2008 Etter et al., 2011 with modifications from Houston et al., 2012 | 300-550 | Illumina HiSeq 2000 | STACKS | 30 reads per locus per individual |
| Baltic sea herring (Clupea harengus) | Corander et al., 2013. Molecular Ecology, 22(11) | Online (SE) | 63,742 | 95 | 63,742 | NA | Baird et al., 2008 Hohenlohe et al., 2010 Emerson et al., 2010 | 200-500 | Illumina HiSeq 2000 | FLORAGENEX unitag assembler v2.0, FLORAGENEX pipeline | 5 reads per locus per individual |
| Spotted gar (Lepisosteus oculatus) | Amores et al., 2011. Genetics, 188(4) | Provided by authors (SE) | 64,483 | 75 | 64,483 | 4,396 | Miller et al., 2007 Baird et al., 2008 Hohenlohe et al., 2010 | 200-500 | Illumina GAIIx | STACKS | Locus sequenced in 85 (90 %) individuals |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gudgeon (Gnathopogon sp.) | Kakioka et al., 2013. BMC Genomics, 14(32) | Online (SE) | 44,109 | 70 | 44,109 | 1,622 | Etter et al., 2011 | 300-500 | Illumina GAIIx/HiSeq 2000 | STACKS | 3 reads per locus per individual |

§ SE: single-end RAD-Seq; PE: paired-end RAD-Seq

* Paired-end RAD sequencing generated contigs of variable length

** 2 files from two families, sequence counts: 70,207 and 70,739. Subsequently combined into one file with 66,073 common sequences

≠ 46 files (one per individual). Sequence count range: 25,840–42,618. Subsequently combined into one file with 31,118 common sequences

‡ Two separate sequencing studies were implemented, resulting in two different read lengths

# All sequences were aligned to the stickleback genome, hence, their genomic locations were known, regardless of whether they contained a mapped SNP marker

### 5.3.3 Identification of cross-species orthologous RAD loci

The ability to correctly infer orthologous genomic regions across species is of major importance when estimating phylogenetic relationships. As such, there is an abundance of literature on best practices for the inference of sequence orthology, based on the availability of published reference genome sequences [e.g. see Dewey (2011), Kristensen *et al* (2011), Schmitt *et al* (2011)]. In the absence of well-assembled and annotated reference genomes for all included species, the inference of the type of sequence orthology observed (e.g. one-to-one/one-to-many orthology, functional orthology, positional orthology, paralogy, etc.) is difficult, since additional information to decipher the type of orthology observed (such as chromosomal location of orthologs i.e. synteny) is not available in most cases. In such cases, sequence similarity is thought to be a reliable way of inferring sequence orthology across species (Rubin *et al*, 2012), with expected improvements in the ability to infer orthologous relationships with increasing sequence length. While RAD-Seq typically generates large numbers of loci dispersed throughout the genome, the individual sequences are short (typically ~100bp) and, therefore, its potential utility for cross-species comparisons in the absence of reference genomes is unclear.

#### 5.3.3.1 Identification of homologous RAD loci between pairs of species

To identify RAD loci conserved across species, pairwise BLASTN analyses of the trimmed and repeat-masked consensus RAD sequences were conducted ['blastn' alignment algorithm; BLAST+ package version 2.2.25+; Zhang *et al* (2000)]. The most significant alignment for each sequence (i.e. 'best hit') was extracted. Two files of best hits were created: (i) within salmonid species only; and (ii) across all ten species (including the salmonid species).

Best hit alignment files were quality-checked based on the following thresholds: (i) within salmonid species only, using 'strict' alignment parameters of ≥ 95% percentage identity, ≥50bp alignment length, and ≤2 base mismatches; and (ii) between all ten species, using more 'relaxed' alignment parameters of ≥ 85% percentage identity, ≥45bp alignment length, and ≤10 base mismatches. Alignment parameters were arbitrarily chosen in an attempt to eliminate the collapsing of

paralogous sequences into a single RAD locus, and to minimise the grouping of conserved regions of different genes into single RAD loci (for example, zinc finger domains, which can be found in a number of genes). Alignment parameters remained constant within each analysis (rather than varying parameters according to the evolutionary distance between species) such that: (i) consistency in parameters across all pairwise alignments was maintained, in order to aid comparisons of the number of loci identified between species of differing relatedness; and (ii) the identification of misleading alignments (for example between sequences corresponding to conserved regions of the same gene family rather than the same RAD locus) is minimised.

The inference of sequence orthology in salmonid species is complicated by the recent (~80–100 MYA) salmonid-specific genome duplication and retention of high sequence similarity across paralogous regions of the genome (>85% in rainbow trout) (Volff, 2005; Berthelot *et al*, 2014). To minimise multiple alignments of paralogous sequences within salmonid species, alignments were further filtered to retain only unique one-to-one alignments (i.e. where the subject sequence was the best hit to a single query sequence).

#### 5.3.3.2 Identification of cross-species orthologous RAD loci

To identify orthologous RAD loci across groups of species of differing levels of evolutionary relatedness, pairwise alignments were clustered, first within the salmonid species only based on the strict pairwise alignments, and second, across all ten species, based on the relaxed alignment parameters. The clustering pipeline was implemented as follows (details and scripts given in Appendix G). Using the two files of filtered pairwise best hits, sequence clusters were inferred if individual RAD locus sequences across all included species all aligned to each other respectively as the most significant and unique match. For example, if SpeciesA_RAD1 significantly aligned to SpeciesB_RAD1, SpeciesC_RAD1, SpeciesD_RAD1 and SpeciesE_RAD1, and these all aligned to each other as the most significant and unique match, then these were inferred as a single candidate cluster. To limit the effect of paralogous sequences on inferring clusters across the salmonid species,

candidate clusters containing sequences which were assigned to multiple clusters were removed. Finally, candidate clusters were filtered to remove those containing more than one RAD locus sequence from a single species. These were taken as true RAD clusters for further analysis.

#### 5.3.3.3 Absence of cross-species orthologous RAD loci in some species

Data obtained from next-generation sequencing platforms is generally heavily filtered to remove low quality sequences and those with a large number of errors, resulting in datasets which may contain missing data in particular samples or genomic regions. Therefore, further filtering or imputation methods to minimise the amount of missing data in the overall dataset is generally conducted. However, due to polymorphic variation in the restriction enzyme cut site, variation in methylation status of the locus, or genomic rearrangements though evolution, not all RAD loci are expected to be present in all species (Poland and Rife, 2012; Arnold *et al*, 2013; Eaton and Ree, 2013; Cruaud *et al*, 2014; Huang and Knowles, 2014). As such, the absence of sequence for a given RAD locus may be informative for evolutionary analyses.

To analyse the effect of incorporating RAD loci which were 'absent' for a given species (i.e. no ortholog identified in the available dataset for that species), clusters were filtered using varying thresholds for minimum taxon coverage per locus. Within the salmonid species strict analysis, clusters containing sequences from all five salmonid species (salmonid dataset 1) and clusters containing sequences from at least three of the five salmonid species (salmonid dataset 2) were retained. Across all ten species in the relaxed analysis, only a single RAD locus cluster was identified. Therefore, downstream analyses were conducted using clusters with a minimum of seven sequences from at least seven different species (all fish dataset 1) or a minimum of five sequences from at least five different species (all fish dataset 2). The proportion of clusters within genic regions of the genome was quantified, based on alignment to the custom-made fish nucleotide gene database (see section 5.3.2). To prevent bias in relationship estimation due to the expected higher ability to detect

sequence orthology between salmonid species, clusters which were specific to salmonid species only were removed.

To test the cluster inferences made against published software packages, sequences were re-clustered using the UCLUST algorithm [USEARCH package, version v7.0.1090; Edgar (2010)]. This was done for salmonid dataset 1 and all fish dataset 2. UCLUST was run using default parameters, changing only the -id option to 0.95 for salmonid dataset 1 or to 0.85 for all fish dataset 2 to reflect the thresholds for percentage identity used in the BLASTN alignments.

### 5.3.4 Reconstructing teleost fish phylogeny using RAD data

To my knowledge, the study with the greatest number of loci sampled across fish species is that by Broughton *et al* (2013), where 21 loci (one mitochondrial, 20 nuclear) were used for relationship estimation. To date, the most comprehensive study of teleost phylogeny is that described in Near *et al* (2012), where relationships amongst 232 fish species were estimated based on nine coding sequences and fossil calibration times. Based on this phylogeny and the salmonid species relationships described in Shedko *et al* (2013), the expected relationships between the ten teleost species in the current study are given in Figure 5.1.

To test the utility of cross-laboratory RAD-Seq data to infer teleost species relationships, cross-species orthologous RAD locus clusters described above were used to construct phylogenetic trees. The salmonid datasets 1 and 2, and all fish datasets 1 and 2 described above were used in the analysis. For each identified RAD locus cluster, sequences for each species within the cluster were extracted. If absence of a RAD locus for a given species was permitted (as in salmonid dataset 2 and all fish datasets 1 and 2), species with no sequence for that locus were assigned a string of 60 * ‘N’. Sequences within a cluster were aligned using the MUSCLE software package [Multiple Sequence Alignment; version 3.8.31, (Edgar, 2004)], and the resulting alignments were investigated for the presence of between-species single nucleotide variants. Alleles for each variant for each species across all RAD loci were concatenated into a single sequence (see Appendix H for shell and Python

scripts written to run the MUSCLE alignment, variant identification and concatenation). Concatenated variant sequence files were converted into the PHYLIP format (Felsenstein, 2005) for input into the RAxML software package [version 8; (Stamatakis, 2014)] (see Appendix I for details on parameters used for relationship estimation in RAxML). RAxML employs a maximum likelihood based algorithm for phylogeny inference, and was chosen since it allows for correction of ascertainment bias which may arise when using variants for relationship estimation. The 'best tree' was obtained using 10,000 bootstrap replicates. By default, RAxML only reports support values for nodes appearing in at least 3% of the bootstrap replicates (i.e. at least 300 trees in this study). Resulting Newick trees were visualised using Phylodendron (http://iubio.bio.indiana.edu/treeapp/treeprint-form.html), T-REX (Boc *et al*, 2012), or Archaeopteryx (Han and Zmasek, 2009).

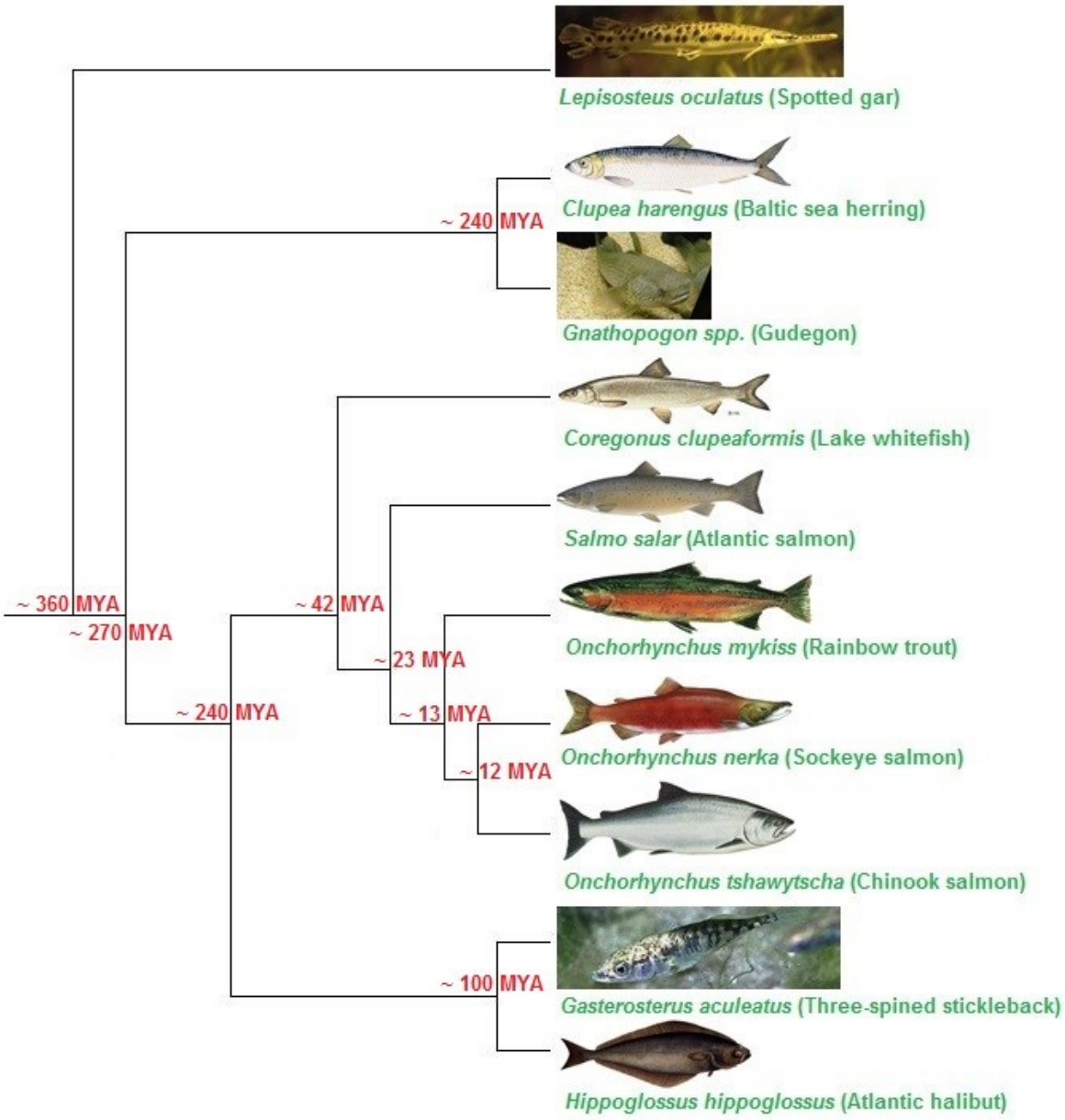



**Figure 5.1: Expected evolutionary relationships and approximate divergence times, as defined by Near et al (2012) and Shedko et al (2013)**

Species images were taken from http://en.wikipedia.org/ or are published for open access use. Divergence times and branch lengths not drawn to scale. Divergence estimates for the non-salmonid teleost fish species were obtained from Near *et al* (2012), and divergence estimated for the salmonid species were obtained from Shedko *et al* (2013).

### 5.3.5 Establishment of large-scale chromosomal orthologous relationships between species pairs

To assess the potential of RAD data for the inference of chromosomal orthologous relationships between species pairs, pairwise orthologous RAD loci (described

above) where both species contained a published mapped SNP were identified (hereafter referred to as 'mapped RAD loci'). This was done using the linkage maps associated with the original study from which the RAD sequences were obtained (Table 5.1). No linkage map was available for Baltic Sea herring (Corander *et al*, 2013), therefore, this species was omitted from this analysis. The chromosomal location of mapped RAD loci in the two species was taken as a potential chromosomal orthologous relationship. The number of occurrences of these same potential orthologous relationships was counted. This analysis was performed using results from both the strict and relaxed alignment parameters.

## 5.4 Results

Although RAD-Seq datasets exist for a diverse number of species, best practices for combining datasets generated across studies have not been fully investigated. In particular, the reproducibility of the same RAD locus across laboratories for the same species, the use of cross-laboratory sequences for the identification of cross-species orthologous RAD loci, and the utility of these loci in the estimation of evolutionary relationships, has not been widely studied.

### 5.4.1 Sharing of RAD loci across populations

To investigate RAD data reproducibility across populations (and studies) within species, RAD loci from two different populations of Atlantic salmon and four different populations of rainbow trout were compared (Table 5.1; see Appendix F for details). A high concordance between RAD loci identified across studies was seen, with 99% of Atlantic salmon and 79% of rainbow trout sequences being shared across the different studies (percentages are given relative to the study with the fewest number of RAD loci). The higher percentage obtained across the two distinct Atlantic salmon populations may be partly due to the data originating from the same laboratory, and, therefore, a higher similarity in library preparation protocols and downstream bioinformatic analyses for data filtering. Overall, the results highlight the ability of RAD-Seq to consistently identify the same RAD loci across studies, despite inevitable variation in sample library preparation, sequencing platforms and downstream filtering pipelines.

### 5.4.2 Sharing of RAD loci across species

The correct inference of sequence orthology across species is of prime importance when estimating evolutionary relationships, and relies heavily on the conservation of sequence similarity across evolution. To test the ability to identify orthologous RAD loci using cross-laboratory datasets, pairwise alignments of consensus RAD sequences across the ten teleost species of varying evolutionary relatedness were conducted. To infer cross-species orthologous RAD loci, pairwise alignments were clustered across salmonid species and then across all ten teleost species.

Despite setting strict alignment parameters for inferring pairwise orthologous RAD loci between salmonid species, a high proportion were identified, ranging from ca. 6,500 (8%) loci between Chinook salmon and lake whitefish to ca. 16,000 (50%) loci between rainbow trout and Chinook salmon (Table 5.2). Under the relaxed alignment parameters, the number of putative orthologous RAD loci identified increased, ranging from ca. 11,000 (34%) between rainbow trout and lake whitefish to ca. 19,500 loci (61%) between rainbow trout and sockeye salmon (Table 5.2). This may be due to the increased ability to infer orthology between RAD loci which lie within less conserved regions of the salmonid species genomes (such as gene introns), although a relaxation of alignment parameters is also likely to increase the chance of false positive orthologies.

Across the salmonid species, a total of 3,050 loci with sequence present for all five species (i.e. cross-species orthologous RAD loci) were identified (Table 5.3). Of these, 2,176 (71%) clusters were independently confirmed using the UCLUST software package, and this concordance validates the approach taken using clusters derived from pairwise BLASTN alignments. To investigate the effect of including RAD loci which appear absent in some species (i.e. no orthologous sequence identified in the dataset), clusters with at least three sequences from three different salmonid species were identified. A total of 22,710 loci were identified, of which 78 were removed due to containing sequences which were assigned to multiple clusters

(potential paralogous regions), leaving 22,632 clusters for further analysis (Table 5.3).

In contrast, the number and proportion of shared RAD loci between the five distantly related non-salmonid species based only on relaxed alignment parameters was much lower, with fewer than 500 (<2%) identified in most of the pairwise comparisons. Of these species pairs, stickleback and Atlantic halibut contained the most orthologous RAD loci (~2,700, 9%), as expected due to their closer evolutionary relationship compared to any other pair of non-salmonids in the study (Near *et al*, 2012; Broughton *et al*, 2013; Berthelot *et al*, 2014).

Across all ten species, only a single orthologous RAD locus was identified, which contained orthologous sequences predicted to originate from Transcription factor 7 (T cell specific, HMG box). Therefore, a less stringent parameter of RAD loci common to at least seven different species was applied, resulting in a total of 137 clusters. The majority of these clusters contained sequences from exactly seven different species (94 clusters, 69%; 43 clusters had sequence for 8, 9 or 10 species).

To investigate the effect of including RAD loci which appear absent in a larger number of species, clusters with at least five sequences from five different species were identified. A total of 4,945 clusters were identified, of which 4,493 were salmonid species specific and were removed from further analysis. Of the remaining 452 clusters, 303 (67%) were also identified by the UCLUST software package. 370 (80%) clusters contained sequences from six or more species. 217 of these clusters contained sequences from all five salmonid species and one additional species (i.e. exactly six sequences within the cluster). In most cases (145 clusters, 67%), the additional species was stickleback. This is likely due to the higher quality of the sequences obtained for stickleback compared to the other nine species, since RAD loci were identified by alignment of RAD-Seq reads to the stickleback reference genome (Roesti *et al*, 2013).

**Table 5.2: Number (percentage) of shared RAD loci identified by pairwise BLASTN alignments**

| | Chinook salmon | Sockeye salmon | Rainbow trout | Atlantic salmon | Lake whitefish | Three-spined stickleback | Atlantic halibut | Baltic sea herring | Spotted gar | Gudgeon |
|---|---|---|---|---|---|---|---|---|---|---|
| **Chinook salmon** | NA | 32,648 (52.4) | 18,606 (58.1) | 24,108 (38.7) | 16,615 (26.7) | 595 (1.9) | 792 (1.3) | 520 (0.8) | 271 (0.4) | 439 (1.0) |
| **Sockeye salmon** | 32,911 (52.9) | NA | 19,625 (61.3) | 25,028 (38.7) | 17,739 (27.5) | 567 (1.8) | 625 (1.0) | 349 (0.6) | 283 (0.4) | 313 (0.7) |
| **Rainbow trout** | 19,338 (60.4) | 16,970 (52.0) | NA | 15,393 (48.1) | 11,026 (34.4) | 317 (1.0) | 493 (1.5) | 331 (1.0) | 187 (0.6) | 226 (0.7) |
| **Atlantic salmon** | 25,935 (41.7) | 27,118 (42.0) | 16,670 (52.0) | NA | 19,457 (29.6) | 609 (2.0) | 892 (1.4) | 459 (0.7) | 413 (0.6) | 472 (1.1) |
| **Lake whitefish** | 15,579 (25.0) | 21,402 (33.1) | 13,433 (41.9) | 18,035 (27.4) | NA | 711 (2.3) | 894 (1.1) | 421 (0.7) | 361 (0.6) | 427 (1.0) |
| **Three-spined stickleback** | NA | NA | NA | NA | NA | NA | 2,704 (8.7) | 368 (1.2) | 197 (0.6) | 228 (0.7) |
| **Atlantic halibut** | NA | NA | NA | NA | NA | NA | NA | 310 (0.5) | 330 (0.5) | 362 (0.8) |
| **Baltic sea herring** | NA | NA | NA | NA | NA | NA | NA | NA | 213 (0.3) | 434 (1.0) |
| **Spotted gar** | NA | NA | NA | NA | NA | NA | NA | NA | NA | 199 (0.5) |
| **Gudgeon** | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |

Percentages (in parentheses) are given relative to the fish with the fewest number of sequences in processed FASTA files. Upper quadrant of matrix shows values obtained from the 'relaxed' analyses. The number and percentage of alignments obtained from the salmonid 'strict' analysis is given in the lower quadrant of the matrix.

**Table 5.3: Number of RAD locus clusters and interspecific variants identified for each analysis**

| Species | Parameters | Analysis pipeline | Minimum taxon coverage | Number of orthologous RAD loci | Number (%) of orthologous RAD loci in genes | Number of variants for relationship estimation | Range of missing information for each included RAD locus |
|---|---|---|---|---|---|---|---|
| **Salmonids (salmonid dataset 1)** | Strict | BLASTN | 5 | 3,050 | 375 (12.3) | 6,959 | NA |
| **Salmonids (salmonid dataset 2)** | Strict | BLASTN | ≥3 | 22,632 | 1407 (6.2) | 39,890 | 3,135–21,480 |
| **Salmonids** | Strict | BLASTN-UCLUST common | 5 | 2,176 | 307 (14.1) | 4,088 | NA |
| **All ten species** | Relaxed | BLASTN | 10 | 1 | 1 (100) | NA | NA |
| **All ten species (all fish dataset 1)** | Relaxed | BLASTN | ≥7 | 137 | 106 (77.4) | 1,440 | 37–745 |
| **All ten species (all fish dataset 2)** | Relaxed | BLASTN | ≥5 | 452 | 321 (71.0) | 4,094 | 371–2,881 |
| **All ten species** | Relaxed | BLASTN-UCLUST common | ≥5 | 303 | 213 (70.3) | 2,476 | 196–1,747 |

### 5.4.3 Relationship estimation

Whilst the application of strict thresholds for filtering of raw RAD-Seq reads often results in the removal of loci or individuals with excess missing data, recent studies suggest that more relaxed thresholds could be favourable in resolving relationships (Eaton and Ree, 2013; Huang and Knowles, 2014). To test this in this study, phylogenetic relationships were estimated using RAD loci with varying proportions of absent sequences, firstly amongst the closely related salmonid species, and secondly amongst all ten teleost species in the analysis. For each level of species relatedness and level of absent RAD sequence data, cross-species orthologous RAD loci were inferred. Multiple alignments of sequences within orthologous RAD loci allowed the identification of interspecific single nucleotide variants, which were concatenated into a single sequence for each species and used to estimate evolutionary relationships (RAxML software package). An example tree obtained in the current analysis is given in Figure 5.2 (all trees are given in Appendix J).

Two sets of orthologous RAD clusters were used to re-construct phylogenetic relationships between the five salmonid species: salmonid dataset 1, where all included RAD loci had complete sequence information across all five salmonid species (3,050 loci, 6,959 variants; Table 5.3); and salmonid dataset 2, where all included RAD loci had complete sequence information for at least three of the five salmonid species (22,632 loci, 39,890 variants; Table 5.3). Both datasets were able to recover expected relationships within the five salmonid species, with the three *Oncorhynchus* species forming a monophyletic group relative to Atlantic salmon and lake whitefish (Appendix J, trees 1 and 2). All nodes were estimated with >96% support. The only difference between trees obtained using salmonid datasets 1 and 2 was an increase in estimated branch lengths when the minimum taxon coverage per locus was reduced.

Across the ten teleost species, evolutionary relationships were estimated using variants derived from RAD loci common to at least seven of the ten species (all fish dataset 1 = 137 loci, 1,440 variants; Table 5.3; Appendix J, tree 3) or to five of the

ten species (all fish dataset 2 = 452 loci, 4,094 variants; Table 5.3; Appendix J, tree 4). In all analyses, trees were consistent with previously published literature (Figure 5.1). Monophyly of the Salmonidae and monophyly of the three *Onchorhynchus* species was predicted with 100% bootstrap support. As above, the use of RAD loci with sequence for seven species compared to five species did not change estimated relationships and tree topology, with improvements in node support and only a slight elongation of branch lengths observed in most cases (Appendix J, trees 3 and 4). However, in some cases (for example in the branch separating the salmonid species from the other five teleost species), branch lengths estimated using all fish dataset 2 (Appendix J, tree 4) were double that estimated using all fish dataset 1 (Appendix J, tree 3). This suggests that a slight change in the proportion of missing sequence per RAD locus is unlikely to affect estimation of evolutionary relationships, but could bias the estimated divergence times between more distantly related species (not estimated in this study).

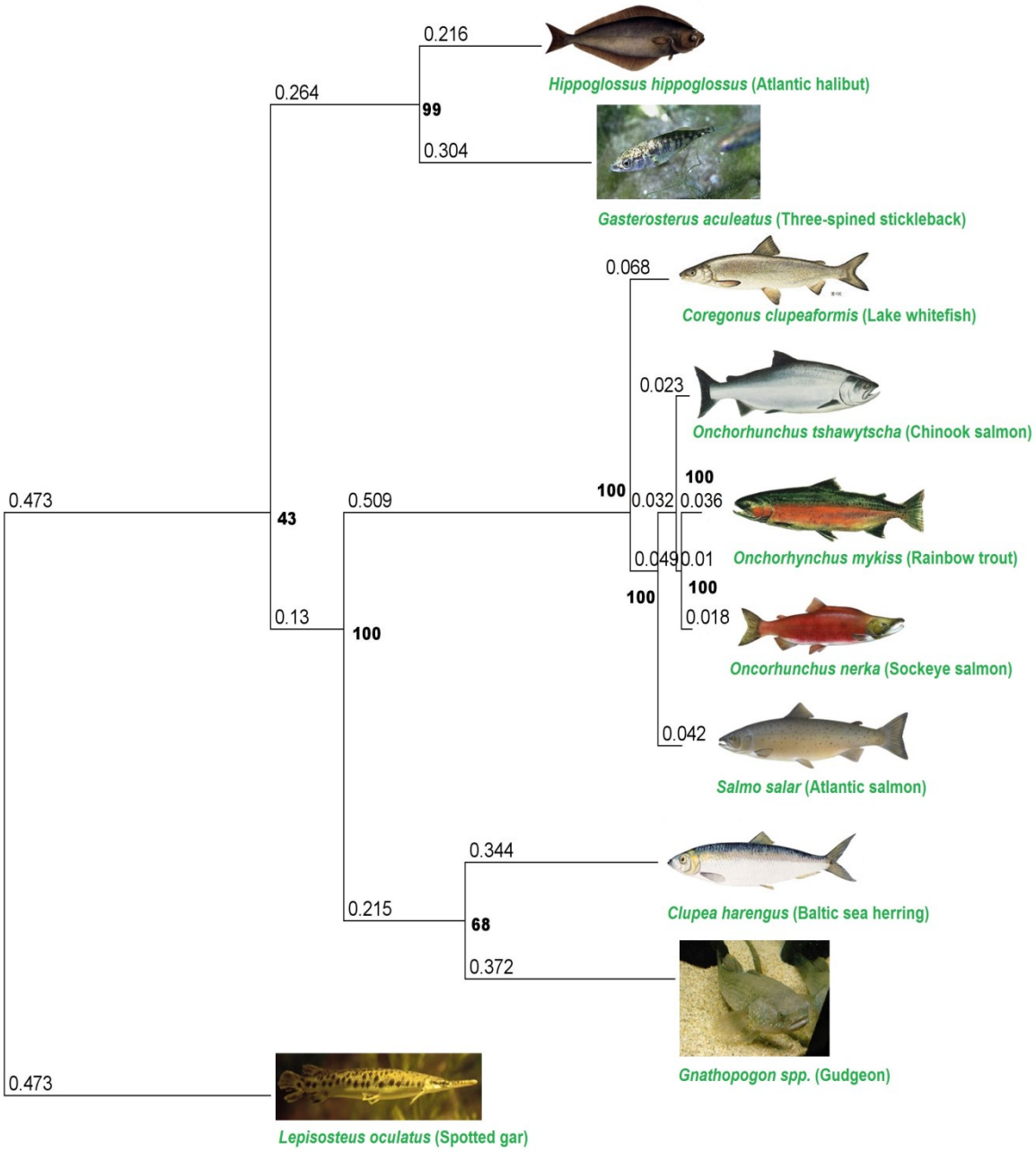



**Figure 5.2: Example tree of all ten fish species obtained in this study using RAxML**
Evolutionary relationships obtained using RAD data in this study were congruent with those of Near *et al* (2012) (teleost species) and Shedko *et al* (2013) (salmonid species) (Figure 1). Parameters – RAD loci present in at least five of ten species; 452 loci, 4,094 between-species variants. Branch lengths estimated in RAxML are given along individual branches, and node bootstrap support values (10,000 bootstrap replicates) are given at individual nodes. Branch lengths are not drawn to scale. This tree is the same as Appendix J, tree 4.

### 5.4.4 Number of genic RAD loci

Given the expected higher evolutionary conservation of coding (i.e. gene) regions (Cooper and Brown, 2008; Bergmiller *et al*, 2012), a valid assumption would be that the majority of cross-species orthologous RAD loci originate from coding regions. In addition, previous studies have suggested that given the GC-rich recognition site of the *Sbf*I restriction enzyme, RAD loci obtained from *Sbf*I RAD-Seq analyses may be biased towards gene-rich regions of the genome (Amores *et al*, 2011; Everett *et al*, 2012; Arnold *et al*, 2013; Bruneaux *et al*, 2013). To test these hypotheses, consensus RAD sequences identified within each species individually, and all inferred cross-species orthologous RAD clusters, were aligned (TBLASTX) to a custom-made database of known fish gene nucleotide sequences (see section 5.3.2). RAD loci originating from putative genic regions of the genome were identified.

Within individual salmonid species, approximately 2% of the RAD loci were identified as originating from genic regions. Across the salmonid species, <15% of the orthologous RAD clusters were estimated to be genic in origin (Table 5.3). For the other, non-salmonid species, a higher percentage of putative genic RAD loci were identified for a given species (ranging from 4–50%), and >70% of cross-species orthologous RAD loci estimated to be genic in origin (Table 5.3).

The large discrepancy in the number of putative genic orthologous RAD locus clusters identified within salmonid species and across all ten fish species may be due to the expected higher genome conservation (both coding and non-coding regions) across the salmonid species due to their closer evolutionary relatedness. As such, both genic and non-genic RAD loci may be represented within the salmonid species RAD locus clusters, and a higher proportion of the RAD loci identified across all ten species would be from coding regions of the genome. Alternatively, this may be due to the absence of salmonid gene sequences in the nucleotide database used for alignment, and the closer evolutionary relationship of the other teleost species with those in the database. In the case of stickleback, which has a high-quality, annotated reference genome and was included in the nucleotide database, ~50% of the RAD sequences were identified as genic. Based on the size of the stickleback genome

[~530Mb; (Jones *et al*, 2012)] and the total known stickleback gene sequences [~192Mb; Ensembl 78, (Flicek *et al*, 2014)], ~36% of the stickleback genome is estimated to be genic. This supports the hypothesis that *Sbf*I RAD-Seq may be biased towards genic regions of the genome (Amores *et al*, 2011; Everett *et al*, 2012; Arnold *et al*, 2013; Bruneaux *et al*, 2013), and suggests that *Sbf*I RAD data is potentially useful for evolutionary and comparative genomics studies.

### 5.4.5 Establishment of large-scale chromosomal orthologous relationships

Using published linkage maps (Table 5.1), mapped RAD loci were identified and used to infer orthologous relationships between fish species chromosomes (see section 5.3.5 and Appendix K). This resulted in the successful identification of previously identified and published orthologies, and many additional ones. For example, eight of the nine previously identified orthologous relationships between the Chinook salmon and Atlantic salmon genomes were supported in the current study (Brieuc *et al*, 2014) (see Appendix K). However, the overall support for inferred orthologies was typically low (range: 1-17 RAD loci). This could be due to the relatively low number of markers on available linkage maps (typically less than 5,000 markers; Table 5.1), due to the lack of a reference genome, and/or difficulty in obtaining a reliable marker order and position for large numbers of SNPs using existing software packages (Recknagel *et al*, 2013b). All of the above reasons could result in fewer RAD loci being assigned to a genomic position. Further, there was a requirement for both RAD loci in the pairwise matches to contain a mapped SNP. Despite this, the recovery of expected orthologous relationships highlights the possible utility of cross-laboratory RAD-Seq data for the inference of large scale chromosomal orthologies across species, particularly in future studies with denser genetic maps.

## 5.5 Discussion

The ability to rapidly obtain reliable sequence information at genome-wide loci has meant that RAD-Seq has become increasingly popular in phylogenetic studies [e.g. Poland and Rife (2012); Eaton and Ree (2013); Cruaud *et al* (2014)]. To my

knowledge, all published RAD-Seq phylogenetic studies first involve the collection and processing of genomic DNA samples for all included species. Orthologous RAD loci are subsequently derived using clustering algorithms similar to those applied for population level RAD-Seq analyses [e.g. the Stacks software package (Catchen *et al*, 2011; Catchen *et al*, 2013)]. Currently, RAD-Seq analyses are being conducted in a variety of different species for a number of different purposes. These datasets (either raw RAD sequence data or RAD consensus sequences) have the potential to be utilised in meta-analyses and for other additional purposes beyond the scope of the original studies. However, best practices for the use of this data for reliably inferring cross-population/cross-species loci remain undefined. This study highlights the potential utility of cross-laboratory RAD-Seq datasets for the identification of cross-population and cross-species homologous RAD loci. In addition, this study tests the utility of these loci in: (i) the inference of cross-species chromosomal orthologies; and (ii) the estimation of phylogenetic relationships, whilst allowing for varying levels minimum taxon coverage per RAD locus.

The high percentage of homologous RAD loci identified between the two different Atlantic salmon populations (~99%) and across the four independent rainbow trout populations (~79%) (Table 5.1), demonstrates that the same loci are being sampled across different populations of the same species. This is despite expected minor differences in library production, sequencing and bioinformatics pipelines implemented across studies. This suggests that the use of RAD-Seq data for cross-population analyses is feasible (for example for investigating differential selection in isolated populations of the same species).

One of the main advantages of the RAD-Seq protocol for phylogenetic studies is that a substantially higher proportion of the genome can be sampled across a large number of individuals compared to traditional methods, which sample a small number of loci across a limited number of individuals (Gontcharov *et al*, 2004; Rubin *et al*, 2012; Hipp *et al*, 2014). The disadvantage of RAD data for phylogenetic analyses could be two-fold. Firstly, although a large amount of sequence data is obtained, these sequences are typically short (~95bp for single-end RAD-Seq, even

shorter for data from much earlier RAD-Seq experiments). Given that the ability to infer sequence orthology improves with increasing sequence length, the short (60bp) sequences used in this study could reduce the number of inferred orthologous loci. Despite this, in this study, hundreds of orthologous RAD loci were identified across distantly related species. In addition, although the ability to identify orthologous loci is expected to decrease with increasing evolutionary distance, the results presented herein demonstrate that even if only a few hundred common RAD loci are identified between groups of species (as was the case for the five non-salmonid species), sufficient between-species sequence variation may be obtained for reliable estimation of evolutionary relationships.

Secondly, although the RAD-Seq protocol samples a large proportion of the genome, the sequences obtained are both genic and non-genic in origin. In many species, only a small proportion of the genome is expected to be coding [e.g. ~3% for rainbow trout (Berthelot *et al*, 2014)]. Therefore, a high proportion of the identified RAD loci may be from less conserved non-coding regions, and potentially less informative for phylogenetic studies (Cooper and Brown, 2008; Bergmiller *et al*, 2012). The high proportion of RAD loci identified as originating from putative genic regions in this analysis supports the hypothesis that *Sbf*I RAD-Seq may be biased towards genic regions (Amores *et al*, 2011; Everett *et al*, 2012; Arnold *et al*, 2013), which would be advantageous for phylogenetic analyses.

When using RAD-Seq data for the identification of cross-species orthologous sequences, the final amount of sequence available for analysis is likely to be much reduced compared to the original dataset. This is because not all loci are expected to have sequence present for every species in the analysis, due to genome mutations, deletions and re-arrangements. Recent studies suggest that these loci may be phylogenetically informative, and that parameters applied to remove loci with an excess of missing data may be too strict (Eaton and Ree, 2013; Huang and Knowles, 2014).

In this study, despite the application of relatively strict filtering parameters for the inference of cross-species RAD locus clusters, thousands of loci still remained. This suggests that the abundance of data obtained from RAD-Seq is potentially robust to these filtering parameters. The inclusion of RAD loci which appear absent in some species did not affect estimated relationships, and an overall improvement in node support was observed, further supporting the hypothesis that such data could be informative in phylogenetic studies. However, a slight elongation in branch lengths was observed, suggesting that care must be taken when incorporating missing sequence data in analyses aiming to estimate divergence times for a given collection of species. Recent studies suggest that an excess of missing sequences at multiple RAD loci may mask any true signals when estimating co-ancestry coefficients across diverging populations (Chattopadhyay *et al*, 2014). Therefore, the proportion of missing RAD data allowed should be carefully considered, and may be influenced by the specific aims of the study.

## 5.6 Conclusion

In this study, RAD-Seq data derived from different laboratories was used to estimate evolutionary relationships across ten aquatic species. Within species and across populations, a large proportion of shared RAD loci were identified, despite potential variation in laboratory techniques and bioinformatic pipelines. As expected, the number of orthologous RAD loci identified across species decreased as the evolutionary distance increased, ranging from ~3,000 between salmonid species to ~450 between all ten species. Multiple alignments of sequences within orthologous RAD loci allowed the identification of interspecific variants. Concatenated sequences of variants were used to estimate evolutionary relationships, which were consistent with previously published phylogenies. Although no difference in tree topologies were observed, overall improvements in node support were seen when parameters were relaxed to include loci for which only a proportion of species had sequence data. The high proportion of putative genic RAD loci identified supports the previous inferences of a bias of *Sbf*I RAD-Seq towards gene-rich regions, which is likely to be useful for the identification of orthologous RAD loci across species in phylogenetic studies. Overall, this study has highlighted the potential utility of cross-

laboratory RAD-Seq datasets. Despite the expected reduction in the ability to identify orthologous loci with increasing evolutionary distance, the results presented herein demonstrate that even if only a few hundred common RAD loci are identified between groups of species, sufficient phylogenetic signal may be obtained for the reliable estimation of evolutionary relationships.

## 5.7 Acknowledgements

I would like to acknowledge the following people for the donation of RAD-Seq data used in this analysis: Dr Pierre-Alexandre Gagnaire (Institut des Sciences de l'Evolution de Montpellier, France), Dr Daniel Berner (Universität Basel, Switzerland), Dr Krista Nichols (Purdue University, US), Dr Matt Hale (Purdue University, US), Dr Ben Hecht (Purdue University, US), Dr Meredith Everett (University of Washington, US), Dr Michaël Bekaert (Institute of Aquaculture, Stirling), Dr Christos Palaiokostas (Institute of Aquaculture, Stirling), Dr Angel Amores (University of Oregon, US), and Dr Julian Catchen (University of Oregon, US).

# Chapter 6

# Discussion

## 6.1 Thesis motivation

Atlantic salmon aquaculture is a rapidly growing industry, providing food security and economic benefits across many countries worldwide (FAO, 2013). Although farming practices are relatively well established, challenges, such as disease outbreaks, can result in reductions in production efficiency, economic losses, as well as fish welfare issues. These diseases do not originate from a farmed environment, and are known to affect wild populations of salmon (Waknitz *et al*, 2002; Waknitz *et al*, 2003; Thorstad *et al*, 2008; Whelan, 2010; Murray *et al*, 2011; Ruane and Jones, 2013). However, the intense rearing conditions on farms is thought to exacerbate the effects of naturally occurring diseases, since interactions and disease transmission is more likely relative to wild conditions. As such, high levels of mortality and morbidity may be observed during infection epidemics in farmed populations (Heuch and Mo, 2001; Bjorn and Finstad, 2002; Turnbull *et al*, 2005; Skilbrei and Wennevik, 2006; Krkosek *et al*, 2007).

In addition, the transfer of farmed salmon to seawater cages for the adult stage of the lifecycle means that farmed salmon are exposed to natural environmental conditions. This increases the potential for interactions and pathogen transmission with wild salmon populations, either through waterborne infections though seawater cages, or through direct contact of wild salmon with farmed escapees (Heuch and Mo, 2001; Naylor *et al*, 2005; Thorstad *et al*, 2008). As such, the correct management of disease and the prevention of disease outbreaks on farms is both economically advantageous, and reduces concerns for wild populations.

Although disease management strategies exist (such as site hygiene, vaccination, limiting fish transfer across sites and testing for disease carriers before breeding),

these are not always fully effective (Akhlaghi *et al*, 1996; Mikalsen *et al*, 2004; FRS, 2007; Murray *et al*, 2011; Jensen *et al*, 2012; Karlsen *et al*, 2012; FAO, 2014b; Graham *et al*, 2014). In agricultural farming, one of the ways of tackling disease is by improving the natural host resistance of farmed individuals through family-based selective breeding programs. However, as discussed throughout this thesis, family-based selective breeding programs have some limitations when selecting for disease resistance. An alternative to family-based selection is to identify and breed from the most resistant individuals. One way of doing this is by incorporating molecular markers into selective breeding programs, either for marker-assisted or genomic selection (Meuwissen and Goddard, 1996; Soller and Andersson, 1998; Nicholas, 2005; Ragimekula *et al*, 2013). This firstly requires the quantification of genetic variation for the trait of interest within the population, followed by linkage and association studies to identify significant marker-QTL associations.

Following the identification and mapping of QTL, studies into further fine-mapping and identification of the causative variant(s) underlying the QTL can be undertaken. This would enable the implementation of selection directly on the causative variants themselves, as well as hinting at biological mechanisms which may be involved in the expression of resistance. For many diseases currently affecting production efficiency on Atlantic salmon farms, resistance to the disease is known to be heritable [e.g. see Kjøglum *et al* (2008), Taylor *et al* (2009), Salte *et al* (2010); summary given in Yáñez *et al* (2014)]. However, the underlying genetic basis for this resistance (i.e. resistance QTL) remains uncharacterised [(FAO, 2014b), see Table 1.1].

The mapping and identification of tight marker-QTL linkage relationships requires a high density of genome-wide markers for the species of interest. In comparison to other aquatic species, genomic resources for Atlantic salmon are relatively abundant, and include several linkage maps, EST databases, BAC and physical libraries, and a dense SNP array (Moen *et al*, 2004a; Rise *et al*, 2004; Ng *et al*, 2005; Davidson *et al*, 2010; Leong *et al*, 2010; Gidskehaug *et al*, 2011; Brenna-Hansen *et al*, 2012; NCBI, 2013a; NCBI, 2013b; Houston *et al*, 2014a). In addition, the Atlantic salmon genome

is being sequenced. The latest draft assembly is available and contains 944,548 contigs ($N_{50}$=34,932bp), although contigs are yet to be assigned to linkage group [ASalBase, http://www.ncbi.nlm.nih.gov/assembly/GCA_000233375.3; Davidson *et al* (2010)]. However, resources are still not as abundant as for livestock species, and there is still much room for improvement.

Whilst this is ongoing, investigations into characterising features of the Atlantic salmon genome, and the identification of causative variants underlying disease resistance QTL, may be undertaken through comparative orthology analyses to related fish species with more abundant genomic resources. Therefore, the identification of model species closely related to Atlantic salmon is paramount, since this increases the confidence in the inferences made.

This thesis describes: (i) the generation of Atlantic salmon genomic resources for use in genetics and genomics research and industry; (ii) the use of these and other resources in the characterisation of the Atlantic salmon genome; and (iii) the potential use of these and other available resources in Atlantic salmon breeding programs to improve resistance to two viral diseases on Atlantic salmon farms.

## 6.2 Thesis objectives

The lack of a high-quality assembled and annotated reference genome, together with the sparse availability of genomic resources, has made the exploration of features of the Atlantic salmon genome a difficult task. This includes investigations into the salmonid-specific genome duplication, differences in recombination rates between the sexes, and in the identification and downstream characterisation of QTL underlying traits of interest.

Whilst the construction of a reference genome and other resources is ongoing, one way of conducting such investigations is through comparative mapping and orthology to other closely related species with high-quality reference genomes. Comparative mapping first requires the generation of sequence information from the genome of interest. For non-model species such as Atlantic salmon, one way of

generating such sequence data is by using next-generation sequencing technologies, such as Restriction site-Associated DNA sequencing (RAD-Seq). however, despite the increasing popularity of RAD-Seq, best practices for the quality control of the data obtained, ability to combine RAD-seq datasets to identify cross-population and/or cross-species homologous RAD loci, and the utility of these loci in downstream analyses (such as the inference of evolutionary relationships), are yet to be investigated. To investigate the utility of short consensus RAD-Seq contigs generated across different laboratories, sequence alignment followed by clustering of loci across populations and species was conducted, using published RAD-Seq datasets from ten teleost fish (chapter five). To investigate the utility of these sequences in the inference of evolutionary relationships, identified cross-species orthologous RAD loci were used in the reconstruction of relationships amongst the ten fish.

In addition to this, these Atlantic salmon RAD-Seq contigs were utilised in the construction of the first Atlantic salmon high-density RAD-Seq SNP linkage map, and in downstream investigations of the architecture of the salmon genome (chapter two). In particular, the salmonid-specific genome duplication event, the differences in recombination rates and patterns between Atlantic salmon males and females, and the chromosomal positions of as yet unassembled Atlantic salmon reference genome contigs were investigated.

Currently, the genetic architecture of host response to many of the diseases on Atlantic salmon farms is unknown. As such, this thesis aimed to use the linkage map constructed herein and other available genomic resources to quantify and/or characterise the underlying genetic variation for host response to two highly problematic viral diseases, infectious pancreatic necrosis (IPN) and pancreas disease (PD), in Atlantic salmon.

For IPN, the major QTL involved in resistance has been previously discovered and mapped to linkage group 21 (Houston *et al*, 2008; Moen *et al*, 2009; Houston *et al*, 2010). Recent reports suggest that the putative causative gene and mutation(s)

underlying resistance to IPN may have been identified (Moen and Ødegård, 2014). However, these have not been released in the public domain, and the underlying resistance mechanisms remain unknown. To investigate the biological mechanisms influencing resistance to IPN and to generate a list of candidate genes within the vicinity of the QTL, next-generation sequencing data generated from within the QTL region and gene expression data between resistant and susceptible individuals after viral challenge were utilised (chapter four). This study was conducted prior to the release of recent reports on the identification of the causative gene and mutation(s) underlying resistance.

For PD, a moderate heritability for resistance was previously estimated ($h^2$=0.21±0.005) (Norris *et al*, 2008). However, the underlying genetic architecture to resistance has not been described. Therefore, the aim of chapter three was to estimate genetic parameters (i.e. heritability) for resistance to PD, and to characterise the underlying genetic basis for variation in resistance though QTL mapping analyses.

## 6.3 Overview of outcomes

Chapter two describes the construction and characterisation of the first Atlantic salmon high-density RAD-Seq SNP linkage map, based on SNPs discovered and genotyped within two reference SalMap families. Overall, approximately 6,500 SNP markers were assigned to 29 Atlantic salmon linkage groups. Of these, ~1,800 male-segregating and ~1,400 female-segregating markers were ordered and positioned within each family. Alignment of mapped SNP-flanking sequences to the Atlantic salmon reference genome contigs enabled the assignment of ~4,000 contigs to a linkage group. 112 of these contigs mapped to two or more linkage groups, suggestive of putative regions of homeology between Atlantic salmon chromosomes. Based on final sex-specific map length comparisons, a recombination ratio of ~1:1.5 (male:female) was estimated, in line with recent ratios obtained from denser Atlantic salmon SNP maps [1:1.38, Lien *et al* (2011)]. Analysis of marker distributions within linkage groups indicated a difference in the distribution of recombination events between males and females, with a higher degree of marker clustering towards

putative centromeric regions and increased marker spacing in putative telomeric regions in males.

Alignment of mapped SNP-flanking RAD contigs to the stickleback reference genome identified chromosomal orthologies between the two species, and enabled the inference of orthologous relationships between Atlantic salmon and rainbow trout chromosomes, based on previously published analyses (Danzmann *et al*, 2005; Phillips *et al*, 2009). The identification of shared orthology to a single stickleback linkage group confirmed many of the salmon genome homeologous relationships identified based on shared reference genome contig assignments. These homeologous relationships had previously been identified in other studies [e.g. Danzmann *et al* (2008)]. In addition, in all cases where a 2:1 relationship between Atlantic salmon and stickleback linkage groups respectively were identified, a 2:1 relationship between rainbow trout and stickleback linkage groups had previously been reported (Danzmann *et al*, 2005; Phillips *et al*, 2009). This provides some support for the salmonid-specific genome duplication.

In chapter three, investigations into the host genetic variation underlying resistance to PD are described. A high heritability for resistance ($h^2$ ~0.5) was estimated using survival data from a large population of Atlantic salmon fry, challenged with the PD-causing virus SAV. QTL mapping analyses identified four QTL influencing resistance within the fry population, on chromosomes 3, 4, 7 and 23. The QTL on chromosome 23 reached genome-wide significance in the sire-based half-sibling linkage analysis. Although this QTL was not significant in the dam-based analysis, two dams showed significant segregation for this QTL, which is suggestive of a QTL of relatively low frequency and of large effect segregating within the population.

The only QTL detected as significant across all analyses (including all half-sibling and sib-pair analyses) was mapped to the distal end of chromosome 3. This QTL was estimated to explain ~10% of the within-family phenotypic variation for resistance when results from both the sire- and dam-based linkage analyses were considered in a combined analysis. In addition, two SNP markers showing population-level

association with resistance were identified on this chromosome, and were estimated to independently explain ~30% and ~2% of the genetic variation for resistance. The large difference in the estimates obtained for these two SNPs could be due to the larger distance between the estimated QTL map position and the map position of the SNP explaining the lower proportion of genetic variation, or due to the selective genotyping strategy implemented in this study (see chapter three for further discussion). Importantly, this QTL has been replicated in an independent study of SAV challenged post-smolts, which validates this QTL and suggests its potential for incorporation into Atlantic salmon PD resistance breeding programs (Gonen *et al*, 2015).

Chapter four outlines the approaches taken to investigate the underlying biological role of the major IPN resistance QTL on linkage group 21. Alignment of QTL-linked sequences to four model teleost reference genomes enabled the identification of two QTL-orthologous regions in each of these four published genomes. Analysis of gene presence and conservation of gene order within these regions identified two groups of orthology across species, orthologous groups (OG) A and B, and suggested high gene order conservation within these regions, particularly for OG A. Analysis of gene expression data from IPN resistant and susceptible individuals after viral challenge and mapping of differentially-expressed genes to the stickleback genome suggested enrichment of differentially-expressed genes on stickleback LG II (OG A).

To further refine the QTL-orthologous region on stickleback LG II, salmon sequences more tightly linked to the QTL region were aligned to the stickleback genome. This identified a 2Mb region on stickleback LG II, and the top three differentially-expressed genes were mapped to within this region. Pathway analysis of all differentially-expressed genes generated a list of pathways which appear to be differentially regulated between resistant and susceptible individuals during the initial infection period. These included cell survival and apoptotic pathways, viral entry and replication mechanisms and cell energy production pathways. Mapping of all genes within the 2Mb region in stickleback to these pathways generated a list of putative candidate genes which may be involved in viral response and genetic

resistance, despite showing no change in expression in challenged individuals. These candidate genes and pathways improve our understanding of the host response to infection, and provide candidates for future investigations.

With improvements in sequencing technologies, investigations of non-model genomes are now becoming much more feasible. Datasets generated using similar sequencing protocols and platforms are being utilised in a variety of investigations within the biological studies. In particular, RAD-Seq is increasingly being utilised, especially within aquatic species. Recently, the use of RAD-Seq data for the purposes of cross-species relationship estimation has been investigated [e.g. Eaton and Ree (2013), Jones *et al* (2013), Hipp *et al* (2014)]. However, despite its popularity and the growing number of datasets, problems of data analysis still exist, and best practices for combining datasets across studies for use in meta-analyses remain unclear. For example, the alignment thresholds to apply to infer cross-species orthologous RAD loci (in the absence of a reference genome), and the impact of RAD locus absence for different species in the inference of evolutionary relationships, have not been fully investigated. In addition, further work is required to determine whether the short RAD contigs retain enough sequence similarity to reliably detect cross-species orthologous RAD loci, and, at the same time, a sufficient number of polymorphisms within reads to distinguish between species.

As yet, the use of RAD-Seq data for the inference of evolutionary relationships amongst teleost species has not been investigated. Chapter five describes the estimation of evolutionary relationships across ten teleost fish, using RAD-Seq data generated across different studies and populations. To identify orthologous RAD loci, pairwise alignments of consensus RAD sequences were conducted. As expected, the number of orthologous RAD loci identified decreased with evolutionary distance, and was influenced by the alignment parameters applied. The identification of salmonid species-specific RAD loci was much more feasible, and approximately six-fold more loci were identified within the salmonid species compared to other species. This was despite the application of stricter salmonid species sequence alignment parameters.

Pairwise alignments were grouped across species, and RAD loci shared across species of differing evolutionary relatedness were identified. As expected, the number of shared RAD loci decreased with increasing evolutionary distance. Overall, estimated relationships were concordant with previously published studies. This suggests that RAD-Seq data obtained from different studies can generate enough sequence data for evolutionary relationship estimation, even between distantly related species. The relaxation of filtering parameters to include RAD loci for which not all species had sequence resulted in improved support for inferred relationships. This suggests that, in the context of RAD-Seq data, the incorporation of loci with absent sequences at some species may be phylogenetically informative, and should be considered in such analyses.

## 6.4 Conclusions and relevance of findings

In the absence of a fully-assembled and annotated reference genome, the generation of high-density linkage maps for non-model organisms can be difficult. Genotyping-by-sequencing approaches such as RAD-Seq are enabling the identification of genetic markers (generally SNPs) for use in the generation of new high-density genomic resources (such as SNP linkage maps and SNP arrays) for non-model organisms [e.g. Baird *et al* (2008), Willing *et al* (2011), Houston *et al* (2014a)]. In addition to the identified SNPs, RAD-Seq generates short SNP-flanking sequences of up to ~500bp which can be used for downstream genome characterisation analyses in non-model species.

Many examples of high-density RAD-Seq linkage maps currently exist for non-model, aquatic species. These maps have been utilised in a variety of population genetic, comparative genomic and evolutionary studies [e.g. Baxter *et al* (2011), Amish *et al* (2012), Houston *et al* (2012), Amores *et al* (2014), Penaloza *et al* (2014)]. Although a high-density SNP linkage map is available (Lien *et al*, 2011), no Atlantic salmon RAD-Seq SNP linkage map exists. In addition to the linkage map, chapter two presents the generation of other Atlantic salmon genomic resources, including a database of putative paralogous sequence variants for future RAD-Seq

data filtering, putative genes associated with mapped SNPs, and the chromosome assignments of ~4,000 salmon reference genome contigs. These will facilitate future quality control of Atlantic salmon RAD data, QTL mapping and fine-mapping analyses, and could prove useful in the assembly of the Atlantic salmon reference genome.

The comparisons of mapped marker positions between males and females presented in chapter two provided further evidence to support the previously described differences in the patterns of recombination events between the sexes (Gilbey *et al*, 2004; Moen *et al*, 2004a; Moen *et al*, 2008; Lien *et al*, 2011). Further, although a small disparity in overall recombination rates between the sexes was noted, this was minor compared to those reported by lower density microsatellite maps, and is in line with that reported by the only other published Atlantic salmon high-density SNP linkage map (Lien *et al*, 2011). This supports previous hypotheses of the improved ability to detect male recombination patterns with increasing marker density at telomeric regions of the chromosome (Moen *et al*, 2004a; Moen *et al*, 2008; Lien *et al*, 2011).

To my knowledge, the results presented in chapter three are the first reported investigations towards improving our understanding of the underlying genetic architecture of PD resistance in farmed Atlantic salmon. The estimated high heritability for resistance suggests that family-based selection is plausible, and could result in improvements in resistance in future generations. Therefore, until additional studies to validate and further characterise the identified QTL are conducted, family-based selection can be applied.

The heritability estimated in this study was almost double that estimated in the only other published study in a natural PD outbreak in post-smolts (Norris *et al*, 2008), and was more similar to that estimated in a population of post-smolt challenged with SAV (Gonen *et al*, 2015). This is likely to be a reflection of the greater control and uniformity of the time of infection in deliberate challenge experiments as compared to natural field data, where pathogen exposure and infection times are not as

synchronised amongst individuals. This may result in increased data noise and lower heritability estimates obtained from natural field outbreaks (Bishop and Woolliams, 2010b; Woo *et al*, 2011).

Of the identified PD resistance QTL in chapter three, the QTL consistently identified as significant across different statistical models, and estimated to explain the most within-family variation for resistance, was mapped to chromosome 3. Recently, this QTL was confirmed in an independent population of Atlantic salmon post-smolts. Common anchor markers between the two linkage maps used to map QTL in this population of post-smolts and in the fry population data used in chapter three positioned this QTL to within the same location on the chromosome. The independent confirmation and co-localisation of the PD resistance QTL identified on chromosome 3 validates this QTL, and highlights its potential for marker-assisted selection. In addition, the identification of the same QTL in both fry and post-smolt SAV challenge experiments suggests that some of the same mechanisms underlie the response to infection in both the juvenile and adult stages of the lifecycle, and gives some insight into the biology behind infection. For example, the adaptive immune system is unlikely to play a role in the initial mechanisms of resistance, since this is not developed in fry (Uribe *et al*, 2011; Grove *et al*, 2013).

The identification of the same QTL across both life stages is useful knowledge for future studies, in enabling more controlled challenge experiments to further our understanding of the basis for host resistance. So far, natural field and farm infections with SAV have only been reported in post-smolts (Rowley *et al*, 1998; Weston *et al*, 1999; McLoughlin *et al*, 2002; Rodger and Mitchell, 2007; Taksdal *et al*, 2007; Kristoffersen *et al*, 2009; Jansen *et al*, 2010b; Hjortaas *et al*, 2013; Jansen *et al*, 2014). However, conducting post-smolt challenge experiments is costly (since fish must be reared to adult stage before conducting the challenge experiment) and labour intensive (since viral exposure is generally though intraperitoneal injection). The ability to conduct challenge experiments in a more controlled environment in the fry stage, as well as the potential utility of estimated genetic parameters in selection

for resistant post-smolt as suggested by this study, will enable further studies into the characterisation of the host genetics to this disease.

For many viral diseases affecting aquaculture farms, resistance at both the juvenile and adult stages of the salmon lifecycle has been attributed to a single or a small number of common QTL [e.g. IPN (Houston *et al*, 2008), infectious salmon anaemia in Atlantic salmon (Moen *et al*, 2007), Rhabdovirus in rainbow trout (Verrier *et al*, 2013), viral haemorrhagic septicaemia in turbot (Rodríguez-Ramilo et al, 2014)]. This may suggest that mechanisms of the innate immune response may play an important role in the response to viral infections in fish. Studies into the characterisation of the immune repertoire of fish are rare, and the extent of the complexity of the adaptive immune response in fish is not well understood (Uribe *et al*, 2011). Published and continuing studies into characterising T cells and adaptive immune genes in fish are filling this knowledge gap (Desvignes *et al*, 2002; Rønneseth *et al*, 2006; Cepeda *et al*, 2011; Xu *et al*, 2012; Grove *et al*, 2013; Herath *et al*, 2013; Ballesteros *et al*, 2014; Collet, 2014).

The identification of QTL influencing resistance to diseases on farms is the first step towards the identification of the actual causative variant(s) underlying the QTL effect. In the absence of a fully-assembled and annotated Atlantic salmon genome, studies into the identification of candidate gene(s) underlying resistance QTL have often adopted a comparative approach, based on orthologous chromosomal relationships with teleost species with high-quality assembled and annotated genomes [e.g. Li *et al* (2011)]. Comparative cross-species analyses based on sequence orthology may not result in the identification of the causative variant(s) underlying resistance QTL. However, when combined with additional resources (such as gene expression data and knowledge of the biological mechanisms underlying the trait), they can further the understanding of the biological processes contributing to the trait, and highlight interesting candidates for future investigation [for example, as described in Hegarty *et al* (2013) for perennial ryegrass].

The results presented in chapter four provide additional knowledge on the IPN resistance mechanisms in the following ways. First, the identification of regions of conserved orthology across assembled and annotated teleost genomes suggests an ancestral origin to the QTL region. These regions were previously found to derive from the same ancestral proto-Actinopterygian linkage groups. Further, the results presented herein suggest conservation of gene order across species within the QTL region, which may be indicative of a region under positive selection for co-localisation of genes involved in the same biological pathways.

Second, analyses of pathways enriched for differentially-expressed genes between resistant and susceptible individuals after viral challenge suggest that apoptotic, cell energy production (such as the fatty acid synthesis pathway), viral entry/replication (for example the RhoA/Rac signalling pathway), and viral stress response pathways (such as the endoplasmic reticulum stress response and the protein ubiquitination pathways) may play a role during the initial stages of infection. These pathways, in particular the apoptotic pathways, have previously been implicated during IPN infection (Hong *et al*, 2002; Chiu *et al*, 2010), or in response to other viral infections.

Third, mapping of differentially-expressed genes to the stickleback genome suggested enrichment of differentially-expressed genes in the QTL-orthologous region on stickleback LG II. The most significantly differentially-expressed genes have previously been implicated in response to viral infections across a large number of studies, and some are known to be involved in the biological pathways described above (O'Brien, 1998; Lin *et al*, 2010; Collet, 2014).

The mapping of the top three most differentially-expressed genes to within the 2Mb QTL-orthologous region of stickleback LG II may be due a common cis-QTL effect for the expression of these genes. Examples of gene order conservation and gene co-localisation are well documented for gene clusters known to be of significant importance across species, and include, for example, the *Hox* gene cluster (Santini *et al*, 2003). The innate immune response is thought to be highly conserved across the tree of life. Therefore, if the genes underlying the IPN QTL are innate immune

related, their conservation across evolution is not unexpected. However, the mapping of differentially-expressed genes to the same region may also be due to a hitchhiking effect, caused by the close family relationships amongst the selected resistant and susceptible individuals for analysis of gene differential expression, and not caused by the resistance mechanisms underlying the QTL. Overall, the results presented herein will be useful in directing future investigations into the IPN resistance-causing mechanisms.

The potential of obtaining large amounts of data and the non-requirement of prior knowledge of the genome of interest are amongst the many advantages of RAD-Seq, which make it particularly attractive for use in non-model species. As such, several species now have RAD-Seq data available across a number of populations. However, despite the potential to combine these datasets into large meta-datasets across species, available datasets are rarely used in more than one study, since the utility of such meta-data in answering important biological questions is unknown. The reproducibility of RAD data across populations has been suggested in some published studies, and a high concordance in the number of shared loci across populations within a given species has been noted (Amish *et al*, 2012; Houston *et al*, 2012; Miller *et al*, 2012; Hale *et al*, 2013). In congruence with these studies, a high proportion of shared RAD loci were identified in chapter five across different populations of Atlantic salmon (99%) and rainbow trout (70%). Furthermore, the identification of many thousands of cross-species orthologous RAD loci and the use of these in the estimation of evolutionary relationships as presented in chapter five, is a step towards demonstrating the reproducibility and potential utility of RAD data across studies and species. This would eliminate the need for costly laboratory sample preparations and re-sequencing of libraries when aiming to identify cross-species orthologous RAD loci.

Previous studies on the use of RAD data for phylogenetic relationship re-construction have suggested that, given the short RAD contig lengths, RAD-Seq data would not be suitable for the estimation of evolutionary relationships between species with more than 100 million years (MY) since their last most recent common

ancestor (Rubin *et al*, 2012). Current published literature on teleost evolutionary relationships suggests that the ten fish included in chapter five shared a last most recent common ancestor ~400 MY (Near *et al*, 2012). Although the number of orthologous RAD loci identified decreased with evolutionary distance, the overall conformity of drawn phylogenies with those previously published suggests that the large amount of sequence data obtained from a single RAD-Seq study, combined with the sampling of many genomic regions achieved through the restriction enzyme digest, provide enough information for relationship inference across distantly related species.

Finally, chapter five also describes investigations into another much debated area of the RAD-Seq protocol, namely, the thresholds to be applied for the filtering of missing sequence at a given RAD locus. The general consensus for filtering of sequences generated from next-generation sequencing platforms is that, given the amplification steps and the known sequencing machine error rates and biases, a high threshold for read filtering should be applied (Everett *et al*, 2011; Koboldt *et al*, 2013; Mardis, 2013; Pavlopoulos *et al*, 2013). This results in the reduction of overall read depth and the amount of data available for subsequent analyses.

In general, given the high volume of data obtained from next-generation sequencers, filtering using strict thresholds does not significantly reduce the amount of data available for analysis. As such, many thousands of shared RAD loci across individuals within a population can be identified [e.g. Hohenlohe *et al* (2011); Lamer *et al* (2014); Zhou *et al* (2014)]. However, with the further expected reduction in data due to the difficulties of detecting orthologous RAD loci across distantly related species, the application of strict filtering criteria could result in the loss of a large amount of data.

In addition to this, recent studies have suggested that the absence of sequence for a given RAD locus in a particular species of interest is useful for phylogenetic inference (Huang and Knowles, 2014). Although the inability to identify a putative RAD locus within a given species could be due to bioinformatic pipelines imposed

on the initial RAD locus discovery steps, missing sequence at a RAD locus could also suggest polymorphisms, mutations, or differences in methylation patterns across species. These are likely to be more pronounced with increasing evolutionary distance (Poland and Rife, 2012; Arnold *et al*, 2013; Eaton and Ree, 2013; Cruaud *et al*, 2014; Huang and Knowles, 2014).

To investigate this, in chapter five, relationships amongst the ten fish species were estimated using RAD loci with varying levels of missing sequence information. In congruence with previous findings, the incorporation of RAD loci with a certain proportion of the species being 'absent' for that RAD locus resulted in improved support for inferred relationships. This demonstrates the importance for consideration of the thresholds applied when attempting to minimise the amount of missing data when inferring cross-species orthologous RAD loci.

## 6.5 Challenges and perspectives for future research

### 6.5.1 Next-generation sequencing technologies in constructing and interpreting linkage maps of non-model species

The SNP linkage map presented in chapter two is a substantial improvement in marker density compared to other linkage maps (i.e. not SNP maps) available in Atlantic salmon, and will be useful in future studies as outlined above. However, despite the improved density, this map is still not as marker populated as those available for some livestock species [e.g. Groenen *et al* (2009); The Bovine HapMap Consortium (2009)]. In addition, although ~6,500 markers were assigned to a linkage group, the ordering and positioning of all linkage group-assigned markers was not possible within the current study. There were two main reasons for this, which stem first from the initial experimental design, and second, from the properties of the linkage mapping software packages currently available. These are discussed below.

In this study, SNP marker identification and genotyping using RAD-Seq was conducted within two reference SalMap families, each consisting of two parents and 46 offspring, to total 96 individuals overall. With the application of filtering steps to remove individuals with excess missing SNP genotypes and/or Mendelian errors, the

number of individuals for final linkage map construction decreased to 77. Many of the currently available linkage mapping software packages (including the OneMap software package used in this study) are only able to construct linkage maps within a single family, and are not able to infer marker orders and positions through combined information across families. Therefore, maps presented in chapter two were constructed for each of the four mapping parents individually. Consequently, the number of available meioses to infer marker order and position was low (36 and 37 offspring in families Br5 and Br6 respectively). Following on from this, the number of informative meioses over all available meiotic events between two markers is going to be even lower, since not all offspring will derive from recombinant gametes (recombination events are required for the inference of marker distances). Therefore, the ability to reliably order and position tightly linked markers is reduced, particularly in the centromeric regions of male linkage groups, since this would require the detection of rare male recombination events. These rare recombination events may be detected in studies with larger numbers of offspring per full-sibling family, which is feasible within aquaculture breeding programs.

### 6.5.2 Utilising linkage maps for QTL detection: Pancreas Disease

At the time of data analysis, to my knowledge, CRI-MAP [Green *et al* (1990); version 2.4, as modified by Xuelu Liu (Monsanto)] was the only freely available software packages able to utilise cross-family information for the construction of population-level linkage maps. The ordering and positioning of SNP markers using CRI-MAP was investigated in this study. However, the CRI-MAP algorithm was written for linkage maps comprised of, at most, a few hundred markers in total. As such, the ordering and positioning of the many thousands of SNPs identified in this study resulted in computational problems (such as limitations in software memory). Despite attempts to overcome this, it was not possible to obtain linkage maps using CRI-MAP.

Since the construction of the final version of the map presented in chapter two, a new algorithm for linkage map construction has been published as part of the Lep-MAP software package, originally written for linkage map construction within *Lepidoptera*

(Rastas *et al*, 2013). As well as the ability to utilise genotypes across families and the capacity to cope with denser marker sets, the Lep-MAP algorithm is able to account for sex-specific differences in recombination rates, and to construct joint maps based on this. Although this software package was not tested using the SNP data from chapter two, it was utilised in the construction of the cross-family linkage maps used for the dam-based PD resistance QTL mapping analyses described in chapter three. The obtained linkage group assignments and order of markers were as expected (given the linkage group assignment and predicted order of markers within the study from which they were originally sourced). Therefore, Lep-MAP could potentially be useful for linkage map construction in Atlantic salmon, given the major differences in sex-specific recombination rates reported.

The PD resistance QTL presented in chapter three, and in particular, the QTL on chromosome 3 which has been independently verified in a population of post-smolts, highlights potential candidates which may be incorporated into selective breeding programs. Although this QTL was consistently identified across studies and the different statistical models implemented, the effect size and proportion of variance explained was not as large as those reported for QTL currently being implemented in Atlantic salmon breeding programs [such as for IPN resistance; (Houston *et al*, 2008)]. This suggests a more polygenic architecture to PD resistance, with the involvement of other non-genetic components in determining resistance (e.g. gene-environment interactions), as has been suggested in previous studies (Rodger and Mitchell, 2007; Kristoffersen *et al*, 2009; Jansen *et al*, 2010b; Jansen *et al*, 2010a; Stene *et al*, 2013; Jansen *et al*, 2014; Taksdal *et al*, 2014). In addition to MAS for the QTL on chromosome 3, the potential of genomic selection for improved PD resistance utilising information from a large number of genome-wide markers should be investigated. For some diseases, such as sea lice (*Lepeophtheirus salmonis*), host genetic resistance has been attributed to many QTL of small effect (i.e. polygenic trait) (Houston *et al*, 2014b). In such cases, genomic rather than marker-assisted selection will be of greater importance in disease selective breeding programs.

Although markers at population-level association with the QTL on chromosome 3 have been identified, these are likely to still be a considerable distance from the QTL (given the low density of the linkage map used for QTL identification). Therefore, the identified marker-QTL linkage relationships may not hold across populations. Validation of the QTL and markers identified in this study in other farmed salmon populations, and the identification of more closely linked markers using higher density genomic resources, is required. This has already partially been achieved for the QTL identified on chromosome 3 in this analysis. Not only was the QTL confirmed in a separate population of post-smolts, the SNP linkage map used for QTL mapping was of higher density compared to that used in chapter three. Therefore, the SNP markers identified as linked to this QTL are more likely to be closer to the QTL region compared to those identified in chapter three.

The identified QTL and linked markers will most likely form part of a larger selective breeding strategy, incorporating multiple methods of improving PD resistance on farms. However, the implementation of these or any of the other identified QTL in marker-assisted selection will take time. Whilst this process is ongoing, the high heritability reported herein suggests that family-based selection could result in substantial improvements in resistance amongst fish stock, and should therefore form a large component of current breeding programs.

### 6.5.3 Next-generation sequences in comparative orthology: Refining QTL in non-model species

In the absence of a fully-assembled and annotated reference genome, the characterisation of QTL of interest and the identification of underlying causative mutations can be difficult. Over the last decade, advancements in next-generation sequencing technologies are making this process much more feasible for non-model organisms. These new technologies enable the generation of sequence data within the region of the QTL of interest, which can be analysed through comparative mapping analyses to related model species.

Chapter four describes the use of RAD and BAC contigs generated from within and around the IPN QTL region in Atlantic salmon in the characterisation of the underlying host genetic basis to resistance. These contigs were generated from within a 10cM QTL confidence interval, which resulted in the identification of large cross-species QTL-orthologous regions. Although a narrower 2cM QTL confidence interval has been identified (Moen *et al*, 2009; Houston *et al*, 2010), sequence information from this region is sparse. The generation of a denser set of sequences from the QTL region, in addition to the completion (i.e. assembly and annotation improvements) of the currently ongoing Atlantic salmon reference genome sequencing project, will make the comparative approach employed in this study more feasible, and enable the identification of causative mutation(s). At the time of writing, reports suggesting that the putative IPN causative mutation(s) has been identified exist (Moen and Ødegård, 2014). However, details of this putative causative gene/variant have not been published.

At the time of analysis, only five assembled and annotated teleost genomes were available. One of these (fugu) was in scaffold form and could not be used in this study (Ensembl release 69, October 2012). No salmonid reference genome was available. At the time of writing, eleven teleost fish genomes were available (Ensembl release 77, October 2014) and, importantly, the first salmonid species assembled and annotated reference genome for rainbow trout was available [Berthelot *et al* (2014); http://www.genoscope.cns.fr/trout-ggb/]. These new resources will substantially improve comparative analyses in Atlantic salmon, and facilitate the identification of candidate and causative variants underlying traits of interest. Together with the knowledge gained from the results presented in this study, future comparative studies using these resources with a denser set of QTL-linked sequences will improve our understanding of the genetic basis of host resistance to IPN.

### 6.5.4 Added value from sequence data: Next-generation sequences in the inference of evolutionary relationships

The major limitation of RAD-Seq contigs in the inference of cross-species sequence orthology, as was implemented in chapters four and five, is that the read lengths generated from the sequencing platforms currently available are quite short. In particular for the analysis in chapter five, all sequences were further trimmed to 60bp. This was done to account for the shortest read length in the study, and in an attempt to avoid the incorrect interpretation of alignment quality parameters (such as E-value) due to variations in sequence lengths. In general, alignment software manuals suggest a minimum of ~20bp (i.e. ~6 amino acids) for sequence alignment purposes (Barrick, 2014; EMBL-EBI, 2014), and previous studies have suggest that a minimum of 50bp is sufficient for unique alignment of reads back to a reference genome (Storvall *et al*, 2013).

For cross-species orthology identification, longer sequence reads would improve the confidence in inferred orthologies, due to the following reasons. First, with increasing evolutionary distance, the accumulation of mutations, even within highly conserved regions of the genome (such as gene coding regions), is highly likely. This could result in an overall reduced sequence similarity (Hardison *et al*, 1997; Brenner *et al*, 2002; Santini *et al*, 2003). This increases the likelihood of false orthologies, and also increases the chances of a single read aligning with equal significance to multiple locations in the genome of the compared species (Kamvysselis, 2003; Koonin and Galperin, 2003).

Second, although digestion of genomic DNA with restriction enzymes in RAD-Seq is effective at sampling the whole genome, sampled loci include unconserved intronic, repetitive, and conserved protein and other coding regions (Miller *et al*, 2007; Baird *et al*, 2008). Given that in most species only a small proportion of the genome is coding [e.g. see Onyango *et al* (2000)], less evolutionarily conserved intronic/repetitive genomic regions are likely to be highly represented amongst the identified RAD loci. However, the results presented in this and other studies suggest that the choice of restriction enzyme could influence which regions of the genome

are more likely to be sampled. For example, the *Sbf*I enzyme could be utilised in studies requiring fragments from gene-rich regions of the genome [e.g. Everett *et al* (2012)]. With the increasing read lengths being obtained from next-generation sequencing platforms, the inference of orthologous relationships between loci originating from intronic regions of the genome may become possible, particularly for more closely related species with lower sequence divergence expected due to closer evolutionary relatedness and fewer chances for mutations within shorter time intervals.

## 6.6 Implications and practical considerations

As well as the methodology for linkage map construction, chapter two outlines the subsequent annotation of the map, and its uses for the characterisation of genomic features of the salmon genome. In addition, the annotation of putative gene-associated mapped SNPs presented herein will facilitate future studies aiming to identify putative candidate genes underlying QTL of interest. The integration of this map with the available microsatellite map and reference genome contigs will provide further resources for QTL mapping and fine-mapping in future studies, and could facilitate the assembly and annotation of the Atlantic salmon reference genome. Although the map presented herein has not been integrated with the only other published Atlantic salmon high-density SNP linkage map (Lien *et al*, 2011), both maps have/are being integrated with the reference genome contigs [currently ongoing for the Lien *et al* (2011) map, see http://web.uvic.ca/grasp/]. This will provide a richer repertoire of genomic resources for Atlantic salmon.

The identification of the PD resistance QTL on chromosome 3 as described in chapter three is a considerable advancement in the current understanding of host resistance to PD. The validation of this QTL in a population of post-smolts in an independent study (Gonen *et al*, 2015), and the confirmation of the co-localisation of the QTL to the same region of chromosome 3 in both studies, suggests that this QTL is a good candidate for incorporation into resistance breeding programs.

In addition, the replication of the QTL in both the fry and post-smolt studies suggests that investigations into the host basis for resistance to PD may be conducted using a fry challenge model, and that until marker-assisted selection can be fully implemented, family-based selection can be undertaken based on resistance at the fry stage, and individuals do not need to be reared to post-smolt age for viral challenge. For research purposes, the identification of the QTL across the juvenile and adult stages suggests that similar biological mechanisms underlie host resistance at both life stages. If the underlying resistance mechanisms are immune related, these are therefore more likely to be part of the general innate rather than an adaptive immune response. This is partially supported by the mapping of QTL affecting other viral diseases to the same chromosome (chromosome 3) in Atlantic salmon [e.g.Gilbey *et al* (2006)]. This knowledge will facilitate the exploration for candidate variant(s) within the QTL region, using similar approaches to those described in chapter four for characterisation of the IPN QTL.

The methodological approaches presented in chapter four of this study have improved our understanding of the host basis to IPN resistance, and the results presented are a substantial advancement towards the discovery of the underlying biology for resistance. A number of putative candidates which warrant future investigations have been highlighted. In particular, the results suggest that gene order is highly conserved within the region, which could mean that denser marker information for further fine-mapping of the QTL, together with a denser set of QTL-linked sequence data, could generate a more concise list of candidate genes through comparative mapping to model species.

The pathway enrichment analyses were able to highlight biological pathways potentially influencing the differences in the initial stages of IPNV infection between resistant and susceptible fish. Analysis of genes within these pathways may detect novel candidates for selection purposes, or suggest novel approaches for vaccine development. The identified resistant pathways, in conjunction with the mapping of the most significantly differentially-expressed genes to a single location within the stickleback genome, suggests that high-density sequence data obtained from gene

expression studies (such as RNA-Seq) may provide a dense enough repertoire of data for, and may benefit from, comparative mapping approaches to published genomes.

Throughout this thesis, the use of RAD data generated within a specific population for a specific study has been described (chapter two for linkage map construction, chapter four for IPN-QTL characterisation). In recent years, a large number of studies have investigated the many applications and uses of RAD-Seq-derived data in a variety of species within specific populations, including for the purposes of estimating evolutionary relationships [e.g. Eaton and Ree (2013); Jones *et al* (2013); Hipp *et al* (2014)].

At the time of writing, the utility of RAD-Seq data for the purposes of relationship estimation amongst finfish has not been published. Furthermore, all phylogenetics studies have involved the sampling of genomic DNA and inference of cross-species RAD loci within the study itself. Therefore, there exists a large amount of freely available RAD data across species and populations, which can be utilised for additional purposes beyond the scope of the original study for which it was intended. In chapter five, the use of cross-laboratory RAD data for the identification of orthologous RAD loci across populations of the same species and across distantly related species was investigated. Overall, this study demonstrates the reproducibility of RAD data across populations of the same species, with positive implications for the sampling of the same loci across species, provided that the same restriction enzyme is used for RAD library preparation. Contrary to other published findings, this study suggests that the use of RAD data across species with evolutionary distances exceeding 100 MY is feasible, and previously published relationships can be recovered. Further, this study addresses the implications of the incorporation of missing sequence data into analyses, and the results obtained suggest that analyses could benefit from relaxing thresholds on missing sequence information with regards to RAD-Seq data in phylogenetics research.

# Conclusion

For non-model organisms, investigations into characterising the genome and understanding the genetic basis of phenotypic variation in traits of interest can be difficult. In Atlantic salmon, this is further complicated by the large and highly-repetitive genomic structure, as well as the retention of genomic signatures of evolutionary events, such as genome duplications. The recent advancements in next-generation sequencing (NGS) technologies have made the generation of genomic resources for the investigation of non-model genomes much more feasible. This thesis describes the use of one such technology, RAD-Seq, for the generation of genomic resources in Atlantic salmon, and the subsequent use of these and other genomic resources in the characterisation of the Atlantic salmon genome and the underlying genetic architecture for resistance to two viral diseases affecting farmed Atlantic salmon populations.

First, this thesis describes the generation of the first Atlantic salmon high-density RAD-Seq SNP linkage map, and demonstrates the utility of RAD-Seq in the identification of novel genetic markers for use in linkage map construction in non-model organisms. This map was then used in the exploration of Atlantic salmon genomic features, such as the difference in recombination rates between males and females and identification of paralogous regions of the genome, and observed patterns were consistent with previously published studies. In addition, the utility of this map and associated sequence data in genome assembly and comparative mapping purposes is demonstrated.

Second, this thesis presents the use of this map and other published maps for the identification of QTL underlying variation in host resistance to pancreas disease (PD), a viral disease currently affecting Atlantic salmon farms. First, a heritability of ~0.5 for resistance to PD was estimated within a large population of Atlantic salmon fry, challenged with the PD-causing virus. Sire- and dam-based linkage mapping analyses identified four QTL potentially influencing resistance within this

population, on chromosomes 3, 4, 7 and 23. The QTL consistently identified as significantly associated with resistance and estimated to explain the most within-family phenotypic variation for resistance was mapped to the distal end of chromosome 3. This QTL was independently confirmed in a population of post-smolts, challenged with the same strain of the disease causing virus. Further, both studies mapped this QTL to the same location on chromosome 3. SNP markers at population-level association with this QTL have been identified in both populations. The independent mapping of this QTL in two independent populations validates this QTL, and demonstrates its potential for use in marker-assisted selection in Atlantic salmon resistance breeding programs.

Third, this thesis describes the use of sequences generated from RAD-Seq and 454 BAC-end sequencing in the characterisation of the major QTL involved in host resistance to another viral disease, infectious pancreatic necrosis (IPN). Comparative mapping of sequences generated from in and around the QTL region to reference genome sequences of model teleost species enabled the identification of two putative QTL-orthologous regions within each fish genome. Comparison of gene presence and order across QTL-orthologous regions of the model teleost fish identified two groups of orthologous chromosomal relationships, with high gene order conservation across regions.

Comparison of gene expression patterns between IPN resistant and susceptible individuals upon challenge with IPNV enabled the identification of putative candidate genes within the vicinity of the IPN QTL, which may be involved in resistance. Pathway enrichment analysis of differentially-expressed genes identified a list of 255 pathways likely to be differentially regulated between resistant and susceptible individuals. Mapping of genes within the QTL-orthologous region of stickleback to these pathways identified genes which are not differentially-expressed but may be involved in resistance. These analyses demonstrate the utility of sequence data in the identification of biological mechanisms underlying a trait of interest in non-model organisms, through comparative mapping to published reference genomes.

Although NGS sequences are being utilised in many studies, bioinformatic pipelines for the filtering of such data to obtain biologically relevant and reliable sequences are still to be standardised. In particular, best practices to combine NGS sequences generated using different pipelines to produce single large datasets are still not clear. The last part of this thesis presents the exploration of such cross-laboratory data, and demonstrates its use in obtaining thousands of reliable cross-population and cross-species orthologous sequences. In addition, the utility of cross-laboratory data for the inference of evolutionary relationships is demonstrated, using data from ten teleost fish species.

Overall, this thesis presents results towards enhancing the understanding of a number of Atlantic salmon genetics and genomics research questions. Furthermore, this thesis provides new genomic resources and pipelines for use in studies aiming to understand and characterise the Atlantic salmon genome, and how variation at the genetic level influences expression of a particular phenotype of interest.

# Appendices

# Appendix A – Tables

**Table A1: Library structure and read depth for the paired-end RAD-sequencing libraries**

| Family Br5 | | | | | | Family Br6 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fish** | **Sex** | **Library** | **Barcode** | **No. raw PE** | **No. mapped reads** | **Fish** | **Sex** | **Library** | **Barcode** | **No. raw PE** | **No. mapped reads** |
| Sire | M | 1 | CTAGG | 9,840,258 | 3,339,336 | Sire | M | 2 | CTAGG | 9,779,574 | 2,817,941 |
| Dam | F | 1 | GAGAT | 10,618,889 | 3,607,996 | Dam | F | 2 | GAGAT | 14,210,782 | 4,263,519 |
| 01 | F | 6 | CTAGG | 906,003 | 483,118 | 01 | F | 3 | CTAGG | 1,465,063 | 482,837 |
| 02 | F | 6 | GAGAT | 1,415,716 | 729,246 | 02 | M | 3 | GAGAT | 1,597,905 | 526,808 |
| 03 | F | 6 | GCGCC | 392,470 | 221,823 | 03 | M | 3 | GCGCC | 484,244 | 176,234 |
| 04 | M | 6 | GTACA | 2,227,594 | 1,132,308 | 04 | F | 3 | GTACA | 1,536,766 | 511,956 |
| 05 | M | 6 | GTGTG | 1,033,311 | 542,110 | 05 | M | 3 | GTGTG | 978,969 | 333,111 |
| 06 | F | 8 | CATGA | 2,645,093 | 1,314,085 | 06 | F | 3 | TAGCA | 2,434,848 | 785,411 |
| 07 | M | 8 | CACAG | 2,994,367 | 1,462,823 | 07 | F | 3 | TCAGA | 1,787,224 | 593,222 |
| 08 | F | 6 | TCGAG | 1,265,152 | 663,088 | 08 | M | 3 | TCGAG | 1,549,776 | 525,250 |
| 09 | F | 6 | TGACC | 2,210,309 | 1,141,938 | 09 | M | 3 | TGACC | 2,051,583 | 687,784 |
| 10 | M | 6 | ACTGC | 2,071,445 | 1,078,076 | 10 | F | 3 | ACTGC | 2,359,457 | 785,385 |
| 11 | F | 6 | ACACG | 2,442,613 | 1,265,728 | 11 | M | 3 | ACACG | 2,049,384 | 706,676 |
| 12 | M | 6 | AGAGT | 2,429,375 | 1,244,671 | 12 | M | 3 | AGAGT | 2,694,450 | 855,237 |
| 13 | F | 6 | ATGCT | 3,168,094 | 1,591,906 | 13 | F | 3 | ATGCT | 2,617,124 | 854,990 |
| 14 | F | 7 | CTAGG | 1,235,965 | 268,147 | 14 | F | 3 | CAGTC | 2,166,291 | 735,312 |
| 15 | F | 7 | GAGAT | 1,469,717 | 313,862 | 15 | F | 3 | CATGA | 341,025 | 127,581 |
| 16 | M | 6 | CAGTC | 2,989,144 | 1,534,393 | 16 | M | 3 | CACAG | 601,453 | 212,743 |
| 17 | M | 6 | CATGA | 475,276 | 264,810 | 17 | F | 4 | CTAGG | 1,572,331 | 484,234 |
| 18 | M | 6 | CACAG | 542,851 | 299,218 | 18 | F | 4 | GAGAT | 1,984,142 | 597,852 |
| 19 | M | 7 | GCGCC | 690,542 | 164,317 | 19 | F | 4 | GCGCC | 430,374 | 148,146 |
| 20 | F | 7 | GTACA | 2,505,144 | 526,591 | 20 | M | 4 | GTACA | 1,925,850 | 589,276 |
| 21 | F | 7 | GTGTG | 1,096,955 | 236,014 | 21 | F | 4 | GTGTG | 1,194,498 | 368,664 |
| 22 | F | 7 | TAGCA | 1,941,267 | 421,905 | 22 | M | 4 | TAGCA | 2,804,243 | 813,251 |
| 23 | F | 7 | TCAGA | 1,831,912 | 393,314 | 23 | M | 5 | CTAGG | 1,278,277 | 754,333 |
| 24 | M | 7 | TCGAG | 1,564,190 | 341,612 | 24 | F | 5 | GAGAT | 1,450,505 | 847,563 |
| 25 | F | 7 | TGACC | 2,716,325 | 589,106 | 25 | M | 4 | TCAGA | 2,094,522 | 634,384 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | M | 7 | ACTGC | 1,910,822 | 417,196 | 26 | M | 4 | TCGAG | 1,683,378 | 522,012 |
| 27 | F | 7 | ACACG | 2,599,942 | 577,493 | 27 | F | 4 | TGACC | 2,914,441 | 897,761 |
| 28 | F | 8 | CTAGG | 1,734,761 | 828,686 | 28 | F | 4 | ACTGC | 1,919,675 | 597,518 |
| 29 | M | 7 | AGAGT | 2,672,124 | 566,127 | 29 | M | 4 | ACACG | 2,211,708 | 698,042 |
| 30 | F | 8 | GAGAT | 2,109,042 | 972,312 | 30 | M | 4 | AGAGT | 3,916,489 | 1,110,022 |
| 31 | F | 8 | GCGCC | 687,604 | 361,519 | 31 | F | 4 | ATGCT | 2,187,938 | 669,231 |
| 32 | M | 7 | ATGCT | 3,038,521 | 636,525 | 32 | M | 4 | CAGTC | 2,214,529 | 689,992 |
| 33 | M | 7 | CAGTC | 2,292,598 | 501,144 | 33 | F | 5 | GCGCC | 714,044 | 443,931 |
| 34 | M | 7 | CATGA | 466,226 | 111,114 | 34 | F | 5 | GTACA | 2,251,290 | 1,299,516 |
| 35 | M | 7 | CACAG | 750,420 | 174,430 | 35 | M | 5 | GTGTG | 1,308,700 | 767,835 |
| 36 | M | 8 | GTACA | 2,491,872 | 1,157,827 | 36 | M | 5 | TAGCA | 2,405,386 | 1,401,132 |
| 37 | M | 8 | GTGTG | 1,507,512 | 718,618 | 37 | F | 5 | TCAGA | 2,752,922 | 1,570,525 |
| 38 | F | 8 | TAGCA | 2,531,598 | 1,209,215 | 38 | M | 5 | TCGAG | 1,749,406 | 1,031,150 |
| 39 | M | 8 | TCAGA | 3,370,670 | 1,556,289 | 39 | M | 5 | TGACC | 2,789,074 | 1,614,461 |
| 40 | F | 8 | TCGAG | 2,652,488 | 1,252,618 | 40 | F | 5 | ACTGC | 2,945,707 | 1,701,816 |
| 41 | M | 8 | TGACC | 3,505,350 | 1,664,949 | 41 | F | 5 | ACACG | 3,268,172 | 1,876,396 |
| 42 | M | 8 | ACTGC | 3,036,980 | 1,453,001 | 42 | M | 5 | AGAGT | 3,030,322 | 1,749,408 |
| 43 | F | 8 | ACACG | 4,186,888 | 1,987,317 | 43 | F | 5 | ATGCT | 4,074,713 | 2,283,001 |
| 44 | F | 8 | AGAGT | 3,677,255 | 1,727,430 | 44 | F | 5 | CAGTC | 3,336,800 | 1,931,108 |
| 45 | M | 8 | ATGCT | 4,126,124 | 1,919,761 | 45 | M | 5 | CATGA | 614,665 | 380,262 |
| 46 | M | 8 | CAGTC | 3,457,228 | 1,654,626 | 46 | M | 5 | CACAG | 696,853 | 426,350 |

* Number of Illumina paired-end reads per individual following demultiplexing of reads

** Number of reads following removal of 'PCR duplicates' (paired-end reads with identical read 1 and read 2) which should approximate the number of unique DNA fragments in the sample (see Davey et al. 2012).

**Table A2: Markers used as anchors in CRI-MAP for assignment of RAD-Seq-derived SNPs to linkage groups, and their corresponding Atlantic salmon linkage groups/chromosomes**

| Atlantic salmon linkage group | Atlantic salmon chromosome | Marker name | Marker type≠ | Marker information§ |
|---|---|---|---|---|
| 1 | 2 | Omy11/1INRA | Microsatellite | 1 |
| | | OmyFGT8/1TUF | Microsatellite | 2 |
| | | Oneμ18 | Microsatellite | U56718.1 |
| | | Ssa202 | Microsatellite | U43695.1 |
| | | Ssa406UOS | Microsatellite | AJ402723.1 |
| | | Ssa-A14/1 | Minisatellite | Unpublished |
| | | Ssa-A15/1 | Minisatellite | Unpublished |
| | | Str4/1INRA | Microsatellite | 3 |
| 2 | 10 | Ogo8 | Microsatellite | AF009780 |
| | | Oneμ5 | Microsatellite | U56704 |
| | | Ssa-A13 | Minisatellite | Unpublished |
| | | Str-A3/2 | Minisatellite | 4 |
| 3 | 14 | Ssa0014ECIG | SNP | 119096977 |
| | | Ssa0033ECIG | SNP | 119096998 |
| | | Ssa0169bECIG | SNP | 119097160 |
| 4 | 6 | Omy27/1INRA | Microsatellite | 1 |
| | | OmyFGT1TUF | Microsatellite | 2 |
| | | OmyRGT30/2TUF | Microsatellite | 2 |
| | | Ssa171 | Microsatellite | U43693.1 |
| | | Ssa-A12 | Minisatellite | Unpublished |
| 5 | 13 | Sfo23 | Microsatellite | 5 |
| | | Ssa420UOS | Microsatellite | AJ402737.1 |
| | | SSsp2201 | Microsatellite | AY081807.1 |
| | | Str-A5 | Minisatellite | 6 |
| | | Str-A8/1 | Minisatellite | Unpublished |
| 6 | 12 | Omy11/2INRA | Microsatellite | 1 |
| | | Omy21INRA | Microsatellite | 1 |
| | | Omy27DU | Microsatellite | 7 |
| | | OmyFGT25TUF | Microsatellite | 2 |
| | | OmyRGT35TUF | Microsatellite | AB087604.1 |
| | | SSsp2210 | Microsatellite | AY081808.1 |
| 7 | 24 | Ocl9 | Microsatellite | AF028698 |
| | | SSsp2215 | Microsatellite | AY081810.1 |
| | | SSsp2216 | Microsatellite | AY081811.1 |
| 8 | 15 | Omy27/2INRA | Microsatellite | 1 |
| | | Omy301UoG | Microsatellite | 8 |
| | | Oneμ9 | Microsatellite | U56709.1 |
| | | Ssa197 | Microsatellite | U43694.1 |
| | | Ssa401UOS | Microsatellite | AJ402718.1 |
| | | Ssa-A60 | Minisatellite | 6 |
| | | Str-A3/1 | Minisatellite | 4 |
| | | Str-A8/2 | Minisatellite | Unpublished |
| | | Str-A8/3 | Minisatellite | Unpublished |
| 9 | 11 | Ssa132 | Microsatellite | U58901.1 |
| | | Ssa408UOS | Microsatellite | AJ402725.1 |
| | | Ssa413UOS | Microsatellite | AJ402730.1 |
| | | SSspG7 | Microsatellite | AY081813.2 |

| | | | | |
|---|---|---|---|---|
| 10 | 9 | MST541INRA | Microsatellite | AB001072 |
| | | Ogo2/2 | Microsatellite | AF009794 |
| | | OmyFGT21TUF | Microsatellite | 2 |
| | | OmyRGT30/1TUF | Microsatellite | 2 |
| | | Oneμ7 | Microsatellite | U56707.1 |
| | | Ssa412UOS | Microsatellite | AJ402729.1 |
| | | Ssa45/2micUOS | Microsatellite | SRX000001 |
| | | Ssa-A33 | Minisatellite | 6 |
| | | Ssa-A34/2 | Minisatellite | 6 |
| | | Str85INRA | Microsatellite | AB001059 |
| 11 | 3 | Oki2 | Microsatellite | AF055428 |
| | | OmyRGT32TUF | Microsatellite | AB087602.1 |
| | | Ssa417UOS | Microsatellite | AJ402734.1 |
| 12 | 5 | Ocl2 | Microsatellite | AF028699 |
| | | Omy272/2UoG | Microsatellite | 8 |
| | | OmyFGT8/2TUF | Microsatellite | 2 |
| | | Ssa-A14/2 | Minisatellite | Unpublished |
| | | Ssa-A15/2 | Minisatellite | Unpublished |
| | | Str15INRA | Microsatellite | AB001058 |
| | | Str4/2INRA | Microsatellite | 3 |
| | | Str-A9/1 | Minisatellite | 6 |
| 13 | 19 | MC4R | SNP | Unpublished |
| | | Ssa289 | Microsatellite | 9 |
| | | Ssa407UOS | Microsatellite | AJ402724.1 |
| | | Ssa422UOS | Microsatellite | AJ402739.1 |
| 14 | 21 | Str-A22/1 | Minisatellite | 6 |
| | | Str-A22/2/1 | Minisatellite | 6 |
| 15 | 27 | Ssa0122aECIG | SNP | 119097105 |
| 16 | 18 | OmyRGT55TUF | Microsatellite | AB031201.1 |
| | | Ssa416UOS | Microsatellite | AJ402733.1 |
| | | Str-A12/1 | Minisatellite | Unpublished |
| | | Str-A9/2 | Minisatellite | 6 |
| 17 | 1 | Ogo3 | Microsatellite | AF009795 |
| | | OmyRGT34TUF | Microsatellite | AB031199.1 |
| | | Ssa14 | Microsatellite | 10 |
| | | Ssa410UOS | Microsatellite | AJ402727.1 |
| | | Str-A12/2 | Minisatellite | Unpublished |
| | | Str-A22/2/2 | Minisatellite | 6 |
| 18 | 23 | OmyFGT16TUF | Microsatellite | 2 |
| | | Ssa85 | Microsatellite | U43692.1 |
| | | SSsp1605 | Microsatellite | AY081812.1 |
| 19 | 8 | Ssa0136ECIG | SNP | 119097123 |
| | | Ssa0158ECIG | SNP | 119097148 |
| 20 | 25 | MEP-2* | Allozyme | 11 |
| 21 | 26 | IDDH-2* | Allozyme | 11 |
| | | OmyRGT44TUF | Microsatellite | AB087611.1 |
| | | Ssa-A45/2/1 | Minisatellite | 6 |
| 22 | 17 | Ssa12 | Microsatellite | U58900 |
| | | Ssa402/2UOS | Microsatellite | AJ402719 |
| | | Ssa404UOS | Microsatellite | AJ402721.1 |
| 23 | 16 | Ssa402/1UOS | Microsatellite | AJ402719 |
| | | Ssa403UOS | Microsatellite | AJ402720.1 |

| | | | | |
|---|---|---|---|---|
| 24 | 7 | Omy14INRA | Microsatellite | 1 |
| | | Ssa418/1UOS | Microsatellite | AJ402735 |
| | | Ssa-A34/1 | Minisatellite | 6 |
| 25 | 20 | Ocl1/1 | Microsatellite | AF028694 |
| | | Oki10 | Microsatellite | AF055435 |
| | | Omy23INRA | Microsatellite | 1 |
| | | OmyFGT14TUF | Microsatellite | 2 |
| | | OmyFGT34TUF | Microsatellite | 2 |
| | | Ssa421UOS | Microsatellite | AJ402738.1 |
| | | Ssa-A10 | Minisatellite | 6 |
| 28 | 4 | Ssa405UOS | Microsatellite | AJ402722.1 |
| 30 | 29 | Ogo4 | Microsatellite | AF009796 |
| | | Ssa-A11 | Minisatellite | 6 |
| 31 | 28 | AAT-4* | Allozyme | 11 |
| | | SSA224 | Microsatellite | AF019168.1 |
| 32 | 22 | Ssa419UOS | Microsatellite | AJ402736.1 |
| | | Ssa-A45/1 | Minisatellite | 6 |

≠SNPs were sourced from Moen et al. 2008.

§GenBank accession number provided if available. Otherwise a numbered reference is provided where applicable. Numbers correspond to the following references:

1) Gharbi et al., 2006. Genetics. 172:2405-2419.
2) Sakamoto, 1996. PhD thesis.
3) Krieg and Guyomard, Unpublished.
4) Prodöhl et al., 1994. Heredity. 73: 556-566.
5) Angers et al., 1995. Journal of Fish Biology. 47:177-185.
6) Taggart et al., 1995. Animal Genetics. 26:13-20.
7) Perry et al., 2001. Cytotechnology. 37:143-151.
8) Jackson et al., 1998. Heredity. 80:143-151.
9) O'Reilly et al., 1996. Canadian Journal of Fisheries and Aquatic Science. 53:2291-2298.
10) McConnell et al., 1995. Canadian Journal of Fisheries and Aquatic Science. 52:1863-1872.
11) Wilson et al., 1995. Heredity. 75: 578-588.

**Table A3: Map length (cM) for each mapping parent and the comparison between the sexes**

| Atlantic salmon linkage group | Atlantic salmon chromosome | Br5 Female (cM) | Br5 Male (cM) | Ratio M:F | Br6 Female (cM) | Br6 Male (cM) | Ratio M:F |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 229 | 85 | 1:2.7 | 85 | 109 | 1.3:1 |
| 2 | 10 | 196 | 44 | 1:4.5 | 143 | 52 | 1:2.8 |
| 3 | 14 | 64 | 65 | 1:1 | 93 | 39 | 1:2.4 |
| 4 | 6 | 132 | 115 | 1:1.1 | 60 | 113 | 1.9:1 |
| 5 | 13 | 107 | 114 | 1.1:1 | 130 | 82 | 1:1.6 |
| 6 | 12 | 126 | 91 | 1:1.4 | 71 | 43 | 1:1.7 |
| 7 | 24 | 38 | 100 | 2.6:1 | 73 | 30 | 1:2.4 |
| 8 | 15 | 85 | 79 | 1:1.1 | 79 | 64 | 1:1.2 |
| 9 | 11 | 128 | 85 | 1:1.5 | 140 | 14 | 1:10 |
| 10 | 9 | 241 | 136 | 1:1.8 | 145 | 40 | 1:3.6 |
| 11 | 3 | 131 | 130 | 1:1 | 136 | 73 | 1:1.9 |
| 12 | 5 | 86 | 121 | 1.4:1 | 71 | 69 | 1:1 |
| 13 | 19 | 66 | 32 | 1:2.1 | 81 | 42 | 1:1.9 |
| 14 | 21 | 67 | 61 | 1:1.1 | 61 | 17 | 1:3.6 |
| 15 | 27 | 71 | 52 | 1:1.4 | 33 | 26 | 1:1.3 |
| 16 | 18 | 55 | 90 | 1.6:1 | 135 | 68 | 1:2 |
| 17 | 1 | 104 | 187 | 1.8:1 | 112 | 80 | 1:1.4 |
| 18 | 23 | 120 | 63 | 1:1.9 | 47 | 37 | 1:1.3 |
| 19 | 8 | 3 | 16 | 5.3:1 | 0 | 3 | NA |
| 20 | 25 | 13 | 22 | 1.7:1 | 26 | 16 | 1:1.6 |
| 21 | 26 | 59 | 17 | 1:3.5 | 45 | 23 | 1:2 |
| 22 | 17 | 65 | 31 | 1:2.1 | 108 | 44 | 1:2.5 |
| 23 | 16 | 114 | 55 | 1:2.1 | 56 | 54 | 1:1 |
| 24 | 7 | 93 | 85 | 1:1.1 | 67 | 80 | 1.2:1 |
| 25 | 20 | 96 | 50 | 1:1.9 | 89 | 75 | 1:1.2 |
| 28 | 4 | 69 | 112 | 1.6:1 | 77 | 90 | 1.2:1 |
| 30 | 29 | 97 | 44 | 1:2.2 | 16 | 17 | 1.1:1 |
| 31 | 28 | 85 | 22 | 1:3.9 | 78 | 0 | NA |
| 32 | 22 | 67 | 67 | 1:1 | 100 | 28 | 1:3.6 |
| **TOTAL** | - | 2,807 | 2,171 | 1:1.3 | 2,357 | 1,428 | 1:1.7 |

**Table A4: Homeologous Atlantic salmon linkage groups with the stickleback, rainbow trout and proto-Actinopterygian linkage groups which they have in common (Danzmann et al. 2008; Phillips et al. 2009)**

| Atlantic salmon linkage groups | Stickleback linkage group | Rainbow trout linkage groups | Proto-Actinopterygian linkage groups |
|---|---|---|---|
| 4/11 | 11 | 2/9 | E |
| 1/12/15 | 20 | 27/31/16 | B |
| 9/21 | 2 | 10/18 | J |
| 3/11 | 3 | 23/13 | M |
| 3/15 | 10 | 3/16 | B |
| 2/23 | 19 | 6/27 | M,J/K |
| 7/25 | 13 | 10/11,19 | I |
| 6/32 | 17 | 29/12 | L |
| 16/17 | 6 | 6/30 | D |
| 17/31 | 5 | 30/17,22 | D/E |
| 14/20 | 16 | 5/31 | C |
| 13/30 | 21 | 19/7 | M |

**Table A5: Sparse SNP panel used in the initial identification of PD resistance QTL (sire- and dam-based linkage analyses)**

| Chromosome | SNP ID from Moen et al. 2008 |
|---|---|
| 1 | Ssa0064ECIG |
| | Ssa0096ECIG |
| 2 | ESTNV_27268_490 |
| | Ssa0114ECIG |
| 3 | Ssa0137ECIG |
| 4 | Ssa0023ECIG |
| | Ssa0192ECIG |
| 5 | Ssa0181ECIG |
| | Ssa0252ECIG |
| 6 | Ssa0065ECIG |
| | Ssa0175bECIG |
| | Ssa0235ECIG |
| 7 | Ssa0167ECIG |
| | Ssa0171aECIG |
| | Ssa0138aECIG |
| 8 | Ssa0136ECIG |
| | Ssa0158ECIG |
| 9 | Ssa0093ECIG |
| | Ssa0126aECIG |
| | Ssa0180ECIG |
| 10 | Ssa0145ECIG |
| | Ssa0239ECIG |
| | Ssa0051ECIG |
| 11 | Ssa0081ECIG |
| | Ssa0203bECIG |
| 12 | Ssa0204ECIG |
| | Ssa0212ECIG |
| | Ssa0148ECIG |
| 13 | Ssa0261ECIG |
| | Ssa0200ECIG |
| | Ssa0211ECIG |
| 14 | Ssa0014ECIG |
| 15 | Ssa0134aECIG |
| | Ssa0130ECIG |
| | Ssa0048cECIG |
| 16 | Ssa0196bECIG |
| | Ssa0045ECIG |
| | Ssa0004ECIG |
| 17 | Ssa0005ECIG |
| | Ssa0085ECIG |
| 18 | Ssa0207ECIG |
| | Ssa0176ECIG |
| | Ssa0213ECIG |
| 19 | Ssa0142ECIG |
| | Ssa0190ECIG |
| | Ssa0076ECIG |
| 20 | Ssa0047bECIG |
| | Ssa0168ECIG |

| | |
|---|---|
| 21 | Ssa0257ECIG<br>Ssa0214ECIG<br>Ssa0069ECIG |
| 22 | Ssa0016ECIG<br>Ssa0111bECIG |
| 23 | Ssa0080ECIG<br>Ssa0245ECIG |
| 24 | Ssa0012ECIG<br>Ssa0194ECIG |
| 25 | Ssa0117bECIG<br>Ssa0109ECIG |
| 26 | RAD010201*<br>Ssa0139ECIG |
| 27 | Ssa0122aECIG<br>Ssa0248ECIG<br>Ssa0255ECIG |
| 28 | Ssa0067ECIG<br>Ssa0087ECIG |
| 29 | Ssa0201ECIG<br>Ssa0172ECIG |

*Not from Moen et al. 2008, identified in a separate study by our group.

**Table A6: Dense SNP panel used to position PD resistance QTL on significant chromosomes (dam-linkage analysis only)**

| Chromosome | SNP ID | Map position cM (Lep-MAP) | Allele frequency | PVE (%) |
|---|---|---|---|---|
| 3 | ESTNV_35628_2366* | 0 | p=0.74 q=0.26 | 0.0042 |
| 3 | consensus60488_36** | 15 | p=0.83 q=0.17 | 0.9370 |
| 3 | consensus6869_92** | 47 | p=0.53 q=0.47 | 0.6470 |
| 3 | consensus118333_53** | 53 | p=0.81 q=0.18 | 7.3100 |
| 3 | ESTNV_29224_109* | 55 | p=0.69 q=0.31 | 1.2700 |
| 3 | consensus46559_56** | 67 | p=0.57 q=0.42 | 1.7000 |
| 3 | consensus110127_55** | 82 | p=0.91 q=0.09 | 29.8000 |
| 3 | consensus36092_53** | 86 | p=0.60 q=0.40 | 0.0605 |
| 3 | ESTNV_32268_419* | 100 | p=0.68 q=0.32 | 1.1600 |
| 3 | Ssa0137ECIG*** | 128 | p=0.56 q=0.45 | 0.1540 |
| 3 | consensus108417_63** | 133 | p=0.79 q=0.21 | 2.3200 |
| 3 | ESTNV_31140_226* | 135 | p=0.58 q=0.41 | 0.0504 |
| 4 | consensus46277_95** | 0 | p=0.80 q=0.20 | 0.2710 |
| 4 | Ssa0192ECIG*** | 31 | p=0.71 q=0.29 | 0.0569 |
| 4 | consensus138050_52** | 47 | p=0.88 q=0.12 | 8.4700 |
| 4 | consensus135080_72** | 56 | p=0.90 q=0.10 | 0.0352 |
| 4 | Ssa0023ECIG*** | 75 | p=0.77 q=0.23 | 3.1800 |
| 4 | consensus55110_37** | 77 | p=0.89 q=0.11 | 0.6590 |
| 7 | consensus65272_39** | 0 | p=0.20 q=0.80 | 0.1710 |
| 7 | Ssa0138aECIG*** | 8 | p=0.62 q=0.38 | 0.0002 |
| 7 | consensus135256_71** | 25 | p=0.70 q=0.30 | 1.4900 |
| 7 | Ssa0171aECIG*** | 27 | p=0.75 | 1.3900 |

| | | | | |
|---|---|---|---|---|
| | | | q=0.25 | |
| 7 | consensus19661_62** | 30 | p=0.68<br>q=0.32 | 0.0050 |
| 7 | consensus26302_48** | 53 | p=0.53<br>q=0.47 | 0.6280 |
| 7 | Ssa0167ECIG*** | 65 | p=0.53<br>q=0.47 | 0.0101 |
| 23 | Ssa0245ECIG*** | 0 | p=0.63<br>q=0.36 | 1.3400 |
| 23 | ESTNV_31110_261* | 18 | p=0.76<br>q=0.24 | 0.1230 |
| 23 | GCR_cBin6663_Ctg1_796* | 30 | p=0.88<br>q=0.12 | 0.0003 |
| 23 | consensus47083_51** | 33 | p=0.84<br>q=0.16 | 3.4200 |
| 23 | ESTNV_35679_632* | 41 | p=0.83<br>q=0.17 | 0.1850 |
| 23 | consensus39563_62** | 48 | p=0.69<br>q=0.31 | 1.2600 |
| 23 | ESTNV_28034_927* | 50 | p=0.90<br>q=0.10 | 0.0788 |
| 23 | consensus80397_29** | 56 | p=0.94<br>q=0.06 | NA |
| 23 | Ssa0080ECIG*** | 59 | p=0.52<br>q=0.48 | 0.3250 |
| 23 | consensus6840_34** | 61 | p=0.87<br>q=0.13 | 8.4900 |
| 23 | ESTV_14834_524* | 71 | p=0.60<br>q=0.40 | 0.3220 |

*Lien et al 2011
**Gonen et al 2014
***Moen et al 2008

**Table A7: Top 100 genes identified as differentially regulated between resistant and susceptible individuals upon challenge with IPNV.**

| Probe Name | P-value | B2GO BLASTX Hit | Probe Sequence |
|---|---|---|---|
| Omy#S15276461 | 6.31E-12 | CASC4 | TGGACTATTCTGAAGACACGATTACAGATTATTCCATGAATCAGTTTGACCTAAAACCAT |
| Ssa#TC77118 | 1.78E-09 | Sorbitol dehydrogenase | GCACATCCTGTCATGCAGACAATCAATCTCCTGTCCATACAATAAATATAAAACAAATTG |
| Ssa#DW561608 | 2.43E-09 | Probable COP9 signalosome complex subunit 7 | AAGATCTCTCTCTCCAAAACCCCCTTTCCTTTGTCAAACTCCCTTGGTTGAATTTTTTGT |
| Ssa#STIR01613 | 1.80E-08 | nedd8 activating enzyme e1 subunit 1 | CTTTTCAGCTGTAGAAAAATCTATTTTCCTCAATGACATGCACTTAACCCCTCTGTAAAA |
| Ssa#STIR18537 | 5.48E-08 | hypothetical protein BRAFLDRAFT_263599 | CAACTCTCTGTTTGCAAAGATGAACTTGTCTTTTTCCACTACACTGTAAACATTTCCAAA |
| Ssa#S35541768 | 2.24E-07 | Low choriolytic enzyme precursor | AATCCTGACTCAGAACATTTTGCCTGGAAAGGAATCCAATTTCAATAAGGTCAACACGAT |
| Omy#TC162230 | 3.02E-07 | PREDICTED: similar to DTW domain containing 1 | ATGTACTTCTACTGTTACCTACACACACTGATCAACAAGGCTAAGACAACCGCTGGGAGA |
| Ssa#S35575706 | 4.06E-07 | High choriolytic enzyme 2 | ATGGCAAAGCAAAGTACTGTGGTTTACTGAAACATTATCCCAATAAACATGTGTAGCAAC |
| Ssa#STIR02869 | 7.07E-07 | membrane-spanning 4-subfamilymember | ACCCTCTCACTAACCAGCAAAAACATGTCTTAAACTTCCATCAACTAAAATATTGTAGGC |
| Ssa#STIR19576 | 2.61E-06 | ---NA--- | TGGGGACAATAACAGACACAACAGTATCATAATACAATGAAGACAGACTGTATTTCAGAT |
| Ssa#STIR21585 | 3.77E-06 | processing of precursorribonuclease p mrp subunit | TGTAGGCACCACAGGAATCCTAGTGCAGGAGATGAAGCACGTCTTCAAAATCATCACAAA |
| Ssa#STIR05057 | 8.98E-06 | 39s ribosomal proteinmitochondrial precursor | TTCGGATCAACACAATTGAAGTTGCTCCCAGATTCACATGATGCTATATGTTTTGAATTG |
| Ssa#STIR23471 | 8.98E-06 | ---NA--- | CCTGTGCTGAACCAATAAATGTGTGGGAAGAAACATGATTTGGTCAGTTTGTACCATTAT |
| Ssa#STIR13689 | 9.82E-06 | ---NA--- | CTCAGAACTACTGTTGTGCTTACTGTTTCTCTGATATATAGTGTAATATGTGTGATTACC |
| Ssa#S35694408 | 2.51E-05 | FAM60A | GGCAGACTACCCCCAAAACCATTCCCTATTAGTTCACTGTGCATTTAGATGGCTAAATGA |
| Ssa#STIR02298 | 2.62E-05 | c-c motif chemokine 13 precursor | ACTGATTTTCACCACAAAGAAAGGGAAGACATTTTGTGTTGGCCCTTCTGAAGCCTGGGT |
| Ssa#STIR04723 | 2.76E-05 | apaf1 interacting protein///APAF1-interacting protein homolog | CAAAAGAAAATGACATCTGTTTGTGATGCCGACAAGGAAAATGGTTCAGAGTCGACGGAG |
| Ssa#CX357114 | 3.02E-05 | Wnt9a | ATAGTGTACCTTTTCAACTTGCCCGGACCGACAACAGCTTATTTCGGACTGACAGGTAAT |
| Ssa#S35663871 | 3.45E-05 | unnamed protein product | ACAATATTGGTGGTGATGTATTTGGGAAACCCTGCAGTAGTGTATGAAAGACCATGCAAG |
| Ssa#STIR26219 | 4.31E-05 | 39s ribosomal proteinmitochondrial precursor | TACTCGGTTATTGTGTAACCACAGACAGTGCTTCGGATCAACACAATTGAAGTTGCTCCC |
| Ssa#S35585444 | 4.38E-05 | PREDICTED: hypothetical protein | CCTCTGTAGTGCACTACATTTGACCATGGGCCTTATGGGCTATTCTATTGAATACAAAAT |
| Ssa#S35583215 | 4.97E-05 | Signal peptide peptidase-like 2A | ATGCAAGGGAAATACTATAGCTCGTCTATCAATCAATAAGGATTTATACTTGGGGAGGGG |
| Ssa#DW470017 | 5.28E-05 | unnamed protein product | ATATAGAGAAAACCTATACATCTAAGTTGGAGAACGCCCGTCAGTCTGCTGACAGGAGCA |
| Ssa#STIR09119 | 6.64E-05 | c-c motif chemokine 13 precursor | AGATTGGCATAGAGTAGGTTCAGTGTAATGTGCTGTGTGTAATATAGTGTAATTTAGGAT |

| | | | |
|---|---|---|---|
| Ssa#STIR13904 | 8.11E-05 | membrane-spanning 4-subfamilymember | GTTGGTAGAACCCTTATATATGATCTATGGAGGGTTTGTAAATGGCTTTCAATTCACCAA |
| Ssa#S35507110 | 8.18E-05 | Membrane-spanning 4-domains subfamily A member 4A | CATTACTTTGCTGAGTTGATCCTGATTCAAACTGCTCTCATCGCTATCTCTATCACGCTG |
| Ssa#S35498807 | 9.00E-05 | FAM60A | TTCTTACTGGAAAAGGCAAAAGGTGTGCTGTGGGATTGTCTACAAAGGGCGTTTTGGTGA |
| Ssa#DY702855 | 1.06E-04 | myotubularin related protein 10-like | GTGAAGCCCATGAAACCAACCTTTAAATCCTACCTTCCGCCTTTACAGACTGATCTGAAA |
| Ssa#STIR15391 | 1.20E-04 | 39s ribosomal proteinmitochondrial precursor | TTCGGATCAACACAATTGAAGTTGCTCCCAGATTCACATGATGCTATATGTTTTGAATTG |
| Ssa#S35507036 | 1.26E-04 | ---NA--- | AACAAAGACTTTATCGATTCCCTCGGTCTCAACCATGAGCAGAACATGGCTAAGATGCGT |
| Ssa#STIR26253 | 1.27E-04 | cancer susceptibility candidate 4 | GGCAGGCATGGCATATTGCTTGAAATTGTCAACATTAATTTGACTTGTCTTTACCTAAAA |
| Ssa#S18888710 | 1.27E-04 | unnamed protein product | TCTGTGTCCTTTCGCCCTCAATCACCATCAGACAGAGTGTTTTCTGAACACAGACCATTT |
| Ssa#S35596794 | 1.44E-04 | PREDICTED: mucin 2, oligomeric mucus/gel-forming | GTGTTTTCTTGTTACGTTTTAATTAAATCTGTCATGCCAGTGTATTTCTACAATCGGCAG |
| Ssa#S35704345 | 1.45E-04 | Transmembrane protein 9B | ATGTGATCATATAGCTCTATGTATTCAGGAGGTGATCATAAAGGGTCTCTACATCAGTAG |
| Ssa#TC65634 | 1.52E-04 | PREDICTED: si:ch211-260p9.1 | TTATGGGTTCCTGCTACAAAACTAAGAAGTTTCTCCTGTCCCTGGCTGAAAACAAGCTGG |
| Ssa#DY693266_S | 1.63E-04 | Bloodthirsty | GTCATCACTCCTATAGGTGTCACTGACTAACTGTTTTATATCAAACATTGAAATACTGGT |
| Ssa#EG766472 | 2.48E-04 | THO complex 1 | TTAATCTCGACAACATCACAGTGTTCAACAAAAATGAGCTAGAGAGCACTCTTGGCCAGA |
| Ssa#S31979448 | 2.57E-04 | hypothetical protein LOC447898 | TATGCAGTGTGGAACATCATAGTGATCTATGCCCTGGCAGGAAAACATCTCACTACTCTG |
| Ssa#S30292522 | 2.68E-04 | Profilin-2 | TGTTAACAGGACATTGTGACAAATCTGTACTTATAAAGACCTGTACCATTCTAGCCAATG |
| Ssa#STIR22493 | 2.69E-04 | zgc:112084 protein | ACATCTGGAAGGGTAGTCCCAGGTCAGATACAGCAGGAGCTGAGTAGCATCACCCTGGCT |
| Omy#S34425539 | 2.75E-04 | glutaminyl-peptide cyclotransferase-like | ACACTCATAGACACATTGATCTGAGAGCAGAACTACCATCAAGTGACCCTGCTGGATGTT |
| Ssa#KSS2134 | 2.80E-04 | PREDICTED: similar to Y51B11A.1 | GGCCTGTTTCTCATCTTGTTCGGTTTCATTTTCCAATTTGTTACACTTTGTATTCGAGGA |
| Ssa#STIR26158 | 2.82E-04 | pepsinogen a1 precursor | TTGTTGTTCTGCCTGTCAAAGATGGCATAGTACTGCCTGATGAAGACATCTCCAAGGATC |
| Omy#TC147965 | 2.95E-04 | hypothetical protein LOC555262 | CTTGACTAATATGACCTTTGCTGTGGGTATTGCCCTGTTTGGTGCCCTTGGCTACTATAT |
| Ssa#KSS2024 | 3.25E-04 | Trafficking protein particle complex subunit 2 | TCCCAACCACTGTATAAAAGCGATTGTAAGCATGGAACATTATTGCATAATGAATGGATG |
| Ssa#TC89862 | 3.29E-04 | arginine N-methyltransferase 3 | CTGTGGAGTTCACTGCATTTTAATGCCATACCTGTTAATAAAGACATTCCCTCATACAGC |
| Ssa#S31979353 | 3.56E-04 | Coiled-coil domain-containing protein 22 | GAAAAAGCAATTGAAAGAGGAAAATGACTACCACAGGGACTTCCTTCCCCCTGTAACTGC |
| Ssa#STIR06520 | 3.62E-04 | dna replication complex gins protein psf2 | GTTAGGTCAGGGGACACCTCATAACCCTAATTTCTACATGTATGCCCCTCTGGATACATA |
| Ssa#STIR07318 | 3.69E-04 | phosphoglycerate mutase 1 | CTGTGCTGAACCAATAAATGTGTGGGAAGAAACATGATTTGGTCATTTTGTACCATTATT |
| Ssa#STIR40246 | 4.05E-04 | ubiquitin specific peptidase 8 | TTTCGCATATATGATGAAGTTCCATGTTTTTGCTTATCCATTCAACACACATCCATGTAC |
| Ssa#S31982592 | 4.58E-04 | NK2 transcription factor related 3 | CATGTGAACGAATAGGTGCCGATGATAATAATGCTTCCTAGCCCTGTTATTTCTGCCTCC |

| | | | |
|---|---|---|---|
| Ssa#STIR40558 | 4.61E-04 | parvalbumin like 1 | TTGTAATCTTCTCTCTGAAGCAAGAGCCCAAAGAAAGGTTTCCGTTCCTCTCTCTTTCTC |
| Ssa#STIR09395 | 4.72E-04 | vesicle-associated membrane protein 8 | CCCTAAACCATCCCAGTAGTGTAACAATGAGGTTAATTTGAGGCTATTTATTAATGTTGT |
| Ssa#S35479868 | 4.77E-04 | Tripartite motif-containing protein 16 | TAGATGGGCAAGCACTCATGGCTTCATATGTATTTGTGATTGATGTCAGATTAATCAAAT |
| Ssa#S32009297 | 4.94E-04 | Cbp/p300-interacting transactivator 3a | TAAAACGTAGCAGTAGCAAAGTCGCTGAACTTTAATACGAGTAGATACGGACAGGAGCCG |
| Ssa#STIR00102_3 | 5.15E-04 | squalene epoxidase | TAAGAAATGCGACGCAGTTTTGGAGTATGCTCACAAAGAGATAATCCTGGCGGCTGTAGT |
| Ssa#STIR00032_3 | 5.45E-04 | 3-hydroxy-3-methylglutaryl-coenzyme a reductase | GGGCAATATACTGTAATCTGTGCAGGATTGGGCAGAATGCAACCCAAATCTTCTGACTAA |
| Ssa#S48426617 | 5.62E-04 | PREDICTED: similar to PYRIN-containing APAF1-like protein 7 | CCGAATAAATGTTGTTGACTGTGATCATGACAGATATTCGTATTCACACCACTAAGTCAA |
| Ssa#STIR22464 | 5.69E-04 | ---NA--- | GGACACATTGCCTAGTAATGCAATGTAAACAGCTCCATGGTCCCATTGAACTATGCATTT |
| Ssa#S18871413 | 5.81E-04 | unnamed protein product | CCTGGTGGCGGCACAATATTTCCCAACTTTGCATAAAAATGCTGTAAGATAATGTAAGAA |
| Ssa#DW574957 | 5.82E-04 | N-acetylglucosamine-6-sulfatase precursor | AGAAAACCAAAGCCTTGATTGGAGATGCTGGAGCCACCTTTTCAAATGCTTTTACGGCGA |
| Ssa#TC96127 | 6.05E-04 | Arsenite methyltransferase | AGTGATGTCTATAGTAGCAGTAGGCTCTCCGATGAGATCAAGAATCACAAAGTCCTGTGG |
| Ssa#S32007191 | 6.64E-04 | Mitochondrial import inner membrane translocase subunit Tim22 | GGACAATTCCATCCTGGTCTTCATCAATCAATCACATTTATTTATAAAGCCCTTCTTACA |
| Omy#CA378710 | 6.64E-04 | potassium channel Kv1.1b | CCAAATATTCTCACCGATCCTTTCTTCATAGTGGAGACCCTGTGTATAATCTGGGTCTCC |
| Ssa#STIR18553 | 6.76E-04 | ---NA--- | CTAGCAAAGCCTAAAATTGACATTTTGGTCAACCAGCCACTGTGGCAGCCATGCAGATTT |
| Ssa#STIR11144 | 6.97E-04 | ce051_danreame: full=upf0600 protein c5orf51 homolog | CCAGTGACAGGAACTTTGGTATTTTGATAGTATGCATTTAAATGCAGATTCTGTGTTCTT |
| Ssa#DY718082 | 7.00E-04 | PABPN1 protein | GACAAATTCTCTGGCCATCCCAAGGGTTTTGCATATATTGAGTTCCATGACAGGGACTCT |
| Ssa#EG940362_S | 7.04E-04 | Gastrula zinc finger protein XLCGF57.1 | CAAGACTGTTAAGAAAGCAAAGCAACTCTGGATCATTCATCTTGTAGGTGTGTCGTTCCA |
| Ssa#S31979306 | 7.97E-04 | Tenascin precursor | CTCCTGATACAAAATGGTTGTGTGCAGTATAAATACTGTCAAAGCCAAATTGATGTGAAA |
| Ssa#STIR17054 | 8.47E-04 | hematopoietic sh2 domain containing | GTTAGACAATGTCCAATGCTTTTACAGTGCTGTGTATGGTCAACTATTTCATATCAAATG |
| Ssa#S35585570 | 8.69E-04 | Stress-associated endoplasmic reticulum protein 1 | TTTTGCTTTTGACGTGTTTTCTGTTTGGTAGAATGACCAAATGCCTTGGTCAGACTGGAA |
| Ssa#STIR17556 | 8.80E-04 | protein | GTTCTCGCCCACTGTTTCCTATCATAGGTTTTTCGGTTTTATTATGGATAACATGTTTTA |
| Ssa#S37959531 | 9.27E-04 | T cell receptor alpha | AGGGGAACCACACTGTCAATACAATCTAGAGAGAAACATGAGCCATCCTACTACACAAGC |
| Ssa#STIR17772 | 9.60E-04 | wd repeat-containing protein 82 | CAACTCCAATCAAGTATGCATTTTAACTGTGGAGGCATGTTTTATTGGGAATTCAATTTG |
| Ssa#S35530688 | 9.79E-04 | F-box/WD repeat-containing protein 11 | ATGGGACAACCGTGTGCTAGTGACCTATTTGTTTTTCTAACTTTTACTTTTTATTTGGGA |
| Ssa#DY728487 | 9.95E-04 | Solute carrier family 13 member 3 | TAATGAGGAACGCTAAACAATGAGGAACCCGTGTTGAAGGTACCACATATCAGCAGTAAT |

| | | | |
|---|---|---|---|
| Ssa#DY726999 | 0.001047393 | Gata6 protein | AAATGTTCTTTGTTGTTTGAAATGAAGAAACCTGTACATAGTGCTATGGATTCCCCCACT |
| Ssa#EG836985 | 0.00104871 | hematopoietic SH2 domain containing | CCAAATAACCCCGCACACATGAACAGTAGTACTCATCTGAAAGTAGCTGAGGACATTCCA |
| Ssa#TC102876 | 0.001069035 | lipase maturation factor 2 | TTTGCTATCAGTGAGGTTCCCTACTCGTCCATGGAACAGGTGTACAGCAATAAGATCCTA |
| Ssa#STIR39598 | 0.001097823 | hypothetical protein LOC100136074 | TCTATATCTTTGGCTGGATGACCAGAATGAACAGGATTCTGATGGGTTCTCTGGTGAGGA |
| Ssa#STIR17250 | 0.001109643 | mannose-6-phosphate protein p76 | ATATGTATACTGTTTACTGCACTGCTGTTTGAGATGGTATAACTCTGTTTGAAATGAGAC |
| Ssa#S30281203 | 0.001110403 | 39S ribosomal protein L28, mitochondrial precursor | AGCTGTTTACCAGAGAGTTGTACAGTGAGATCCTCAACCACAAGTTCACCATCACCGTAA |
| Ssa#S31993322 | 0.001117369 | Rho GTPase activating protein 25 | AAGTAAATGAGCTCCCACTGAACCAAGATGACCCAGGGAAATTCCTGTTTGAGATCATGC |
| Ssa#STIR19423 | 0.00112427 | ---NA--- | AAAGGAGAATGGTACACTCATCATACCCAACCGCGCTACTTGTTCATGTGGATGGTTTGT |
| Ssa#STIR02179 | 0.00112524 | telomeric repeat-binding factor 2 | AACATCAAGGATCGGTGGAGAACCATGAAAAAACTCAAGATGGTCTGACGCCACCAAACT |
| Ssa#S30283201 | 0.001127956 | Mitochondrial 28S ribosomal protein S34///mitochondrial ribosomal protein s34 | TTAGATAGATCATAGGCTGCGTTCTGTTTCTCAAAGTATTGAATTGGTGTAAGCATGACG |
| Ssa#S30241020 | 0.001176163 | Very long-chain acyl-CoA synthetase | TAAAGGTGAGACAGGACTGTTGGTGTCCAAGATCACAGACATCGCTCCTTTTGTTGGATA |
| Ssa#S35477759 | 0.001183382 | unnamed protein product | AGTTTTTACTCCCCCTCTGTCTCTTCTGATGACCTACTTTGCAATGTCATCATGTATATC |
| Omy#CA375780 | 0.001186436 | PREDICTED: similar to EBF1 | ATTGCTAACGGACTGGCAGATCAGTCTTTTGTGGACTCTAGCAAGTACTCCTCCTCCAGC |
| Ssa#S30293706 | 0.001189368 | RUN and FYVE domain containing 3 | AATTACTTAGGGGGATGTTTCGGGATTTTGGCAATGAGGCCATGTGTCTACTTCTCTGGA |
| Ssa#STIR27646 | 0.001207285 | Solute carrier family 35 member B1 | ACAATTACCCGAGGACAGTATGGTGAGGGGGAGAAGAAAGAGAAATTTGTTTATGCCACA |
| Ssa#STIR09088 | 0.001227192 | ---NA--- | TCAAATGTTGCTAAATGACCTTATGACTGGATGAATCTCTAGTCAAACACAACTATGGGT |
| Ssa#CK875715 | 0.001228724 | guanylate cyclase activating protein 2 | TGACAAAGATGGAAGTGGTTGCATTGACAAGACAGAGCTGCTGGAGATTGTAGAGTCCAT |
| Omy#S22244812 | 0.001229734 | ---NA--- | GAGTAAAATATCCAAGTGTATTTATTTTCTGGATTGGATATAAATGGTCTCCAGTGCTGC |
| Ssa#STIR18046 | 0.001231643 | calreticulin | GCATAGTGCATTTGTTCTCCCCTTAATTGTTTTTGTAGATATTTGTTCCCTTATTTGGGA |
| Ssa#STIR18088 | 0.001278006 | atg3 autophagy related 3 homolog | GGACATGAACATGAACTCTGGTGCTAGTGGAAACGGAGATGATGATGATGATGAAGACG A |
| Ssa#S30120550 | 0.001300319 | Coiled-coil domain-containing protein 49 | TATAACCAGATTTGCTCAGCAGACTGGGGCTCTAAAGAAAAAAGATGACCGGTTGGACTG |
| Ssa#STIR00027_3 | 0.001307439 | ---NA--- | TTGTTGTCTAGACATTACTATGACTTGATTGTGTATCCTGTGGGACCTTTTTAACTATGT |
| Ssa#S35688162 | 0.001357483 | Malate dehydrogenase, cytoplasmic | AATGCACAATAGCTCTACTGTACTCACACGTTATCCAACAATAAAAACAGGTACAACTGA |
| Ssa#STIR20895 | 0.001361684 | annexin a3 | CTTTTCAGGGTGACTTGGATCTCCTTTTGGATTCACCTTCCTCTTTGACATTAACGTCCA |

# Appendix B – R script to run the OneMap software package for linkage map construction

```
#!/usr/bin/Rscript

# R script to run OneMap
# run on each linkage group separately, based on SNP
assignment from CRI-MAP

# so first we take the input file from the command line:

args = commandArgs(TRUE)
inputfile <- args[1]
pdffile <- args[2]


# load OneMap
library(onemap,lib.loc='/nfs_netapp/vlsgonen/.R/onemap')

# read input file
print("Reading input file")
family <- read.outcross(file=inputfile)
print("-------------------------------------------------------
-------------------------------------------------------
")

# calculate twopoints
print("Calculating twopoint scores")

twopts <- rf.2pts(family)

print("-------------------------------------------------------
-------------------------------------------------------
")

# make the LGs

print("Making linkage groups")
mark.all <- make.seq(twopts, "all")
lgs <- group(mark.all, LOD=4)

print("These are the linkage groups")
lgs
print("-------------------------------------------------------
-------------------------------------------------------
")
```

```
# get my LG
print("Getting largest LG")

groups <- lgs$n.groups

sizes <- c()

for (i in 1:groups){
    linkage_group <- make.seq(lgs, i)
    number_of_markers <- length(linkage_group$seq.num)
    sizes <- append(sizes, number_of_markers)
}

wanted_LG <- which.max(sizes)
my_lg <- make.seq(lgs, wanted_LG)

my_lg

print("-------------------------------------------------------
-------------------------------------------------------
")

# make the map
print("Now to make the map by ordering the markers!")

my_lg.ord <- order.seq(my_lg, n.init=5, THRES=4,
draw.try=FALSE)

print("The order for the five anchor markers and possible
locations for the other markers are:")

my_lg.ord

print("-------------------------------------------------------
-------------------------------------------------------
")
print("Let's force make the order:")

my_lg.all <- make.seq(my_lg.ord, "force")

my_lg.all

print("-------------------------------------------------------
-------------------------------------------------------
")
```

```
# check for alternate orders

print("Run ripple to see if there is a better order")

ripple_my_lg <- ripple.seq(my_lg.all, ws=7, LOD=4)

ripple_my_lg

my_lg.all

print("------------------------------------------------------------
------------------------------------------------------------
")

# print your LG
print("Printing pdf...")

pdf(pdffile)
draw.map(my_lg.all, names=TRUE, grid=FALSE, cex.mrk=0.7)
dev.off()

q()
```

# Appendix C – Protocol: Purification of Total DNA from Animal Tissues (DNeasy 96 Protocol)

**Taken from DNeasy® Blood & Tissue Handbook, July 2006 version, pages 35-40**

This protocol is designed for high-throughput purification of total DNA from animal tissues, including rodent tails.

**Important points before starting**
If using the DNeasy 96 Blood & Tissue Kit for the first time, read “Important Notes” (page 15).
All centrifugation steps are carried out at room temperature (15–25°C).
Optional: RNase A may be used to digest RNA during the procedure. RNase A is not provided in the DNeasy 96 Blood & Tissue Kit (see “Copurification of RNA”, page 19).

**Things to do before starting**
Buffer AL should be premixed with ethanol before use. Add 90 ml ethanol (96–100%) to the bottle containing 86 ml Buffer AL or 260 ml ethanol to the bottle containing 247 ml Buffer AL and shake thoroughly. Mark the bottle to indicate that ethanol has been added. (Please note that, for purification of DNA from animal blood, Buffer AL must be used without ethanol. Buffer AL can be purchased separately if the same kit will be used for purification of DNA from animal blood.)
Buffer AW1 and Buffer AW2 are supplied as concentrates. Before using for the first time, add the appropriate amount of ethanol (96–100%) as indicated on the bottle to obtain a working solution.
Buffer ATL and Buffer AL may form precipitates upon storage. If necessary, warm to 56°C for 5 min until the precipitates have fully dissolved.
Mix Buffer AW1 before use by inverting several times.
Preheat an incubator to 56°C for use in step 4.
If using frozen tissue, equilibrate the sample to room temperature. Avoid repeated thawing and freezing of samples since this will lead to reduced DNA size.

**Procedure**
**1. Cut up to 20 mg tissue (up to 10 mg spleen) into small pieces. For rodent tails, place one (rat) or two (mouse) 0.4–0.6 cm lengths of tail into a collection microtube. Earmark the animal appropriately. Use a 96-Well-Plate Register (provided) to identify the position of each sample.**

Ensure that the correct amount of starting material is used (see “Starting amounts of samples”, page 15). For tissues such as spleen with a very high number of cells for a given mass of tissue, no more than 10 mg starting material should be used.

We strongly recommend to cut the tissue into small pieces to enable more efficient lysis. If desired, lysis time can be reduced by disrupting the sample using a bead mill, such as the QIAGEN TissueLyser (see page 56 for ordering information), before addition of Buffer ATL and proteinase K. A supplementary protocol for simultaneous disruption of up to 48 tissue samples using the TissueLyser can be obtained by contacting QIAGEN Technical Services (see back cover).

For rodent tails, a maximum of 1.2 cm (mouse) or 0.6 cm (rat) tail should be used. When purifying DNA from the tail of an adult mouse or rat, it is recommended to use only 0.4–0.6 cm.

Store the samples at –20°C until a suitable number has been collected (up to 192 samples). Samples can be stored at –20°C for several weeks to months without any reduction in DNA yield. DNA yields will be approximately 10–30 µg, depending on the type, length, age, and species of sample used (see “Expected yields”, page 22).

Keep the clear covers from the collection microtube racks for use in step 3.

**2. Prepare a proteinase K–Buffer ATL working solution containing 20 µl proteinase K stock solution and 180 µl Buffer ATL per sample, and mix by vortexing. For one set of 96 samples, use 2 ml proteinase K stock solution and 18 ml Buffer ATL. Immediately pipet 200 µl working solution into each collection microtube containing the tail sections or tissue samples. Seal the microtubes properly using the caps provided.**

**MODIFICATION:** To ensure that enough solution would be available to fill all wells, volumes were increased to 2.2 ml proteinase K and 19.8 ml Buffer ATL.

**Note:** Check Buffer ATL for precipitate. If necessary, dissolve the precipitate by incubation at 56°C for 5 min before preparing the working solution.

**IMPORTANT:** After preparation, the proteinase K–Buffer ATL working solution should be dispensed immediately into the collection microtubes containing the tail or tissue samples. Incubation of the working solution in the absence of substrate for >30 min reduces lysis efficiency and DNA purity.

**3. Ensure that the microtubes are properly sealed to avoid leakage during shaking. Place a clear cover (saved from step 1) over each rack of collection microtubes, and mix by inverting the rack of collection microtubes. To collect any solution from the caps, centrifuge the collection microtubes. Allow the centrifuge to reach 3000 rpm, and then stop the centrifuge. It is essential that the samples are completely submerged in the proteinase K–Buffer ATL working solution after centrifugation.**

If the proteinase K–Buffer ATL working solution does not completely cover the sample, increase the volume of the solution to 300 µl per sample (additional reagents are available separately; see page 56 for ordering information). Do not increase volumes above 300 µl as this will exceed the capacity of the collection microtubes in subsequent steps.

Keep the clear covers from the collection microtube racks for use in step 5.

**4. Incubate at 56°C overnight or until the samples are completely lysed. Place a weight on top of the caps during the incubation. Mix occasionally during incubation to disperse the sample, or place on a rocking platform.**

Lysis time varies depending on the type, age, and amount of tail or tissue being processed. Lysis is usually complete in 1–3 h or, for rodent tails, 6–8 h, but optimal results will be achieved after overnight lysis.

After incubation the lysate may appear viscous, but should not be gelatinous as it may clog the DNeasy 96 membrane. If the lysate appears very gelatinous, see the “Troubleshooting Guide”, page 47, for recommendations.

**Note:** Do not use a rotary- or vertical-type shaker as continuous rotation may release the caps. If incubation is performed in a water bath make sure that the collection microtubes are not fully submerged and that any remaining water is removed prior to centrifugation in step 5.

**5. Ensure that the microtubes are properly sealed to avoid leakage during shaking. Place a clear cover over each rack of collection microtubes and shake the racks vigorously up and down for 15 s. To collect any solution from the caps, centrifuge the**

**collection microtubes. Allow the centrifuge to reach 3000 rpm, and then stop the centrifuge.**

**IMPORTANT:** The rack of collection microtubes must be vigorously shaken up and down with both hands to obtain a homogeneous lysate. Inverting the rack of collection microtubes is not sufficient for mixing. The genomic DNA will not be sheared by vigorous shaking.

Keep the clear covers from the collection microtube racks for use in step 7.

Ensure that lysis is complete before proceeding to step 6. The lysate should be homogeneous following the vigorous shaking. To check this, slowly invert the rack of collection microtubes (making sure that the caps are tightly closed) and look for a gelatinous mass. If a gelatinous mass is visible, lysis needs to be extended by adding another 100 µl Buffer ATL and 15 µl proteinase K, and incubating for a further 3 h. It is very important to ensure that samples are completely lysed to achieve optimal yields and to avoid clogging of individual wells of the DNeasy 96 plate.

**Optional:** If RNA-free genomic DNA is required, add 4 µl RNase A (100 mg/ml). Close the collection microtubes with fresh caps, mix by shaking vigorously, and incubate for 5 min at room temperature. To collect any solution from the caps, centrifuge the collection microtubes. Allow the centrifuge to reach 3000 rpm, and then stop the centrifuge. Remove the caps, and continue with step 6.

Transcriptionally active tissues such as liver and kidney contain high levels of RNA, which will copurify with genomic DNA. For tissues that contain low levels of RNA, such as rodent tails, or if residual RNA is not a concern, RNase A digestion is usually not necessary.
6. Carefully remove the caps. Add 410 µl premixed Buffer AL–ethanol to each sample.

**Note:** Ensure that ethanol has been added to Buffer AL prior to use (see “Buffer AL”, page 18).

**Note:** A white precipitate may form upon addition of Buffer AL–ethanol to the lysate. It is important to apply all of the lysate, including the precipitate, to the DNeasy 96 plate in step 9. This precipitate does not interfere with the DNeasy procedure or with any subsequent application.

If the volumes of Buffer ATL and proteinase K were increased in steps 3 or 5, increase the volume of Buffer AL and ethanol accordingly. For example, 300 µl proteinase K–Buffer ATL working solution will require 615 µl Buffer AL–ethanol.

**7. Ensure that the microtubes are properly sealed to avoid leakage during shaking. Place a clear cover over each rack of collection microtubes and shake the racks vigorously up and down for 15 s. To collect any solution from the caps, centrifuge the collection microtubes. Allow the centrifuge to reach 3000 rpm, and then stop the centrifuge.**

Do not prolong this step.

**IMPORTANT:** The rack of collection microtubes must be vigorously shaken up and down with both hands to obtain a homogeneous lysate. Inverting the rack of collection microtubes is not sufficient for mixing. The genomic DNA will not be sheared by vigorous shaking. The lysate and Buffer AL–ethanol should be mixed immediately and thoroughly to yield a homogeneous solution.

**8. Place two DNeasy 96 plates on top of S-Blocks (provided). Mark the DNeasy 96 plates for later sample identification.**

**9. Remove and discard the caps from the collection microtubes. Carefully transfer the lysate (maximum 900 µl) of each sample from step 7 to each well of the DNeasy 96 plates.**

Take care not to wet the rims of the wells to avoid aerosols during centrifugation.

Do not transfer more than 900 µl per well.

**Note:** Lowering pipet tips to the bottoms of the wells may cause sample overflow and cross-contamination. Therefore, remove one set of caps at a time, and begin drawing up the samples as soon as the pipet tips contact the liquid. Repeat until all the samples have been transferred to the DNeasy 96 plates.

**Note:** If the volume of proteinase K–Buffer ATL working solution was increased in steps 3 or 5, transfer no more than 900 µl of the supernatant from step 7 to the DNeasy 96 plate. Larger amounts will exceed the volume capacity of the individual wells. Discard any remaining supernatant from step 7 as this will not contribute significantly to the total DNA yield.

**10. Seal each DNeasy 96 plate with an AirPore Tape Sheet (provided). Centrifuge for 10 min at 6000 rpm.**
**MODIFICATION:** Centrifuges at The Roslin Institute had a maximum rpm of 4000/3700, thus the centrifugation step was modified to 20 minutes at 4000 rpm or 30 minutes at 3700 rpm.

AirPore Tape prevents cross-contamination between samples during centrifugation.

After centrifugation, check that all of the lysate has passed through the membrane in each well of the DNeasy 96 plates. If lysate remains in any of the wells, centrifuge for a further 10 min.

**MODIFICATION:** The centrifugation step was modified to 20 minutes at 4000 rpm or 30 minutes at 3700 rpm, in order to account for the different maximum rpm values available on the centrifuges at The Roslin Institute.

**11. Remove the tape. Carefully add 500 µl Buffer AW1 to each sample.**

**Note:** Ensure that ethanol has been added to Buffer AW1 prior to use.

It is not necessary to increase the volume of Buffer AW1 if the volume of proteinase K–Buffer ATL working solution was increased in steps 3 or 5.

**12. Seal each DNeasy 96 plate with a new AirPore Tape Sheet (provided). Centrifuge for 5 min at 6000 rpm.**

**MODIFICATION:** The centrifugation step was modified to 10 minutes at 4000 rpm or 15 minutes at 3700 rpm, in order to account for the different maximum rpm values available on the centrifuges at The Roslin Institute.

**13. Remove the tape. Carefully add 500 µl Buffer AW2 to each sample.**

**Note:** Ensure that ethanol has been added to Buffer AW2 prior to use.

It is not necessary to increase the volume of Buffer AW2 if the volume of proteinase K–Buffer ATL working solution was increased in steps 3 or 5.

**14. Centrifuge for 15 min at 6000 rpm.**

Do not seal the plate with AirPore Tape.

The heat generated during centrifugation ensures evaporation of residual ethanol in the sample (from Buffer AW2) that might otherwise inhibit downstream reactions.

**MODIFICATION:** The centrifugation step was modified to 30 minutes at 4000 rpm, in order to account for the different maximum rpm values available on the centrifuges at The Roslin Institute.

**15. Place each DNeasy 96 plate in the correct orientation on a new rack of Elution Microtubes RS (provided).**

**16. To elute the DNA, add 200 µl Buffer AE to each sample, and seal the DNeasy 96 plates with new AirPore Tape Sheets (provided). Incubate for 1 min at room temperature (15–25°C). Centrifuge for 2 min at 6000 rpm.**

200 µl Buffer AE is sufficient to elute up to 75% of the DNA from each well of the DNeasy 96 plate.

Elution with volumes less than 200 µl significantly increases the final DNA concentration of the eluate but may reduce overall DNA yield. For samples containing less than 1 µg DNA, elution in 50 µl Buffer AE is recommended.

**MODIFICATION:** The centrifugation step was modified to 3 minutes at 4000 rpm or 5 minutes at 3700 rpm, in order to account for the different maximum rpm values available on the centrifuges at The Roslin Institute.

**17. Recommended: For maximum DNA yield, repeat step 16 with another 200 µl Buffer AE.**

A second elution with 200 µl Buffer AE will increase the total DNA yield by up to 25%. However due to the increased volume, the DNA concentration is reduced. If a higher DNA concentration is desired, the second elution step can be performed using the 200 µl eluate from the first elution. This will increase the yield by up to 15%.

Use new caps (provided) to seal the Elution Microtubes RS for storage.

**MODIFICATION:** The second elution was conducted by placing the 200µl eluate from the first elution step back in to the elution tube and re-eluting the same volume. This was done in order to prevent dilution of the DNA.

# Appendix D – Figures

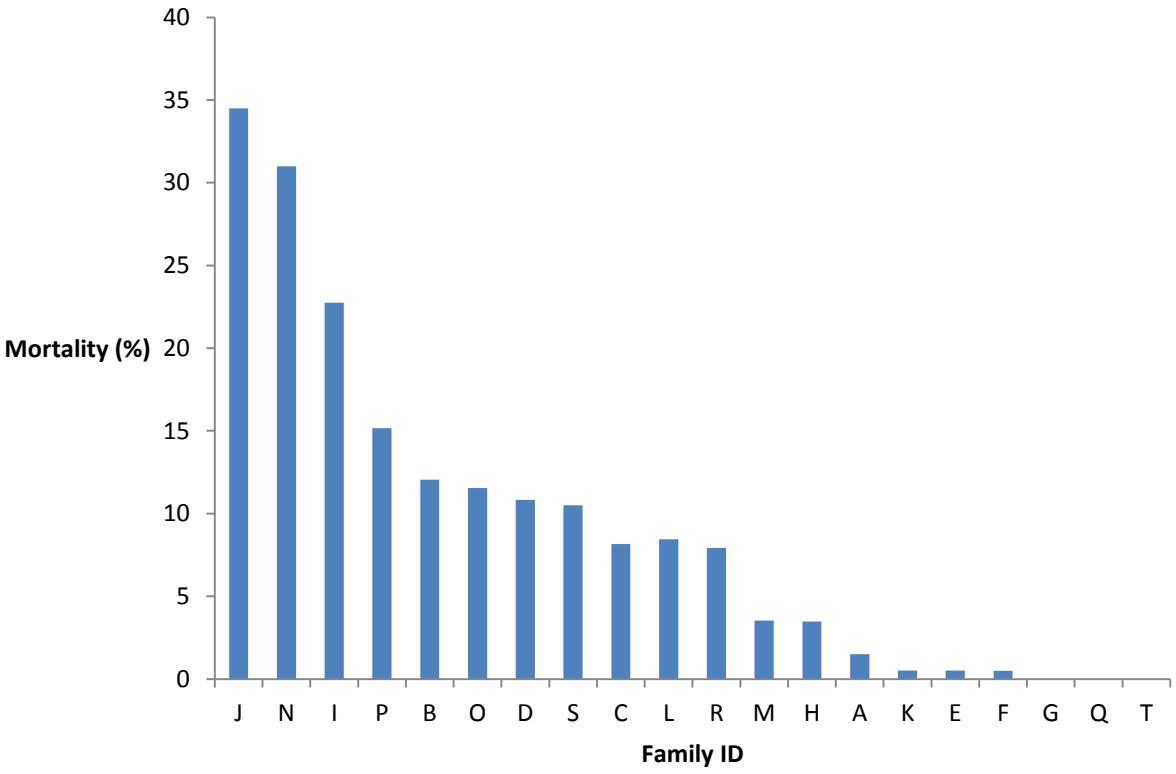



**Figure D1: Mortality levels across the twenty families challenged with IPNV in the experiment described in Houston *et al* (2010).**
Percentage mortalities were averages across the two replicate tanks for each family.

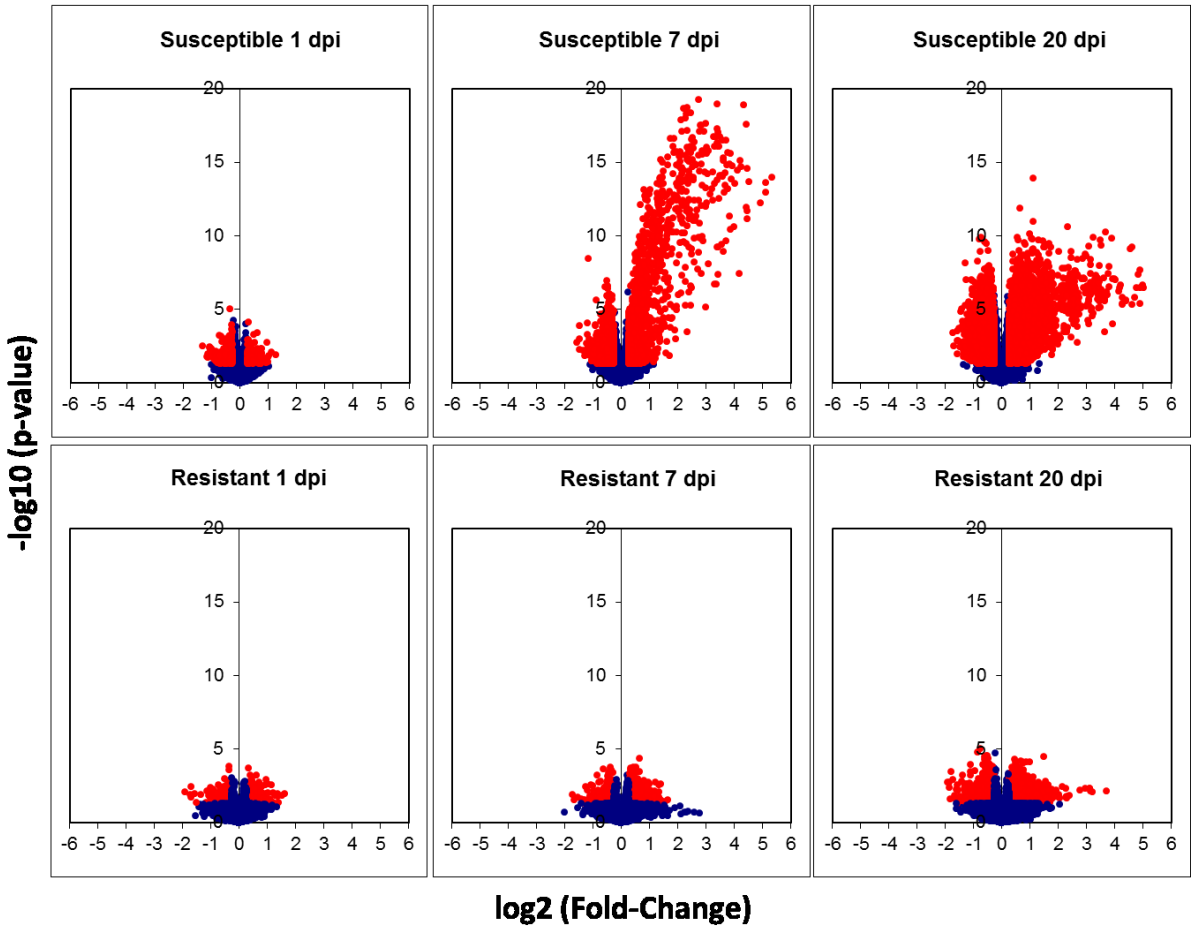



**Figure D2: Volcano plots showing up- and down-regulation of probe sequences (representing genes) on the Atlantic salmon Agilent Oligo Array (Martin et al, 2007) in IPNV resistant and susceptible families at 1, 7 and 20 days post-challenge.**
Row one represents expression levels in susceptible individuals and row two represents expression levels in resistant individuals. Red points represent probes sequences with showing significant levels of hybridisation of fry RNA sequences (nominal P-value<0.05).

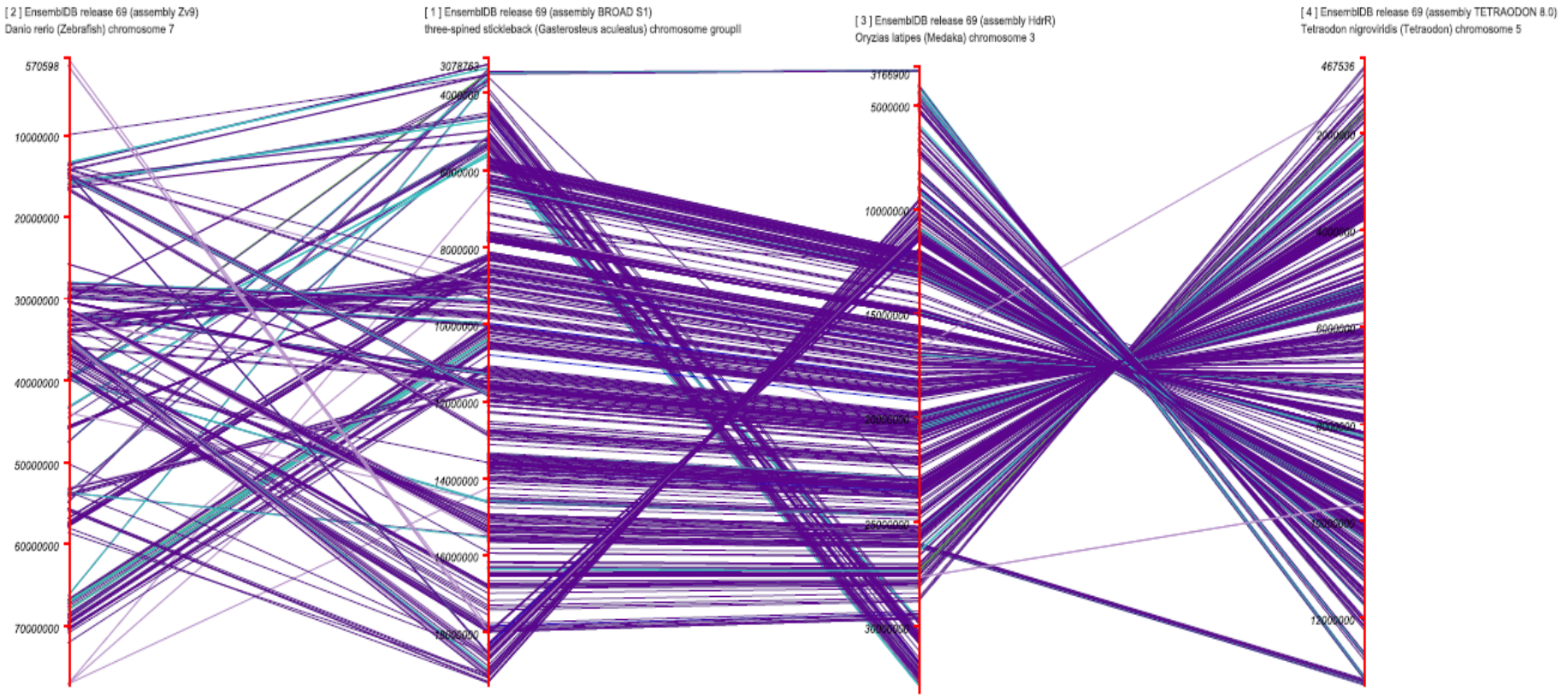



**Figure D3: Orthologous group A gene order conservation**
Orthologous group A consists of (from left to right): zebrafish chromosome 7, stickleback linkage group II, medaka chromosome 3, and green spotted puffer fish chromosome 5. Diagram was drawn relative to the stickleback IPN QTL orthologous region. Gene order is highly conserved between stickleback, medaka and green spotted puffer fish. Diagram was drawn using the ArkDraw software package (http://bioinformatics.roslin.ed.ac.uk/arkmap/).

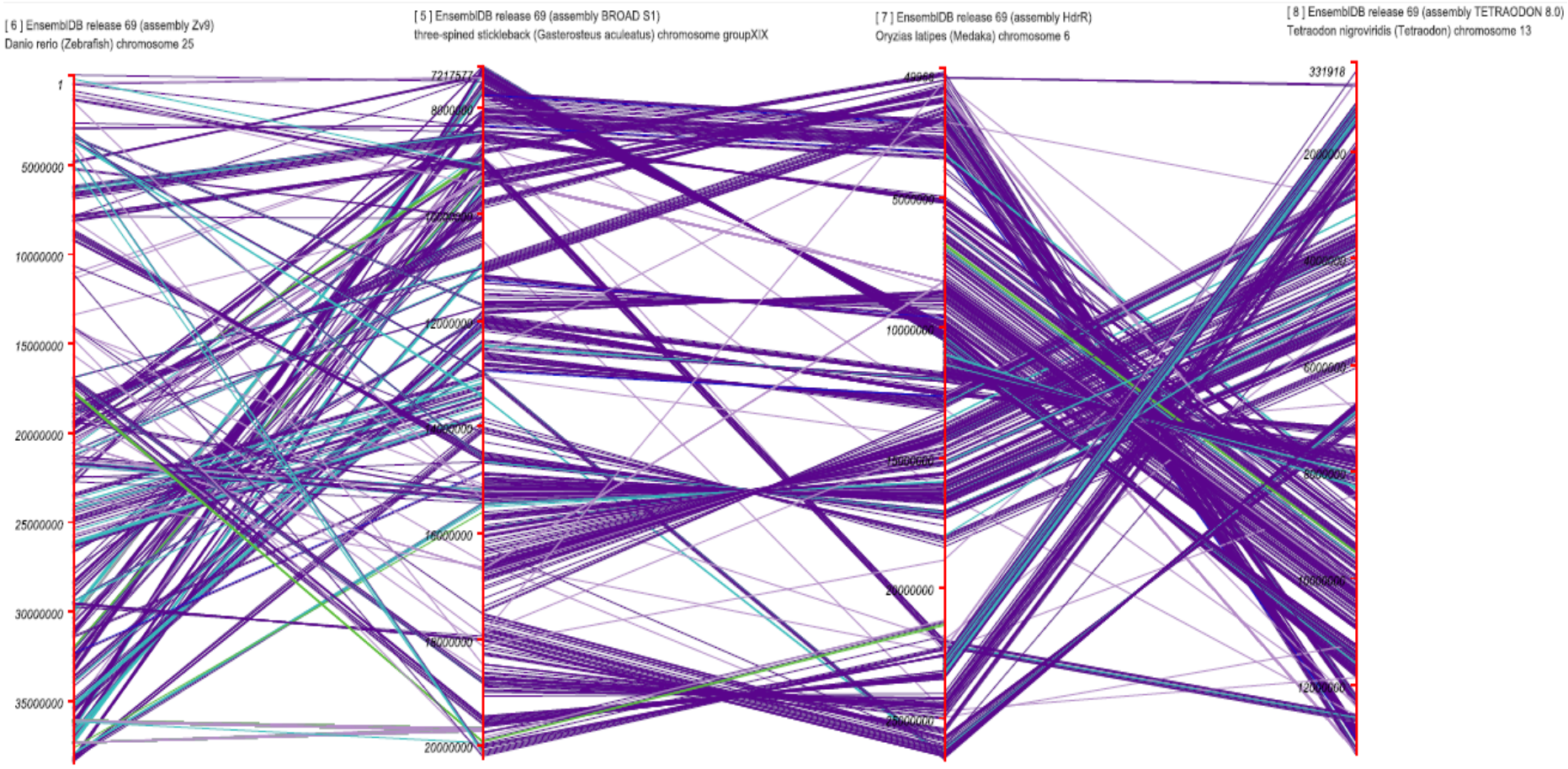



**Figure D4: Orthologous group B gene order conservation**

Orthologous group B consists of (from left to right): zebrafish chromosome 25, stickleback linkage group XIX, medaka chromosome 6, and green spotted puffer fish chromosome 13. Diagram was drawn relative to the stickleback IPN QTL orthologous region. Gene order is not as conserved as orthologous group A (Figure D1). Diagram was drawn using the ArkDraw software package (http://bioinformatics.roslin.ed.ac.uk/arkmap/).

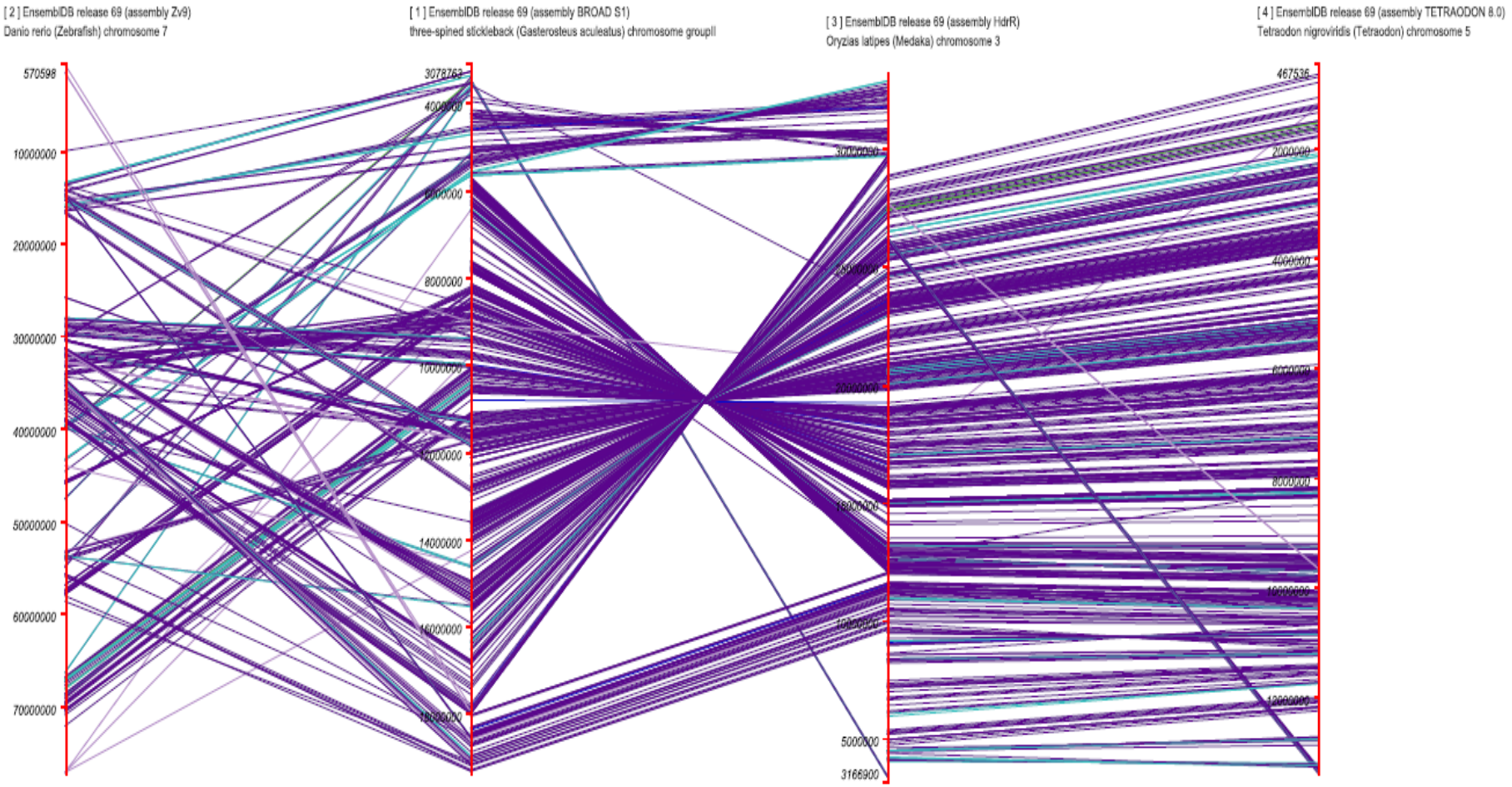



**Figure D5: Orthologous group A, with the medaka assembly inverted to show inversion of central portion relative to the stickleback assembly**
This same inversion is seen when comparing the stickleback and green spotted puffer fish regions, thus this central inversion is either a mis-assembly in the Ensembl stickleback chromosome, or a true inversion in the stickleback genome. Diagram was drawn using the ArkDraw software package (http://bioinformatics.roslin.ed.ac.uk/arkmap/).

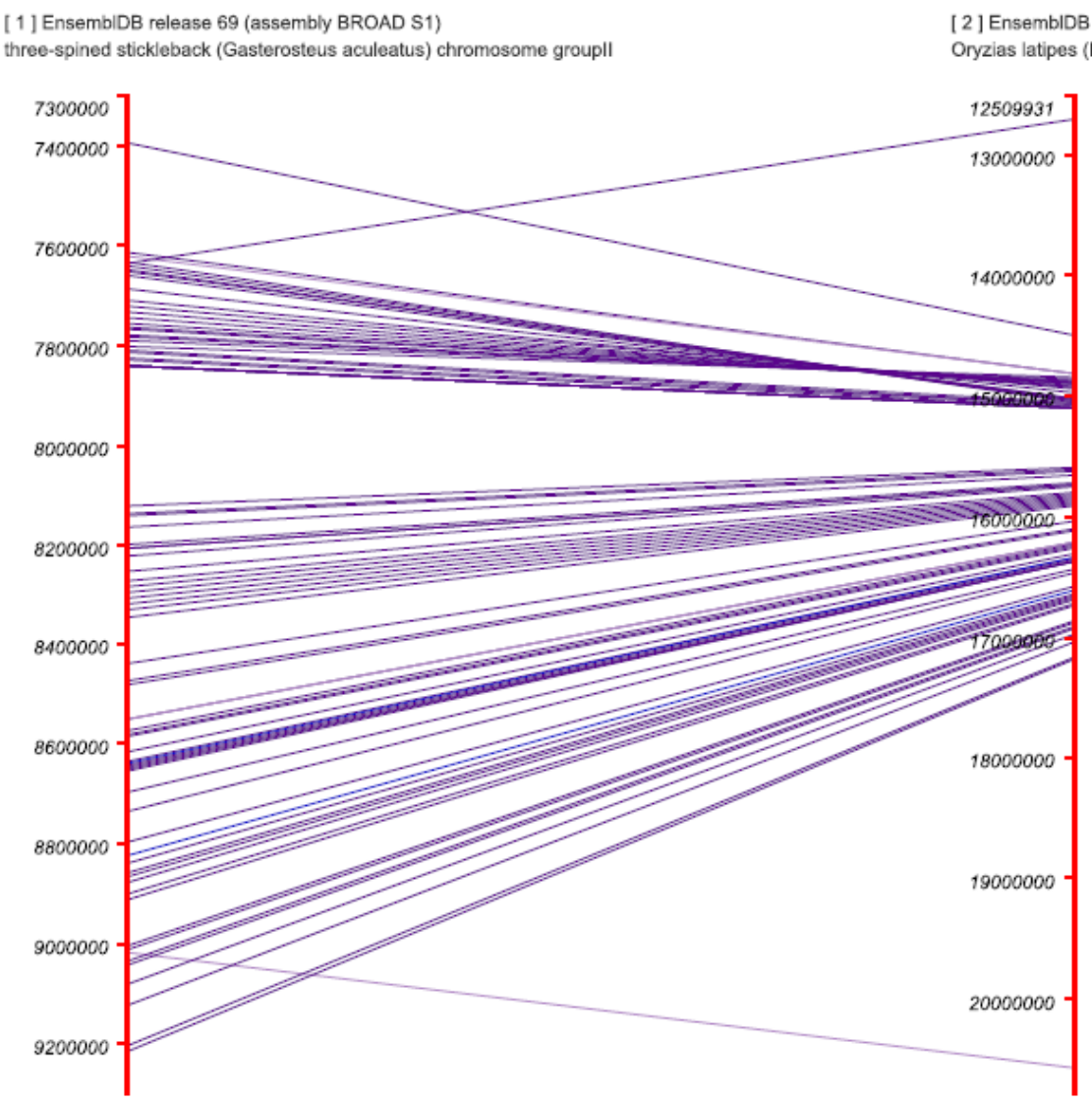



**Figure D6: Orthologous relationship between the 2Mb IPN QTL-orthologous region of stickleback and medaka chromosome 3**

Gene order is largely conserved between these two species in the QTL-orthologous region. The inversion seen in the larger regions in Figure D1 does not affect this region.

# Appendix E – Ingenuity Pathway Analysis: Settings

The pathway analysis of differentially-expressed genes between IPNV resistant and susceptible individuals was run using the IPA software package. IPA utilises information from published literature and publically-available databases across many different species. Therefore, it is enriched for information from species in which many studies have been conducted (such as human, mouse and rat). Although biological pathways are generally consistent across species, some variation in the specific roles of pathways and the molecules which map to these pathways is expected. As such, results based on cross-species comparative pathway inference should be interpreted with caution.

IPA was run with the following settings in the 'Create Core Analysis' option:

- Use reference set - Ingenuity Knowledge Base (genes only)
- Consider direct and indirect relationships
- Generate networks as part of this analysis–with 70 molecules per network and 25 networks per analysis maximum
- Use all available data sources
- Only use information for which there is experimental evidence
- Use all species information available and all the information on different tissues and cell lines and mutations

# Appendix F – Processing and combining consensus RAD sequences within species

## Rainbow trout

Four FASTA files from four different studies (Hohenlohe et al, 2011; Hecht et al, 2012; Hale et al, 2013; Hecht et al, 2013; Hohenlohe et al, 2013) were obtained for use in this analysis (details of the sequences from each study are given in Table 5.1). To obtain consensus sequences across all populations, a custom-written clustering pipeline was applied. First, sequences across all four populations were combined into a single file (total number of sequences: 407,332). A BLASTN nucleotide database of all sequences in this file was created, and all sequences were aligned (BLASTN) to this database (i.e. self-alignment) [BLASTN version 2.2.25+, (Altschul et al, 1990)]. Alignments were quality filtered to retain only those with a minimum percentage identity of 95%, and ≤2 base mismatch.

Homologous cross-population RAD loci were recovered as follows. For each sequence, the top match within each population for that query sequence was identified. For example, SEQ1_POP1 would align first to itself, then potentially to SEQX_POP2, SEQY_POP3 and SEQZ_POP4, and these were assigned to a single common RAD locus cluster. To reduce the inclusion of repetitive elements, sequences with high quality alignments to multiple clusters were removed, as were the clusters which they belonged to. Finally, clusters were filtered to retain those with a minimum of three and a maximum of four sequences. A total of 32,027 clusters were identified. For each cluster, a representative sequence was obtained, and this was used in all downstream analyses.

The Python script used to conduct this analysis is given below.

```
__author__ = 'Serap_Gonen'

# START DATE: 20/03/14
# END DATE: 20/03/14

# SCRIPT DESCRIPTION
# script was written to group rainbow trout loci across the four
# populations of RAD data
# criteria for grouping:
#       > 95% identity
#       < 3 mismatches
# criteria for a valid locus:
#       must have at least 3 population matches (including
#       itself) so is allowed to not match one population
#       out of four

# INPUT FILE FORMAT:
# blastn -outfmt 7 output file (tab del)
# Fields: query id, subject id, query length, subject length,
# % identity, alignment length, mismatches, gap opens,
# q. start, q. end, s. start, s. end, evalue, bit score

# OUTPUT FILE FORMAT:
# Each row contains IDs of matched loci, tab del

# HOW TO RUN SCRIPT:
# Requirements: Python 2.6 and above. Not compatible with python3
#             : Libraries : optparse
# Run on the command line as:
# python rtrout_common_loci_post_blastn.py
#  --blastn_file <input_file>

#######################################################################

# SCRIPT

#######################################################################
```

```
# NECESSARY IMPORTS

# if script is main script being run
if __name__ == "__main__":
    # use optparse to specify input file
    from optparse import OptionParser
    parser = OptionParser()
    parser.add_option("--blastn_file", dest = "blastn_file")
    (options, args) = parser.parse_args()

#################################################################

# OBJECT ORIENTED CODE

# class Locus
class Locus(object):

    def __init__(self):
        # holds the hits assigned to this locus
        self.hit_list = []
        # assures that only one seq per pop is assigned to this locus
        self.populations = ["MILLER", "HOHENLOHE", "HECHT", "HALE"]

    # method addHit
    # adds subject_id to hit_list if a hit for that population hasn't
    # been assigned before
    def addHit(self, subject_id):
        for population in self.populations:
            if population in subject_id:
                self.populations.remove(population)
                self.hit_list.append(subject_id)

# class FileParser
class FileParser(object):

    def __init__(self):
        self.blastn_file =\
                        open(options.blastn_file,'r').read().splitlines()
        # keeps track of queries/subjects already assigned to a locus
        self.seen = []
        # locus object dictionary
        self.query2locusobj = {}

    # method assignHit
    # checks how good alignment is
    # if good then sends subject to be assigned to locus
    # if a hit for that population has not
    # been assigned to that locus
    # then checks if that subject_id has been assigned to the locus
    # if it has then it adds it to the self.seen list
    def assignHit(self, locus, subject_id, percentage_identity,\
                  mismatches):
        if subject_id not in self.seen:
            if float(percentage_identity) > 95.0 and \
                                            int(mismatches) < 3:
                locus.addHit(subject_id)
                if subject_id in locus.hit_list:
                    self.seen.append(subject_id)
```

```
  # method parseLines
    # to actually read the file
    def parseLines(self):
        for line in self.blastn_file:
            # split line into columns
            query_id, subject_id, query_length, subject_length, \
            percentage_identity, alignment_length, mismatches, \
            gap_opens, q_start, q_end, s_start, s_end, evalue, \
            bit_score = line.split("\t")
            # if this is a new query
            if query_id not in self.seen:
                # make a new locus object
                locus = Locus()
                locus.addHit(query_id)
                # say that we have now seen the query_id
                self.seen.append(query_id)
                # see if subject_id belongs to this locus
                self.assignHit(locus, subject_id, \
                               percentage_identity, mismatches)
                # add locus to the dictionary
                self.query2locusobj[query_id] = locus
            # if we have seen query_id and it is actually a locus
            elif query_id in self.query2locusobj.keys():
                # get the object
                locus = self.query2locusobj[query_id]
                # see if subject_id belongs to this locus
                self.assignHit(locus, subject_id, \
                                    percentage_identity, mismatches)
            # else we have seen this query_id, however we saw it

            # as a subject_id and not a query id

# class Combiner
class Combiner(object):

    def __init__(self):
        self.file_object = FileParser()

    def runScript(self):
        self.file_object.parseLines()
        # print hits!
        for locusobj in self.file_object.query2locusobj.values():
            # locus seen in > 2 populations (ie 3, one of
            # which is a match to itself)
            if len(locusobj.hit_list) > 2:
                print "\t".join(locusobj.hit_list)

######################################################################

# PROCEDURAL CODE

Combiner().runScript()

# DONE

######################################################################
```

## Atlantic salmon

Two sets of RAD sequences were obtained from two different Atlantic salmon populations. The first set [SET1, (Houston et al, 2012)] was from a single-end RAD sequencing study conducted in two families [labelled as B and C in Houston et al (2012)], where RAD loci had been inferred separately within each family. Therefore, the first step in this analysis was the identification of common RAD loci across the two families. First, a BLASTN nucleotide database of the 337,315 RAD loci identified in family C was created. The 559,823 RAD locus sequences identified in family B were aligned (BLASTN) to this database. Alignments were filtered to retain those with high quality, based on a minimum percentage identity of 95%, no base mismatches, and an E-value of 1e-30. These thresholds were determined by preliminary BLASTN alignments using simulated sequences of 95 base pairs (bp) in length, since this was the length of the sequences in both families. To eliminate RAD loci originating from repetitive regions, alignments where one or both of the sequences showed significant alignment to multiple sequences were removed. The final number of common RAD loci across the two families was 66,073.

The second set of RAD sequences [SET2, (Gonen et al, 2014)] was derived from paired-end RAD-sequencing, and was a mixture of 366,219 single- and 116,328 paired-end sequences (total: 482,547). For the purposes of this study, only the single-end sequences were utilised. A BLASTN nucleotide database of these sequences was created, and the 66,073 representative sequences from SET1 were aligned (BLASTN) to this database. As above, alignment significance was determined based on a minimum percentage identity of 95%, no base mismatches, and an E-value of 1e-30, and filtering for RAD locus clusters originating from putative repetitive/duplicate regions of the genome was conducted based on the identification and removal of clusters containing sequences which mapped to multiple clusters. A total of 65,758 (99.5%) shared RAD loci were identified across the two sets.

The Python script for processing of the resulting BLASTN file is given below:

```
__author__ = 'Serap_Gonen'

# START DATE: 10/03/14
# END DATE: 10/03/14

# SCRIPT DESCRIPTION
# Script to identify the best hits using a blast -outfmt 7
# output tabular file
# For matching within population - to get 1:1 correspondence
# between loci
# the following unix command would have already been run on the
# blastn tabular output file:
# cat FamilyB_vs_C.blastn | grep "^AS" | cut -f 1,2,11 | \
# awk 'BEGIN {FS = "\t"; query = ""}; {if($1!=query) \
# {print $1"\t"$2"\t"$3; query=$1} else{}} ' | sort -k2 > \
# query_hit_evalue.txt

# INPUT FILE FORMAT:
# query \t hit \t evalue

# OUTPUT FILE FORMAT:
# loci_from_file_1 \t loci_from_file_2

# HOW TO RUN SCRIPT:
# Requirements: Python 2.6 and above. Not compatible with python3
#             : Libraries

#######################################################################
###
# SCRIPT
```

```
# NECESSARY IMPORTS

from sys import argv

script, blastn_file = argv

########################################################################

# OBJECT ORIENTED CODE

class FileParser(object):

    def __init__(self, blastn_file):
        self.blastn_file = open(blastn_file, 'r').read().splitlines()
        self.hit2evalue = {}
        self.hit2query = {}

        for hit,query in self.hit2query.items():
            print "{q}\t{h}".format(q=query, h=hit)

    def parseLines(self):
        for line in self.blastn_file:
            query, hit, evalue = line.split("\t")
            if hit not in self.hit2evalue:
                self.hit2evalue[hit] = evalue
                self.hit2query[hit] = query
            else:
                last_evalue = self.hit2evalue[hit]
                if float(last_evalue) > float(evalue):
                    self.hit2evalue[hit] = evalue
                    self.hit2query[hit] = query

########################################################################

# PROCEDURAL CODE
FileParser(blastn_file).parseLines()

# DONE

########################################################################
```

## Three-spined stickleback

Sequences from 46 stickleback originating from populations in Vancouver Island, British Columbia, Canada were kindly donated for this study by Dr Daniel Berner (Universität Basel, Zoologisches Institut, Switzerland) (Roesti et al, 2012; Roesti et al, 2013). Since sequences originated from two independent sequencing experiments/technologies, read lengths across individuals were different, whereby ten individuals had sequence lengths of 138bp, and the remaining 36 had sequence lengths of 64bp. The number of sequences across all individuals ranged from 25,840–42,618.

Sequences across all individuals were combined into a single FASTA file containing 1,668,843 sequences. A BLASTN nucleotide database of this file was produced and aligned (BLASTN) to itself. Alignments were quality filtered, based on a minimum of 98% match identity, maximum of 2 mismatches and alignment length (minimum of 64bp if analysing the shorter reads, 138bp otherwise). Filtered alignments were clustered into common RAD loci across individuals. If a single sequence was significantly mapped to multiple different clusters, this sequence, and the clusters it was assigned to, were removed from further analyses. The remaining clusters containing uniquely assigned sequences were filtered to retain those with a minimum of 20 sequences from 20 different individuals and a maximum of 50 sequences overall (to filter for repeats). A total of 31,118 clusters (i.e. shared RAD loci) were identified. A single representative sequence was selected and used in all downstream analyses.

The Python script used to implement this clustering pipeline is given below:

```
__author__ = 'Serap_Gonen'

# START DATE: 13/03/14
# END DATE: 15/03/14

# SCRIPT DESCRIPTION
# this script was written to parse blastn tabular output
# written specifically for the stickleback RAD data in
# order to match RAD loci between the 46 individuals
# however it can of course be adapted to other blastn comparsions
# and output, just by changing a few things

# INPUT FILE FORMAT:
# blastn output file in -outfmt 7 format. Columns:
# query id, subject id, % identity, alignment length,
# mismatches, gap opens, q. start, q. end, s. start, s. end,
# evalue, bit score

# OUTPUT FILE FORMAT:
# Locus 1
# defline of matching sequences as a column
# Locus 2
# defline of matching sequences as a column
# etc....

# HOW TO RUN SCRIPT:
# Requirements: Python 2.6 and above. Not compatible with python3
#             : Libraries: optparse
# Run on the command line as:
# python blastn_output_parser.py
# --blastn_input all_46_individuals.blastn
# --max_num_hits 50
# --min_num_hits 20
# --percentage_identity 100
# --alignment_length 64,138
# --max_num_mismatches 2
# > common_loci.txt
```

```
#####################################################################

# SCRIPT

#####################################################################

# OBJECT ORIENTED CODE

class CheckLine(object):

    def __init__(self, line):
        self.line = line

    def check(self):
         if "hits" in self.line:
            number_of_hits = int(self.line.split(" ")[1])
            if number_of_hits > int(options.max_num_hits) or \
                        number_of_hits < int(options.min_num_hits):
                return False
            else:
                return True

class Locus(object):

    def __init__(self, name):
        self.name = name
        self.hit_list = []

    def addHit(self, hit):
        self.hit_list.append(hit)

    def validLocus(self):
        if len(self.hit_list) > int(options.min_num_hits):
            return True

class AlignmentType(object):

    def __init__(self, percentage_identity, mismatches, \
                                                alignment_length):
        self.percentage_identity = percentage_identity
        self.mismatches = mismatches
        self.alignment_length = alignment_length

    def perfect(self):
        if float(self.percentage_identity) == \
            float(options.percentage_identity) \
            and self.mismatches == "0" \
            and self.alignment_length in \
            options.alignment_length.split(","):
            return True
        else:
            return False
```

```
    def good(self):
        if int(self.mismatches) < 3:
            return True
        else:
            return False

class FileParser(object):

    def __init__(self):
        self.blastn_input = open(options.blastn_input, \
                                 'r').read().splitlines()
        self.locus_dict = {}
        self.seen = []
        self.mismatch1_hit2query = {}
        self.mismatch2_hit2query = {}

    def updateMismatches(self, subject_id, query_id, mm_count):
        mm_dict = False
        if mm_count == "1":
            mm_dict = self.mismatch1_hit2query
        elif mm_count == "2":
            mm_dict = self.mismatch2_hit2query
        if mm_dict:
            if subject_id not in mm_dict:
                mm_dict[subject_id] = []
            mm_dict[subject_id].append(query_id)

    def updateLocusDict(self, locus):
        if locus.name not in self.locus_dict.keys():
            self.locus_dict[locus.name] = locus

    def getLocusObject(self, query_id):
        if query_id not in self.seen:
            self.seen.append(query_id)
            return Locus(query_id)
        elif query_id in self.locus_dict.keys():
            return self.locus_dict[query_id]
        else:
            return False

    def lineParser(self):
        want = False
        for line in self.blastn_input:
            if line.startswith("#"):
                want = CheckLine(line).check()
            else:
                if want:
                    query_id, subject_id, percentage_identity,\
                    alignment_length, mismatches, gap_opens, \
                    query_start, query_end, subject_start, \
                    subject_end, evalue, bitscore = \
                    line.split("\t")
                    locus = self.getLocusObject(query_id)
```

```
                    if locus:
                        if subject_id not in self.seen:
                            alignment = \
                            AlignmentType(percentage_identity, \
                            mismatches, alignment_length)
                            if alignment.perfect():
                                locus.addHit(subject_id)
                                self.seen.append(subject_id)
                            elif alignment.good():
                                self.updateMismatches( \
                                subject_id, query_id, \
                                mismatches)
                            self.updateLocusDict(locus)

class AssignMismatches(object):

    def __init__(self, locus_dict, mismatch_dict, seen):
        self.locus_dict = locus_dict
        self.mismatch_dict = mismatch_dict
        self.seen = seen

    def addMismatchHits(self):
        for hit, query_list in self.mismatch_dict:
            if len(query_list) == 1:
                if hit not in self.seen:
                    self.locus_dict[query_list[0]].addHit(hit)
                    self.seen.append(hit)


class TrueLocus(object):

    def __init__(self, locus_dict):
        self.locus_dict = locus_dict

    def locusParser(self):
        for locus_id, locus_object in self.locus_dict.items():
            if locus_object.validLocus():
                print locus_id, "\t", \
                                '\t'.join(locus_object.hit_list)

class Combiner(object):

    def __init__(self):
        self.file_object = FileParser()
```

```
    def parseMismatches(self):
        self.locus_dict = self.file_object.locus_dict
        self.mismatch1_hit2query = \
                                self.file_object.mismatch1_hit2query
        self.mismatch2_hit2query = \
                                self.file_object.mismatch2_hit2query
        AssignMismatches(self.locus_dict, \
                    self.mismatch1_hit2query, self.file_object.seen)
        AssignMismatches(self.locus_dict, \
                    self.mismatch2_hit2query, self.file_object.seen)

    def printValidLoci(self):
        TrueLocus(self.locus_dict).locusParser()

    def scriptRunner(self):
        self.file_object.lineParser()
        self.parseMismatches()
        self.printValidLoci()

#########################################################################

# PROCEDURAL CODE

Combiner().scriptRunner()

# DONE

#########################################################################
```

# Appendix G – Cross-species orthologous RAD locus identification

Pairwise BLASTN alignments were clustered into cross-species orthologous RAD loci as follows.

1. Identify 'best hits' for each pairwise alignment.
2. Filter best hits to identify only unique top alignments (i.e., one-to-one alignments)
3. Filter best hits based on sequence similarity parameters:
    a. Strict analysis, within salmonid species only:
        i. 95% sequence similarity
        ii. ≤2 base mismatch
        iii. Minimum 50bp alignment
    b. Relaxed analysis, across all ten species
        i. 85% sequence similarity
        ii. ≤10 base mismatch
        iii. Minimum 45bp alignment
4. Generate a concatenated file of all filtered pairwise alignments across all species.
5. Group pairwise alignments into putative RAD clusters. E.g. within salmonid species only, if Atlantic_salmon_RAD_1 significantly aligned to Sockeye_salmon_RAD_1, Chinook_salmon_RAD_1, Lake_whitefish_RAD_1 and Rainbow_trout_RAD_1, and these all aligned to each other respectively, then these were inferred as a single cluster. Python script written for this is given below.
6. Identify sequences assigned to more than one cluster. Remove all clusters containing these sequences.
7. Filter clusters to remove those with more than one sequence originating from a given species.
8. Filter clusters for a minimum number of species sequences (e.g. minimum of 7 of the 10 species must have sequence etc.).
9. In the across teleost species analysis, identify and remove salmonid-specific clusters.

Python script (identify_common_loci.py) written to identify cross-species RAD locus clusters:

```
__author__ = 'Serap_Gonen'

# START DATE: 07/05/14
# END DATE: 07/05/14

# SCRIPT DESCRIPTION
# This script uses blastn one to one matches to assign groups
# of loci to clusters, so that common loci across multiple
# populations can be identified using pairwise alignments and
# not having to do one massive blastn alignment across
# multiple populations!

# INPUT FILE FORMAT:
# two input files:
# 1) blastn onetoone match file which I made earlier of all
#    onetoone matches to consider tab del blastn output file,
#    number of rows is irrelevant as long as first and second
#    column are query and subject ID. Actually, it may even
#    be a two column tab del file of query and subject,
#    none of the other columns are mentioned or used
# 2) single column file of sequences aligning to multiple
#   (2/3/4 therefore present in 3/4/5) populations

# OUTPUT FILE FORMAT:
# sequence_id\tcluster

# HOW TO RUN SCRIPT:
# Requirements: Python 2.6 and above.
#             : Not compatible with python3
#             : Libraries : argv only (sys)
# Run on the command line as:
# python ~/PyCharm_Projects/RAD_seq/identify_common_loci.py \
# <blastn_file> <match_file> > <outfile>

#############################################################

# SCRIPT

#############################################################

# NECESSARY IMPORTS

from sys import argv
script, blastn_file, match_file = argv

#############################################################
```

```
# PROCEDURAL CODE

# file of sequence headers matching across multiple pops
match_list = open(match_file, 'r').read().splitlines()
# file of blastn one to one matches
blastn_list = open(blastn_file, 'r').read().splitlines()
# dict for storing sequence id to the allocated cluster
sequence2cluster = {}
count = 0   # count to assign new cluster number

for line in blastn_list:
    elements = line.split("\t")
    if (elements[0] in match_list) and \
       (elements[1] in match_list):
        # elements[0]==query, elements[1]==sequence
        if elements[0] not in sequence2cluster:
            if elements[1] not in sequence2cluster:
                # new rad locus, assign new cluster
                sequence2cluster[elements[0]] = count
                sequence2cluster[elements[1]] = count
                count += 1 # increment count for new cluster
            else:
                sequence2cluster[elements[0]] = \
                                sequence2cluster[elements[1]]
        else:
            if elements[1] not in sequence2cluster:
                sequence2cluster[elements[1]] = \
                                sequence2cluster[elements[0]]

# print print print!
for sequence, cluster in sequence2cluster.items():
    print sequence, "\t", cluster

# DONE

#################################################################
```

# Appendix H – Between-species variant identification

MUSCLE alignment of homologous RAD locus sequences across species, identification of cross-species variants, and concatenation of variants into one sequence per fish species. Scripts are:

Shell script–automates the whole procedure.

Python script (muscle_parser_find_variants.py)–identifies cross-species variants per locus based on MUSCLE alignment output file.

Python script (variant_concatenator_post_muscle.py)–concatenates variants across all RAD loci for each species.

Shell script

```
#!/bin/sh

#$ -N muscle-variant-concatenate
#$ -cwd
#$ -o rediout_msc
#$ -e redierror_msc

# script variables
muscle="/usr/local/shared_bin/muscle/3.8.31/muscle3.8.31_i86linux64"
variant_script="/nfs_netapp/vlsgonen/PyCharm_Projects/RAD_seq/muscle_parser_find_variants.py"
concatenator_script="/nfs_netapp/vlsgonen/PyCharm_Projects/RAD_seq/variant_concatenator_post_muscle.py"
#### change per script ####
current_directory="/groups2/houston_grp/SERAP/TRIMMED_ANALYSIS/Trimmed_sequences/All_alignments/Common_Loci/85/2014_05_BLASTN_CLUSTERS"
fasta_files=`ls $current_directory | grep Locus`
#### change per script ####
final_concatenated_variants_file="2014_05_28_All_variants_concatenated.txt"

# script

# muscle
for fasta_file in $fasta_files
    do
        $muscle -in $fasta_file -out $fasta_file.afa
    done

# variant identification
alignment_files=`ls $current_directory | grep .afa`
for alignment_file in $alignment_files
    do
        filename=`echo $alignment_file | sed 's/.afa//g'`
        python $variant_script --alignment_file \
                        $alignment_file > $filename.variants
    done

# variant concatenation into one file
python $concatenator_script --directory_path \
$current_directory --file_extension .variants > \
$final_concatenated_variants_file
```

muscle_parser_find_variants.py

```
__author__ = 'Serap_Gonen'

# START DATE: 19/05/14
# END DATE: 20/05/14

# SCRIPT DESCRIPTION
# takes in fasta files from muscle software alignments
# and calls variants based on the alignments
# script must be called individually for each file
# the essence of the program is as follows:
# 1) make a dict of deflines to sequences
# 2) for each base position, make a dict of defline to base
#    and interrogate that base position to see if it is
#    variant across the 4 fish (ie more than one allele
#    present)
# 3) if yes then store this as a variant and add to the dict
#    of defline to variants
# 4) for each fish, print defline and a concatenated sequence
#    of all variants

# INPUT FILE FORMAT:
# sequence of deflines goes over multiple files so script must
# parse file accordingly
# .afa file from muscle software which looks like this:
# >CL_R05330_MILLER
# CCTGCAGGTCGAAGAACTTGCTGTAGCGCCTGAAGATGACCTCTGTGCTGCCATCAGACC
# --------
# >Lakewhitefish_137535
# --------TCGAAGAACTTGCTGTAGCGCCTGAAGATGACCTCTGTGCTGCCATCAGACC
# AGGACACT
# >47432{HX13F;18709;47
# --------TCGAAGAACTTGCTGTAGCGCCTGAAGATGACCTCAGTGCTGCCATCAGACC
# AGGACAC-
# >AS_FamilyB_51928_1
# --TGCAGGTCGAAGAACTTGCTGTAGCGCCTGAAGATGACCTCAGTGCTGCCATCAGACC
# AG------
# >Ot000084
# --TGCAGGTCGAAGAACTTGCTGTAGCGCCTGAAGATGACCTCAGTGCTGCCATCAGACC
# AG------
# one sequence per fish

# OUTPUT FILE FORMAT:
# just the bases for the variants identified by the program
# >CL_R05330_MILLER
# T
# >Lakewhitefish_137535
# T
# >47432{HX13F;18709;47
# A
# >AS_FamilyB_51928_1
# A
# >Ot000084
# A
```

```
# HOW TO RUN SCRIPT:
# Requirements: Python 2.6 and above.
#             : Not compatible with python3
#             : Libraries : optparse for command line parsing
# Run on the command line as:
# [vlsgonen@ris-lx01 delete_locus_testing]$ python \
# ~/PyCharm_Projects/RAD_seq/muscle_parser_find_variants.py
# --alignment_file Locus_1.afa> Locus_1.variants

#################################################################

# SCRIPT

#################################################################

# NECESSARY IMPORTS

if __name__ == "__main__":
    from optparse import OptionParser
    parser = OptionParser()
    # .afa file from muscle
    parser.add_option("--alignment_file", dest = \
                                            "alignment_file")
   (options, args) = parser.parse_args()


#################################################################

# OBJECT ORIENTED CODE

# class FileParser
# parses input file
# makes defline2seq dictionary which is just a dict
# representation of input file
class FileParser(object):

    def __init__(self):
        self.alignment_input = open(options.alignment_file, \
                                    'r').read().splitlines()
        # dict representation of input file
        self.defline2seq = {}
        # to retain order of sequences as read from file
        self.defline_list = []
        # need to know sequence length to give restriction on
        # the while loop which interrogates
        # the sequence base by base for variants -otherwise we
        # would be here all day...
        self.seqlength = 0
```

```
# method parseLines
    # is the file reader
    def parseLines(self):
        defline = ""
        sequence = ""
        for line in self.alignment_input:
            if line.startswith(">"):
                self.defline_list.append(line)
                if defline != "":
                    self.defline2seq[defline] = sequence
                defline = line # assigns to new defline
                sequence = "" # resets
            else:
# to account for sequence being over multiple lines
                sequence = sequence + line
# add last sequence to dictionary
        self.defline2seq[defline] = sequence
# get last sequence length
        self.seqlength = len(sequence)


# class Base
# takes in a dict of defline to base for that defline at a
# given position and assess base for a variant
# returns T/F i.e. whether base is variant across species
class Base(object):

    def __init__(self, def2variant):
        self.defline2variant = def2variant
        # unique the bases across the fish.
        # Bases are dict values - obtain as a list by doing:
        # self.defline2variant.values
        # and unique by calling set() on the list
        self.bases = set(self.defline2variant.values())
        # originally I thought I could work out whether a base
        # was a variant by just returning true if
        # len(self.bases) > 1. But if some are [ATCG] and
        # others are "-" for alignment gaps then
        # this would return true when I don't want it to,
        # therefore I need to check that this is a good
        # base and has no gaps. Hence the method below
        self.good_base = self.noGaps()

    # method noGaps
    # see explanation and reasoning above, where it is called
    def noGaps(self):
        if "N" in self.bases:
            self.bases.remove("N")
        if "-" not in self.bases:
            return True

    # method isVariant
    def isVariant(self):
        # if the set length > 1 and there are no gaps
        if (len(self.bases) > 1) and self.good_base:
            return True
        else:
            return False
```

```
# class SequenceParser
# loop through base, makes dict defline2base, passes to Base
# class if variant, append base to defline for fish in
# self.defline2variants dict
class SequenceParser(object):

    def __init__(self, def2seq, seqlength):
        self.defline2sequence = def2seq
        self.defline2variants = {}
        self.seqlength = seqlength

    # method addVariant
    # if variant, this method is called by parseSeqs method
    # to append base to the list of variants for each defline
    def addVariant(self, defline2base):
        for defline, base in defline2base.items():
            if defline not in self.defline2variants:
                self.defline2variants[defline] = []
            self.defline2variants[defline].append(base)

    def parseSeqs(self):
        count = 0 # keep track of base position.
        while count < self.seqlength:
            defline2base = {}
            for defline, sequence in \
                            self.defline2sequence.items():
                # get base at that position
                base = sequence[count]
                defline2base[defline] = base
            if Base(defline2base).isVariant():
               self.addVariant(defline2base)
               count += 1 # next base!

# class Combiner
class Combiner(object):

    def __init__(self):
        self.file_object = FileParser()

    def getObjects(self):
        self.file_object.parseLines() # parse input file
        self.defline2seq = self.file_object.defline2seq
# get defline_list for maintaining order of seqs from in->out
        self.defline_list = self.file_object.defline_list
        self.seqlength = self.file_object.seqlength
        self.sequence_parser_object = /
        SequenceParser(self.defline2seq, self.seqlength)
        self.sequence_parser_object.parseSeqs()
        self.defline2variants = \
                 self.sequence_parser_object.defline2variants
```

```
    def runScript(self):
        self.getObjects()
        if self.defline2variants != {}:
            for defline in self.defline_list:
                print defline
                print ''.join(self.defline2variants[defline])

######################################################################

# PROCEDURAL CODE
Combiner().runScript()

# DONE


######################################################################
```

variant_concatenator_post_muscle.py

```
__author__ = 'Serap_Gonen'

# START DATE: 20/05/14
# END DATE: 20/05/14

# SCRIPT DESCRIPTION
# takes in directory path and file extension
# looks in the directory and parses all files with a given
# extension
# concatenates variants identified for each fish across
# multiple loci, each of which are in separate files, and
# prints file with sequence of concatenated variants per fish

# INPUT FILE FORMAT:
# each parsed file looks something like this:
# eg locus 1, 1 variant
# [vlsgonen@ris-lx01 delete_locus_testing]$
#                                         cat Locus_1. variants
# >CL_R05330_MILLER
# T
# >Lakewhitefish_137535
# T
# >47432{HX13F;18709;47
# A
# >AS_FamilyB_51928_1
# A
# >0t000084
# A
#
# eg locus 3, 3 variants
# [vlsgonen@ris-lx01 delete_locus_testing]$
#                                       cat Locus_3. variants
# >9438{HX13F;85241;12
# GAA
# >Lakewhitefish_129860
# AGG
# >AS_FamilyB_4730_1
# AGG
# >CL_R37048_MILLER
# AGA
# >0t020573
# AGA
```

```
# OUTPUT FILE FORMAT:
# based on locus 1 and 3 only and in that order:
# >Chinook_Salmon_variants
# AAGA
# >Atlantic_Salmon_variants
# AAGG
# >Rainbow_Trout_variants
# TAGA
# >Lake_Whitefish_variants
# TAGG
# >Sockeye_Salmon_variants
# AGAA

# HOW TO RUN SCRIPT:
# Requirements: Python 2.6 and above.
#             : Not compatible with python3
#             : Libraries : os for parsing files in a given
#                           directory path
#                         : optparse to get OptionParser for
#                           parsing command line inputs
# Run on the command line as:
# [vlsgonen@ris-lx01 delete_locus_testing]$ python \
# ~/PyCharm_Projects/RAD_seq/
# \variant_concatenator_post_muscle.py \
# --directory_path /nfs_netapp/vlsgonen/delete_locus_testing \
# --file_extension .variants

#################################################################

# SCRIPT

#################################################################

# NECESSARY IMPORTS
if __name__ == "__main__":
    import os
    from optparse import OptionParser
    parser = OptionParser()
    # needs full path from root
    parser.add_option("--directory_path", dest = "path")
    # all files to be parsed must have same extension
    parser.add_option("--file_extension", \
                                    dest = "file_extension")
    (options, args) = parser.parse_args()

#################################################################
```

```
# OBJECT ORIENTED CODE

# class Identify_Fish
# takes in a defline, identifies which fish it belongs to
# based on patters in the defline for each fish, returns fish
# notice that trout is not in dict - it is the else: because
# seqs come from different populations
class Identify_Fish(object):

    def __init__(self):
        # dict of pattern to fish
        # fish matches key in dict in Combiner object so that
        # sequence concatenation is made easy
        self.fish_pattern2fish_name = \
                            {"Ot":">Chinook_Salmon_variants",
                    Lakewhitefish":">Lake_Whitefish_variants",
                     "AS_FamilyB":">Atlantic_Salmon_variants",
                          "(HX13F":">Sockeye_Salmon_variants",
                       "Halibut":">Atlantic_Halibut_variants",
                       "RADid":">Baltic_Sea_Herring_variants",
                            "rgn":">Gudgeon_variants",
                            "Gar":">Spotted_Gar_variants",
                            "chr":">Stickleback_variants"
                                    }

    # method returnFish
    # takes in a defline
    # identify which pop it belongs to- pattern matching
    # returns fish if in dict
    # otherwise it returns default which is trout, SO CAUTION
    # REQUIRED - SCRIPT IS VERY SPECIFIC
    def returnFish(self, fish_defline):
        fish_return = ">Rainbow_Trout_variants"
        for pattern, fish in \
                            self.fish_pattern2fish_name.items():
            if pattern in fish_defline:
                fish_return = fish
        return fish_return

# class FileParser
# takes in files one at a time as well as a global dictionary
# which contains all variants from all files
class FileParser(object):

    def __init__(self, file, fish2variantseq):
        self.file = open(file, 'r').read().splitlines()
        self.fish2variantseq = fish2variantseq
```

```
    def parseLines(self):
        if self.file:
            fish = Identify_Fish()
            current_fish = ""
            for line in self.file:
                if line.startswith(">"):
                # get the fish that the variant belongs to
                    current_fish = fish.returnFish(line)
                else:
                  # add that variant to the fish before
                   self.fish2variantseq[current_fish]. \
                                            append(line)
        return self.fish2variantseq


# class PrintSequences
# does the printing, via the doPrinting method
class PrintSequences(object):

    def __init__(self, fish2variantseq):
        self.fish2variantseq = fish2variantseq

    def doPrinting(self):
        for fish, sequence in self.fish2variantseq.items():
            print fish
            print ''.join(sequence)
# concatenate all variants into one output file
# order very much dependent on order that files are parsed

# class Combiner
# responsible for running the script
class Combiner(object):

    def __init__(self):
        self.path_to_directory = options.path
        self.fish2variantseq = {
                            ">Rainbow_Trout_variants" : [],
                            ">Sockeye_Salmon_variants" : [],
                            ">Atlantic_Salmon_variants" : [],
                            ">Chinook_Salmon_variants" : [],
                            ">Lake_Whitefish_variants" : [],
                            ">Atlantic_Halibut_variants" : [],
                           ">Baltic_Sea_Herring_variants" : [],
                            ">Gudgeon_variants" : [],
                            ">Spotted_Gar_variants" : [],
                            ">Stickleback_variants" : []
                            }
    # method runScript
    # runs script by looping over files in the directory
    # with the correct extension given on command line
    def runScript(self):
        for file in os.listdir(self.path_to_directory):
            if file.endswith(options.file_extension):
                self.fish2variantseq = FileParser(file, \
                            self.fish2variantseq).parseLines()
        PrintSequences(self.fish2variantseq).doPrinting()
```

```
#################################################################

# PROCEDURAL CODE

# go script!
Combiner().runScript()

# DONE

#################################################################
```

# Appendix I – Parameters for phylogenetic tree construction using RAxML V 8.1.13

Reference: Stamatakis, A., 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics. 30(9):1312–1313.
Version: RAxML 8.1.13, released by Alexandros Stamatakis on 16th December 2014. The command line version of RAxML was used in this analysis. The steps followed for tree construction, and a brief description of input parameters, is given below.

**Steps:**

| Step | Command | Output files |
|---|---|---|
| 1) Obtain most parsimonious tree from the input data | /path/to/raxml<br>-m ASC_GTRGAMMA<br>-p 123456<br>-s file_input.phylip<br>-n Tree1<br>-N 10000<br>-f o<br>--asc-corr=lewis | a) One output file for each run<br>b) RAxML_bestTree.Tree1 file, which is the tree with the maximum likelihood, given the input data<br>c) RAxML_info.Tree1 file, with logs of output to terminals |
| 2) Bootstrap to obtain estimates of tree confidence | /path/to/raxml<br>-m ASC_GTRGAMMA<br>-b 123456<br>-p 123456<br>-s file_input.phylip<br>-n Tree2<br>-N 10000<br>-f o<br>--asc-corr=lewis<br>-k | a) RAxML_bootstrap.Tree2 which contains the values from the bootstrap runs<br>b) RAxML_info.Tree2 |
| 3) Check if a sufficient number of bootstraps were performed | /path/to/raxml<br>-m ASC_GTRGAMMA<br>-z RAxML_bootstrap.Tree2<br>-I autoMRE<br>-n Tree3<br>--asc-corr=lewis<br>-p 123456 | Screen output only |
| 4) Write tree support values from bootstrapping on to most parsimonious tree from step 1 | /path/to/raxml<br>-m ASC_GTRGAMMA<br>-p 123456<br>-t RAxML_bestTree.Tree1<br>-z RAxML_bootstrap.Tree2<br>-n Tree4<br>-N 10000<br>-f b<br>--asc-corr=lewis | a) RAxML_bipartitions.Tree4 which contains the node supports<br>b) RAxML_bipartitionsBranchLabels.Tree 4 which contains support values on nodes and branches |

**Input parameter descriptions:**

| Parameter | Option | Description |
|---|---|---|
| -m | ASC_ GTRGAMMA | Model for estimating tree |

| | | |
|---|---|---|
| | | parameters |
| -p | 123456 | This can take any value. It is a way of ensuring that parameter estimations start from the same value (to make results reproducible) |
| -b | 123456 | Specifies the requirement for bootstrapping. This can take any value and allows reproducibility of runs using the input integer as a starting value |
| -f | b | Used to draw bipartitions on an input tree specified using -t, using the bootstrap tree parameters file specified in -z |
| | o | Specifies use of older (and slower) algorithm to obtain log likelihoods. Estimates obtained using this algorithm are thought to be typically better |
| -N | 10000 | Executes 10,000 maximum likelihood searches, using 10,000 different starting trees |
| -n | TreeX | Output file name extension |
| -t RAxML_bestTree.Tree1 | | Most parsimonious tree from the input data produced in step 1. A user specified tree can also be given as input using this parameter, resulting in bootstrap results from step two being used to obtain support for the user specified tree instead |
| -z RAxML_bootstrap.Tree2 | | Output file with bootstrap statistics |
| --asc-corr | lewis | Standard Lewis correction for ascertainment bias correction due to use of between species variants |
| -k | N/A | Bootstrapped trees will be printed with branch lengths |
| -l | autoMRE | Option to check for bootstrap convergence in step 3 |

Trees in the resulting tree in the RAxML_bipartitionsBranchLabels.Tree4 was drawn using the Phylodendron (http://iubio.bio.indiana.edu/treeapp/treeprint-form.html), T-REX (Boc et al, 2012), or Archaeopteryx (Han and Zmasek, 2009) software packages.

Further details for running analyses using RAxML can be found in the manual: http://sco.h-its.org/exelixis/resource/download/NewManual.pdf.

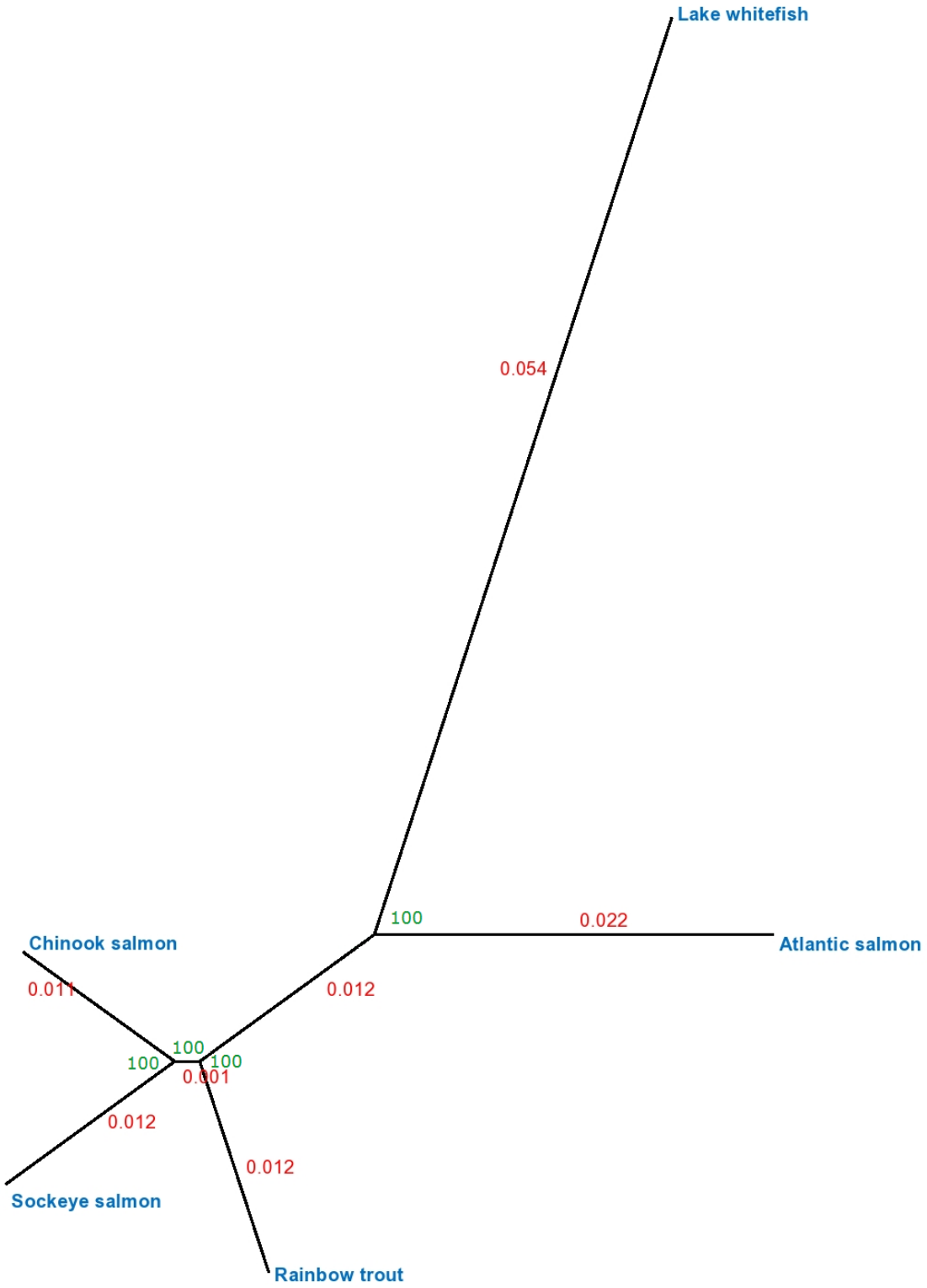
# Appendix J – Phylogenetic trees

Evolutionary relationships amongst ten teleost fish species were reconstructed based on RAD-Seq data, using the RAxML software package (version 8; see Appendix I).

Trees were visualised using one of the three following software packages:

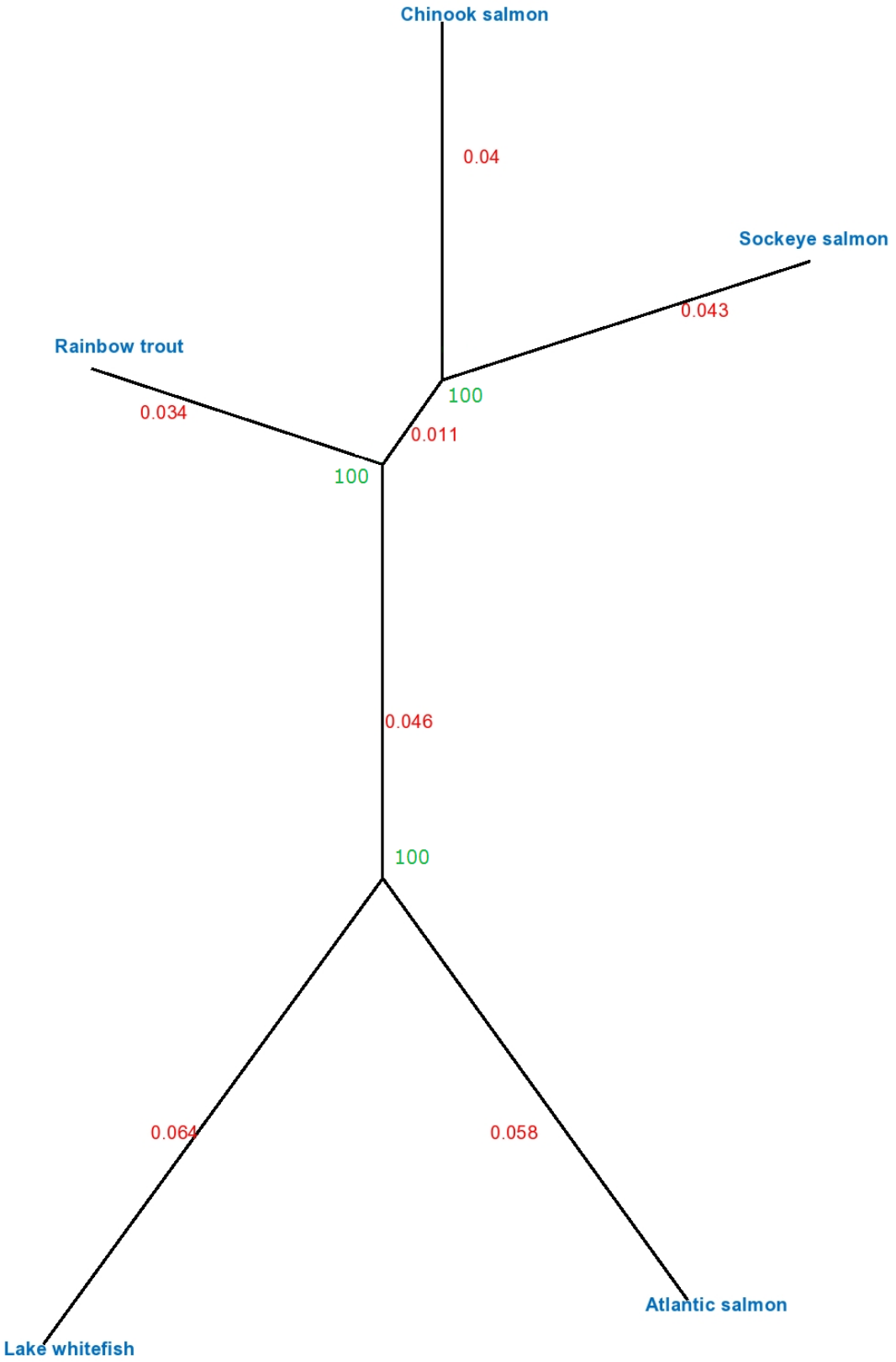
- Phylodendron (http://iubio.bio.indiana.edu/treeapp/treeprint-form.html)
- T-REX (Boc et al, 2012)
- Archaeopteryx (Han and Zmasek, 2009)

**Salmonid species - all RAD loci have complete information** **TREE 1**

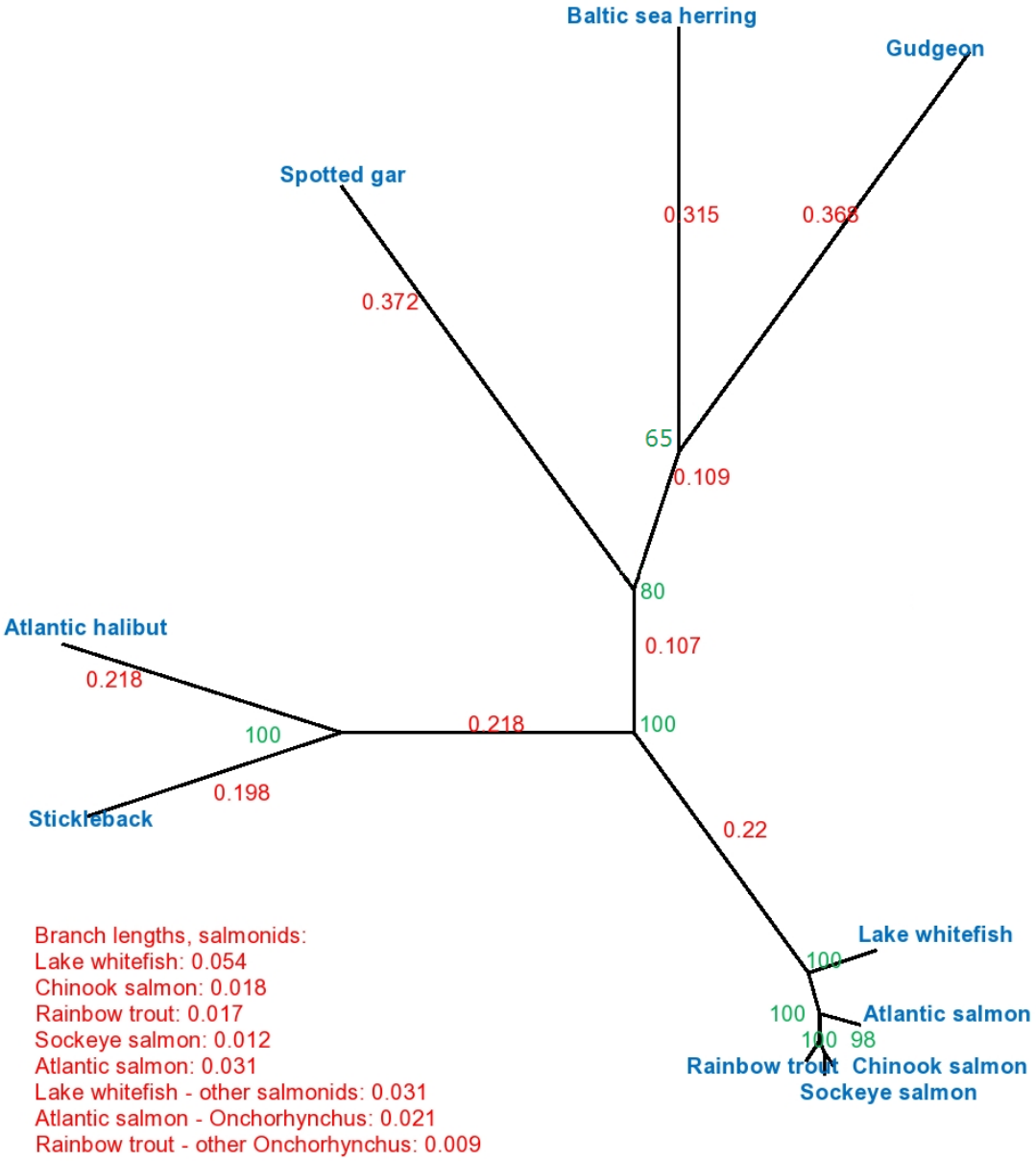


**Salmonid species - all RAD loci have complete information for at least three of five species**

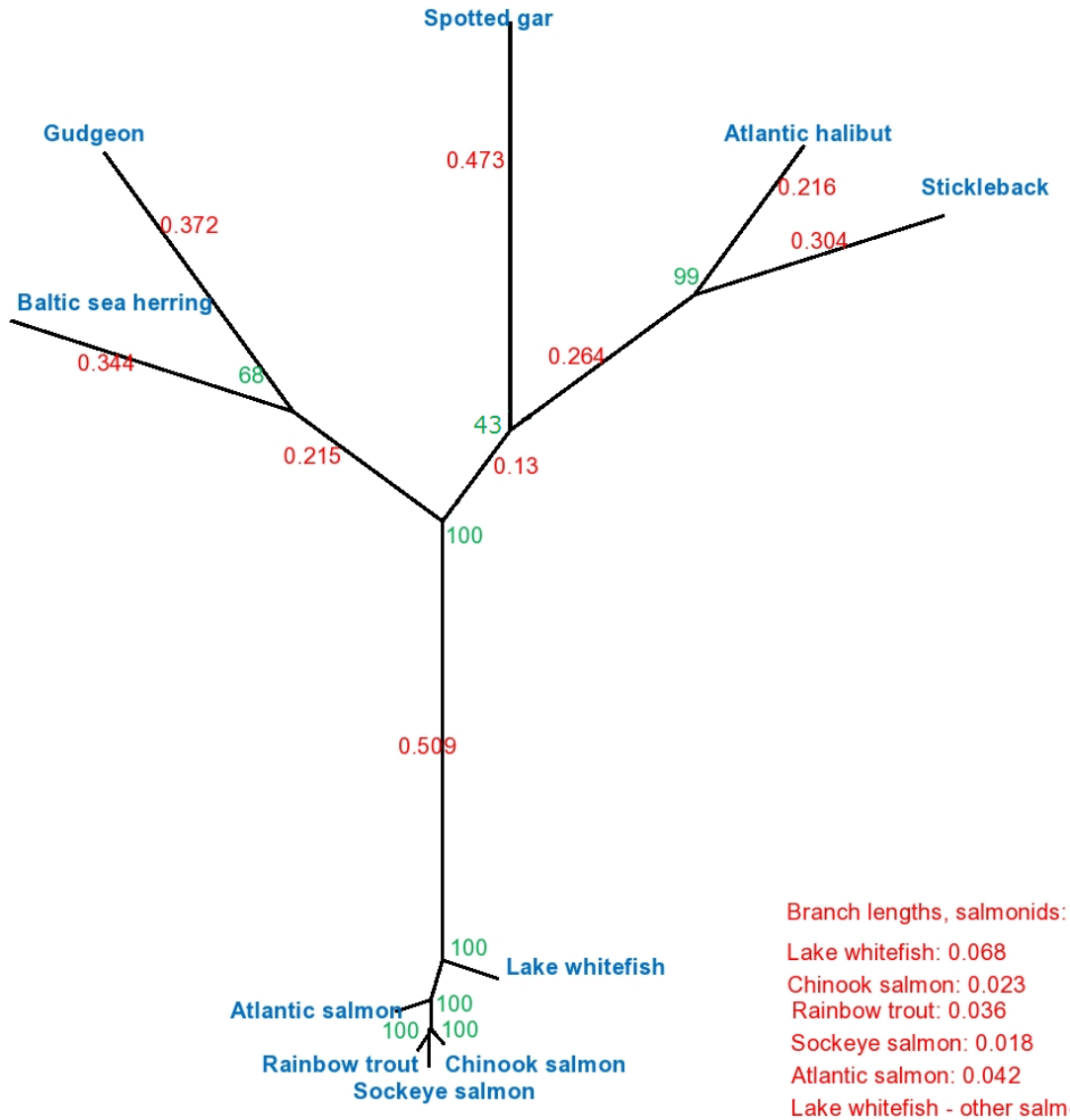
**TREE 2**



**All fish (unrooted) - BLASTN clusters - RAD loci have complete information for at least 7/10 fish** **TREE 3**



**All fish (unrooted) - BLASTN clusters - RAD loci have complete information for at least 5/10 fish** **TREE 4**



# Appendix K – Synteny tables

## Salmonids only (strict analysis parameters)

**Table J1: Chinook salmon & Lake whitefish**

| Chinook salmon | Lake whitefish | Number of RAD loci |
|---|---|---|
| 30 | 8 | 3 |

**Table J2: Chinook salmon & Atlantic salmon**

| Chinook salmon | Atlantic salmon | Number of RAD loci |
|---|---|---|
| 1 | 17 | 2 |
| 1 | 18 | 2 |
| 2 | 6 | 2 |
| 3 | 1 | 3 |
| 7 | 32 | 2 |
| 8 | 10 | 3 |
| 10 | 3 | 4 |
| 13 | 15 | 3 |
| 14 | 10 | 2 |
| 18 | 4 | 2 |
| 20 | 17 | 3 |
| 26 | 14 | 5 |
| 28 | 11 | 3 |
| 29 | 30 | 4 |
| 30 | 28 | 2 |

**Table J3: Sockeye salmon & Atlantic salmon**

| Sockeye salmon | Atlantic salmon | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J4: Atlantic salmon & Lake whitefish**

| Atlantic salmon | Lake whitefish | Number of RAD loci |
|---|---|---|
| 2 | 6 | 2 |
| 21 | 4 | 2 |

**Table J5: Chinook salmon & Sockeye salmon**

| Chinook salmon | Sockeye salmon | Number of RAD loci |
|---|---|---|
| 4 | 1 | 2 |
| 8 | 4 | 2 |

**Table J6: Sockeye salmon & Lake whitefish**

| Sockeye salmon | Lake whitefish | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J7: Rainbow trout & Chinook salmon**

| Rainbow trout | Chinook salmon | Number of RAD loci |
|---|---|---|
| 8 | 5 | 2 |
| 11 | 12 | 2 |

**Table J8: Rainbow trout & Atlantic salmon**

| Rainbow trout | Atlantic salmon | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J9: Rainbow trout & Sockeye salmon**

| Rainbow trout | Sockeye salmon | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J10: Rainbow trout & Lake whitefish**

| Rainbow trout | Lake whitefish | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

## All fish (relaxed analysis parameters)

**Table J11: Chinook salmon & Sockeye salmon**

| Chinook salmon | Sockeye salmon | Number of RAD loci |
|---|---|---|
| 4 | 1 | 2 |
| 8 | 4 | 3 |
| 29 | 28 | 2 |

**Table J12: Chinook salmon & Lake whitefish**

| Chinook salmon | Lake whitefish | Number of RAD loci |
|---|---|---|
| 4 | 34 | 2 |
| 5 | 36 | 2 |
| 5 | 40 | 2 |
| 6 | 5 | 2 |
| 6 | 11 | 2 |
| 8 | 16 | 2 |
| 21 | 29 | 2 |
| 22 | 8 | 2 |
| 25 | 15 | 2 |
| 27 | 18 | 2 |
| 30 | 37 | 5 |
| 31 | 24 | 3 |
| 34 | 9 | 2 |

**Table J13: Chinook salmon & Atlantic salmon**

| Chinook salmon | Atlantic salmon | Number of RAD loci |
|---|---|---|
| 1 | 17 | 3 |
| 1 | 18 | 3 |
| 2 | 6 | 8 |
| 3 | 1 | 5 |
| 3 | 20 | 2 |
| 4 | 7 | 4 |
| 5 | 8 | 4 |
| 6 | 16 | 7 |
| 6 | 23 | 3 |
| 7 | 22 | 3 |
| 7 | 23 | 3 |
| 7 | 32 | 4 |
| 8 | 10 | 7 |
| 9 | 5 | 7 |
| 9 | 11 | 2 |
| 10 | 3 | 4 |
| 11 | 17 | 2 |
| 12 | 8 | 2 |
| 12 | 9 | 4 |
| 13 | 15 | 5 |
| 14 | 10 | 2 |
| 15 | 24 | 3 |
| 16 | 8 | 3 |
| 16 | 13 | 2 |
| 18 | 4 | 3 |
| 19 | 2 | 2 |
| 20 | 17 | 6 |
| 22 | 5 | 5 |
| 23 | 12 | 2 |
| 25 | 13 | 2 |
| 25 | 31 | 2 |
| 26 | 14 | 7 |
| 27 | 4 | 2 |
| 28 | 11 | 2 |
| 29 | 30 | 8 |
| 30 | 28 | 3 |
| 31 | 3 | 2 |
| 32 | 1 | 3 |
| 32 | 6 | 2 |
| 33 | 9 | 4 |

**Table J14: Rainbow trout & Chinook salmon**

| Rainbow trout | Chinook salmon | Number of RAD loci |
|---|---|---|
| 8 | 5 | 2 |
| 11 | 12 | 2 |
| 14 | 31 | 2 |

**Table J15: Sockeye salmon & Atlantic salmon**

| Sockeye salmon | Atlantic salmon | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J16: Atlantic salmon & Lake whitefish**

| Atlantic salmon | Lake whitefish | Number of RAD loci |
|---|---|---|
| 5 | 13 | 2 |
| 9 | 4 | 2 |
| 9 | 33 | 2 |
| 10 | 28 | 2 |
| 10 | 31 | 2 |
| 19 | 9 | 2 |
| 21 | 4 | 4 |
| 23 | 11 | 3 |

**Table J17: Rainbow trout & Atlantic salmon**

| Rainbow trout | Atlantic salmon | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J18: Rainbow trout & Sockeye salmon**

| Rainbow trout | Sockeye salmon | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J19: Rainbow trout & Lake whitefish**

| Rainbow trout | Lake whitefish | Number of RAD loci |
|---|---|---|
| 25 | 16 | 2 |

**Table J20: Sockeye salmon & Lake whitefish**

| Sockeye salmon | Lake whitefish | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J21: Three-spined stickleback & Chinook salmon**

| Three-spined stickleback | Chinook salmon | Number of RAD loci |
|---|---|---|
| 2 | 12 | 2 |
| 5 | 24 | 3 |
| 11 | 9 | 3 |

**Table J22: Three-spined stickleback & Atlantic salmon**

| Three-spined stickleback | Atlantic salmon | Number of RAD loci |
|---|---|---|
| 2 | 8 | 2 |
| 11 | 11 | 4 |

**Table J23: Three-spined stickleback & Sockeye salmon**

| Three-spined stickleback | Sockeye salmon | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J24: Three-spined stickleback & Lake whitefish**

| Three-spined stickleback | Lake whitefish | Number of RAD loci |
|---|---|---|
| 7 | 8 | 2 |
| 10 | 24 | 2 |
| 12 | 27 | 2 |

**Table J25: Three-spined stickleback & Rainbow trout**

| Three-spined stickleback | Rainbow trout | Number of RAD loci |
|---|---|---|
| NO MATCHES | | |

**Table J26: Chinook salmon & Gudgeon**

| Chinook salmon | Gudgeon | Number of RAD loci |
|---|---|---|
| NO MATCHES | | |

**Table J27: Chinook salmon & Atlantic halibut**

| Chinook salmon | Atlantic halibut | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J28: Gudgeon & Atlantic halibut**

| Gudgeon | Atlantic halibut | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J29: Three-spined stickleback & Gudgeon**

| Three-spined stickleback | Gudgeon | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J30: Gudgeon & Atlantic salmon**

| Gudgeon | Atlantic salmon | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J31: Atlantic salmon & Atlantic halibut**

| Atlantic salmon | Atlantic halibut | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J32: Sockeye salmon & Gudgeon**

| Sockeye salmon | Gudgeon | Number of RAD loci |
|---|---|---|
| NO MATCHES | | |

**Table J33: Sockeye salmon & Atlantic halibut**

| Sockeye salmon | Atlantic halibut | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J34: Rainbow trout & Gudgeon**

| Rainbow trout | Gudgeon | Number of RAD loci |
|---|---|---|
| NO MATCHES | | |

**Table J35: Rainbow trout & Atlantic halibut**

| Rainbow trout | Atlantic halibut | Number of RAD loci |
|---|---|---|
| NO MATCHES | | |

**Table J36: Lake whitefish & Gudgeon**

| Lake whitefish | Gudgeon | Number of RAD loci |
|---|---|---|
| NO MATCHES | | |

**Table J37: Lake whitefish & Atlantic halibut**

| Lake whitefish | Atlantic halibut | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J38: Gudgeon & Spotted gar**

| Gudgeon | Spotted gar | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J39: Three-spined stickleback & Atlantic halibut**

| Three-spined stickleback | Atlantic halibut | Number of RAD loci |
|---|---|---|
| 1 | 12 | 5 |
| 1 | 17 | 3 |
| 2 | 24 | 5 |
| 3 | 18 | 9 |
| 4 | 16 | 3 |
| 4 | 19 | 7 |
| 5 | 2 | 12 |
| 6 | 15 | 8 |
| 7 | 6 | 2 |
| 7 | 23 | 6 |
| 8 | 9 | 12 |
| 9 | 10 | 14 |
| 10 | 3 | 8 |
| 11 | 21 | 8 |
| 12 | 11 | 5 |
| 13 | 7 | 8 |
| 14 | 13 | 4 |
| 15 | 20 | 11 |
| 16 | 1 | 5 |
| 17 | 14 | 17 |
| 19 | 4 | 7 |
| 20 | 22 | 7 |
| 21 | 8 | 6 |

**Table J40: Chinook salmon & Spotted gar**

| Chinook salmon | Spotted gar | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J41: Atlantic salmon & Spotted gar**

| Atlantic salmon | Spotted gar | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J42: Sockeye salmon & Spotted gar**

| Sockeye salmon | Spotted gar | Number of RAD loci |
|---|---|---|
| ALL MATCHES SUPPORTED BY A SINGLE MAPPED RAD LOCUS | | |

**Table J43: Rainbow trout & Spotted gar**

| Rainbow trout | Spotted gar | Number of RAD loci |
|---|---|---|
| NO MATCHES | | |

**Table J44: Lake whitefish & Spotted gar**

| Lake whitefish | Spotted gar | Number of RAD loci |
|---|---|---|
| NO MATCHES | | |

**Table J45: Three-spined stickleback & Spotted gar**

| Three-spined stickleback | Spotted gar | Number of RAD loci |
|---|---|---|
| 12 | 25 | 2 |

**Table J46: Atlantic halibut & Spotted gar**

| Atlantic halibut | Spotted gar | Number of RAD loci |
|---|---|---|
| NO MATCHES | | |

# Acknowledgements

So I guess this section is mine, to thank those who have supported me throughout my four years as a PhD student. There's always a whole plethora of people to thank, and my acknowledgements list is no different. I genuinely couldn't have completed this thesis without the help and support of all the people mentioned, and I am so lucky and grateful to have all of them in my life.

First, I would like to thank my primary and secondary supervisors, **Dr Ross Houston** and **Professor Stephen Bishop**. Without both of their huge and extraordinary (definitely not ordinary) scientific minds and patience in teaching me new things, I wouldn't have gotten very far based on my undergraduate knowledge alone. They taught me patience (although I am still quite impatient sometimes, so maybe they failed), and dealt with my tantrums (they weren't that frequent...). When they couldn't teach me anymore (or when they wanted to be rid of me for a while), they sent me on training courses and conferences. I think I am one of the rare luck PhD students who has been all over the world (sorry for the negative budget)! They never tired of reading, re-reading, and re-reading again the drafts of manuscripts, chapters and presentations that I sent them, and I am eternally grateful for their time.

As well as my supervisors, I also have to thank my self-appointed mentor and friend, **Dr Andy Law**, without whom I would have failed my PhD, or quit and gone home. First, for tolerating my annoying visits and for not sending me away too often! Second, for teaching me to write code in Python and Linux, for guiding me to work it out for myself, and for not giving up until I understood. For listening and encouraging, and providing me with a space to sit quietly. He is my scientific idol, and I wish I could think, analyse, write and design code as he can. I've always felt indebted, and I will never be able to repay it. Remember Andy, "Linux is your friend".

I also owe a huge thank you to my thesis committee members **Dr Bob Dalziel** and **Professor Alan Archibald** for their input to the project, for listening to my presentations at committee meetings, for reading my reports (even if it was just as you were walking in to the meeting Alan), and for not giving me too much of a hard time! I also owe Bob a huge thank you for his help outside of committee meetings, for his advice on PhD related and other private discussions, and for always making me smile and laugh.

Similarly, not directly related to my project, but always ready to save a damsel in distress, **Dr John Taggart** (Institute of Aquaculture, Stirling) was always on hand, offering his help and support whether I required it or not. He is one of the legends on my short list. I was always made to feel welcome at Stirling, and loved my visits there!

I also need to thank the IT crowd at Roslin: **Matt**, **Anthony**, **Barry**, **Laurie** and **Stuart**, for bending over backwards to provide me with the resources that I needed finish this PhD, and for fixing my computing problems! Sorry for the number of times I broke my computer/KOed the system... Thank you for the banter and friendship.

Throughout my PhD I got to visit many places for conferences and courses, none of which would have been possible without our funding bodies. I would like to thank and

acknowledge funding from the **BBSRC** and **Roslin Institute** for my PhD project. I would also like to thank the following funders for travel awards: 1) **BSAS** travel award, 2) **Workshop on Genomics** travel award, 3) **Birrell Grey** vet school travel award.

Lastly (I saved the best for last!), I have a huge list of friends and family that I need to thank, both in London and in Edinburgh.

First, my dearest **Ozzie Matika**, for the laughter, the encouragement, and for putting up with my miserable attitude, particularly towards the end of my PhD. For all the help in ASReml. But most importantly for the fatherly hugs, which I will miss the most.

My baby homefry, **Andrew Mason**. It took me less than a day to like him, and even less to trust him. I genuinely thought at one point: if I fail my PhD, it'll be because this homefry won't stop making me laugh. Thank you for making the last year and a bit of my PhD the best part. For the help in programming, software use, presentation edits, and project ideas. For the support and encouragement , the fun times (there are too many to list), for the drunk face (you know the one), for teasing me and for showing me that genuinely good hearted people still exist in the world, even if they are in Yorkshire... But most of all, I thank him for his honesty.

My family away from home, **Valentina Riggio** and **Ricardo Pong-Wong**. I've always looked up to them both and taken them as role models (ok maybe not Ricardo...). For the constant love, care and support, the morning and random daytime chats, for making me laugh, and for always confirming that I am not the weird one at Roslin, it's everyone else! Special thanks to Ricardo for shredding my WCGALP presentation – I needed the honesty and appreciated the improvements, even though I didn't say so at the time! For being true friends.

**William Ho** and **Stacey Human**, thank you for the amazing friendship we shared together, and for staying on my case and not leaving me alone. **Niruja Balaskandan**, for being my oldest friend, and the only person that continuously kept in contact after college. **Jason Ioannidis**, my Greek homie, for the friendship, support and sound advice, and for helping me train for my half-marathons and making me eat properly! My three girls in London, **Erica Lam**, **Christina Sakellariou** and **Upekha Karunarathna**. We might meet up once a year but it's like we never left each other. We laughed and cried through our undergraduate years. They say your best friends are found at university, and it's true. Please always be in my life!

Finally, the three shining stars of my life.

My grandmother **Muzehher Salahi**, for always calling me up to check how I am and whether the cold weather in Edinburgh has killed me yet! I thank her for being there for us as our second mother, for taking care of us, and for being so patient with me during my studies. For letting me go and tolerating my weekend visits when I worked rather than spent time with the family – for that, I am sorry.

My amazing sister **Sevda Gonen**, my rock, my secret keeper and my best friend. I thank her for the support through the tough and fun times, and encouragement in everything that I did. I am so lucky to have her in my life. I thank her for putting up with my disorganisation and for sorting things out in London when I was away, for always putting a smile on my face and for checking up on me and forcing me to talk, and then

listening and giving me advice. The most gorgeous chick, talented performer, and rising star that I have ever known.

Lastly, my mother, **Fatma Gonen**. I thank her for the sacrifices that she made to make me who I am today. For the care, attention, and constant support. For being understanding and for listening, and for giving me the chance to make my dreams come true. For the advice, for making me smile and for the old grandma medicinal remedies that always seem to make me recover from whatever illness I may have. Most importantly, I thank her for the unconditional love and safety throughout my life. And for the amazing food, that I dreamt about every night of my PhD! I don't know what I would do without her.

**Annem, I hope that I have made you proud, and that your efforts in bringing me up have not gone to waste. It's not much, but this thesis is dedicated to you...**

# Bibliography


Aas TS, Grisdale-Helland B, Terjesen BF, Helland SJ (2006). Improved growth and nutrient utilisation in Atlantic salmon (*Salmo salar*) fed diets containing a bacterial protein meal. *Aquaculture* 259**:** 365-376.

Agostino M (2012). Introduction to the BLAST Suite and BLASTN reference. In: *Practical Bioinformatics*. New York: Garland Science. 47-71.

Akhlaghi M, Munday BL, Rough K, Whittington RJ (1996). Immunological aspects of amoebic gill disease in salmonids. *Dis Aquat Org* 25**:** 23-31.

Allen J, Scott D, Illingworth M, Dobrzelecki B, Virdee D, Thorn S *et al* (2012). CloudQTL: Evolving a bioinformatics application to the cloud. *Digital Research 2012, September 10-12, 2012 Oxford, UK*.

Allendorf FW, Danzmann RG (1997). Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. *Genetics* 145**:** 1083-1092.

Altschul SF, Gish W Fau - Miller W, Miller W Fau - Myers EW, Myers Ew Fau - Lipman DJ, Lipman DJ (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215**:** 403-410.

Amish SJ, Hohenlohe PA, Painter S, Leary RF, Muhlfeld C, Allendorf FW *et al* (2012). RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays. *Molecular Ecology Resources* 12**:** 653-660.

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011). Genome evolution and meiotic maps by massively parallel DNA sequencing: Spotted gar, an outgroup for the teleost genome duplication. *Genetics* 188**:** 799-808.

Amores A, Catchen J, Nanda I, Warren W, Walter R, Schartl M *et al* (2014). A RAD-Tag genetic map for the Platyfish (*Xiphophorus maculatus*) reveals mechanisms of karyotype evolution among teleost fish. *Genetics* 197**:** 625-641.

Anderson EC (2010). Computational algorithms and user-friendly software for parentage-based tagging of Pacific salmonids. Final report submitted to the Pacific Salmon Commission's Chinook Technical Committee (US Section). 46 p

Andreassen R, Lunner S, Hoyheim B (2010). Targeted SNP discovery in Atlantic salmon (*Salmo salar*) genes using a 3' UTR-primed SNP detection approach. *Bmc Genomics* 11.

AquaGen (2013). *Use of QTL-eggs results in an IPN reduction for the whole of Norway.* [online] Available at: <http://aquagen.no/wp-content/uploads/2013/06/01-2013-Use-of-QTL-eggs-results-in-an-IPN-reduction-for-the-whole-of-Norway.pdf> [Accessed 29 October 2014].

Arakawa Y, Cordeiro JV, Way M (2007). F11L-mediated inhibition of RhoA-mDia signaling stimulates microtubule dynamics during vaccinia virus infection. *Cell Host & Microbe* 1**:** 213-226.

Argue BJ, Arce SM, Lotz JM, Moss SM (2002). Selective breeding of Pacific white shrimp (*Litopenaeus vannamei*) for growth and resistance to Taura Syndrome Virus. *Aquaculture* 204**:** 447-460.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology* 22**:** 3179-3190.

Aslam ML, Bastiaansen JWM, Crooijmans RPMA, Vereijken A, Megens H-J, Groenen MAM (2010). A SNP based linkage map of the turkey genome reveals multiple intrachromosomal rearrangements between the turkey and chicken genomes. *Bmc Genomics* 11.

Aspenström P, Fransson A, Saras J (2004). Rho GTPases have diverse effects on the organization of the actin filament system. *Biochemical Journal* 377**:** 327-337.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA *et al* (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3.

Ballesteros NA, Rodríguez Saint-Jean S, Pérez-Prieto SI, Aquilino C, Tafalla C (2014). Modulation of genes related to the recruitment of immune cells in the digestive tract of trout experimentally infected with infectious pancreatic necrosis virus (IPNV) or orally vaccinated. *Developmental & Comparative Immunology* 44**:** 195-205.

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

Barchi L, Lanteri S, Portis E, Acquadro A, Vale G, Toppino L *et al* (2011). Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *Bmc Genomics* 12.

Barrick JE (2014). *breseq 0.25 documentation*. [online] Available at: <http://barricklab.org/twiki/pub/lab/ToolsBacterialGenomeResequencing/documentation/methods.html> [Accessed 3rd December 2014].

Barthes P, Buard J, de Massy B (2011). Epigenetic factors and regulation of meiotic recombination in mammals. In: Rousseaux S and Khochbin S (eds) *Epigenetics and Human Reproduction*, pp 119-156.

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD *et al* (2011). Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE* 6**:** e19315.

Benraiss A, Chmielnicki E, Lerner K, Roh D, Goldman SA (2001). Adenoviral Brain-Derived Neurotrophic Factor Induces Both Neostriatal and Olfactory Neuronal Recruitment from Endogenous Progenitor Cells in the Adult Forebrain. *The Journal of Neuroscience* 21**:** 6718-6731.

Bergmiller T, Ackermann M, Silander OK (2012). Patterns of evolutionary conservation of essential genes correlate with their compensability. *PLoS Genet* 8.

Bergstrom DA, Penn BH, Strand A, Perry RLS, Rudnicki MA, Tapscott SJ (2002). Promoter-Specific Regulation of MyoD Binding and Signal Transduction Cooperate to Pattern Gene Expression. *Molecular Cell* 9**:** 587-600.

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B *et al* (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun* 5.

Biering E, Villoing S, Sommerset I, Christie KE (2005). Update on viral vaccines for fish. *Developments in biologicals* 121**:** 97-113.

Birnbaum M, Shainberg A, Salzberg S (1993). Infection with Moloney Murine Sarcoma Virus Inhibits Myogenesis and Alters the Myogenic-Associated (2′-5′)Oligoadenylate Synthetase Expression and Activity. *Virology* 194**:** 865-869.

Bishop SC, Woolliams JA (2010a). Understanding field disease data. *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Germany, August 1–6, 2010*.

Bishop SC, Woolliams JA (2010b). On the genetic interpretation of disease data. *PLoS ONE* 5**:** e8940.

Bishop SC, Woolliams JA (2014). Genomics and disease resistance studies in livestock. *Livestock Science* 166**:** 190-198.

Bjorn PA, Finstad B (2002). Salmon lice, *Lepeophtheirus salmonis* (Kroyer), infestation in sympatric populations of Arctic char, *Salvelinus alpinus* (L.), and sea trout, *Salmo trutta* (L.), in areas near and distant from salmon farms. *ICES J Mar Sci* 59**:** 131-139.

Blake S, Ma JY, Caporale DA, Jairath S, Nicholson BL (2001). Phylogenetic relationships of aquatic birnaviruses based on deduced amino acid sequences of genome segment A cDNA. *Dis Aquat Org* 45**:** 89-102.

Boc A, Diallo AB, Makarenkov V (2012). T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Research* 40**:** W573-W579.

Bootland LM, Dobos P, Stevenson RMW (1991). The IPNV carrier state and demonstration of vertical transmission in experimentally infected brook trout *Dis Aquat Org* 10**:** 13-21.

Brenna-Hansen S, Li J, Kent MP, Boulding EG, Dominik S, Davidson WS *et al* (2012). Chromosomal differences between European and North American Atlantic salmon discovered by linkage mapping and supported by fluorescence in situ hybridization analysis. *Bmc Genomics* 13.

Brenner S, Venkatesh B, Yap WH, Chou C-F, Tay A, Ponniah S *et al* (2002). Conserved regulation of the lymphocyte-specific expression of lck in the Fugu and mammals. *Proceedings of the National Academy of Sciences* 99**:** 2936-2941.

Brieuc MSO, Waters CD, Seeb JE, Naish KA (2014). A dense linkage map for Chinook salmon (*Oncorhynchus tshawytscha*) reveals variable chromosomal divergence after an ancestral whole genome duplication event. *G3: Genes|Genomes|Genetics* 4**:** 447-460.

Broughton RE (2010). Phylogeny of teleosts based on mitochondrial genome sequences. In: Nelson, J. S., Schultze, H. -P. and Wilson, M. V. H. (ed.) *Origin and Phylogenetic Interrelationships of Teleosts.* Germany: Verlag Dr. Friedrich Pfeil.

Broughton RE, Betancur-R R, Li C, Arratia G, Ortí G (2013). Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. *PLoS Currents* 5**:** ecurrents.tol.2ca8041495ffafd8041490c8092756e75247483e.

Bruneaux M, Johnston SE, Herczeg G, Merila J, Primmer CR, Vasemagi A (2013). Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach. *Molecular Ecology* 22**:** 565-582.

Busnadiego I, Maestre AM, Rodriguez D, Rodriguez JF (2012). The Infectious Bursal Disease Virus RNA-Binding VP3 polypeptide inhibits PKR-mediated apoptosis. *Plos One* 7.

CanadianAquaculture (2012). *Aquaculture in Canada - species.* [online] Available at: <http://www.aquaculture.ca/files/species.php> [Accessed 13 November 2014].

Cano I, Joiner C, Bayley A, Rimmer G, Bateman K, Feist SW *et al* (2014). An experimental means of transmitting pancreas disease in Atlantic salmon *Salmo salar* L. fry in freshwater. *Journal of Fish Diseases* doi: 10.1111/jfd.12310.

Cariou M, Duret L, Charlat S (2013). Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecology and Evolution* 3**:** 846-852.

Caron E (2003). Rac signalling: a radical view. *Nat Cell Biol* 5**:** 185-187.

Carr IM, Markham AF (1995). Molecular-genetic analysis of the human sorbitol dehydrogenase gene. *Mammalian Genome* 6**:** 645-652.

Castresana J (2007). Topological variation in single-gene phylogenetic trees. *Genome Biology* 8**:** 216-216.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology* 22**:** 3124-3140.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011). Stacks: Building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* 1**:** 171-182.

Cepeda V, Cofre C, Gonzalez R, MacKenzie S, Vidal R (2011). Identification of genes involved in immune response of Atlantic salmon (*Salmo salar*) to IPN virus infection, using expressed sequence tag (EST) analysis. *Aquaculture* 318**:** 54-60.

Chattopadhyay B, Garg K, Ramakrishnan U (2014). Effect of diversity and missing data on genetic assignment with RAD-Seq markers. *BMC Research Notes* 7**:** 841.

Chen YZ, Liu WY, McPhie DL, Hassinger L, Neve RL (2003). APP-BP1 mediates APP-induced apoptosis and DNA synthesis and is increased in Alzheimer's disease brain. *Journal of Cell Biology* 163**:** 27-33.

Cheng RH, Kuhn RJ, Olson NH, Rossmann MG, Choi HK, Smith TJ *et al* (1995). Nucleocapsid and glycoprotein organization in an enveloped virus. *Cell* 80**:** 621-630.

Chiu C-L, Wu J-L, Her G-M, Chou Y-L, Hong J-R (2010). Aquatic birnavirus capsid protein, VP3, induces apoptosis via the Bad-mediated mitochondria pathway in fish and mouse cells. *Apoptosis* 15**:** 653-668.

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L *et al* (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* 2**:** 65-73.

Churchill GA, Doerge RW (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138**:** 963-971.

Claiborne Stephens J, Nei M (1985). Phylogenetic analysis of polymorphic DNA sequences at the Adh locus in *Drosophila melanogaster* and its sibling species. *J Mol Evol* 22**:** 289-300.

Cole RK (1964). Strain difference in response to the JM leucosis virus. *Poultry Science* 43 1308-1309.

Cole RK (1969). Breeding for resistance to Marek's disease. *Poultry Science*.

Collet B, Urquhart K, Noguera P, Larsen KH, Lester K, Smail D *et al* (2013). A method to measure an indicator of viraemia in Atlantic salmon using a reporter cell line. *J Virol Methods* 191**:** 113-117.

Collet B (2014). Innate immune responses of salmonid fish to viral infections. *Developmental & Comparative Immunology* 43**:** 160-173.

Cooper GM, Brown CD (2008). Qualifying the relationship between sequence conservation and molecular function. *Genome Research* 18**:** 201-205.

Corander J, Majander KK, Cheng L, Merila J (2013). High degree of cryptic population differentiation in the Baltic Sea herring *Clupea harengus*. *Molecular Ecology* 22**:** 2931-2940.

Cordeiro JV, Guerra S, Arakawa Y, Dodding MP, Esteban M, Way M (2009). F11-Mediated inhibition of RhoA signalling enhances the spread of Vaccinia Virus *in vitro* and *in vivo* in an intranasal mouse model of infection. *Plos One* 4**:** e8506.

Coulthard MG, Morgan M, Woodruff TM, Arumugam TV, Taylor SM, Carpenter TC *et al* (2012). Eph/Ephrin signaling in injury and inflammation. *American Journal of Pathology* 181**:** 1493-1503.

Crête-Lafrenière A, Weir LK, Bernatchez L (2012). Framing the Salmonidae family phylogenetic portrait: A more complete picture from increased taxon sampling. *PLoS ONE* 7**:** e46662.

Cruaud A, Gautier M, Galan M, Foucaud J, Sauné L, Genson G *et al* (2014). Empirical assessment of RAD sequencing for interspecific phylogeny. *Molecular Biology and Evolution*.

CSIRO (2012). *Breeding better salmon.* [online] Available at: < http://www.csiro.au/Outcomes/Food-and-Agriculture/Breeding-better-salmon.aspx> [Accessed 17 November 2014].

Danzmann RG, Cairney M, Davidson WS, Ferguson MM, Gharbi K, Guyomard R *et al* (2005). 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* 48**:** 1037-1051.

Danzmann RG, Davidson EA, Ferguson MM, Gharbi K, Koop BF, Hoyheim B *et al* (2008). Distribution of ancestral proto-Actinopterygian chromosome arms within the genomes of 4R-derivative salmonid fishes (Rainbow trout and Atlantic salmon). *Bmc Genomics* 9.

Darvasi A, Soller M (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theor Appl Genet* 85**:** 353-359.

Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ *et al* (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487**:** 94-98.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12**:** 499-510.

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

Davisson MT, Wright JE, Atherton LM (1973). Cytogenetic analysis of pseudolinkage of LDH loci in the teleost genus *Salvelinus*. *Genetics* 73**:** 645-658.

Deschamps S, Llaca V, May GD (2012). Genotyping-by-sequencing in plants. *Biology* 1**:** 460-483.

Desvignes L, Quentel C, Lamour F, Le Ven A (2002). Pathogenesis and immune response in Atlantic salmon (*Salmo salar* L.) parr experimentally infected with salmon pancreas disease virus (SPDV). *Fish Shellfish Immunol* 12**:** 77-95.

Devlin RH, McNeil BK, Groves TDD, Donaldson EM (1991). Isolation of a Y-chromosomal DNA probe capable of determining genetic sex in Chinook salmon (*Oncorhynchus tshawytscha*). *Canadian Journal of Fisheries and Aquatic Sciences* 48**:** 1606-1612.

Dewey CN (2011). Positional orthology: putting genomic evolutionary relationships into context. *Briefings in Bioinformatics* 12**:** 401-412.

Di Polo A, Aigner LJ, Dunn RJ, Bray GM, Aguayo AJ (1998). Prolonged delivery of brain-derived neurotrophic factor by adenovirus-infected Müller cells temporarily rescues injured retinal ganglion cells. *Proc Natl Acad Sci U S A* 95**:** 3978-3983.

Drangsholt TMK, Gjerde B, Odegard J, Finne-Fridell F, Evensen O, Bentsen HB (2011). Quantitative genetics of disease resistance in vaccinated and unvaccinated Atlantic salmon (*Salmo salar* L.). *Heredity* 107**:** 471-477.

Duncan R, Mason CL, Nagy E, Leong J-A, Dobos P (1991). Sequence analysis of infectious pancreatic necrosis virus genome segment B and its encoded VP1 protein: A putative RNA-dependent RNA polymerase lacking the Gly-Asp-Asp motif. *Virology* 181**:** 541-552.

Eaton DAR, Ree RH (2013). Inferring phylogeny and introgression using RADseq data: An example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology* 62**:** 689-706.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32**:** 1792-1797.

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26**:** 2460-2461.

Egea J, Klein R (2007). Bidirectional Eph-ephrin signaling during axon guidance. *Trends in Cell Biology* 17**:** 230-238.

EMBL-EBI (2014). *Blast Search Guidelines*. [online] Available at: <http://www.ebi.ac.uk/ipd/imgt/hla/blast.html> [Accessed 3rd December 2014].

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011). Local *de novo* assembly of RAD paired-end contigs using short sequencing reads. *Plos One* 6.

Everett MV, Grau ED, Seeb JE (2011). Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources* 11**:** 93-108.

Everett MV, Miller MR, Seeb JE (2012). Meiotic maps of sockeye salmon derived from massively parallel DNA sequencing. *Bmc Genomics* 13.

Falconer DS, Mackay TFC (1996). *Introduction to quantitative genetics.* 4th ed. London: Longmann & Co.

FAO (2012). *Species Fact Sheets Salmo salar*. [online] Available at: <http://www.fao.org/fishery/species/2929/en> [Accessed 19 April 2012].

FAO (2013). *Feeding the world*. [online] Available at: <http://www.fao.org/docrep/018/i3107e/i3107e03.pdf> [Accesssed 26 September 2013].

FAO (2014a). *State of world aquaculture*. [online] Available at: <http://www.fao.org/fishery/topic/13540/en> [Accessed 4 December 2014].

FAO (2014b). *Cultured Aquatic Species Information Programme. Salmo salar (Linnaeus, 1758)*. [online] Available at: <http://www.fao.org/fishery/culturedspecies/Salmo_salar/en#tcNA00B1> [Accessed 19 October 2014].

FAO (2014c). *Cultured Aquatic Species Information Programme. Penaeus vannamei (Boone, 1931)*. [online] Available at: <http://www.fao.org/fishery/culturedspecies/Litopenaeus_vannamei/en> [Accessed 4 December 2014].

FDA (2013). *An overview of Atlantic salmon, its natural history, aquaculture, and genetic engineering*. [online] Available at: <http://www.fda.gov/AdvisoryCommittees/CommitteesMeetingMaterials/VeterinaryMedicineAdvisoryCommittee/ucm222635.htm#Section_1:_Salmon> [Accessed 4 December 2014].

Felsenstein J (2005). PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

Fjalestad KT, Gjedrem T, Gjerde B (1993). Genetic improvement of disease resistance in fish: an overview. *Aquaculture* 111**:** 65-74.

Fjalestad KT, Fevolden S-E, Jørstad K, Olesen I (2014). *Breeding and genetics – new species.* [online] Available at <http://www.forskningsradet.no/servlet/Satellite?blobcol=urldata&blobheader=application%2Fpdf&blobheadername1=Content-Disposition%3A&blobheadervalue1=+attachment%3B+filename%3D%2205SelectiveBreeding.pdf%22&blobkey=id&blobtable=MungoBlobs&blobwhere=1274460351746&ssbinary=true> [Accessed 17 November 2014].

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S *et al* (2014). Ensembl 2014. *Nucleic Acids Research* 42**:** D749-D755.

Fraser A, Davey JW (2012). *radseq.info: UK RAD Sequencing Wiki.* [online] Available at: <https://www.wiki.ed.ac.uk/display/RADSequencing/Home>.

Fringuelli E, Rowley HM, Wilson JC, Hunter R, Rodger H, Graham DA (2008). Phylogenetic analyses and molecular epidemiology of European salmonid alphaviruses (SAV) based on partial E2 and nsP3 gene nucleotide sequences. *Journal of Fish Diseases* 31**:** 811-823.

Frost P, Ness A (1997). Vaccination of Atlantic salmon with recombinant VP2 of infectious pancreatic necrosis virus (IPNV), added to a multivalent vaccine, suppresses viral replication following IPNV challenge. *Fish Shellfish Immunol* 7**:** 609-619.

FRS (2007). *Infectious Haematopoietic Necrosis (IHN)*. [online] Available at: <http://www.thefishsite.com/diseaseinfo/4/infectious-haematopoietic-necrosis-ihn> [Accessed 3 December 2014].

Fuji K, Hasegawa O, Honda K, Kumasaka K, Sakamoto T, Okamoto N (2007). Marker-assisted breeding of a lymphocystis disease-resistant Japanese flounder (*Paralichthys olivaceus*). *Aquaculture* 272**:** 291-295.

Fujioka Y, Tsuda M, Nanbo A, Hattori T, Sasaki J, Sasaki T *et al* (2013). A Ca2+-dependent signalling circuit regulates influenza A virus internalization and infection. *Nat Commun* 4.

Fuller SD (1987). The T=4 envelope of sindbis virus is organized by interactions with a complementary T=3 capsid. *Cell* 48**:** 923-934.

Gabbay KH (1975). Hyperglycemia, polyol metabolism, and complications of diabetes-mellitus. *Annual Review of Medicine* 26**:** 521-536.

Gadan K, Sandtrø A, Marjara IS, Santi N, Munang'andu HM, Evensen Ø (2013). Stress-induced reversion to virulence of infectious pancreatic necrosis virus in naïve fry of Atlantic salmon (*Salmo salar* L.). *Plos One* 8**:** e54656.

Gaedigknitschko K, Schlesinger MJ (1990). The Sindbis virus 6k protein can be detected in virions and is acylated with fatty-acids. *Virology* 175**:** 274-281.

Gagnaire PA, Normandeau E, Pavey SA, Bernatchez L (2013). Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (*Coregonus clupeaformis*). *Molecular Ecology* 22**:** 3036-3048.

Gibson JR, Bishop SC (2005). Use of molecular markers to enhance resistance of livestock to disease: a global approach. *Rev Sci Tech Off Int Epizoot* 24**:** 343-353.

Gidskehaug L, Kent M, Hayes BJ, Lien S (2011). Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array. *Bioinformatics* 27**:** 303-310.

Gilbey J, Verspoor E, McLay A, Houlihan D (2004). A microsatellite linkage map for Atlantic salmon (*Salmo salar*). *Animal Genetics* 35**:** 98-105.

Gilbey J, Verspoor E, Mo TA, Sterud E, Olstad K, Hytterod S *et al* (2006). Identification of genetic markers associated with *Gyrodactylus salaris* resistance in Atlantic salmon *Salmo salar*. *Dis Aquat Org* 71**:** 119-129.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009). ASReml User Guide Release 3.0 *VSN International Ltd, Hemel Hempstead, HP1 1ES, UK*.

Gjedrem T (1985). Improvement of productivity through breeding schemes. *GeoJournal* 10**:** 233-241.

Gjedrem T, Baranski M (2009). *Selective Breeding in Aquaculture: An Introduction*. *London:Springer*.

Gjerde B, Sonesson A, Storset A, Rye M (2014). *Selective Breeding and Genetics – Atlantic Salmon.* [online] Available at <http://www.forskningsradet.no/servlet/Satellite?blobcol=urldata&blobheader=application%2Fpdf&blobheadername1=Content-Disposition%3A&blobheadervalue1=+attachment%3B+filename%3D%2205SelectiveBreeding.pdf%22&blobkey=id&blobtable=MungoBlobs&blobwhere=1274460351746&ssbinary=true> [Accessed 17 November 2014].

Gjøen HM, Bentsen HB (1997). Past, present, and future of genetic improvement in salmon aquaculture. *ICES Journal of Marine Science: Journal du Conseil* 54**:** 1009-1014.

Glass EJ, Baxter R, Leach RJ, Jann OC (2012). Genes controlling vaccine responses and disease resistance to respiratory viral pathogens in cattle. *Vet Immunol Immunopathol* 148**:** 90-99.

Glover KA, Otterå H, Olsen RE, Slinde E, Taranger GL, Skaala Ø (2009). A comparison of farmed, wild and hybrid Atlantic salmon (*Salmo salar* L.) reared under farming conditions. *Aquaculture* 286**:** 203-210.

Goddard ME, Hayes BJ, Meuwissen THE (2010). Genomic selection in livestock populations. *Genetics Research* 92**:** 413-421.

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

Gonen S, Baranski M, Thorland I, Norris A, Grove H, Arnesen P *et al* (2015). Mapping and validation of a major QTL affecting resistance to pancreas disease (salmonid alphavirus) in Atlantic salmon (Salmo salar). *Heredity*.

Gontcharov AA, Marin B, Melkonian M (2004). Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and rbcL sequence comparisons in the Zygnematophyceae (Streptophyta). *Molecular Biology and Evolution* 21**:** 612-624.

Gradi A, Svitkin YV, Imataka H, Sonenberg N (1998). Proteolysis of human eukaryotic translation initiation factor eIF4GII, but not eIF4GI, coincides with the shutoff of host protein synthesis after poliovirus infection. *Proc Natl Acad Sci U S A* 95**:** 11089-11094.

Graham DA, Wilson C, Jewhurst H, Rowley H (2008). Cultural characteristics of salmonid alphaviruses - influence of cell line and temperature. *Journal of Fish Diseases* 31**:** 859-868.

Graham DA, Fringuelli E, Rowley HM, Cockerill D, Cox DI, Turnbull T *et al* (2012). Geographical distribution of salmonid alphavirus subtypes in marine farmed Atlantic salmon, *Salmo salar* L., in Scotland and Ireland. *Journal of Fish Diseases* 35**:** 755-765.

Graham DA, Rowley HR, Frost P (2014). Cross-neutralization studies with salmonid alphavirus subtype 1-6 strains: results with sera from experimental studies and natural infections. *Journal of Fish Diseases* 37**:** 683-691.

Green P, Falls K, Crooks S (1990). Documentation for CRI-MAP, version 2.4. *Washington School of Medicine St Louis, MO*.

Greseth MD, Traktman P (2014). *De novo* fatty acid biosynthesis contributes significantly to establishment of a bioenergetically favorable environment for Vaccinia virus infection. *PLoS Pathog* 10**:** e1004021.

Griffiths R, Tiwari B (1993). The isolation of molecular genetic markers for the identification of sex. *Proceedings of the National Academy of Sciences* 90**:** 8324-8326.

Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens HJ, Crooijmans R *et al* (2009). A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Research* 19**:** 510-519.

Grove S, Austbo L, Hodneland K, Frost P, Lovoll M, McLoughlin M *et al* (2013). Immune parameters correlating with reduced susceptibility to pancreas disease in experimentally challenged Atlantic salmon (*Salmo salar*). *Fish Shellfish Immunol* 34**:** 789-798.

Guryev V, Koudijs MJ, Berezikov E, Johnson SL, Plasterk RHA, van Eeden FJM *et al* (2006). Genetic variation in the zebrafish. *Genome Research* 16**:** 491-497.

Guy DR, Bishop SC, Brotherstone S, Hamilton A, Roberts RJ, McAndrew BJ *et al* (2006). Analysis of the incidence of infectious pancreatic necrosis mortality in pedigreed Atlantic salmon, *Salmo salar* L., populations. *Journal of Fish Diseases* 29**:** 637-647.

Guy DR, Bishop SC, Woolliams JA, Brotherstone S (2009). Genetic parameters for resistance to Infectious Pancreatic Necrosis in pedigreed Atlantic salmon (*Salmo salar*) post-smolts using a Reduced Animal Model. *Aquaculture* 290**:** 229-235.

Guyomard R, Boussaha M, Krieg F, Hervet C, Quillet E (2012). A synthetic rainbow trout linkage map provides new insights into the salmonid whole genome duplication and the conservation of synteny among teleosts. *BMC Genet* 13.

Hale MC, Thrower FP, Berntson EA, Miller MR, Nichols KM (2013). Evaluating adaptive divergence between migratory and nonmigratory ecotypes of a Salmonid fish, *Oncorhynchus mykiss*. *G3-Genes Genomes Genetics* 3**:** 1273-1285.

Haley CS, Visscher PM (1998). Strategies to utilize marker-quantitative trait loci associations. *J Dairy Sci* 81**:** 85-97.

Han M, Zmasek C (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* 10**:** 356.

Handa Y, Durkin Charlotte H, Dodding Mark P, Way M (2013). Vaccinia Virus F11 promotes viral spread by acting as a PDZ-containing scaffolding protein to bind myosin-9A and inhibit RhoA signaling. *Cell Host & Microbe* 14**:** 51-62.

Hardison RC, Oeltjen J, Miller W (1997). Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Research* 7**:** 959-966.

Harmon B, Ratner L (2008). Induction of the Gαq signaling cascade by the Human Immunodeficiency Virus envelope is required for virus entry. *Journal of Virology* 82**:** 9191-9205.

Haseman JK, Elston RC (1972). Investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2**:** 3-19.

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

Hayes BJ, Macleod IM, Baranski M (2009). Sampling strategies for whole genome association studies in aquaculture and outcrossing plant species. *Genetics Research* 91**:** 367-371.

Heaton NS, Perera R, Berger KL, Khadka S, LaCount DJ, Kuhn RJ *et al* (2010). Dengue virus nonstructural protein 3 redistributes fatty acid synthase to sites of viral replication and increases cellular fatty acid synthesis. *Proceedings of the National Academy of Sciences* 107**:** 17345-17350.

Hecht BC, Thrower FP, Hale MC, Miller MR, Nichols KM (2012). Genetic architecture of migration-related traits in rainbow and steelhead trout, *Oncorhynchus mykiss*. *G3-Genes Genomes Genetics* 2**:** 1113-1127.

Hecht BC, Campbell NR, Holecek DE, Narum SR (2013). Genome-wide association reveals genetic basis for the propensity to migrate in wild populations of rainbow and steelhead trout. *Molecular Ecology* 22**:** 3061-3076.

Hegarty M, Yadav R, Lee M, Armstead I, Sanderson R, Scollan N *et al* (2013). Genotyping by RAD sequencing enables mapping of fatty acid composition traits in perennial ryegrass (*Lolium perenne* (L.)). *Plant Biotechnology Journal* 11**:** 572-581.

Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT *et al* (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources* 11**:** 123-136.

Henryon M, Berg P, Olesen NJ, Kjær TE, Slierendrecht WJ, Jokumsen A *et al* (2005). Selective breeding provides an approach to increase resistance of rainbow trout ( Onchorhynchus mykiss ) to the diseases, enteric redmouth disease, rainbow trout fry syndrome, and viral haemorrhagic septicaemia. *Aquaculture* 250**:** 621-636.

Herath TK, Thompson KD, Adams A, Richards RH (2013). Interferon-mediated host response in experimentally induced salmonid alphavirus 1 infection in Atlantic salmon (*Salmo salar* L.). *Vet Immunol Immunopathol* 155**:** 9-20.

Heringstad B, Klemetsdal G, Ruane J (2000). Selection for mastitis resistance in dairy cattle: a review with focus on the situation in the Nordic countries. *Livestock Production Science* 64**:** 95-106.

Heuch PA, Mo TA (2001). A model of salmon louse production in Norway: effects of increasing salmon production and public management measures. *Dis Aquat Org* 45**:** 145-152.

Hill W (2013). On estimation of genetic variance within families using genome-wide identity-by-descent sharing. *Genet Sel Evol* 45**:** 32.

Hipp AL, Eaton DAR, Cavender-Bares J, Fitzek E, Nipper R, Manos PS (2014). A framework phylogeny of the American Oak clade based on sequenced RAD data. *PLoS ONE* 9**:** e93975.

Hjortaas MJ, Skjelstad HR, Taksdal T, Olsen AB, Johansen R, Bang-Jensen B *et al* (2013). The first detections of subtype 2–related salmonid alphavirus (SAV2) in Atlantic salmon, *Salmo salar* L., in Norway. *Journal of Fish Diseases* 36**:** 71-74.

Hodneland K, Bratland A, Christie KE, Endresen C, Nylund A (2005). New subtype of salmonid alphavirus (SAV), Togaviridae, from Atlantic salmon *Salmo salar* and rainbow trout *Oncorhynchus mykiss* in Norway. *Dis Aquat Org* 66**:** 113-120.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* 6.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources* 11**:** 117-122.

Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, Boyer MC *et al* (2013). Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology* 22**:** 3002-3013.

Holtzer H, Biehl J, Yeoh G, Meganathan R, Kaji A (1975). Effect of oncogenic virus on muscle differentiation. *Proc Natl Acad Sci U S A* 72**:** 4051-4055.

Hong J-R, Gong H-Y, Wu J-L (2002). IPNV VP5, a novel anti-apoptosis gene of the Bcl-2 family, regulates Mcl-1 and viral protein expression. *Virology* 295**:** 217-229.

Houston RD, Haley CS, Hamilton A, Guy DR, Tinch AE, Taggart JB *et al* (2008). Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic salmon (*Salmo salar*). *Genetics* 178**:** 1109-1115.

Houston RD, Haley CS, Hamilton A, Guy DR, Mota-Velasco JC, Gheyas AA *et al* (2010). The susceptibility of Atlantic salmon fry to freshwater infectious pancreatic necrosis is largely explained by a major QTL. *Heredity* 105**:** 318-327.

Houston RD, Davey JW, Bishop SC, Lowe NR, Mota-Velasco JC, Hamilton A *et al* (2012). Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. *Bmc Genomics* 13.

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

Houston RD, Bishop SC, Guy DR, Tinch AE, Taggart JB, Bron JE *et al* (2014b). Genome wide association analysis for resistance to sea lice in Atlantic salmon: application of a dense SNP array. *Proceedings of the 10th World Congress of Genetics Applied to Livestock Production*.

Huang H, Knowles LL (2014). Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. *Systematic Biology*.

Hurst LD, Pal C, Lercher MJ (2004). The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* 5**:** 299-310.

IcelandicFisheries (2014). *Selective breeding - The key to sustainability in aquaculture.* [online] Available at <http://www.fisheries.is/aquaculture/slective-breeding/> [Accessed 17 November 2014].

Imajoh M, Yagyu K-i, Oshima S-i (2003). Early interactions of marine birnavirus infection in several fish cell lines. *Journal of General Virology* 84**:** 1809-1816.

Imajoh M, Hirayama T, Oshima S-i (2005). Frequent occurrence of apoptosis is not associated with pathogenic infectious pancreatic necrosis virus (IPNV) during persistent infection. *Fish Shellfish Immunol* 18**:** 163-177.

Jaffe AB, Hall A (2005). RHO GTPASES: Biochemistry and Biology. *Annual Review of Cell and Developmental Biology* 21**:** 247-269.

Jansen MD, Wasmuth MA, Olsen AB, Gjerset B, Modahl I, Breck O *et al* (2010a). Pancreas disease (PD) in sea-reared Atlantic salmon, *Salmo salar* L., in Norway; a prospective, longitudinal study of disease development and agreement between diagnostic test results. *Journal of Fish Diseases* 33**:** 723-736.

Jansen MD, Taksdal T, Wasmuth MA, Gjerset B, Brun E, Olsen AB *et al* (2010b). Salmonid alphavirus (SAV) and pancreas disease (PD) in Atlantic salmon, *Salmo salar* L., in freshwater and seawater sites in Norway from 2006 to 2008. *Journal of Fish Diseases* 33**:** 391-402.

Jansen MD, Jensen BB, Brun E (2014). Clinical manifestations of pancreas disease outbreaks in Norwegian marine salmon farming – variations due to salmonid alphavirus subtype. *Journal of Fish Diseases* doi: 10.1111/jfd.12238.

Jensen BB, Kristoffersen AB, Myr C, Brun E (2012). Cohort study of effect of vaccination on pancreas disease in Norwegian salmon aquaculture. *Dis Aquat Org* 102**:** 23-U33.

Jessri M, Farah CS (2014). Next generation sequencing and its application in deciphering head and neck cancer. *Oral Oncol* 50**:** 247-253.

Jia Q, Liang F, Ohka S, Nomoto A, Hashikawa T (2002). Expression of brain-derived neurotrophic factor in the central nervous system of mice using a poliovirus-based vector. *Journal of NeuroVirology* 8**:** 14-23.

Joerger Andreas C, Wilcken R, Andreeva A (2014). Tracing the Evolution of the p53 Tetramerization Domain. *Structure (London, England:1993)* 22**:** 1301-1310.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J *et al* (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484**:** 55-61.

Jones JC, Fan SH, Franchini P, Schartl M, Meyer A (2013). The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Molecular Ecology* 22**:** 2986-3001.

Kakioka R, Kokita T, Kumada H, Watanabe K, Okuda N (2013). A RAD-based linkage map and comparative genomics in the gudgeons (genus Gnathopogon, Cyprinidae). *Bmc Genomics* 14.

Kamvysselis MK (2003). *Computational comparative genomics: genes, regulation, evolution*. PhD thesis. Massachusetts Institute of Technology.

Karlsen M, Tingbo T, Solbakk IT, Evensen O, Furevik A, Aas-Eng A (2012). Efficacy and safety of an inactivated vaccine against Salmonid alphavirus (family Togaviridae). *Vaccine* 30**:** 5688-5694.

Karnoub AE, Der CJ (2000). Rho Family GTPases and cellular transformation. In: Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience; 2000-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK6594/.

Kitano T, Matsuoka N, Saitou N (1997). Phylogenetic relationship of the genus *Oncorhynchus* species inferred from nuclear and mitochondrial markers. *Genes Genet Syst* 72**:** 25-34.

Kjøglum S, Henryon M, Aasmundstad T, Korsgaard I (2008). Selective breeding can increase resistance of Atlantic salmon to furunculosis, Infectious Salmon Anaemia and Infectious Pancreatic Necrosis. *Aquac Res* 39**:** 498-505.

Knott SA, Elsen JM, Haley CS (1996). Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theor Appl Genet* 93**:** 71-80.

Knott SA, Haley CS (1998). Simple multiple-marker sib-pair analysis for mapping quantitative trait loci. *Heredity* 81**:** 48-54.

Koboldt Daniel C, Steinberg Karyn M, Larson David E, Wilson Richard K, Mardis ER (2013). The next-generation sequencing revolution and its impact on genomics. *Cell* 155**:** 27-38.

Koonin EV, Galperin MY (2003). *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Boston: Kluwer Academic.

Koop BF, Davidson WS (2008). Genomics and the genome duplication in salmonids. *Fisheries for Global Welfare and Environment, 5th World Fisheries Congress 2008*.

Korpelainen H, Kostamo K Fau - Virtanen V, Virtanen V (2007). Microsatellite marker identification using genome screening and restriction-ligation. *BioTechniques* 42**:** 479–486.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011). Computational methods for gene orthology inference. *Briefings in Bioinformatics* 12**:** 379-391.

Kristoffersen AB, Viljugrein H, Kongtorp RT, Brun E, Jansen PA (2009). Risk factors for pancreas disease (PD) outbreaks in farmed Atlantic salmon and rainbow trout in Norway during 2003–2007. *Preventive Veterinary Medicine* 90**:** 127-136.

Krkosek M, Ford JS, Morton A, Lele S, Myers RA, Lewis MA (2007). Declining wild salmon populations in relation to parasites from farm salmon. *Science* 318**:** 1772-1775.

Kullander K, Mather NK, Diella F, Dottori M, Boyd AW, Klein R (2001). Kinase-dependent and kinase-independent functions of EphA4 receptors in major axon tract formation in vivo. *Neuron* 29**:** 73-84.

Kumari J, Bøgwald J, Dalmo RA (2013). Vaccination of Atlantic salmon, *Salmo salar* L., with *Aeromonas salmonicida* and infectious pancreatic necrosis virus (IPNV) showed a mixed Th1/Th2/Treg response. *Journal of Fish Diseases* 36**:** 881-886.

Lamer JT, Sass GG, Boone JQ, Arbieva ZH, Green SJ, Epifanio JM (2014). Restriction site-associated DNA sequencing generates high-quality single nucleotide polymorphisms for assessing hybridization between bighead and silver carp in the United States and China. *Molecular Ecology Resources* 14**:** 79-86.

Lebarbenchon C, Poitevin F, Arnal V, Montgelard C (2010). Phylogeography of the weasel (*Mustela nivalis*) in the western-Palaearctic region: combined effects of glacial events and human movements. *Heredity* 105**:** 449-462.

Leitch IJ, Greilhuber J, Dolezel J, Wendel JE (2013). Plant genome diversity volume 2: physical structure, behaviour and evolution of plant genomes. *London: Springer-Verlag Wien*.

Leon KA (1975). Improved growth and survival of juvenile Atlantic salmon (*Salmo salar*) hatched in drums packed with a labyrinthine plastic substrate. *The Progressive Fish-Culturist* 37**:** 158-163.

Leong JS, Jantzen SG, von Schalburg KR, Cooper GA, Messmer AM, Liao NY *et al* (2010). *Salmo salar* and *Esox lucius* full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome. *Bmc Genomics* 11.

Levings CS, Alexander DE (1966). Double reduction in autotetraploid maize. *Genetics* 54**:** 1297-1305.

Lhorente JP, Gallardo JA, Villanueva B, Araya AM, Torrealba DA, Toledo XE *et al* (2012). Quantitative genetic basis for resistance to *Caligus rogercresseyi* sea lice in a breeding population of Atlantic salmon (*Salmo salar*). *Aquaculture* 324–325**:** 55-59.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25**:** 2078-2079.

Li J, Boroevich KA, Koop BF, Davidson WS (2011). Comparative genomics identifies candidate genes for Infectious Salmon Anemia (ISA) resistance in Atlantic Salmon (*Salmo salar*). *Marine Biotechnology* 13**:** 232-241.

Liang X, Draghi NA, Resh MD (2004). Signaling from integrins to Fyn to Rho Family GTPases regulates morphologic differentiation of oligodendrocytes. *The Journal of Neuroscience* 24**:** 7140-7149.

Lien S, Gidskehaug L, Moen T, Hayes BJ, Berg PR, Davidson WS *et al* (2011). 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* 12**:** 615.

Lin JJ, Milhollen MA, Smith PG, Narayanan U, Dutta A (2010). NEDD8-targeting drug MLN4924 elicits DNA rereplication by stabilizing Cdt1 in S phase, triggering checkpoint activation, apoptosis, and senescence in cancer cells. *Cancer Research* 70**:** 10310-10320.

Loewy A, Smyth J, Vonbonsdorff CH, Liljestrom P, Schlesinger MJ (1995). The 6-kilodalton membrane-protein of Semliki Forest Virus is involved in the budding process. *Journal of Virology* 69**:** 469-475.

Lubieniecki KP, Jones SL, Davidson EA, Park J, Koop BF, Walker S *et al* (2010). Comparative genomic analysis of Atlantic salmon, *Salmo salar*, from Europe and North America. *BMC Genet* 11.

Lukacs MF, Harstad H, Bakke HG, Beetz-Sargent M, McKinnel L, Lubieniecki KP *et al* (2010). Comprehensive analysis of MHC class I genes from the U-, S-, and Z-lineages in Atlantic salmon. *Bmc Genomics* 11.

Lunter G, Goodson M (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* 21**:** 936-939.

Luo ZW, Zhang RM, Kearsey MJ (2004). Theoretical basis for genetic linkage analysis in autotetraploid species. *Proc Natl Acad Sci U S A* 101**:** 7040-7045.

Lusa S, Garoff H, Liljestrom P (1991). Fate of the 6k membrane-protein of Semliki Forest Virus during virus assembly. *Virology* 185**:** 843-846.

Lynch M (1999). The age and relationships of the major animal phyla. *Evolution* 53**:** 319-325.

Macqueen DJ, Johnston IA (2014). *A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification*, Vol 281.

Maddison WP, Knowles LL (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* 55**:** 21-30.

Madhun AS, Karlsbakk E, Isachsen CH, Omdal LM, Eide Sørvik AG, Skaala Ø *et al* (2014). Potential disease interaction reinforced: double-virus-infected escaped farmed Atlantic salmon, *Salmo salar* L., recaptured in a nearby river. *Journal of Fish Diseases*n/a-n/a.

Mardis ER (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry* 6**:** 287-303.

Margarido GRA, Souza AP, Garcia AAF (2007). OneMap: software for genetic mapping in outcrossing species. *Hereditas* 144**:** 78-79.

MarineHarvest (2014). Salmon Farming Industry Handbook.

Marjara IS, Bain N, Evensen O (2011). Naive Atlantic salmon (*Salmo Salar* L.) surviving a lethal challenge with infectious pancreatic necrosis virus (IPNV) shows upregulation of antiviral genes in head-kidney, including Vig-2. *Aquaculture* 318**:** 300-308.

Martín-Acebes MA, Blázquez A-B, Jiménez de Oya N, Escribano-Romero E, Saiz J-C (2011). West Nile Virus replication requires fatty acid synthesis but is independent on Phosphatidylinositol-4-Phosphate lipids. *Plos One* 6**:** e24970.

Martin SAM, Taggart JB, Seear P, Bron JE, Talbot R, Teale AJ *et al* (2007). Interferon type I and type II responses in an Atlantic salmon (*Salmo salar*) SHK-1 cell line by the salmon TRAITS/SGP microarray. *Physiol Genomics* 32**:** 33-44.

Masoudi AA, Uchida K, Yokouchi K, Miyadera K, Ogawa H, Sugimoto Y *et al* (2007). Marker-assisted selection for forelimb-girdle muscular anomaly of Japanese Black cattle. *Anim Sci J* 78**:** 672-675.

McKay SJ, Trautner J, Smith MJ, Koop BF, Devlin RH (2004). Evolution of duplicated growth hormone genes in autotetraploid salmonid fishes. *Genome* 47**:** 714-723.

McLoughlin MF, Nelson RN, McCormick JI, Rowley HM, Bryson DB (2002). Clinical and histopathological features of naturally occurring pancreas disease in farmed Atlantic salmon, *Salmo salar* L. *Journal of Fish Diseases* 25**:** 33-43.

McLoughlin MF, Graham DA, Norris A, Matthews D, Foyle L, Rowley HM *et al* (2006). Virological, serological and histopathological evaluation of fish strain susceptibility to experimental infection with salmonid alphavirus. *Dis Aquat Org* 72**:** 125-133.

McLoughlin MF, Graham DA (2007). Alphavirus infections in salmonids – a review. *Journal of Fish Diseases* 30**:** 511-531.

MERCK (2014). *Infectious pancreatic necrosis - External & internal signs.* [online] Available at: <http://aqua.merck-animal-health.com/diseases/infectious-pancreatic-necrosis/ProductAdditional_127_113315.aspx> [Accessed 12 November 2014].

Meuwissen T, Goddard M (1996). The use of marker haplotypes in animal breeding schemes. *Genet Sel Evol* 28**:** 161-176.

Miedaner T, Korzun V (2012). Marker-assisted selection for disease resistance in wheat and barley breeding. *Phytopathology* 102**:** 560-566.

Mikalsen AB, Torgersen J, AlestrÃ¶m P, Hellemann A-L, Koppang E-O, Rimstad E (2004). Protection of Atlantic salmon *Salmo salar* against infectious pancreatic necrosis after DNA vaccination. *Dis Aquat Org* 60**:** 11-20.

Milhollen MA, Narayanan U, Soucy TA, Veiby PO, Smith PG, Amidon B (2011). Inhibition of NEDD8-Activating Enzyme induces rereplication and apoptosis in human tumor cells consistent with deregulating CDT1 turnover. *Cancer Research* 71**:** 3042-3051.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* 17**:** 240-248.

Miller MR, Brunelli JP, Wheeler PA, Liu SX, Rexroad CE, Palti Y *et al* (2012). A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology* 21**:** 237-249.

Misztal I (2006). Challenges of application of marker assisted selection – a review. *Animal Science Papers and Reports* 24**:** 5-10.

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

Moen T, Fjalestad KT, Munck H, Gomez-Raya L (2004b). A multistage testing strategy for detection of quantitative trait loci affecting disease resistance in Atlantic salmon. *Genetics* 167**:** 851-858.

Moen T, Sonesson AK, Hayes B, Lien S, Munck H, Meuwissen THE (2007). Mapping of a quantitative trait locus for resistance against infectious salmon anaemia in Atlantic salmon (*Salmo salar*): comparing survival analysis with analysis on affected/resistant data. *BMC Genet* 8.

Moen T, Hayes B, Baranski M, Berg PR, Kjoglum S, Koop BF *et al* (2008). A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. *Bmc Genomics* 9**:** 223.

Moen T, Baranski M, Sonesson AK, Kjoglum S (2009). 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* 10**:** 368.

Moen T, Ødegård J (2014). Genomics in selective breeding of Atlantic salmon. *Proceedings of the 10th World Congress of Genetics Applied to Livestock Production*.

Morais A, Terzian A, Duarte D, Bronzoni R, Madrid M, Gavioli A *et al* (2013). The eukaryotic translation initiation factor 3 subunit L protein interacts with Flavivirus NS5 and may modulate yellow fever virus replication. *Virology Journal* 10**:** 205.

Morais S, Taggart J, Guy D, Bell J, Tocher D (2012). Hepatic transcriptome analysis of inter-family variability in flesh n-3 long-chain polyunsaturated fatty acid content in Atlantic salmon. *Bmc Genomics* 13**:** 410.

Morris-Desbois C, Réty S, Ferro M, Garin J, Jalinot P (2001). The Human Protein HSPC021 Interacts with Int-6 and Is Associated with Eukaryotic Translation Initiation Factor 3. *Journal of Biological Chemistry* 276**:** 45988-45995.

Moser TS, Schieffer D, Cherry S (2012). AMP-Activated Kinase restricts rift valley fever virus infection by inhibiting fatty acid synthesis. *PLoS Pathog* 8**:** e1002661.

Munro ALS, Ellis AE, McVicar AH, McLay HA, Needham EA (1984). An exocrine pancreas disease of farmed Atlantic salmon in Scotland. *Helgol Meeresunters* 37**:** 571-586.

Murray AG, Munro LA, Wallace IS, Peeler EJ, Thrush MA (2011). Bacterial kidney disease: assessment of risk to Atlantic salmon farms from infection in trout farms and other sources. *Scottish Marine and Freshwater Science* 2.

Naqvi N, Bonman JM, Mackill D, Nelson R, Chattoo B (1995). Identification of RAPD markers linked to a major blast resistance gene in rice. *Mol Breeding* 1**:** 341-348.

Nascimento R, Costa H, Parkhouse RME (2012). Virus manipulation of cell cycle. *Protoplasma* 249**:** 519-528.

Naylor R, Hindar K, Fleming IA, Goldburg R, Williams S, Volpe J *et al* (2005). Fugitive salmon: assessing the risks of escaped fish from net-pen aquaculture. *Bioscience* 55.

NCBI (2012). *CASC4 cancer susceptibility candidate 4 [ Homo sapiens ]*. [online] Available at: <http://www.ncbi.nlm.nih.gov/gene/113201> [Accessed 17 December 2012].

NCBI (2013a). *dbEST: database of "Expressed Sequence Tags"*. [online] Available at: <http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html> [Accessed 29 March 2013].

NCBI (2013b). Unigene <http://www.ncbi.nlm.nih.gov/unigene> [Accessed 29 March 2013].

Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP *et al* (2012). Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences* 109**:** 13698-13703.

Ng SHS, Artieri CG, Bosdet IE, Chiu R, Danzmann RG, Davidson WS *et al* (2005). A physical map of the genome of Atlantic salmon, *Salmo salar*. *Genomics* 86**:** 396-404.

Nicholas FW (2005). Animal breeding and disease. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360**:** 1529-1536.

Nimnual AS, Taylor LJ, Bar-Sagi D (2003). Redox-dependent downregulation of Rho by Rac. *Nat Cell Biol* 5**:** 236-241.

Nirea K, Sonesson A, Woolliams J, Meuwissen T (2012). Strategies for implementing genomic selection in family-based aquaculture breeding schemes: double haploid sib test populations. *Genet Sel Evol* 44**:** 30.

Norris A, Foyle L, Ratcliff J (2008). Heritability of mortality in response to a natural pancreas disease (SPDV) challenge in Atlantic salmon, *Salmo salar* L., post-smolts on a West of Ireland sea site. *Journal of Fish Diseases* 31**:** 913-920.

O'Brien V (1998). Viruses and apoptosis. *Journal of General Virology* 79**:** 1833-1845.

Ogden R, Gharbi K, Mugue N, Martinsohn J, Senn H, Davey JW *et al* (2013). Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. *Molecular Ecology* 22**:** 3112-3123.

Oh JJ, Grosshans DR, Wong SG, Slamon DJ (1999). Identification of differentially expressed genes associated with HER-2/neu overexpression in human breast cancer cells. *Nucleic Acids Research* 27**:** 4008-4017.

Oldenbroek (ed) K (2007). *Utilisation and conservation of farmed animal genetic resources*. The Netherlands: Wageningen Academic Publishers.

Olsen GJ, Pace NR, Nuell M, Kaine BP, Gupta R, Woese CR (1985). Sequence of the 16S rRNA gene from the thermoacidophilic archaebacterium *Sulfolobus solfataricus* and its evolutionary implications. *J Mol Evol* 22**:** 301-307.

Onyango P, Miller W, Lehoczky J, Leung CT, Birren B, Wheelan S *et al* (2000). Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain. *Genome Research* 10**:** 1697-1710.

Orpetveit I, Gjoen T, Sindre H, Dannevig BH (2008). Binding of infectious pancreatic necrosis virus (IPNV) to membrane proteins from different fish cell lines. *Archives of Virology* 153**:** 485-493.

Ortega F, Lopez-Vizcon C (2012). Application of molecular marker-assisted selection (MAS) for disease resistance in a practical potato breeding programme. *Potato Res* 55**:** 1-13.

Palaiokostas C, Bekaert M, Khan MGQ, Taggart JB, Gharbi K, McAndrew BJ *et al* (2013). Mapping and validation of the major sex-determining region in Nile Tilapia (*Oreochromis niloticus* L.) using RAD sequencing. *Plos One* 8**:** e68389.

Palti Y, Gao G, Liu S, Kent MP, Lien S, Miller MR *et al* (2014). The development and characterization of a 57K SNP array for rainbow trout. *Molecular Ecology Resources*n/a-n/a.

Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigo R (2003). Comparative gene prediction in human and mouse. *Genome Research* 13**:** 108-117.

Pavlopoulos GA, Oulas A, Iacucci E, Sifrim A, Moreau Y, Schneider R *et al* (2013). Unraveling genomic variation from next generation sequencing data. *BioData Min* 6.

Penaloza C, Bishop SC, Toro J, Houston RD (2014). RAD Sequencing reveals genome-wide heterozygote deficiency in pair crosses of the Chilean mussel *Mytilus* spp. *Proceedings of the 10th World Congress of Genetics Applied to Livestock Production*.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *Plos One* 7.

Phillips RB, Keatley KA, Morasch MR, Ventura AB, Lubieniecki KP, Koop BF *et al* (2009). Assignment of Atlantic salmon (*Salmo salar*) linkage groups to specific chromosomes: Conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (*Oncorhynchus mykiss*). *BMC Genet* 10.

Polakowski N, Terol M, Hoang K, Nash I, Laverdure S, Gazon H *et al* (2014). HBZ Stimulates Brain-Derived Neurotrophic Factor/TrkB Autocrine/Paracrine Signaling To Promote Survival of Human T-Cell Leukemia Virus Type 1-Infected T Cells. *Journal of Virology* 88**:** 13482-13494.

Poland JA, Rife TW (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* 5**:** 92-102.

Pous J, Chevalier C, Ouldali M, Navaza J, Delmas B, Lepault J (2005). Structure of birnavirus-like particles determined by combined electron cryomicroscopy and X-ray crystallography. *Journal of General Virology* 86**:** 2339-2346.

Quinn NL, Boroevich KA, Lubieniecki KP, Chow W, Davidson EA, Phillips RB *et al* (2010). Genomic organization and evolution of the Atlantic salmon hemoglobin repertoire. *Bmc Genomics* 11.

Ragimekula N, Varadarajula NN, Mallapuram SP, Gangimeni G, Reddy RK, Kondreddy HR (2013). Marker assisted selection in disease resistance breeding. *2013* 1**:** 20.

Ramstad A, Midtlyng PJ (2008). Strong genetic influence on IPN vaccination-and-challenge trials in Atlantic salmon, *Salmo salar* L. *Journal of Fish Diseases* 31**:** 567-578.

Rastas P, Paulin L, Hanski I, Lehtonen R, Auvinen P (2013). Lep-MAP: Fast and accurate linkage map construction for large SNP datasets. *Bioinformatics* 29**:** 3128-3134.

Recknagel H, Elmer KR, Meyer A (2013a). A hybrid genetic linkage map of two ecologically and morphologucally divergent midas cichlid fishes (*Amphilophus* spp.) obtained by massively parallel DNA sequencing (ddRADSeq). *G3: Genes, Genomes, Genetics* 3**:** 65-74.

Recknagel H, Elmer KR, Meyer A (2013b). A hybrid genetic linkage map of two ecologically and morphologically divergent midas cichlid fishes (*Amphilophus* spp.) obtained by massively parallel DNA sequencing (ddRADSeq). *G3-Genes Genomes Genet* 3**:** 65-74.

Richerioux N, Blondeau C, Wiedemann A, Rémy S, Vautherot J-F, Denesvre C (2012). Rho-ROCK and Rac-PAK signaling pathways have opposing effects on the cell-to-cell spread of Marek's disease virus. *Plos One* 7**:** e44072.

Riener M-O, Stenner F, Liewen H, Soll C, Breitenstein S, Pestalozzi BC *et al* (2009). Golgi Phosphoprotein 2 (GOLPH2) expression in liver tumors and its value as a serum marker in hepatocellular carcinomas. *Hepatology* 49**:** 1602-1609.

Rise ML, von Schalburg KR, Brown GD, Mawer MA, Devlin RH, Kuipers N *et al* (2004). Development and application of a salmonid EST database and cDNA microarray: Data mining and interspecific hybridization characteristics. *Genome Research* 14**:** 478-490.

Roberts RJ, Pearson MD (2005). Infectious pancreatic necrosis in Atlantic salmon, *Salmo salar* L. *Journal of Fish Diseases* 28**:** 383-390.

Rodger H, Mitchell S (2007). Epidemiological observations of pancreas disease of farmed Atlantic salmon, *Salmo salar* L., in Ireland. *Journal of Fish Diseases* 30**:** 157-167.

Rodríguez-Ramilo S, De La Herrán R, Ruiz-Rejón C, Hermida M, Fernández C, Pereiro P *et al* (2014). Identification of quantitative trait loci associated with resistance to Viral Haemorrhagic Septicaemia (VHS) in turbot (*Scophthalmus maximus*): a comparison between bacterium, parasite and virus diseases. *Marine Biotechnology* 16**:** 265-276.

Roesti M, Hendry AP, Salzburger W, Berner D (2012). Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology* 21**:** 2852-2862.

Roesti M, Moser D, Berner D (2013). Recombination in the threespine stickleback genome patterns and consequences. *Molecular Ecology* 22**:** 3014-3027.

Rognes T (2001). ParAlign: a parallel sequence alignment algorithm for rapid and sensitive database searches. *Nucleic Acids Research* 29**:** 1647-1652.

Rønneseth A, Pettersen EF, Wergeland HI (2006). Neutrophils and B-cells in blood and head kidney of Atlantic salmon (*Salmo salar* L.) challenged with infectious pancreatic necrosis virus (IPNV). *Fish Shellfish Immunol* 20**:** 610-620.

Rosindell J, Harmon LJ (2012). OneZoom: A fractal explorer for the tree of life. *PLoS Biol* 10**:** e1001406.

Rowe HC, Renaut S, Guggisberg A (2011). RAD in the realm of next-generation sequencing technologies. *Molecular Ecology* 20**:** 3499-3502.

Rowley HM, Doherty CE, McLoughlin MF, Welsh MD (1998). Isolation of salmon pancreas disease virus (SPDV) from farmed Atlantic salmon, *Salmo salar* L., in Scotland. *Journal of Fish Diseases* 21**:** 469-471.

Ruane N, Rodger H, Graham D, Foyle L, Norris A, Ratcliff J *et al* (2005). Research on pancreas disease in Irish farmed salmon 2004/2005 – current and future initiatives. *Marine Environment and Health*

Ruane N, Geoghegan F, Cinneide MÓ (2007). Infectious pancreatic necrosis virus and its impact on the Irish salmon aquaculture and wild fish sectors. *Marine Environment & Health Series* 30.

Ruane NM, Jones SRM (2013). Amoebic Gill Disease (AGD) of farmed Atlantic salmon (*Salmo salar* L.). *ICES Identification Leaflets for Diseases and Parasites of Fish and Shellfish* Leaflet No. 60**:** 6 pp.

Rubin BER, Ree RH, Moreau CS (2012). Inferring phylogenies from RAD sequence data. *PLoS ONE* 7**:** e33394.

Ryynanen HJ, Primmer CR (2006). Single nucleotide polymorphism (SNP) discovery in duplicated genomes: intron-primed exon-crossing (IPEC) as a strategy for avoiding amplification of duplicated loci in Atlantic salmon (*Salmo salar*) and other salmonid fishes. *Bmc Genomics* 7.

Sadasiv EC (1996). Immunological and pathological responses of salmonids to infectious pancreatic necrosis virus (IPNV). *Annual Review of Fish Diseases* 5**:** 209-223.

Sakamoto T, Danzmann RG, Gharbi K, Howard P, Ozaki A, Khoo SK *et al* (2000). A microsatellite linkage map of rainbow trout (*Oncorhynchus mykiss*) characterized by large sex-specific differences in recombination rates. *Genetics* 155**:** 1331-1345.

Salte R, Bentsen HB, Moen T, Tripathy S, Bakke TA, Odegard J *et al* (2010). Prospects for a genetic management strategy to control *Gyrodactylus salaris* infection in wild Atlantic salmon (*Salmo salar*) stocks. *Canadian Journal of Fisheries and Aquatic Sciences* 67**:** 121-129.

Salvatori G, Lattanzi L, Coletta M, Aguanno S, Vivarelli E, Kelly R *et al* (1995). Myogenic conversion of mammalian fibroblasts induced by differentiating muscle cells. *Journal of Cell Science* 108**:** 2733-2739.

Santini S, Boore JL, Meyer A (2003). Evolutionary conservation of regulatory elements in vertebrate *Hox* gene clusters. *Genome Research* 13**:** 1111-1122.

Sanz MA, Madan V, Carrasco L, Nieva JL (2003). Interfacial domains in Sindbis virus 6K protein - Detection and functional characterization. *Journal of Biological Chemistry* 278**:** 2051-2057.

Saravanan K, Nilavan SE, Sudhagar SA, Naveenchandru V (2013). Diseases of mariculture finfish species: a review *The Israeli Journal of Aquaculture* 65.

Schmitt T, Messina DN, Schreiber F, Sonnhammer ELL (2011). Letter to the Editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Briefings in Bioinformatics* 12**:** 485-488.

Schwartz M (2004). Rho signalling at a glance. *Journal of Cell Science* 117**:** 5457-5458.

Seeb LW, Waples RK, Limborg MT, Warheit KI, Pascal CE, Seeb JE (2014). Parallel signatures of selection in temporally isolated lineages of pink salmon. *Molecular Ecology* 23**:** 2473-2485.

Shedko SV, Miroshnichenko IL, Nemkova GA (2013). Phylogeny of salmonids (Salmoniformes: Salmonidae) and its molecular dating: Analysis of mtDNA data. *Russ J Genet* 49**:** 623-637.

Skilbrei OT, Wennevik V (2006). Survival and growth of sea-ranched Atlantic salmon, *Salmo salar* L., treated against sea lice before release. *ICES J Mar Sci* 63**:** 1317-1325.

Skjesol A, Skjaeveland I, Elnaes M, Timmerhaus G, Fredriksen B, Jorgensen S *et al* (2011). IPNV with high and low virulence: host immune responses and viral mutations during infection. *Virology Journal* 8**:** 396.

Slavov GT, Nipper R, Robson P, Farrar K, Allison GG, Bosch M *et al* (2014). Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass *Miscanthus sinensis*. *New Phytologist* 201**:** 1227-1239.

Smail DA, Bruno DW, Dear G, McFarlane LA, Ross K (1992). Infectious Pancreatic Necrosis (IPN) virus sp serotype in farmed Atlantic salmon, *Salmo salar* L, post-smolts associated with mortality and clinical disease. *Journal of Fish Diseases* 15**:** 77-83.

Smail DA, McFarlane L, Bruno DW, McVicar AH (1995). The pathology of an IPN-Sp sub-type (Sh) in farmed Atlantic salmon, *Salmo salar* L, post-smolts in the Shetland Isles, Scotland. *Journal of Fish Diseases* 18**:** 631-638.

Smit A, Hubley R, Green P (1996-2010). *RepeatMasker Open-3.0.*

Soller M, Andersson L (1998). Genomic approaches to the improvement of disease resistance in farm animals. *Revue Scientifique Et Technique De L Office International Des Epizooties* 17**:** 329-345.

Sonesson A, Meuwissen T (2009). Testing strategies for genomic selection in aquaculture breeding programs. *Genet Sel Evol* 41**:** 37.

Sonesson AK (2005). A combination of walk-back and optimum contribution selection in fish: a simulation study. *Genet Sel Evol* 37**:** 587-599.

Spiering D, Hodgson L (2011). Dynamics of the Rho-family small GTPases in actin regulation and motility. *Cell Adhesion & Migration* 5**:** 170-180.

Stamatakis A (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30**:** 1312-1313.

Stead SM (2002). *The Handbook of Salmon Farming*. UK: Springer Science & Business Media, Nature.

Steine T (1998). Realized effect of selection for mastitis resistance. Interbull meeting, Rotorua, New Zealand, January 18–19, 3pp.

Stene A, Bang Jensen B, Knutsen Ø, Olsen A, Viljugrein H (2013). Seasonal increase in sea temperature triggers pancreas disease outbreaks in Norwegian salmon farms. *Journal of Fish Diseases* doi: 10.1111/jfd.12165.

Storset A, Strand C, Wetten M, Sissel K, Rarnstad A (2007). Response to selection for resistance against infectious pancreatic necrosis in Atlantic salmon (*Salmo salar* L.). *Aquaculture* 272**:** S62-S68.

Storvall H, Ramsköld D, Sandberg R (2013). Efficient and comprehensive representation of uniqueness for next-generation sequencing by minimum unique length analyses. *PLoS ONE* 8**:** e53822.

Sunadome K, Yamamoto T, Ebisuya M, Kondoh K, Sehara-Fujisawa A, Nishida E (2011). ERK5 Regulates Muscle Cell Fusion through Klf Transcription Factors. *Developmental Cell* 20**:** 192-205.

Surawska H, Ma PC, Salgia R (2004). The role of ephrins and Eph receptors in cancer. *Cytokine & Growth Factor Reviews* 15**:** 419-433.

Swords RT, Kelly KR, Smith PG, Garnsey JJ, Mahalingam D, Medina E *et al* (2010). Inhibition of NEDD8-activating enzyme: a novel approach for the treatment of acute myeloid leukemia. *Blood* 115**:** 3796-3800.

Tacken MGJ, Thomas AAM, Peeters BPH, Rottier PJM, Boot HJ (2004). VP1, the RNA-dependent RNA polymerase and genome-linked protein of infectious bursal disease virus, interacts with the carboxy-terminal domain of translational eukaryotic initiation factor 4AII. *Archives of Virology* 149**:** 2245-2260.

Taggart JB (2007). FAP: An exclusion-based parental assignment program with enhanced predictive functions. *Mol Ecol Notes* 7**:** 412-415.

Taksdal T, Ramstad A, Stangeland K, Dannevig BH (1998). Induction of infectious pancreatic necrosis (IPN) in covertly infected Atlantic salmon, *Salmo salar* L., post smolts by stress exposure, by injection of IPN virus (IPNV) and by cohabitation. *Journal of Fish Diseases* 21**:** 193-204.

Taksdal T, Olsen AB, Bjerkas I, Hjortaas MJ, Dannevig BH, Graham DA *et al* (2007). Pancreas disease in farmed Atlantic salmon, *Salmo salar* L., and rainbow trout, *Oncorhynchus mykiss* (Walbaum), in Norway. *Journal of Fish Diseases* 30**:** 545-558.

Taksdal T, Bang Jensen B, Böckerman I, McLoughlin MF, Hjortaas MJ, Ramstad A *et al* (2014). Mortality and weight loss of Atlantic salmon, *Salmon salar* L., experimentally infected with salmonid alphavirus subtype 2 and subtype 3 isolates from Norway. *Journal of Fish Diseases*n/a-n/a.

Taylor RS, Kube PD, Muller WJ, Elliott NG (2009). Genetic variation of gross gill pathology and survival of Atlantic salmon (*Salmo salar* L.) during natural amoebic gill disease challenge. *Aquaculture* 294**:** 172-179.

Thorstad EB, Fleming IA, McGinnity P, Soto D, Wennevik V, Whoriskey F (2008). Incidence and impacts of escaped farmed Atlantic salmon *Salmo salar* in nature. *NINA Special Report 36*110 pp.

Toranzo AE, Magariños B, Romalde JL (2005). A review of the main bacterial fish diseases in mariculture systems. *Aquaculture* 246**:** 37-61.

Turnbull J, Bell A, Adams C, Bron J, Huntingford F (2005). Stocking density and welfare of cage farmed Atlantic salmon: application of a multivariate analysis. *Aquaculture* 243**:** 121-132.

Uribe C, Folch H, Enriquez R, Moran G (2011). Innate and adaptive immunity in teleost fish: a review. *Veterinarni Medicina* 56**:** 486-503.

USDA (2014a). *Alternative Crops and Plants.* [online] Available at: <http://afsic.nal.usda.gov/alternative-crops-and-plants> [Accessed 17 November 2014].

USDA (2014b). *Aquaculture and Soilless Farming.* [online] Available at: <http://afsic.nal.usda.gov/aquaculture-and-soilless-farming/aquaculture> [Accessed 17 November 2014].

Vandeputte M, Mauger S, Dupont-Nivet M (2006). An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion. *Mol Ecol Notes* 6**:** 265-267.

Verrier ER, Dorson M, Mauger S, Torhy C, Ciobotaru C, Hervet C *et al* (2013). Resistance to a Rhabdovirus (VHSV) in rainbow trout: Identification of a major QTL related to innate mechanisms. *PLoS ONE* 8**:** e55302.

Visendi P, Batley J, Edwards D (2013). Next generation characterisation of cereal genomes for marker discovery. *Biology* 2**:** 1357-1377.

Visscher PM, Thompson R, Haley CS (1996). Confidence intervals in QTL mapping by bootstrapping. *Genetics* 143**:** 1013-1020.

Vogel RH, Provencher SW, Vonbonsdorff CH, Adrian M, Dubochet J (1986). Envelope structure of Semliki Forest Virus reconstructed from cryoelectron micrographs. *Nature* 320**:** 533-535.

Volff JN (2005). Genome evolution and biodiversity in teleost fish. *Heredity* 94**:** 280-294.

Voorrips RE (2002). MapChart: Software for the graphical presentation of linkage maps and QTLs. *J Hered* 93**:** 77-78.

Waknitz FW, Tynan TJ, Nash CE, Iwamoto RN, Rutter LG (2002). Review of potential impacts of Atlantic salmon culture on Puget Sound chinook salmon and Hood Canal summer-run chum salmon evolutionarily significant units. *US Dept Commer, NOAA Tech Memo* NMFS-NWFSC-53**:** 83 p.

Waknitz FW, Iwamoto RN, Strom MS (2003). Interactions of Atlantic salmon in the Pacific Northwest: IV. Impacts on the local ecosystems. *Fisheries Research* 62**:** 307-328.

Wang S, Chi X, Wei H, Chen Y, Chen Z, Huang S *et al* (2014). Influenza A Virus-Induced Degradation of Eukaryotic Translation Initiation Factor 4B Contributes to Viral Replication by Suppressing IFITM3 Protein Expression. *Journal of Virology* 88**:** 8375-8385.

Wang XQ, Zhao L, Eaton DAR, Li DZ, Guo ZH (2013). Identification of SNP markers for inferring phylogeny in temperate bamboos (Poaceae: Bambusoideae) using RAD sequencing. *Molecular Ecology Resources* 13**:** 938-945.

Weng JK, Tanurdzic M, Chapple C (2005). Functional analysis and comparative genomics of expressed sequence tags from the lycophyte *Selaginella moellendorffii*. *Bmc Genomics* 6.

Weston JH, Welsh MD, McLoughlin MF, Todd D (1999). Salmon pancreas disease virus, an alphavirus infecting farmed Atlantic salmon, *Salmo salar* L. *Virology* 256**:** 188-195.

Whelan K (2010). A review of the impacts of the salmon louse, *Lepeophtheirus salmonis*, on wild salmonids *Atlantic Salmon Trust*.

Whittington AC, Moerland TS (2012). Resurrecting prehistoric parvalbumins to explore the evolution of thermal compensation in extant Antarctic fish parvalbumins. *The Journal of Experimental Biology* 215**:** 3281-3292.

Willing EM, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011). Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics* 27**:** 2187-2193.

Wolf K, Bradford AD, Quimby MC (1963). Egg-associated transmission of IPN virus of trouts. *Virology* 21**:** 317-&.

Wolf K, Quimby MC (1971). Salmonid viruses: Infectious pancreatic necrosis virus. *Archiv f Virusforschung* 34**:** 144-156.

Woo PTK, Leatherland JF, Bruno DW (2011). *Fish Diseases and Disorders Volume 3*. 2nd ed. London: CABI.

Wyckoff EE, Hershey JW, Ehrenfeld E (1990). Eukaryotic initiation factor 3 is required for poliovirus 2A protease-induced cleavage of the p220 component of eukaryotic initiation factor 4F. *Proc Natl Acad Sci U S A* 87**:** 9529-9533.

Wyckoff EE, Lloyd RE, Ehrenfeld E (1992). Relationship of eukaryotic initiation factor 3 to poliovirus-induced p220 cleavage activity. *Journal of Virology* 66**:** 2943-2951.

Xu C, Guo TC, Mutoloki S, Haugland O, Evensen O (2012). Gene expression studies of host response to Salmonid alphavirus subtype 3 experimental infections in Atlantic salmon. *Vet Res* 43.

Yamaguchi A, Tazuma S, Nishioka T, Ohishi W, Hyogo H, Nomura S *et al* (2005). Hepatitis C Virus core protein modulates fatty acid metabolism and thereby causes lipid accumulation in the liver. *Dig Dis Sci* 50**:** 1361-1371.

Yáñez JM, Houston R, Newman S (2014). Genetics and genomics of disease resistance in salmonid species. *Frontiers in Genetics* 5.

Yang WC, Francis DM (2005). Marker-assisted selection for combining resistance to bacterial spot and bacterial speck in tomato. *J Am Soc Hortic Sci* 130**:** 716-721.

Zeng L, Zhang Q, Sun R, Kong H, Zhang N, Ma H (2014). Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat Commun* 5.

Zhang Z, Schwartz S, Wagner L, Miller W (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7**:** 203-214.

Zhou TL, Guo JT, Nunes FA, Molnar-Kimber KL, Wilson JM, Aldrich CE *et al* (2000). Combination therapy with lamivudine and adenovirus causes transient suppression of chronic woodchuck hepatitis virus infections. *Journal of Virology* 74**:** 11754-11763.

Zhou X, Zheng Y (2013). Cell type-specific signaling function of RhoA GTPase: Lessons from mouse gene targeting. *Journal of Biological Chemistry*.

Zhou XJ, Xia YL, Ren XP, Chen YL, Huang L, Huang SM *et al* (2014). Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). *Bmc Genomics* 15.

Zhou Y, Li L, Hu L, Peng T (2011). Golgi phosphoprotein 2 (GOLPH2/GP73/GOLM1) interacts with secretory clusterin. *Molecular Biology Reports* 38**:** 1457-1462.

Zhu C, Tong J, Yu X, Guo W, Wang X, Liu H *et al* (2014). A second-generation genetic linkage map for bighead carp (*Aristichthys nobilis*) based on microsatellite markers. *Animal Genetics* 45**:** 699-708.