Victoria University of Wellington
School of Biological Sciences

A thesis submitted to the Victoria University of Wellington in fulfilment of the
requirements for the degree of Doctor of Philosophy

The genomics of growth and blue spots in a cultured population
of Australasian snapper *Chrysophrys auratus*

Doctoral Dissertation of:
David T. Ashton

Primary Supervisor:

Dr. Peter Ritchie

Secondary Supervisor:

Dr. Maren Wellenreuther

February 2019

Abstract

Characterizing the genome and understanding how it influences phenotypic variation is a central goal for studies on evolution. The findings of genomic research are applicable to a wide range of human endeavours, including predicting disease risk, supporting selective breeding programmes, and understanding adaptive variation in natural populations. One industry that could particularly benefit from this knowledge is Aquaculture. In recent years, aquaculture production has been increasing to offset the production limits of wild fisheries. Genomics can be used in aquaculture to quantify variation of captive populations, reconstruct pedigrees, and improve the gains from selective breeding programs. The overall goal of this thesis research was to generate a genome-wide genotyping dataset and investigated several key traits for Australasian snapper (*Chrysophrys auratus* or *Pagrus auratus*). The findings will be used to establish one of the first genomics-informed New Zealand aquaculture programmes and provide a better understanding of the genotype-phenotype relationships in this teleost species.

The first two chapters of this thesis provide a review of the literature and establish the background information and context for the research in subsequent data chapters. A brief overview of genomics, fisheries and aquaculture, and the intersection of these two fields are provided in the **Chapter 1**. An in-depth quantitative review of 146 Quantitative Trait Loci (QTL) studies in teleost fish was then carried out in **Chapter 2**.

Chapter 3 provides details about the study population and the collection of genotyping data. Genotyping-By-Sequencing (GBS) was used to generate 11K Single Nucleotide Polymorphism (SNP) markers for individuals in the three generation pedigree. Together with phenotypic data the genotyping was used to reconstruct the pedigree, measure inbreeding, and estimate heritability for a range of traits. Parents were identified for 93% of the offspring and successful pedigree reconstruction indicated highly uneven contributions of each parent to the subsequent generations. The average inbreeding level did not change between generations, but significantly different inbreeding levels were observed between offspring from the two founding cohorts and as a result full and half sibling crosses within the group spawning teleost species. Heritability was estimated for a range of traits using both a pedigree relatedness matrix and a genomic relatedness matrix.

Chapter 4, uses the genotyping and phenotyping data to generate a linkage map and carry out a scan for quantitative trait loci (QTLs) associated with growth rate. The linkage map reported in this thesis is one of the highest density maps for any Sparidae species at the time of writing. It contained 24 linkage groups, which represent the 24 snapper chromosomes. Growth QTLs were found on three linkage groups and a scan of available genome data identified three candidate growth genes nearby on the linkage groups.

Chapter 5, uses the genotyping data and images collected during the study to characterize snappers blue spots and search for QTLs associated with spot numbers. QTLs were found on 12 of the 24 linkage groups, of which one was consistent between

two QTL methods applied. A scan of available genome data identified the tyrosinase gene in the middle of the putative QTL region, which is a causal gene for iridophore cell numbers that form blue spots in other fish species.

Chapter 6, discuss the implications, future directions, and application of this research to the snapper breeding programme.

Acknowledgments

I would like to say a big THANKYOU to all my friends, family, and colleagues for their support while completing this thesis. While there are some specific people I would like to mention, you have all been important in making my pursuit of a PhD a reality.

Firstly, I would like to thank my primary supervisor Dr. Peter Ritchie (VUW). You have been an amazing supervisor throughout both my MSc and now my PhD thesis. I appreciate your calm and steady input on my thesis research and that you were always available to discuss it. I would also like to thank my secondary supervisor Dr. Maren Wellenreuther (PFR). I was very lucky that you started work at Plant and Food in time to supervise my PhD and I particularly appreciate your awesome management of our team at PFR.

This research would not have been possible without the support of many people from Plant & Food Research. First, I would like to acknowledge our Nelson science group leader, Alistair Jerrett, who has provided ongoing support for my switching between student and staff throughout my study. Thanks to all the staff in Nelson. In particular, Warren Fantham, Therese Wells, and everyone else who has helped with breeding, feeding, and maintaining all the fish. You guys do an awesome job. A big thanks to Dr. Elena Hilario (PFR) for your help with the molecular work in Auckland. You made the daunting task of preparing all the sequencing libraries achievable. Thanks to everyone that has had input into this research from the PFR bioinformatics and biometrics teams, including Dr. Peter Jacksons, Dr. Ross Crowhurst, Dr. Cecilia Deng, and many more.

A big shout out to my previous office mate and fellow PhD student Dr. Karen Middlemiss. Our discussions about study and everything else in life were lots of fun. I look forward to many more fun outings, dinners, and discussions. I would also like to acknowledge my current PFR lab group. You guys are an awesome group to work with. Also, thanks Dr. Peter Morrison-Whittle for all your help with formatting this thesis in LaTeX.

I would like to acknowledge all the financial support which has made this research possible. This project was funded through the New Zealand Ministry of Business, Innovation, and Employment Endeavour Fund "Accelerated breeding for enhanced seafood production" (C11X1603). Throughout my PhD I was also employed at Plant & Food Research, where I was responsible for running this project. The support from PFR throughout my study has been above and beyond and I appreciate it very much.

Finally, I would like to acknowledge my family for their support throughout all my study. It's been a long time from homeschooling and catching yabbies on the farm to completing my PhD thesis, but totally worth it! I would not be me without you. The biggest thanks of all goes to my partner Ellie. Each stage of my study has coincided with a new, different, and fun patch in our lives. Love you heaps and thanks for being so supportive during my study. Also, I am super proud of your study achievements in the last few years!

DAVID T. ASHTON

June 2018

External contributions

This work is part of the project "Accelerated breeding for enhanced seafood production" (C11X1603) funded by the New Zealand Ministry of Business, Innovation, and Employment (MBIE) Endeavour Fund. All fish used in this study were supplied by the New Zealand Institute for Plant & Food Research. All laboratory work, analysis and writing were conducted by the author with the help of supervisors Dr. Peter Ritchie and Dr. Maren Wellenreuther. Several other very important contributions are mentioned below:

The Genotyping-by-Sequencing method applied in this thesis was supervised by Dr. Elena Hilario from Plant & Food Research in Auckland.

The initial snapper genome assembly was supplied by Ross Crowhurst at Plant & Food Research in Auckland. This genome is currently under development and will be published at a later date.

Many of the statistics carried out in this thesis were reviewed and advised by Dr. Peter Jaksons at Plant & Food Research in Lincoln. In particular, initial heritability models were prepared by him and subsequently modified as needed for this research. The heritability models were also reviewed and improved by Fiona Hely at the agribusiness consulting company ABACUSBIO.

Maintenance of fish populations throughout this thesis was largely carried out by staff at Plant & Food Research in Nelson.

Contents

1	General Introduction	1
1.1	A brief history of genetics	1
1.2	The "genomics era"	3
1.3	Complex traits and the missing heritability problem	4
1.4	Phenomics	6
1.5	Integrating results between studies	7
1.6	Computerization	7
1.7	Application of genomics to fisheries and aquaculture	8
1.8	Selective breeding in aquaculture	8
1.9	Study species: Australasian snapper	9
1.10	Aims and thesis structure	12
1.10.1	Specific objectives of this research	13
1.10.2	Thesis structure	13
2	Factors affecting QTL detection	15
2.1	Abstract	16
2.2	Introduction	16
2.3	Methods	18
2.3.1	Study selection	18
2.3.2	Temporal trends and basic study info	18
2.3.3	Factors affecting QTL detection	19
2.4	Results	19
2.4.1	Species and pedigree information	19
2.4.2	Temporal trends	21
2.4.3	Factors affecting QTL detection and fine-mapping	24
2.5	Discussion	26
2.5.1	Reviewed study designs	26
2.5.2	Reviewed study results	28
2.5.3	Data reporting and software limitations	29
2.5.4	Epigenetic markers	30
2.5.5	Applications to selective breeding	31

2.6	Future outlook and conclusions	32
2.7	Implications for current thesis	33
3	Genetic diversity and trait heritability in Australasian snapper	35
3.1	Abstract	36
3.2	Introduction	36
3.3	Methods	38
3.3.1	Study population	38
3.3.2	Genotyping by Sequencing	39
3.3.3	Processing of sequencing data	39
3.3.4	Molecular pedigree reconstruction	40
3.3.5	Calculation of inbreeding levels	40
3.3.6	Phenotyping and trait correlations	40
3.3.7	Trait heritability and selection potential	42
3.4	Results	42
3.4.1	Quantity and quality of sequencing data	42
3.4.2	Pedigree reconstruction based on molecular relatedness values	43
3.4.3	Inbreeding in the pedigree	44
3.4.4	Trait values and phenotypic correlations	44
3.4.5	Trait heritability and selection potential	47
3.5	Discussion	48
3.5.1	Future directions	52
4	Linkage map and QTLs for growth in Australasian snapper	53
4.1	Abstract	54
4.2	Introduction	54
4.3	Methods	56
4.3.1	Study population	56
4.3.2	Phenotyping	56
4.3.3	Genotyping by Sequencing	56
4.3.4	Linkage map construction	57
4.3.5	QTL identification	57
4.3.6	Candidate genes and their location	58
4.4	Results	58
4.4.1	Phenotyping	58
4.4.2	Genotyping by Sequencing	60
4.4.3	Pedigree structure and linkage map	60
4.4.4	QTL mapping	62
4.4.5	Candidate genes	64
4.5	Discussion	64
4.6	Future directions	67

5	Genetic basis of blue spots in Australasian snapper	69
5.1	Abstract	70
5.2	Introduction	70
5.3	Methods	72
5.3.1	Study population	72
5.3.2	Phenotyping of iridophore pattern and number	72
5.3.3	Genotyping by Sequencing	73
5.3.4	Genetic basis of iridophore patterns	73
5.3.5	QTL mapping for iridophore number	73
5.3.6	Searching for candidate genes in the QTL region	74
5.4	Results	75
5.4.1	Iridophore pattern and number	75
5.4.2	Genetic basis of iridophore patterns	75
5.4.3	QTL mapping of blue spot numbers	75
5.4.4	Candidate gene search in the QTL region	76
5.5	Discussion	79
5.6	Future directions	81
6	General Discussion	83
6.1	Overview	84
6.2	Major findings	84
6.2.1	Reconstruction of a detailed pedigree	84
6.2.2	Quantification of inbreeding	85
6.2.3	Trait heritability and potential gains from selective breeding	86
6.2.4	Linkage map and recombination	89
6.2.5	Identification of QTLs and candidate genes	89
6.3	Future research and applications	91
6.3.1	Snapper selective breeding programme	91
6.3.2	Snapper genome assembly	94
6.3.3	Refining the genetic architecture of traits and gene discovery	95
6.3.4	Exploration of polygenic and phenomic methods	95
6.4	Conclusion	97
A	Supplementary material	99
A.2	Chapter 2	100
A.3	Chapter 3	102
A.4	Chapter 4	106
A.5	Chapter 5	109
B	Appendix 1: Genomic resources for the marine finfish Australasian snapper <i>Chrysophrys auratus</i>	113
B.1	Abstract	114

B.2	Introduction	114
B.3	Material and methods	115
B.3.1	Snapper populations	115
B.3.2	Genotyping by Sequencing	116
B.3.3	Genome sequencing	117
B.3.4	Phenotyping and QTL mapping	117
B.4	Results	118
B.5	Discussion	118
B.6	Conclusions	119
B.6.1	Country specific: New Zealand	119
B.7	Acknowledgements	120
	Bibliography	121

1.1 A brief history of genetics

Evolution by natural selection is one of the most foundational concepts in biology. Proposed by Charles R. Darwin and Alfred R. Wallace in 1858 (Darwin and Wallace, 1858), it caused a flurry of discussion and debate in both the scientific world and broader society. In their separate work both naturalists presented the theory that life evolved through the process of mutation and subsequent natural selection. Parallel to this, Gregor Mendel published important discoveries in 1866 about the basic rules governing inheritance of traits (Mendel, 1866). Mendel conceived the idea of heritable units, which he referred to as factors (now known to be genes), and developed fundamental laws for how these affect the inheritance of traits. Since Mendel's work, there has been growing recognition that many or indeed most of these heritable units are quantitative (i.e. explain part of the variation of a trait) rather than qualitative, as tested by Mendel. In 1942, the theory of evolution and laws of Mendelian inheritance were unified through the development of the modern synthesis. The term "modern synthesis" was coined by Julian Huxley (Huxley, 1943) and refers to a mathematical model capable of merging both the theory of evolution and mendelian inheritance. Since its development in the first half of the 20th century, the modern synthesis has formed the basis of evolutionary investigations, from large-scale population investigations to small-scale breeding programs.

Fundamental to understanding the functional basis of biological evolution have been efforts to understand the molecule responsible for storing heritable biological information – deoxyribonucleic acid (DNA). DNA was first discovered by Friedrich Miescher in 1869 (Dahm, 2005). However, it was almost 75 years later in 1944 that DNA was recognised as the hereditary material of the cell (Avery et al., 1944). Since then efforts have been ongoing to characterize this molecule, its variation, and how this variation affects biological evolution. In 1953, James Watson and Francis Crick built on the work of Rosalind Franklin

and published the first structure of DNA (Watson and Crick, 1953). Crick et al. followed this several years later by publishing how the nucleotides within this structure code for the amino acids used in the construction of proteins (Crick et al., 1961).

The last half of the 20th century saw a heavy focus placed on the development of ever-more powerful technologies, which could be used for determining DNA sequence variation. A large variety of molecular marker technologies were developed from early Restriction Fragment Length Polymorphism (RFLP) through to currently used Single Nucleotide Polymorphisms (SNPs) (Grover and Sharma, 2016). The power of these technologies is predominantly measured by the number of positions within the genome that they can interrogate and the sample sizes which can be cost effectively analysed. The most recent iteration of this goal has been the development of high-throughput DNA sequencing technology, which can generate information about every base-pair position within the genome. Using this technology the complete genomes of many organisms are now being sequenced (Braasch et al., 2015) and reduced representation sequencing provides access to this technology while allowing a compromise between power and price (Elshire et al., 2011). Transcriptome sequencing can complement genome sequencing by providing data on gene expression (Grover and Sharma, 2016). The collective effect of these technological advances has led to the recognition of the current era of genetic research as the "genomics era" (Figure 1.1).

Figure 1.1: Data volumes (Trillion base-pairs) in GenBank and WGS databases from 1983 to 2018. A near exponential accumulation of data is occurring, with no indications that it will slow in the near future (NCBI, 2018).

1.2 The "genomics era"

The genome encompasses all DNA of an individual. Developing an understanding of a species' genome and the processes connecting it to visible phenotypes, promises to shed light on many fundamental questions in biology. Knowledge gained from genomics is applicable to a wide variety of human endeavours, such as predicting disease risks (Lehner, 2013), supporting selective breeding programmes (Dekkers, 2012), and identifying adaptive variation in natural populations (Savolainen et al., 2013). Importantly, genomic change, either through new mutations or allele frequency changes within a population, have long-term consequences for a species or population. In some cases these modifications may be intentional (e.g. selective breeding and targeted genetic modification) and in others they may be unintentional (e.g. lost genetic variation in wild populations due to population size reductions). Understanding the positive and negative effects of these changes requires a thorough understanding of genomic variation and the processes through which mutations affect the individual organism.

One of the primary resources needed in genomic research is a road-map of the genome (e.g. linkage map or high-quality genome assembly), which can provide information about the position of specific genetic variants. Historically, linkage maps have been the primary resource used to do this. Linkage maps are sets of markers which have been arranged according to the order in which they appear on the genome. Compared with high-quality genome assemblies, construction of a linkage map requires relatively few resources. More recently, the development of high-throughput sequencing has allowed the development of high-quality genome assemblies for even non-model species (Braasch et al., 2015). When completed these genomes contain a set of sequences representing each chromosome. However, most genomes (especially in non-model species) have not been fully constructed and instead contain a large number of DNA sequences (scaffolds) which are yet to be connected. Although they are much less precise than high-quality genome assemblies, linkage maps can serve the dual purpose of bridging the resource gap before a genome is developed and providing useful information to improve the arrangement of scaffolds during the genome assembly process (Fierst, 2015).

A key focus of genetics has always been to identify the genetic variation that is responsible for variation in phenotypic traits (Bettembourg et al., 2017; Gutierrez et al., 2015; Pértille et al., 2017). In the genomics era, two approaches are typically used: 1) Quantitative Trait Loci (QTL) mapping, and 2) the analysis of transcriptomic data. The aim of QTL mapping is to locate the specific loci or genome regions affecting a trait based on a correlation between allelic variation at loci and the observable variation of the trait (Lynch and Walsh, 1998). When a region of the genome, which is significantly correlated to a trait, has been detected it can also be further characterized to identify potential candidate genes which may be influencing the trait (Bettembourg et al., 2017). Alternatively, transcriptomic analysis can be used to identifying whether variation in the expression level of specific genes is correlated with variation in phenotypic traits (Lowe et al., 2017).

Although both QTL mapping and transcriptomics can be used to identify important functional genetic variation, the two methods have some essential differences. Firstly, transcriptomic data does not contain the DNA sequences that flank a gene so cannot be used to identify causative mutations outside the coding region. However, although QTL mapping might contain this information most studies rarely reach a level of precision that sufficient to find a specific mutation. Secondly, a quantitative analysis of gene transcript representation in a transcriptome can be used to build networks of how genes are regulated and interconnected (Lowe et al., 2017). QTL mapping is very limited for understanding gene function. Thirdly, transcriptomic data can provides a snapshot in time of the transcripts present within a cell (Lowe et al., 2017), which can be used when investigating the short term impacts of an environment (e.g. temperature, disease, or age) on gene expression. In contrast, QTL mapping compares the current trait with the static genomic variation of individuals. Comparing the trait with genomic variation does not allow the same type of investigation into the influence of environmental factors. However, mapping does allow variants to be identified which may have had a historical effect on the trait, even if those interactions no longer affects gene expression. These and many more differences between the two approaches mean the choice between transcriptomics or QTL mapping for a study is largely dependent on the key objectives and funding. For example, QTL mapping datasets may be used to improve selective breeding efforts (Arruda et al., 2016) and transcriptomics may be used for characterizing the gene networks involved with a trait (Lowe et al., 2017) (Figure 1.2). This thesis focuses on QTL mapping, because funding was not available for collecting transcriptomic data and some objectives of the research (e.g. pedigree reconstruction and development of a linkage map) could not be achieved using transcriptomic data.

1.3 Complex traits and the missing heritability problem

Efforts to characterize the genetic variation that affects phenotypic traits have had mixed results – with much of the heritability for complex traits often not explained by known causative variation (often referred to as the “missing heritability problem”) (Benfey and Mitchell-Olds, 2008; Manolio et al., 2009; Rockman, 2011). One highly cited example which highlights the problem and solutions to this problem is human body height (Yang et al., 2010, 2015). Until recently, this trait was indicated to have a heritability of 0.8, but for which only 0.45 is explained by one of the most effective studies (Yang et al., 2010) and only 0.05 was explained by all previous studies combined (Yang et al., 2010). More recently, an even more powerful study was carried which indicated that there was negligible missing heritability after considering the potential overestimation of heritability in family-based studies (heritability was closer to 0.6-0.7) and using a sample size ($n = 44,126$) large enough to capture variation at rare variants (Yang et al., 2015). Importantly for studies with smaller sample sizes, the causative variants, which are successfully identified, often have relatively large effects on the target trait, but may not be representative of the remaining

causative variation (Rockman, 2011). The remaining variation is likely to have a very small effect on the traits and consequently be difficult to detect (Gibson, 2011; Rockman, 2011).

Figure 1.2: An example of a complex trait with a polygenic basis. Skin colour, which is determined heavily by the melanin pathway (shown), is shared between a wide range of species (fish, mice, and humans). The many interacting components necessitate polygenic investigation methods to fully understand these complex traits. Figure from (D’Mello et al., 2016).

Some of the solutions for the ‘missing heritability problem’ include increased sample sizes and marker density, using family studies to investigate rare low-frequency variants, improving the accuracy of heritability estimates, and using polygenic or gene-gene interaction methods to improve the power of individual analysis (Manolio et al., 2009; Yang et al., 2010; Rockman, 2011; Yang et al., 2015). Increasing sample size can increase power when detecting low frequency causative alleles. However as noted by Gibson (2011), under certain models (e.g. where hundreds of genetic variants affect a trait) the sample size may have to be larger than the population to detect the rare causative variants. Increased marker density could increase detection of causative variants by improving coverage of the genome. Utilizing family studies could improve power when causative variants with low frequency

in the wider population have higher frequency within the focal family. Lastly, the application of polygenic methods may help identify causative alleles by accounting for gene-gene interactions (Wellenreuther and Hansson, 2016). However, the sample sizes required for polygenic methods can be large (e.g. 1000s of individuals) and as noted by Wellenreuther and Hansson (2016), despite the advantages of polygenic analysis, application of these methods has been slow. Implementation of all these solutions will likely be key to addressing the problem of missing heritability.

1.4 Phenomics

The identification of genetic variation, which is responsible for variation observed in phenotypic traits, requires input from both genomic and phenotypic data (Benfey and Mitchell-Olds, 2008; Houle et al., 2010). Recently, it has been highlighted that while genomic methods have advanced significantly in the last decades, advances in phenotyping have been much slower to develop (Houle et al., 2010). In short, our ability to characterize an organism’s phenome lags behind our ability to characterize its genome. High-dimensional phenotyping (many traits with known interactions) is important as it allows researchers to start to disentangle the interactions between phenotypes, between phenotypes and the environment, and phenotypes included as covariates can also enhance the power when searching for genome regions affecting traits of interest (Benfey and Mitchell-Olds, 2008; Houle et al., 2010).

One of the primary limitations for more complex phenotyping is the lack of high-throughput phenotyping methods. Some progress has been made towards developing high throughput image-based trait analysis systems. For example, measuring wing traits in *Drosophila* through the construction of a wing machine, which non-lethally sucks the wing of an individual fly into position and captures an image (Houle et al., 2003). In plants, the software programme “LeafAnalyser” was developed to simplify the measurement of leaf shape variation. To do this it placed a large number of evenly distributed landmarks along leaf margins and recorded the position of each point automatically (Weight et al., 2007). In fish, software was recently developed that utilise image-based data input to automatically extract morphometric traits from still images (Navarro et al., 2016). Despite these advances, much work is needed to apply high-throughput phenotyping more broadly. Houle et al. (2010) noted three general areas that are critical for the development of phenomics, namely the development of the information capture technology, improved statistical and analytical capabilities, and better methods for integrating results between different types of data. In the short term, studies will be likely to continue to focus on relatively simple phenotyping, but in the medium to long term significant gains could be made as the tools for phenomic data collection and analysis advance.

1.5 Integrating results between studies

Although most studies may not find all the causative genetic variation underlying their respective traits, they do contribute to the body of knowledge about how genetic variation affects phenotypic traits. Specifically, studies can identify candidate genes and other genetic variants or pathways that may be involved in forming a specific trait (Gutierrez et al., 2015) or generate data which can be used in future analysis. As this data accumulates it can be individually informative (e.g. comparison of candidate genes between studies) or could be used to build more complex cross species models, which incorporate biological knowledge from a wide range of genomic and phenotypic investigations (Ashbrook et al., 2015; Mueller et al., 2017). While candidate gene usage across studies and species is relatively common, the development of more complex models is less common. However, the increased availability of data and improved computer software may change this in the future. Merging the results from multiple studies or species could help overcome the limitations of individual studies (e.g. insufficient sample sizes to offset multiple testing corrections, Ashbrook et al. (2015)) and arguably will be a necessary step for genomics to completely understand the effect of specific genetic changes on phenotypic variation.

The integration of results and data is significantly enhanced by high-throughput DNA sequencing technologies, which can facilitate the sharing of genomic results between studies, methodologies, and even among species (Braasch et al., 2015; Sutherland et al., 2016; Riddell and Crewther, 2017). Sequencing data can be aligned between related species to compare results found in otherwise disconnected studies. Cross-species sharing is possible and useful because many of the genes and molecular pathways are similar between closely and sometimes even distantly related individuals including mice and humans (Johnson et al., 2011), *Drosophila* through to humans (Reiter et al., 2001), and teleost fish compared with mammals (Braasch et al., 2007). Although the variation which causes the observed phenotypic difference may be specific to the study population, the information it provides about potential candidate genes and molecular interactions may not be. Together these studies help to build a more complete model of genomic processes, which can then inform individual studies.

1.6 Computerization

With high throughput DNA sequencing technology producing increasingly large volumes of raw data, the bottleneck for genomic research is increasingly shifting from the laboratory (data collection) to the computer (data storage and analysis) (Goecks et al., 2010; Nekrutenko and Taylor, 2012; Stephens et al., 2015). Future integration of phenomics with genomic investigations is likely to further inflate this problem (Houle et al., 2010). Genomic research will benefit from general advances occurring in the Information Technology (IT) sector (e.g. cloud computing, increased computation speeds and data storage capacity). However, more specific to genomics, there has been an explosion of available genomic ana-

lytic software; for example there are 8917 genomic tools in one software database (Omics, 2018). While these tools provide many opportunities for advancing genomic research, they are not without their problems. Individual software tools often have their own set of requirements (e.g. input data formats, parameters, and software environments), and if these requirements are not met they can result in software behaving differently than expected. Overall, the increased availability of genomic data and software is presenting new opportunities and challenges for genomic research. While often not at the forefront of genetic publications it is one of the more complex and important determinants of the types of genomic work that is done and which questions can be answered.

1.7 Application of genomics to fisheries and aquaculture

Wild fisheries and aquaculture are two industries that stand to gain substantial advances and insights from a more widespread application of genomic research (see for example Table 1.1, Bernatchez et al. (2017)). Both industries are in a state of change: wild fisheries are under increasing pressure from diminished fish stocks and aquaculture is growing at an accelerated pace to keep up with the increasing demand for fish protein (FAO, 2016). There is an opportunity to develop and use genomics tools to inform management and production practices in both industries. Applications of genomics for fisheries include, delineation of populations and management units (Laikre et al., 2005), estimating effective population sizes (Hollenbeck et al., 2016), investigating adaptive divergence (Sandoval-Castillo Jonathan et al., 2018), and traceability and species identification (Helyar et al., 2014). In aquaculture, applications for genomics include reconstructing pedigrees (Liu et al., 2012), monitoring genetic diversity and inbreeding (Wang et al., 2002), and developing genetically estimated breeding values (Vallejo et al., 2017). The application of genomic tools often overlaps between aquaculture and fisheries, because resources developed in one (e.g. a reference genome) are often crucial to the application of genomic tools in the other. Also, in a number of cases there is an overlap between populations which are used for aquaculture and wild fisheries, for example, the restocking of wild stocks (Gonzalez et al., 2015), and collection of new aquaculture broodstock from the wild (Dahle et al., 2006). In these cases, it is important to have genomic data from both cultured and wild fisheries, when attempting to predict what the consequences of these interactions may be. With the world's human population continuing to increase in size and most of the world's wild fisheries at maximum production capacity, aquaculture will be under continued pressure to meet the demand for fish products (FAO, 2016; Bernatchez et al., 2017).

1.8 Selective breeding in aquaculture

Selective breeding programs can be used to increase production from aquaculture systems; for example increased growth in *Pagrus major* (Murata et al., 1996). The extent of selective breeding effort and the application of genomic tools to aquaculture varies among species.

Table 1.1: Examples of some applications for genomic research in fisheries and aquaculture (Bernatchez et al., 2017).

Fisheries	Aquaculture
Stock connectivity	Biosecurity
Management units	Parentage-based tagging
Mixed-stock analysis	Monitoring of pathogens
Population size estimates	Marker assisted selection
Adaptive divergence	Genotype-phenotype associations
Traceability	-
Species identification	-

For some species there has been relatively high breeding efforts (e.g. Atlantic salmon, *Salmo salar*), for others there has been much less (e.g. Carp species), but for most species there are no current selective breeding programs (e.g. white Amur bream, *Parabramis pekinensis*) (Hulata, 2001). However, several examples have shown that genetic enhancement through selective breeding can lead to significant increases in aquaculture production (for example doubling of growth rates in *Pagrus major* Murata et al. (1996)). Furthermore, studies have indicated that selective breeding programs based on Marker Assisted Selection (MAS) may make additional gains of 10-20% per generation, dependant on the specific scenario (Meuwissen and Goddard, 1996). Genomic selection, is an improved version of MAS, which uses information from markers throughout the genome (Arruda et al., 2016). Despite the advantages of genomic tools, the uptake of this technology by commercial aquaculture has been relatively slow (suggested to be 10 to 20 years from experimental establishment to industry application by Hulata (2001)). Consequently, there is still a largely untapped opportunity for fisheries and aquaculture to benefit from more widespread application of genomics-enabled technologies (Bernatchez et al., 2017).

1.9 Study species: Australasian snapper

The study species for this thesis research was the Australasian snapper (*Chrysophrys auratus* also *Pagrus auratus*) - referred to as "snapper" in this thesis. In New Zealand the Maori name for this species is *Tamure* for adults and *Karati* for juvenile individuals. Snapper is a teleost species from the family Sparidae. Historically there has been much confusion about the exact phylogenetic position of this species, which has resulted in it receiving over 20 different names (Paulin, 1990). Most recently, it has been referred to as either *Chrysophrys auratus* or *Pagrus auratus*. I have followed the suggestion of several researchers based on recent phylogenetic results to refer to it as *C. auratus* (Parsons et al., 2014). Snapper is closely related to several well-studied species including *Pagrus major* in Japan and *Sparus aurata* in the Mediterranean. The species boundary between snapper and *P. major* is still a point of discussion – with many unofficial sources (e.g. wikipedia) combining statistics and discussions about these two closely related sister species (Wikipedia, 2018).

Figure 1.3: Estimated historical biomass of Australasian snapper (*Chrysophrys auratus*) in the Hauraki Gulf (2013). The stock recently recovered from below 20% of original biomass following the introduction of the Quota Management System (QMS) in 1986. The QMS only applies to fisheries within the New Zealand Exclusive Economic Zone (EEZ). Figure modified from (MPI, 2013)

Snapper is found throughout much of the Pacific Ocean including, around the coasts of Australia, New Zealand, and a number of other Pacific Islands (Paulin, 1990). Like most other species in the family Sparidae, snapper are predominantly an inshore species (down to 200m, Parsons et al. (2014)), which can be found from the tropics through to the warmer regions of the temperate zone. New Zealand represents the southern-most end of the geographical range for this species, with no known breeding populations south of the Tasman Bay in the South Island. The occurrence of snapper further south is limited by water temperature, with breeding disrupted at temperatures below 18°C and feeding disrupted below 10°C (Parsons et al., 2014). Snapper can grow to as large as 15 kgs, although size does vary among populations and most are much smaller (Paul and Tarring, 1980). Their physical appearance is silver-red in colour – with a large number of blue spots along the top half of their body and intermittent dark colouration sometimes present on their belly and tail. The skin pigment of snapper can be affected by both environment and diet (Doolan et al., 2008a). Snapper are protogynous hermaphrodites, which start life as immature females and then around 2-5 years of age transition into sexually mature males and females (Parsons et al., 2014).

Snapper is valued as a commercial and recreational fisheries resource in many Pacific countries. On globally comparisons, New Zealand and Australia have the highest catches

of snapper for any countries. New Zealand has a total allowable catch (TAC) of 10,591 tonnes, which was reached in most recent years (MPI, 2016). In Australia, approximately 1,100 tonnes were caught in 2015 (excluding confidential catch) (FRDC, 2018). A range of methods have been used to estimate the volume of snapper caught recreationally in New Zealand including, phone surveys, aerial surveys, and tag recovery. The most recent of these reports suggested that around 4,567 tonnes of snapper or 43% of the TAC were caught by recreational fishers in the year 2011-2012 (MPI, 2016). Stock assessment models indicate a significant reduction in the snapper population between 1900 and 2013 (for example, 20

Since the early 1990s, snapper has been proposed as a possible species for marine aquaculture (Battaglene and Talbot, 1992). One of the primary reasons snapper has been suggested is because its close sister species, *P. major*, has been successfully cultured in Japan and a more distant relative, *Sparus aurata*, is cultured in the Mediterranean. As early as 1985 aquaculture production of *P. major* was producing 10% more than the wild fisheries (Foscarini, 1988). One key to the success of *P. major* was the relative ease and consistency with which fertilized eggs could be collected. This is also true of snapper which have been successfully produced in hatcheries in both Australia and New Zealand. Another component to the success of *P. major* was its relatively fast growth rate, which has been subsequently enhanced through a selective breeding programme started in 1964 (Murata et al., 1996). Comparatively, snapper have a much slower growth rate; original wild-sourced *P. major* took 3 years to reach 1kg while snapper take over 5 years (Paul and Tarring, 1980; Murata et al., 1996). This slower growth means that if snapper were to be used as a commercial species then its growth rate will likely need to be improved (e.g. via a selective breeding program).

Investigation of colour in snapper is another area of active interest for aquaculture and understanding the basic biology of the species. Snapper are most often recognised for their red colouration. However, on closer inspection individuals also have a large number of blue spots along the length of their body. In cultured populations, such as those held at Plant & Food Research (PFR), individuals may also vary in darkness (external black colouration). For snapper aquaculture, colouration is important because light red colouration is considered desirable by most markets. The blue spots have also gained interest within the PFR breeding programme for their use as visual identifiers of specific individuals (MPI, 2015), their potential use for creating unique colour morphs (e.g. a visually different colour morph for selectively breed snapper strain), and related research about their underlying biology and how to maintain their blue colouration after death (important for presentation of a fresh product in the whole fish market). Red colouration in snapper is thought to be the product of the same astaxanthin pigment documented in *P. major* (Doolan et al., 2008b). Astaxanthin is a carotenoid that fish cannot synthesis but instead get from their diet (Doolan et al., 2008b). The genetic and molecular basis of the blue spots and variation in black colouration is not known. Blue colouration in all except two fish species, the mandarin fish (*Synchiropus splendidus*) and psychedelic fish (*Synchiropus picturatus*), is

the result of iridophores (small crystal structures which reflect light) rather than pigments (Schaefer et al., 2014). While it is not documented for certain, it is likely that snapper's blue spots are iridophores because of the the rarity of blue pigments. The potential genetic basis of variation in dark colouration of the PFR snapper has also not been investigated. This trait has anecdotally been suggested to be the result of environmental conditions (e.g. stress, lighting, and density), so clarifying if this trait has a genetic basis could begin to untangle this question.

Although snapper is commercially valuable there has been limited genetic research reported for this species. Several studies have been carried out using lower powered genetic tools to investigate the genetic structure of wild populations around the coast of New Zealand (Smith et al., 1978; Hauser et al., 2002; Bernal-Ramírez et al., 2003). In 1978, allozyme analysis indicated distinct stocks on the east and west side of the North Island of the New Zealand distribution (Smith et al., 1978). In 2003, mitochondrial and microsatellite DNA analysis indicated weak levels of differentiation between populations at the north and southern end and the eastern and western sides of the New Zealand distribution (Bernal-Ramírez et al., 2003). In addition to investigating the population structure, microsatellite data showed a drop in heterozygosity in the Tasman Bay population since 1950, indicative of a recent genetic bottleneck (Hauser et al., 2002). No genomic or QTL mapping studies have been carried out in this species to date.

A domesticated snapper population is being developed at PFR in Nelson, New Zealand, for research purposes and to investigate their potential as an aquaculture species. The first founding cohort was formed in 1994 and included approximately 25 individuals, which had been collected from around the Tasman Bay and Marlborough Sounds in the northern part of the South Island (Lat: -41.0457625, Long: 173.2934086). These individuals were used to develop the initial methods needed to culture snapper in a tank-based aquaculture system. In 2006, a second founding cohort of approximately 25 individuals were collected from the Tasman Bay and added to the breeding programme. Prior to the start of this thesis the offspring from both founding cohorts were combined into a single population and had produced a third generation. At the time this thesis was started a total of three generations had been produced at the Nelson hatchery and the full life-cycle had been closed within an aquaculture system. All offspring within the aquaculture system had been produced from group spawning where individuals within a population could all contribute. Consequently, for each offspring the parent individuals were not known. A selective breeding programme for the snapper was initiated in 2016, which included the work carried out in this thesis research. The goal of this project was to use genomic and other available tools to improve the domestication process of this new Australasian snapper population.

1.10 Aims and thesis structure

This PhD research was carried out using the snapper population held at Plant & Food Research in Nelson, New Zealand. The research was funded by the Ministry of Business, In-

novation, and Employment (MBIE) programme “Accelerated breeding for enhanced seafood production” (#C11X1603). The first general aim of this PhD research was to investigate the pedigree structure, genetic diversity, and genetic basis of key traits in snapper populations in the breeding programme. A second general aim of the thesis was to produce a range of genetic resources that could contribute to ongoing genome investigations in this and related fish species.

1.10.1 Specific objectives of this research

- Determine the pedigree structure of the existing population and what has occurred during previous breeding seasons.
- Develop a high-density linkage map, which can be used to support the genome assembly and positioning of markers during QTL mapping
- Calculate the heritability of key traits within the breeding program.
- Run a genome-wide scan for QTLs affecting growth rate and number of blue spots.
- Locate potential candidate genes within QTL regions.
- Identify the implications of the genetic results for the breeding programme.

1.10.2 Thesis structure

Each chapter of this thesis (except Chapters 1 and 6) has its own specific objectives and is divided into sections including introduction, methods, results, and discussion. Chapters 2, 3, 4, and 5 have been prepared as manuscripts for publication and are in varying stages of being submitted. Because Chapters 3, 4, and 5 are based on a single large dataset each of these chapters briefly describe the study population and relevant data collection.

Chapters 1 and 6 present an introduction and final discussion to tie together the results of all the other chapters. Specifically they discuss the broad implications of the data chapters and how the results can be applied to the snapper breeding and management. The objective of **Chapter 2** was to conduct a literature review of QTL mapping studies investigating traits in fish species over the 15 years preceding the start of this research. This laid the foundation and theoretical underpinnings and considerations for QTL mapping in Chapters 4 and 5. One of the main reasons for needing this chapter is that there is a wide diversity of methods and software available for QTL mapping and before proceeding there needed to be a better understanding of the subject area.

Chapter 3 is the first data chapter. This chapter provides the groundwork for the following chapters by conducting a review of the pedigree structure and history of the Plant & Food snapper population. It describes in detail the Genotyping-by-Sequencing (GBS) method used to generate the genotype dataset which is used in the other data chapters. It

also investigates the heritability for phenotypic traits relevant to the breeding program. **Chapter 4** is the second data chapter. This chapter investigates growth rate traits and constructs a linkage map. QTL mapping was carried out for growth rate traits. The location of 11 candidate growth genes was then compared to putative QTLs using the linkage map and available genome scaffolds. **Chapter 5** is the third data chapter. This chapter investigates the blue spots. QTL mapping was carried out for the number of blue spots. The similarity of blue spot patterns between related individuals was investigated. Finally, a BLAST search was carried out in putative QTL regions for candidate genes, which may be responsible for the observed phenotypes. **Chapter 6** is a general discussion of the major findings of this PhD research, the implications for the PFR breeding programme, and some future research directions.

Authors

David T. Ashton, Peter A. Ritchie, & Maren Wellenreuther

Author Contributions

Conceived and designed the experiments: DTA, PR, and MW; Data collection: DTA; Data analysis: DTA; Wrote the primary manuscript: DTA; Prepared final manuscript: DTA, PR, and MW

Publication status

This chapter contains a literature review of QTL mapping studies in fishes over the 15 years preceding the start of this thesis research. This literature review was conducted to help guide methodology decisions in the Chapters 4 and 5. Specifically, investigating the number of genetic markers, available software packages, and sample sizes needed to locate putative QTLs. Some small changes have been made to this published article to fit into the thesis format.

This chapter was published in *Molecular Ecology* – Ashton D.T., Ritchie, P. A, Wellenreuther, M. (2017) Fifteen years of quantitative trait loci studies in fish: challenges and future directions. *Molecular Ecology*, 26(6): 1465-1476 - DOI:10.1111/mec.13965

2.1 Abstract

Understanding the genetic basis of phenotypic variation is a major challenge in biology. In this study, 146 Quantitative Trait Loci (QTL) studies on teleost fish over the last 15 years were used to investigate 1) temporal trends and 2) factors affecting QTL detection and fine-mapping. The number of fish QTL studies per year increased over the review period and identified a cumulative number of 3632 putative QTLs. Most studies used linkage-based mapping approaches and were conducted on non-model species with limited genomic resources. A gradual and moderate increase in the size of the mapping population and a sharp increase in marker density from 2011 onwards were observed; however, the number of QTLs and variance explained by QTLs changed only minimally over the review period. Based on these findings, the causative factors for QTL mapping effectiveness and how larger sample sizes, phenomics, comparative genomics, epigenetics, and software development could improve both the quantity and quality of QTLs in future genotype-phenotype studies are discussed. Given that the technical limitations on DNA sequencing have mostly been overcome in recent years, a renewed focus on these and other study design factors will likely lead to significant improvements of QTL studies in the future.

2.2 Introduction

A major challenge in biology is to understand the genetic basis of phenotypic trait variation. Most phenotypic variation is thought to be caused by quantitative genetic variation that results from the segregation of alleles at multiple Quantitative Trait Loci (QTL) and is influenced by and sensitive to the environment (Mackay, 2001). Insights into this complex genotype-phenotype map, including environmental effects, promises to yield important knowledge for predicting disease risks (Lehner, 2013), supporting selective breeding programmes (Dekkers, 2012), and identifying adaptive variation in natural populations (Savolainen et al., 2013). It is no surprise that numerous studies over recent years have attempted to dissect the genotype-phenotype connections in a wide range of species.

Genotype-phenotype relationships can be mapped to identify the genomic regions controlling phenotypic trait variants, with the ultimate, yet often time-consuming, goal to locate the causal genes or nucleotide mutations underlying the trait in question (Rockman, 2011). Linkage, association (linkage disequilibrium), and combined linkage and linkage disequilibrium (LDLA) methods have been developed to more precisely map areas containing quantitative trait loci. The general principle of these mapping methods is that they try to account for phenotypic trait variation by measuring its correlation with markers that segregate in a Mendelian fashion. Of these, linkage and association methods are most frequently used. These methods differ in several ways. The most noteworthy difference is that association studies are typically more precise at locating QTL regions but require many more markers to achieve the high level of precision (Ott et al., 2011). Another difference is that association studies can be conducted using samples from wild or captive populations with

limited familial information, whereas linkage methods require some form of a relatedness matrix, which is usually obtained by the controlled breeding of a managed population.

Historically, the lack of abundant polymorphic markers has limited the application of these QTL mapping methods. However, rapid advances in DNA sequencing technologies and downstream analyses since the 1990s have enabled the discovery of a significantly larger number of markers. In particular, methods such as Restriction site Associated DNA (RAD) sequencing (e.g. Genotyping-by-Sequencing), have made it possible to generate genome-wide marker coverage in less time and cheaper than ever before (Elshire et al., 2011; Hilario, 2015). The promise of the ‘genomics era’ is that it will bring a significant increase in the power and precision of QTL studies and an enhanced ability to finely map the genetic basis of phenotypic traits. While some have been critical of this claim (Ioannidis and Kavvoura, 2006; Rockman, 2011), others are optimistic about both what has been achieved so far and what the future of genotype-phenotype investigations will accomplish (McCarthy et al., 2008; Visscher et al., 2012). Indeed, over the last few years a vast number of new QTLs and associated genes have been identified across a wide range of invertebrate and vertebrate species (Stranger et al., 2011; Visscher et al., 2012). However, important QTL mapping issues have repeatedly been highlighted, including the lack of power to detect QTLs that have small effects on highly polygenic traits (missing heritability problem), problems around the biological relevance of QTL mapping studies given that effect sizes of QTL reaching significance tend to be overestimated (e.g. “Beavis effect”), and difficulties incorporating epistatic, environmental and complex phenotypic trait interactions into QTL mapping strategies (Mackay et al., 2009; Rockman, 2011; Wellenreuther and Hansson, 2016).

Most work investigating genotype-phenotype relationships has been carried out on model species such as yeast (*Saccharomyces cerevisiae*), fruit flies (*Drosophila melanogaster*), mouse (*Mus musculus*) and zebrafish (*Danio rerio*) (see available information held in public database PhenomicDB, Kahraman et al. (2005)). The aforementioned advances in high-throughput DNA sequencing technologies are, however, now enabling an expansion of research on non-model species which have previously been woefully lagging behind model species because of limited funding and resources (Ellegren, 2014; Elshire et al., 2011). Teleost fish is one such group that contains many non-model organisms. They are a particularly interesting group for investigation because they 1) form the largest group of vertebrates (around 33,000 species, i.e. ~50% of all vertebrates), 2) exhibit high levels of morphological diversity, and 3) aquaculture is currently the fastest growing primary industry.

Despite several good reviews investigating some aspects of genotype-phenotype research in teleost fish (Hemmer-Hansen et al., 2014; Tong and Sun, 2015), a general review that thoroughly investigates all teleost QTL studies over the last years has not been conducted. To address this gap, a systematic survey of the published QTL studies on teleost fish was conducted in the current study covering the last 15 years and extracted QTL information from each study, including factors concerning the experimental design and data analysis. Temporal trends in QTL mapping studies were investigated using this data, including changes

over time to study design factors (e.g. number of markers and sample size) and outcomes of studies (e.g. number of QTLs found). Second, correlations between study design factors and outcomes of studies was investigated to identify what factors might be influencing QTL mapping success. Finally, the findings are discussed and the factors that will potentially improve the power of genotype-phenotype studies in the future are highlighted.

2.3 Methods

2.3.1 Study selection

Fish QTL studies were selected using the Web of Science search engine with the “Topic” search terms “QTL” and “fish” or “Quantitative trait loci” and “fish” for the period between the 1st of January 2000 and the 31st of December 2015. This search yielded a total of 715 individual published papers, of which 563 were removed because they were not available in English, did not represent QTL identification studies, were not conducted on teleost fish, or were based (fully or in part) on previously identified QTLs. Studies using previously identified QTLs were removed to reduce bias in QTL detection and mapping precision due to previous information. Only studies that used de novo information were included. The final data set used in our review comprised 146 published studies. A second Web of Science search was carried out using the “Topic” search terms “GWAS” and “fish” or “Genome wide association” and “fish” and restricted to the period between the 1st of January 2000 and the 31st of December 2015. This search yielded a total of 222 individual papers, of which there were an additional 11 Genome Wide Association Studies (GWAS) not found in the first search. These 11 studies were used to look at the number of studies using linkage and association methods, but were excluded from all other parts of the review.

Record information for each newly described QTL in the first Web of Science search included publication date, species name, number of individuals sampled at each generation in the pedigree, breeding design used for the pedigree (e.g. F_2 backcross, intercross), genetic marker type and number of markers, QTL identification software, type of QTL detection method (linkage or association), individual QTL statistics (e.g. LOD, ANOVA, significance level, genome wide versus linkage group wide), target trait, trait distribution (discrete/quantitative), percentage variance explained (PVE), and peak, start and end location in centiMorgans (cM) of QTL regions. Additional information about the target species was also included in the data file, which included genome size (based on the C-value, Gregory 2016) and the species’ primary habitat (freshwater, seawater, or euryhaline).

2.3.2 Temporal trends and basic study info

All data analyses were carried out in the R statistical environment (version: 3.2.3) (R Core Team, 2013). Information about experimental setup of QTL studies was reported including the numbers of parents and progeny in the mapping populations, major influences in the construction of the mapping populations (e.g. backcrossing, inter-species crosses),

the most common types of species investigated, and frequency of linkage and association study designs. The temporal trends of studies was then investigated, including how study design elements (software, sample size, and marker density) and results of QTL studies have changed over time. Three measures of QTL study results were used, including the number of QTLs found per trait, PVE for QTLs, and QTL region width (cM). The y-axis for temporal graphs were placed on a log scale for all variables (except Studies (n)).

2.3.3 Factors affecting QTL detection

A correlation matrix was calculated using Spearman's correlation coefficient for eight variables, namely number of QTLs (QTLs), PVE of QTLs (PVE), width of QTL regions (Width), number of traits investigated (Traits), number of genetic markers used (Markers), size of mapping population (Sample), size of genome based on C-values (Genome), and marker density (Markers/Genome) (Table 2.1). Correlations were calculated for all studies and for a subset containing only salmonid studies to reduce noise caused by the wide range of species. For all variables the average value for each study was used. Sample size of the mapping population and PVE were graphed on a scatter plot. A regression line was placed using a general linear model (with 95% CI) and a log 10 transformation of the x and y axis. Significant difference of the slope from 0 was tested using bootstrapping (10000x).

2.4 Results

The initial Web of Science search identified 712 individual papers. Of these, studies were removed if they were not investigating new QTLs (411), conducted on teleost fish (67), based on previously identified QTLs (50), written in English (25), and other miscellaneous reasons (13) (Supplementary figure 2.1). The remaining 145 studies included 128 linkage studies, 14 association studies, and three studies using both linkage and association to identify QTLs and detected a total of 3627 putative QTLs. One extreme outlier study was removed from the majority of analyses because the number of markers was over 300 times higher than the next closest study (Ayllon et al., 2015). A second search identified 11 additional GWAS studies, which were only included in Figure 3.2 of this review.

2.4.1 Species and pedigree information

QTL data from 49 fish species were reported in the studies. The three most commonly studied species were *Oncorhynchus mykiss* (15%), *Salmo salar* (10%), and *Cyprinus carpio* (8%). Considering all of the studies together, 66% were carried out on freshwater species, 23% on marine species, and 11% on euryhaline species. The majority of studies (60%) produced their mapping population from a single set of parents (i.e. all progeny in the mapping population were full siblings) (Supplementary figure 2.2). The remaining studies ranged up to a maximum of 400 individuals. Of the different breeding designs approximately half used

Figure 2.1: The distribution of genetic marker types (A), software used for QTL identification (B), and average number of markers and sample size (C) for studies over the review period. Average number of markers is a count of markers used for each study averaged for each year. Average sample size denotes the number of individuals in the mapping populations.

inter-strain or inter-species crosses with ($\sim 66\%$) or without ($\sim 33\%$) a secondary backcrossing step. Only five studies were conducted on wild (outbred) populations.

Box 1: Experimental design considerations

Many factors are important when designing a QTL mapping study including but not limited to, the number of genetic markers, type and size of the mapping population, type of statistics used, and phenotyping and environmental measurements.

With a significant increase in genotyping power over the recent years, due to lower cost and higher throughput sequencing, it may be timely to focus more attention on the limitations of other experimental factors. Reduced limitations on genotyping are likely to have an indirect effect of making larger sample sizes more achievable. As discussed by Hong and Park (2012), the sample size required for sufficient power in a QTL mapping study can be strongly influenced by the number of markers used, choice of linkage or association approaches, level of linkage disequilibrium, and effect size of QTLs. Hong and Park (2012) found that testing a single genetic marker required 248 cases (doubled when including controls) to reach 80% detection power for weak effect QTLs (5% disease prevalence, 5% minor allele frequency, and complete linkage disequilibrium). The number of cases increased to 1,206 when testing 500,000 markers. A sufficient sample size is important for being able to detect weak-effect alleles and avoid biasing our understanding of quantitative genetic variation by leading us to believe that most QTLs have large effects ("Beavis effect") (Slate, 2013).

Mapping population design is also an important consideration. The first decision is whether to use a natural population (unrelated group of individuals) or family groups. As discussed by Ott et al. (2011) one of the primary advantages of family designs is that they can be used to control for population stratification. However, a challenge presented by family-based designs is the level of resources required to construct and maintain them, which can be considerable. Investigations in natural populations have the advantage of providing insights into the fitness effects of traits under real world circumstances (Slate, 2005). A family-based design requires careful consideration for the type of pedigree and crosses used, because this influences the power for detecting QTLs and the complexity of the analytical approach. A biparental cross is the simplest design and compatible with most software packages, and indeed, is also the pedigree design that has been most often used in fish QTL studies (see Supplementary figure 2.2). As the pedigrees become more complicated more sophisticated analyses are needed; for example mixed models, Zhou and Stephens (2012). Some methods such as recombinant inbred lines or extreme phenotype selection enhance several aspects of QTL mapping power, however if a diverse range of phenotypes (see Box 3) are of interest these population designs, which are typically selected for a single trait, would be problematic (Risch and Zhang, 1995; Rockman and Kruglyak, 2008).

2.4.2 Temporal trends

A large increase in the number of QTL studies was observed over time (Figure 2.1A). Prior to 2010 the number of studies per year averaged around five, but this increased from 2010

Figure 2.2: Distribution over the review period of linkage and association studies. The results of an additional literature search for fish GWAS studies were also included.

onwards to a maximum of 31 in 2015. Before 2015 simple sequence repeats (SSRs) were the most commonly used marker type (Figure 2.1A). A noticeable shift to Single Nucleotide Polymorphisms (SNPs) started to occur in 2011 and by 2015 almost all studies used SNPs. The current study found that a total of 24 different software packages were used for QTL identification. The most common packages were GridQTL, MapQTL, and various R packages (Figure 2.1B). In particular, GridQTL came into use from 2009 onwards, MapQTL was used across almost the entire survey period, and different R packages were primarily used from 2011 onwards. A consequence of the shift to SNPs from 2011 and onwards (Figure 2.1A) was a significant increase in the average number of markers used in each study. Prior to 2011 an average of 166 markers were reported in studies, but this increased to 1171 from 2011 onwards, and then to 2461 in 2015 (Figure 2.1C). Unlike the rapid increase in marker density, the size of the mapping population increased more moderately. The mapping size increased from 122 in 2000 to 525 in 2008 and remained subsequently relatively stable from then on, with an average of 334 after 2008 (Figure 2.1C). The average sample size of mapping populations across the survey period was 294. Linkage studies were found to consistently outnumber association studies across all years (Figure 2.2). This remained

Figure 2.3: Distribution of the number of QTLs found per trait (A), Percentage Variance Explained (PVE) for identified QTLs (B) and QTL region width in centimorgans (C) over the review period. Results are shown as average per study, with the number of studies in each plot and correlation coefficient also shown.

Figure 2.4: A moderate negative correlation was observed between size (number of individuals) in the mapping population used by QTL studies and the average Percentage Variance Explained (PVE) for QTLs in those studies. A regression line with 95% confidence intervals was placed using a generalized linear model. Pearson's correlation coefficient (r) with P-value for difference from 0 based on bootstrapping is also shown (see also Table 1 for Spearman correlation coefficient of this relationship).

unchanged even after including studies resulting from a second search that was specifically targeted at finding GWAS studies on teleost fish. Minimal change over time was found when investigating the number of QTLs per trait, the PVE of identified QTLs or the width of QTL regions (Figure 2.3).

2.4.3 Factors affecting QTL detection and fine-mapping

A Spearman's correlation matrix was calculated for eight variables using the total dataset and a subset of studies using only salmonid species, to reduce noise caused by the wide variety of species (Table 2.1). The correlation matrix indicated that the number of QTLs found per study and the number of traits measured were strongly positively correlated ($r = 0.63$), suggesting that the studies with a higher number of phenotypic trait measurements found more QTLs compared to studies with fewer measurements. A moderate negative correlation ($r = -0.48$) was observed between sample size and PVE of QTLs; indicating that

Table 2.1: Results for Spearman's correlation between eight study design factors and study outcomes including, number of QTLs found (QTLs), average PVE of QTLs (PVE), width of QTL regions (Width), number of traits investigated (Traits), number of markers (Markers), size of mapping population (Sample), genome size based on C-values (Genome), and marker density (Markers/Genome). Correlations were calculated for all studies (below diagonal) and only salmonid studies (above diagonal). The number of studies used for each correlation are shown in brackets next to the correlation coefficient. Correlations in bold are those with correlation coefficients greater than >0.3 in both total and salmonid only datasets.

-	QTLs	PVE	Width	Traits	Markers	Sample	Genome	Marker Density
QTLs	-							
PVE	-0.41 (145)	-0.41 (36)	0.45 (10)	0.43 (51)	0.16 (51)	0.05 (51)	0.04 (51)	0.15 (51)
Width	0.11 (145)	-	0.38 (36)	-0.27 (36)	0.37 (36)	-0.41 (36)	-0.33 (36)	0.41 (36)
Traits	0.63 (145)	-0.19 (120)	-	-0.46 (10)	0.84 (10)	-0.29 (10)	-0.71 (10)	0.83 (10)
Markers	0.10 (145)	-0.31 (120)	0.00 (57)	-	-0.15 (51)	0.22 (51)	0.17 (51)	-0.18 (51)
Sample	0.29 (145)	0.13 (120)	-0.01 (57)	-0.07 (145)	-	-0.20 (51)	-0.23 (51)	0.99 (51)
Genome	0.21 (145)	-0.48 (120)	0.45 (57)	0.34 (145)	-0.06 (145)	-	0.17 (51)	-0.22 (51)
Marker Density	0.04 (145)	-0.06 (120)	0.25 (57)	-0.08 (145)	0.00 (145)	0.04 (145)	-	-0.33 (51)
		0.10 (120)	-0.08 (57)	0.00 (145)	0.85 (145)	-0.04 (145)	-0.44 (145)	-

with more samples the PVE of QTLs declined. Plotting the PVE of QTLs versus sample size with a linear regression and transformed axis indicated that the negative regression line was significantly different from zero (Figure 2.4, $P\text{-value} < 0.001$). The PVE of putative QTLs also declined as the number of QTLs increased ($r = -0.41$). The strongest correlation for the width of QTL regions was sample size ($r = 0.45$), which suggested QTL width increased along with increasing sample size. However, this correlation was not supported in the salmonid only subset ($r = -0.29$). The remaining correlations were mostly weak (< 0.3 in either the total data set or the salmonid subset).

2.5 Discussion

Insights into the genotype-phenotype map are of central interest in many fields of biology and also provide important information for applied research (Dekkers, 2012; Lehner, 2013; Savolainen et al., 2013). The ease with which polymorphic markers can now be discovered, even in non-model species, has led to significant advances in our ability to conduct QTL mapping studies. Contemporary studies now routinely use hundreds to thousands of markers. This new statistical power enables them to overcome the key limitation of previous work. Given the rise in our ability to search for QTLs in a true genome-wide manner, it may be expected that both the detection and mapping precision of QTLs would have increased over time. Indeed, such an increase in mapping precision has been suggested in a previous study evaluating the future prospects of genotype-phenotype mapping efforts in the ‘genomics era’ (Mackay, 2001). The aim of our review was to test these predictions and to assess the current state of QTL mapping in teleost fish by examining QTL studies from the last 15 years. The current study concentrated on teleost QTLs because fish are the largest vertebrate group, they are commonly subjected to QTL mapping studies, and have a significant level of economic importance in a range of countries.

2.5.1 Reviewed study designs

Our analysis revealed a steady increase in the number of QTL mapping studies over the past 15 years (Figure 2.1A). Overall, the studies were diverse and included a wide range of target species ($n = 48$), genetic markers (e.g. AFLPs, SNPs, SSRs), and study designs (e.g. two-stage linkage mapping, GWA). Interestingly, while a recent review by Ott et al. (2011) stated that linkage mapping has lost predominance in favour of association mapping, this trend was not obvious when evaluating fish studies (Figure 2.2). The absence of this trend is most likely the consequence of many studies being carried out on non-model species, which typically have fewer genetic markers and therefore favour linkage-based approaches. There was some variation in the types of mapping populations used by studies, although the majority of studies (60%) utilized a biparental cross and it was also common for studies (50%) to use interstrain or interspecies crosses (Supplementary figure 2.2, see Box 1 for some basic information about the advantages of different population designs). There has been an

increase in marker density for fish studies over the last 15 years (Figure 2.1C), with a steep increase since 2011, the latter being primarily driven by the development of Restriction site-Associated DNA (RAD) sequencing approaches (e.g. Genotyping-by-Sequencing, Elshire et al. (2011) and Hilario (2015)) coupled with decreasing sequencing costs. This increasing use of RAD sequencing is still working its way through the research community, with > 50% of published papers in 2015 still not using the technology. One possible reason for this is the relatively high level of technical and bioinformatic skills that this method requires. However, given its singular advantage over other methods it is expected that higher marker densities produced by RAD sequencing will become increasingly common for fish studies.

Box 2: Comparative genomics and data sharing

Transferring knowledge (e.g. about QTLs, candidate genes, and genetic markers) between studies and species provides a fruitful way for obtaining more widespread use of candidate genes and QTLs that have previously been identified. One of the more basic ways this might be done is by transferring information about candidate genes between species. One example of this was a study which compared growth hormone genes in a range of fish species while investigating a QTL in the promoter region that affects growth (Almuly et al. 2005). Another approach which can be used is meta-QTL analysis techniques which can be carried out by collating the results from multiple studies which share a set of genetic markers (e.g. Lanaud et al. 2009). Two programs which have been developed to facilitate this are BioMercator (Arcade et al., 2004) and cMap (Fang et al., 2003). With the recent development of high-quality reference genome assemblies a range of new comparative genomic techniques have become available (Sarropoulou et al., 2008; Sutherland et al., 2016). Sarropoulou et al. (2008) describes a comparative genomics approach where genome-wide EST markers from two species were aligned to the reference genome of a third species. This enabled orthologous linkage groups to be easily identified. To make this approach easier the software MapComp has been built to identify identical and proximal markers between linkage maps and use a reference genome of a related species as an intermediate. When a likely set of markers have been located, MapComp allows transfer of QTLs between closely related species, even when their marker sets are quite different (Sutherland et al., 2016). The application of comparative genomics in teleost fish is an important new development, because a wide diversity of non-model species are studied, which means that the overall research effort is spread much more thinly compared to land-based production species (e.g. cattle, corn, and wheat). Therefore, open sharing of fish reference genomes and maps will be highly beneficial to researchers, because these will facilitate data integration and iterative approaches to QTL discovery. The growth of comparative-based research will be fuelled by more high-quality genome assemblies, the reconstruction of clear evolutionary and phylogenetic relationships, and perhaps most importantly, open and effective data sharing.

2.5.2 Reviewed study results

In addition to reporting on the types of methods used by the reviewed studies, the current study attempted to evaluate the relative success of QTL studies. The ultimate goal of genotype-phenotype investigations, including QTL mapping, is to identify the genetic variants which collectively explain the heritable components of a given trait. However, a persistent problem is the inability to detect small effect QTLs, which are implicated in the “missing heritability problem” and “Beavis effect” (Eichler et al., 2010; Slate, 2013) and are thought to be responsible for much, if not most, of the heritable variation (Rockman 2012). A wide array of potential factors contribute to this QTL detection weakness including, a simple lack of statistical power (e.g. insufficient samples or markers, Hong and Park (2012), choice of study design (e.g. population design), genetic architecture for a given trait, and potential epigenetic and uncontrolled environmental factors (Sham and Purcell, 2014). A detailed quality assessment of the state of QTL mapping in teleosts was problematic because of the inherent variation in study designs, different genetic architectures among species, and the other factors previously mentioned. However, the general expectation is that as the power of studies increases a greater number of small effect QTLs should become detectable. This in turn should improve the amount of explained heritability and the total number of QTLs detected (see Figure 2 in Visscher et al. (2012)).

Using two basic measures (namely, number of QTLs found per trait and percentage variance explained) there appears to have only been minimal change to study results over the review period (Figure 2.3). There may be a very small increase in number of QTLs detected (Figure 2.3A) and a small reduction in percentage variance for identified QTLs (Figure 2.3B), but these changes are not beyond contestation. It should be noted that all studies used a minimum QTL quality limit of 95% confidence with multiple sampling correction, which is important if the number of QTLs is used as a measure of QTL identification success. Higher sample sizes and marker densities are two of the most commonly discussed experimental design factors influencing QTL detection and mapping accuracy (Hong and Park, 2012; Hu and Xu, 2008; Massault et al., 2008). The results of our review suggest that there was only a weak relationship between these factors and the number of QTLs detected (sample size: $r=0.29$, genetic markers: $r=0.10$, Table 1). For sample size, the weakness of this correlation was most likely influenced by the relatively small range of sample sizes (min = 30, max = 3297) among the published studies. Visscher et al. (2012) plotted similar statistics for a number of complex traits from human studies and found strong correlations using a much larger range of sample sizes (min = 2000, max > 175K). Their finding suggests that a large increase in sample size (1000s of individuals) is needed to improve QTL mapping for teleosts. Larger sample sizes would also be beneficial as studies continue to use higher marker densities, which require more stringent multiple sampling corrections to reduce false positives. Another correlation of note was a moderate negative correlation between sample size and PVE (Figure 2.4, table 2.1, $r = -0.48$), which supports the idea that sample size in the current studies may be important for reducing the bias towards the detection of only

large effect alleles (see Box 1) (see also "Beavis effect" discussion by Slate (2013)). Although there was limited correlation between the number of genetic markers and QTLs found, higher marker densities are necessary for fine-mapping and GWA studies. Ultimately the highest level of genome coverage would be best, but this should always be balanced against the cost and the relative importance of other study design factors (e.g. sample size). Lastly, the width of QTL regions was also investigated as a potential measure of QTL mapping success, but the inconsistency of how this was reported made inferences from this data difficult (see next section). Overall, limited improvement to the average width of QTL regions was observed over time.

2.5.3 Data reporting and software limitations

The reviewed studies employed a wide diversity of methods in their search for QTL regions, particularly when reporting QTL regions for linkage studies. For example, less than half of the studies (63 out of 146) reported QTL regions in cMs (Figure 2.3C) and those that did record the width of the region employed a number of different methods. Methods used included the allele drop-off (Le Bras et al., 2011), bootstrapping (Kirschner et al., 2012), Bayesian credibility intervals (Sauvage et al., 2012), set LOD limit from the peak of the QTL (Gagnaire et al., 2013), or distance between markers above the significance level cut-off method (Jin et al., 2012). The finding that QTL regions are described using a wide array of methods is surprising, given that simulation studies have clearly indicated that some approaches are more accurate than others (Manichaikul et al., 2006; Visscher et al., 1996). In particular, simulations have shown that Bayesian credibility intervals are more accurate than bootstrapping, which itself is more accurate than the allele drop-off method (Manichaikul et al., 2006; Visscher et al., 1996). One possible reason for this wide diversity in reporting is variation among available software packages. For example, some, such as R/QTL, include Bayesian credibility interval functions, while others, such as GridQTL, do not. Additionally, software packages are often restricted to data that was obtained following a particular experimental design. For example, R/QTL requires single full-sib families, whereas GridQTL allows half-sib family designs. An ongoing focus on software development and implementation of best practices will be important as QTL studies continue to employ more complex methods and handle more complex genotyping, phenotyping, epigenetic, and environmental data. Also, development of open source software that can be updated, added to, and has an active user/developer base would be of pronounced benefit.

Box 3: Improved phenotyping approaches

Whereas genotyping technologies have undergone a revolution in the last decades, advances in phenotyping have been slow. Indeed, despite awareness that the genotype-phenotype map is inaccessible without detailed phenotypic data our ability to characterize phenomes lags greatly behind our ability to characterize genomes (see discussion by Houle et al. (2010)). Improved high-dimensional phenotyping (many traits with known interactions) is important as it allows researchers to start to disentangle the interactions between phenotypes and between phenotypes and the environment, thereby reducing background variation and enhancing the power of QTL mapping studies (Benfey and Mitchell-Olds, 2008; Houle et al., 2010). Mixed model approaches looking at within and between trait variation are one example of the move towards a phenomics-based approach (Korte et al., 2012). Understanding the full range of phenotype and environmental interactions is also important for determining the biological relevance of genetic variants once they have been identified (Benfey and Mitchell-Olds, 2008; Grigorenko, 2005; Houle et al., 2010); for example understanding how those genetic variants produce a given phenotype under real world circumstances and environments.

One of the main limitations to the use of more complex phenotyping is the lack of high-throughput phenotyping methods (Houle et al., 2010). Individual phenotypes typically require unique measurement processes, which make collection of the large amounts of data necessary for phenomics-based approaches difficult or even impossible. However, noteworthy progress is starting to be made towards high-throughput phenotyping in some fields. For example, measuring wing traits in *Drosophila* through the construction of a wing machine (Houle et al. 2003). Likewise, in plants the software LeafAnalyser was developed to simplify the measurement of leaf shape variation by placing a large number of evenly distributed landmarks along leaf margins and by recording the position of each automatically (Weight et al., 2007). In fish, some recent progress has also been made, with some procedures being developed that utilise image-based data input to automatically extract traits from still images or video footage (Navarro et al., 2016; Silva et al., 2015; Viazzi et al., 2015). However these methods are often highly species-specific or require a specific technological set-up, as indicated by the absence of such methods in most if not all of the reviewed studies. Houle et al. (2010) noted three general areas that are critical for the development of phenomics, namely the development of technology (e.g. high-quality imaging), improved statistical and analytical capabilities, and better methods for integrating study results (see also Box 2). A lot of progress will come from methods that allow automated phenotyping technologies that are well integrated with databases, particularly from those technologies that can be applied to a suite of species and set-ups, rather than to a single scenario.

2.5.4 Epigenetic markers

While epigenetics was not investigated in any of the papers reviewed, the role that epigenetics plays in modulating gene expression and the genotype-phenotype process is increasingly

being recognised as well as its potential application in selective breeding (Zhang and Hsieh, 2013). Epigenetic mechanisms typically modify gene expression and these effects can vary depending on the genetic variation in and around regulatory elements. Locating epigenetic markers could help explain more precisely the role of genome modifications in producing a range of phenotypes. One good example of the power of epigenetics in QTL mapping is Cortijo et al. (2014), which recently reported several methylated regions that acted as QTLs and accounted for 60-90% of the heritable variation for flowering time and primary root length in *Arabidopsis*. The important role that epigenetics can play in fish has been highlighted by Baerwald et al. (2016), who identified differential methylation patterns between rainbow trout individuals with different migratory behaviour. To facilitate the future use of epigenetics in research, high-throughput sequencing technology can be used to detect both DNA sequence variation and certain epigenetic markers and modifications (e.g. bisulfite sequencing), which will make information such as the level of DNA methylation and histone modification relatively straightforward to add to genomic-based QTL study (Li and Tollefsbol, 2011).

2.5.5 Applications to selective breeding

Selective breeding is one of the most important applications that can benefit from QTL mapping results. Traditional selection techniques use information from a known pedigree and known phenotypes to produce estimated breeding values (e.g. BLUPs, Henderson (1984)). Moving from these traditional breeding approaches into programs applying Marker Assisted Selection (MAS) allows information from QTLs to be incorporated into the selection process (see discussion by Rezende et al. (2012)). Based on simulations MAS has the potential to increase genetic gains by 8-38% over traditional methods (Meuwissen and Goddard, 1996). These gains significantly increased if selection occurred before the trait was measured and if the trait had low heritability. A more recently developed method is Genomic Selection (GS) (Meuwissen et al., 2001), which simultaneously employs thousands of genetic markers that cover the entire genome (Goddard and Hayes, 2007) and with this can improve the genetics-driven approach to selective breeding even further. One of the main features/advantages of GS (as noted by Desta and Ortiz (2014)) is that it avoids the need for independent significance tests for each loci (used for QTL mapping and MAS) by accounting for all trait loci simultaneously while random effects from non-trait loci tend towards zero. As a result of this GS can more effectively use high density genome wide genotyping data sets than MAS and will likely become the gold standard of selective breeding in the near future. Currently, genomic selection has been used extensively in animal breeding (see Samorè and Fontanesi (2016)), to a minimal level in crop breeding (see Desta and Ortiz (2014)), and little at all in teleosts. While epigenetics may add another important layer to QTL studies (see previous section), it has been suggested that epigenetics would only provide limited additional gains to existing GS methods (Goddard and Whitelaw, 2014). The primary reason for this is that

if epi-mutations are stable then they will be in linkage disequilibrium with genetic markers, which means they will already be accounted for through existing GS methods (Goddard and Whitelaw, 2014). Alternatively, if they are unstable they will not be of much use for informing selective breeding decisions. As a final note, if genetic variation is being investigated for use in selective breeding then a genomic selection approach is best, but if the interest is for more thorough characterization of a specific underlying variant then a QTL mapping approach designed specifically to characterize causal QTNs might be best (Clop et al., 2006; Cohen-Zinder et al., 2005). For most studies in this review the focus was on QTL identification and the results would probably be best suited for use in MAS programs or further characterization with a second round of fine-mapping.

2.6 Future outlook and conclusions

The number of QTL mapping studies in fish has increased significantly over the review period and most were carried out on non-model species. Consequently, many studies used fewer than 1000 markers and employed linkage-based mapping approaches. Accordingly, only few of the studies were able to fine-map QTLs sufficiently to identify candidate genes or QTNs. However, the capability of studies has been increasing in recent times owing to the development of new approaches to variant discovery, such as RAD sequencing. Genotyping methods and the bioinformatics resources will continue to improve and these will be more widely adopted by the research community. Because the process of investigating genetic variation underlying traits is typically iterative, the studies conducted so far could be seen as representative of the first step towards identification of candidate genes and QTNs or as identifying QTLs suitable for informing MAS programs.

Future improvements to teleost genotype-phenotype investigations should involve a stronger move towards higher genome coverage of markers and a larger number of individuals in study groups (see Box 4). The ongoing construction of high-quality genome assemblies will help accelerate more widespread use of comparative genomics approaches; allowing knowledge about genes and QTLs to be transferred between species informed by a phylogenetic framework. Fish studies would greatly benefit from these comparative genomics approaches, which could help make up for the lack of concentrated resources that has been a feature of land-based production species. Areas such as phenomics and epigenetics are currently not widely used, however, are likely to take a more central role in future studies. An important aspect uniting these future areas is the ongoing need for better and more user-friendly data analysis software. This is particularly important for high-throughput genotyping, sequencing, and phenotyping technologies because the volume of information produced can only be managed with automation and high-performance computing. For phenotyping, well-designed software can provide an effective method for reducing the burden of collecting large volumes of phenotypic data and thereby allow researchers to focus on the key biological questions that are of interest.

In conclusion, over the last fifteen years there has been a large increase in the num-

ber of studies that managed to successfully identify QTLs in teleost fish. Many of these studies are underpowered compared to research that has been conducted on model species, but an ever increasing level of genomic resources is also becoming available for many of the non-model species. Additionally, the development of high-quality genome assemblies offers many options for comparative genomics and promotes a data-sharing community of researchers. Future genotype-phenotype investigations have a very optimistic outlook with the rise of more sophisticated methods and affordable solutions that can be applied to fish.

Box 4: Recommendations for future studies

With genotyping limitations being largely overcome in recent years, other areas that are likely to be important for future studies carrying out genotype-phenotype mapping include:

1. Increased sample size of the mapping populations to help account for the "Beavis effect", facilitate detection of weak effect QTLs and provide power needed for GWAS and polygenic analyses.
2. Polygenic analyses to allow identification of the interaction among loci and better reflect the polygenic nature of variation underlying phenotypic variation.
3. Advanced data analysis methods (e.g. mixed models) to facilitate use of increasingly complex phenomic, epigenetic, and environmental data in QTL investigations.
4. Comparative genomics and data sharing to facilitate more widespread use of known markers, QTLs and candidate genes in non-model species characterised by few genomic resources.
5. Development of high quality genome assemblies to facilitate comparative genomic and meta-QTL approaches.
6. Advanced phenotyping (phenomics) to help disentangle the interactions between phenotypes, genotypes, and the environment, thereby reducing background variation and increasing the power of QTL studies.
7. Epigenetic approaches to capture the interplay between environment and genes in affecting QTL expression.
8. Improved software to help researchers efficiently and accurately carry out a wide range of analyses while using best and most up to date methods. Advanced software and technology can also help with efficient phenotype data collection.

2.7 Implications for current thesis

The literature review carried out in this chapter provided some information that was useful for the current thesis (specifically Chapter 4 and 5). However, many of the decisions in genomic projects (e.g. sample size, population structure, software and methods), such as

this thesis, are determined by funding and logistical considerations. For example, this thesis investigated a pre-existing population so it was not possible to select the type of sample structure for QTL mapping (e.g. bi-parental pairs, unrelated, or complex). The choice of sample structure then determined which software could be used (e.g. had to support a complex pedigree design). The software which was used then determined the type of outputs that could be generated. The current thesis attempted to increase the sample size and marker numbers above the average sample size in the final year of this review (samples = >400, markers = >1,700). This literature review helped with finding software which could be used for QTL mapping in the current thesis (e.g. GridQTL which is used in Chapters 4 and 5).

Genetic diversity and heritability of economically important traits in Australasian snapper (*Chrysophrys auratus*) 3

Authors

David T. Ashton, Elena Hilario, Peter Jaksons, Peter A. Ritchie, & Maren Wellenreuther

Author Contributions

Conceived and designed the experiments: DTA, PR, and MW; Data collection: DTA; Laboratory work: DTA; Data analysis: DTA; Wrote the primary manuscript: DTA; Prepared final manuscript: DTA, PR, and MW; Advised laboratory or data analysis: EH and PJ

Publication status

This chapter is currently under review for publication in Aquaculture.

3.1 Abstract

Aquaculture is the fastest growing animal production sector in New Zealand, but low species diversity is a barrier to long-term growth. Australasian snapper (*Chrysophrys auratus* also *Pagrus auratus*) has been identified as a promising candidate for aquaculture development and an initial population has been established at the New Zealand Institute for Plant & Food Research Limited. The aim of this chapter was to combine high-throughput Genotyping-by-Sequencing (GBS) and trait data from this population to reconstruct the pedigree, measure the degree of inbreeding across generations, and determine the heritability of 11 traits of interest within the breeding programme, in particular growth-related traits. Molecular relatedness values showed that the pedigree consisted of a complex mixture of full and half-sibling individuals, with skewed contributions across parents. Average inbreeding did not change significantly between generations, but dramatic inbreeding differences were detected between F_2 descendants from the two independent founding (F_0) cohorts and between F_2 offspring from either full-sibling, half-sibling, or unrelated F_1 parents. The heritability for different traits ranged from 0.08 to 0.63. Heritability for growth rate averaged 0.26 and 0.40 in year one (when selective breeding would occur), based on the pedigree and genomic relatedness matrix, respectively. A selective breeding programme using the top 10% of individuals as parents could result in approximately 15.7 to 24.2% gains per generation for weight.

3.2 Introduction

Aquaculture has a fundamental role in meeting current and future global food needs (Bernatchez et al., 2017). In contrast to agricultural animals, many aquaculture species are not domesticated and have not been genetically enhanced through selective breeding programmes. Consequently, selective breeding programmes in aquaculture have the potential to yield significant gains (e.g. faster growth, greater disease resistance). This is particularly relevant for marine finfish, because there are a large number of novel species being explored for commercial aquaculture. In fresh water, an array of carp species represent the most extensively produced aquaculture fishes (Bostock et al., 2010). However, no equivalent species have been found for the marine environment. The development of marine fish species, which are suitable for aquaculture, is a strong focus in the South Pacific area around Australia and New Zealand. This is primarily because these countries have large coastal areas, but a limited number of domesticated marine species for marine aquaculture (Camara and Symonds, 2014; Gentry et al., 2017).

Breeding programmes can benefit greatly from the insights provided by genetic information. Genetic approaches can be used to reconstruct pedigrees, determine the contributions of individual parents, measure inbreeding and genetic diversity of a population, and identifying traits with a suitable genetic basis for enhancement through selective breeding. Until fairly recently, the high cost of genetic methods has limited the application of these tools

beyond model species (Gjedrem et al., 2012). However, the recent development of high throughput sequencing is beginning to remove this limitation. Large volumes of genetic data can now be generated for species with no or limited prior information at much lower costs than were previously possible (e.g. Genotyping-by-Sequencing, Elshire et al. (2011)). This is opening the opportunity for these genetic tools that were previously limited to model species to be applied to a wide array of new species (Ellegren, 2014).

For long-term breeding programmes it is important to understand the pedigree structure of the population and the likely effects this will have on genetic diversity. Loss of genetic diversity can lead to inbreeding, which can adversely affect important phenotypic traits and the long-term suitability of the population for aquaculture (Wang et al., 2002). Three key components to review in aquaculture populations are the number of individuals in the population, relatedness among individuals, and the contribution of each individual to subsequent generations (Falconer and Mackay, 1996). Contribution distortion is especially important to monitor in species that participate in group or otherwise undocumented breeding behaviours, as it can increase the rate at which genetic diversity is lost. It is also important to note that all genetic variation enters a breeding programme through the founding individuals. Once this variation is lost it can only be replaced by the introduction of new founding individuals, which significantly reduces the effectiveness of selective breeding programs.

Understanding the heritability of commercially relevant phenotypic traits is another important application of genetic tools to breeding programmes (Wan et al., 2017). Heritability can include either broad-sense heritability (phenotypic variation due to all genetic variation) or narrow-sense heritability (phenotypic variation due to additive genetic values). In the case of selective breeding programmes, narrow sense heritability is the most important because selection response depends on this (Wray and Visscher, 2008). Traits with a higher narrow sense heritability are better candidates for enhancement through selective breeding, while those with lower heritability are more suitable for enhancement via other factors (e.g. environment, feed, or other manipulations). While heritability estimates are population specific (Wray and Visscher, 2008), the heritability of many traits is often consistent across populations (Visscher et al., 2008).

In this study, we applied genetic tools to a new cultured population of the marine finfish Australasian snapper (*Chrysophrys auratus* also *Pagrus auratus*), which is being developed for aquaculture and research purposes. Snapper is a valuable commercial and recreational fish species located around the coasts of Australia, New Zealand, and several Pacific Islands. It is closely related to *Pagrus major*, a major aquaculture species in Japan (Murata et al., 1996). Despite snapper's recreational and commercial importance in New Zealand, genetic investigations on this species have been relatively scarce. Most of what is known comes from relatively low-powered studies carried out in the wild populations (Adcock et al., 2000; Hauser et al., 2002; Smith et al., 1978), but almost no genetic work has been conducted so far on the captive population. The specific goals of this chapter were to 1) reconstruct the

pedigree for the population using molecular relatedness, 2) investigate inbreeding within each generation, and 3) finally, estimate the heritability for a number of target traits, to determine their potential for enhancement through a selective breeding programme. The primary trait of interest for the breeding programme is growth rate. However there are a number of other traits which have also been considered (e.g. colour traits) and are investigated in this chapter.

3.3 Methods

3.3.1 Study population

The pedigree investigated in this study was a new population of the snapper that is being developed as part of a finfish breeding programme at the Maitai Seafood Research Facility in Nelson, New Zealand (41°25'44.96"S, 173°28'11.46"E). The Seafood Research Facility is located on the seaward side of Port Nelson and seawater is pumped into the facility from an underground bore. The water flowing to the early life stage section of the hatchery is further filtered using mesh filters and UV treatment. The original population was founded from wild sourced F_0 individuals ($n = 50$), the wild F_0 individuals were originally captured in two cohorts: the first cohort in 1994 and 1995 ($n = 25$) and the second in 2006 ($n = 25$). The first cohort of F_0 individuals (w1) were caught from several sites around the Tasman Bay and Marlborough Sounds, New Zealand (41°03'33.3"S, 173°15'01.3"E). The second cohort of F_0 individuals (w2) were caught from a single site within the Tasman Bay, New Zealand (41°03'33.3"S, 173°15'01.3"E). At the time of this study only the F_0 individuals from the second 2006 cohort were alive and available for sampling. An F_1 generation was produced over multiple years from either the wild F_0 individuals caught in 1996 (F_1 year classes: 2004) or the wild-caught individuals caught in 2006 (F_1 year classes: 2006, 2008, 2009, and 2010). F_1 individuals ($n = 70$) were combined into a single population in 2013 and subsequently produced an F_2 generation ($n = 577$ included in this thesis research). All breeding events were tank-based spawning, with equal sex ratios and all individuals able to mate freely with other individuals in the population. Prior to the spawning season parents were fed a specialized diet containing fresh fish and oil supplements. Fertilized eggs were collected from the tank outlet during the first two weeks of November. Individuals were observed to be spawning during the evenings and eggs were collected early in the morning and placed in hatchery tanks. Fertilized eggs were collected on a single day were used for each F_1 year class, but fertilized eggs collected over five consecutive days were used for the F_2 generation. Eggs from the five days were placed into four tanks with identical lighting, oxygen, water flow, and tank setups. The larvae were fed a combination of rotifers (*Brachionus rubens*) and artemia (*Artemia salina*). At 1 month old, the juveniles from all four tanks were combined into a single tank. The juveniles were then fed a combination of Artemia, dry crumb, and wet diet (minced fresh fish). At 1 year old, the juveniles were split into four tanks with identical lighting, oxygen, water flow, and tank setups. After 6

months of age, all fish were fed a combination of dry pellets and wet diet (fish mince or chunks of fresh fish). All research carried out in this study was approved by the Animal Ethics Committee of Victoria University of Wellington: Application number 2014R19.

3.3.2 Genotyping by Sequencing

Samples of fin tissue were collected from each individual at the beginning of the study (sample sizes; $F_0 = 25$, $F_1 = 70$, and $F_2 = 577$). Each sample was placed directly into chilled 96% ethanol, heated to 80°C for 5 minutes within 1 h of collection, and then stored at -20°C until needed. DNA was isolated using a modified salt extraction method (Aljanabi and Martinez, 1997). Quantification of DNA was carried out using Hoescht 33258 fluorescent dye. Fragmentation of the extracted DNA was checked by gel electrophoresis. Samples with moderate (25%) amounts of fragments below 10 kbp were re-extracted and if needed fresh samples were recollected.

Genotyping libraries were prepared for each sample following a modified Genotyping by Sequencing (GBS) protocol (Elshire et al., 2011). To build one library, one microgram of genomic DNA was double digested with the restriction enzymes PstI and MspI. The adaptor ligation step was done after digestion, without drying out the DNA/adaptor mixture. The barcoded adaptors designed by Deena Bioinformatics (van Gurp, 2011) were associated with the PstI cut sites. Adaptors were annealed according to the method of Ko et al. (2003). A high fidelity enzyme was used for amplifications (AccuPrime Taq DNA polymerase High Fidelity, Life Technologies). Each library was amplified separately and its efficiency assessed by capillary electrophoresis (Fragment Analyzer, Advanced Analytical). The GBS libraries were prepared in parallel in plates. Duplicate or triplicate samples were prepared for each of the parent and grandparents and single samples for each of the offspring (except three individuals with poorer DNA quality, for which duplicate samples were prepared). Each plate containing 96 individual libraries were pooled, cleaned up, quantified and sent to the Australian Genome Research Facility (AGRF) in Melbourne, Australia, for sequencing. Each pool was sequenced on a single lane from the Illumina HiSeq 2500 platform in Single End (SE) mode, with read length of 100 bases. In total, eight pools of libraries were sequenced in eight lanes.

3.3.3 Processing of sequencing data

Single Nucleotide Polymorphisms (SNPs) were extracted from the sequencing data using a combination of tools. FastQC was used to conduct an initial check of the sequencing data quality. Sequences were then de-multiplexed and cleaned. Adapters and primers were removed and the sequencing data were cleaned using Fastq-mcf in the ea-utils package (Aronesty, 2011). Genotyping was carried out on the cleaned datasets using the STACKs pipeline (Catchen et al., 2013). The samples were demultiplexed from the eight sequencing libraries using the process_radtags module. Sequencing reads for the duplicate or triplicate samples were concatenated into a single file, after which the reads were trimmed using

Fastq-mcf (minimum sequence length = 50, quality threshold causing base removal = 33). Bowtie 1.0 was used to align the GBS data to an initial snapper genome (version 1, number of scaffolds = 5,998, N50 = 1.5Mb, total bp = 738Mb, Wellenreuther et al. unpublished) being developed at Plant & Food Research (allowed mismatches = 3, reported alignments = 10). The pstacks module was then run with 8x or more sequence coverage required to call each SNP (this level of sequence coverage is indicated to provide >99% accuracy when SNP calling, Song et al. (2016)), followed by cstacks and sstacks; pre-set parameters were used for all these modules. The population module was used to output the data to a Genepop file (minimum minor allele frequency = 0.05, allowed missing data = 0.25, additional commands = `-write_random_snp`).

3.3.4 Molecular pedigree reconstruction

The parents for each individual were determined using Cervus v3.0.7 (Kalinowski et al., 2007) and a subset of SNPs ($n = 2174$) that were present in >98% of individuals. The parents for each individual were selected as the two closest matches, which passed the 95% confidence limit for assignment based on Cervus's built in simulation procedure. The two parents were designated as mother or father based on sex information collected previously for the broodstock populations. A network displaying the pedigree structure was constructed based on these relatedness scores using custom code in the R statistical environment v3.2.3 (R Core Team, 2013).

3.3.5 Calculation of inbreeding levels

A subset of SNP markers ($n = 6441$) that had been successfully placed on the linkage map and were present in >80% of the individuals were used to calculate a method-of-moments F coefficient (F_H) for inbreeding for each individual. This statistics is calculated as (observed homozygotes – expected homozygotes) / (total observations – expected homozygotes) (Kardos et al., 2015) and is equal to Nei's F statistic, but is calculated using a different formula. Inbreeding (F_H) was calculated for each individual which had contributed offspring or was part of the final generation using the `-het` command in the software package PLINK v1.9 (Purcell et al., 2007). The distribution of inbreeding values was then visualized using ggplot2 library in the R statistical environment v3.2.3 (R Core Team, 2013; Wickham, 2009). A Welch two-sample t-test was used to compare the inbreeding values for the offspring from the first and second F_0 cohort. Pairwise t-tests with Bonferroni correction were used to compare the inbreeding values for the F_2 individuals subdivided by grandparent type (w1-w1, w1-w2, and w2-w2) and parent type (full-sibling, half-sibling, and unrelated).

3.3.6 Phenotyping and trait correlations

A total of 11 phenotypic traits were measured for the F_2 individuals including fork length, peduncle length, weight, height at 0.25, 0.50, and 0.75 along the peduncle length from the

Figure 3.1: The 11 phenotypic traits measured in the Australasian snapper (*Chrysophrys auratus*). Nostrils, sex, and weight were measured manually and all the others were measured using custom image analysis scripts. The ruler in picture was used to convert all the length measurements to mm. Height measurements were measured relative to length. Skin darkness was measured by comparing the average pixel values along the edge of the fish in comparison with the background.

nose, number of nostrils, sex, survival from one to three years old, number of spots, and external skin darkness (Figure 3.1). External images of the left side of each individual were taken in the F₂ generation at 436-487 days (hereafter referred to as ‘year one’ fish) and 1045-1131 days (hereafter referred to as ‘year three’ fish). Images were collected with a Panasonic DMC-GH4 camera (16 megapixels) under natural lighting inside a large tunnel-house constructed of light diffusing material. The distance between the camera and fish varied between photos so a ruler was included in each image to provide a scale. All phenotypes, except sex and weight, were extracted from these images either manually or using custom Python 3.0 scripts and the image analysis library OpenCV 2.0. Survival from year one and year three was determined by the presence or absence of an individual in the first and second set of images. The sex of each fish was determined by checking if it was producing milt or eggs during the middle of the breeding season (January - February) after the individuals had reached three years of age in 2016. Individuals that were not obviously producing milt or eggs were assumed to be female, because stripping eggs from females is more difficult. Weight was collected manually by placing the fish on a Mettler-Toledo SB8001 scale with 0.1g accuracy. Fish were anaesthetized with Aqui-S™ to allow handling of fish when collecting all data. The correlations between individual traits was measured using

a Pearson's correlation matrix which was constructed using all phenotypic measurements (year one and three) in Python 3.0 using the Numpy library (McKinney, 2010).

3.3.7 Trait heritability and selection potential

The narrow sense heritability for each of the phenotypic traits was estimated using linear mixed animal models and restricted maximum likelihood methods with ASREML (Gilmour et al., 2009) in the R statistical environment v3.2.3 (R Core Team, 2013). Heritability for each trait was calculated with both a pedigree relatedness matrix (PRM) (year one and year three) and a genomic relatedness matrix (GRM) (year one). For each model the target trait was predicted using a fixed intercept effect and the tank and relatedness matrix (PRM or GRM) as random effects. All traits were fitted as continuous, after testing indicated this provided higher log likelihood scores than using a logit function for the binomial traits. Sex and origin of the F_0 populations were tested as fixed effects in the PRM model, but did not have a significant effect on the results. Maternal effects were tested in year one with the GRM model using a GRM model without maternal effects (GRM), a GRM model with a maternal effects (GRM+M), and a model including only the maternal effect (M-only). The potential gains for selective breeding of the top 10% of parents was calculated for each continuous trait based on the heritability and trait distribution - using the selection response formula ($R = h^2S$). In this formula h^2 is narrow sense heritability and S is the difference between the mean trait value for parent generation and the mean trait value of individuals above the 90% confidence interval for the target trait (i.e. those used for selective breeding). The trait distributions from the F_2 individuals were used for all calculations. The response to selection values (R) from this formula were converted to percentage gains per generation by dividing by the mean trait value from the non-selected parent group and multiplying by 100.

3.4 Results

3.4.1 Quantity and quality of sequencing data

A total of 1.6 billion reads were generated from the eight sequencing lanes with approximately two, four, or six million reads generated for each single, duplicate, or triplicate library, respectively (Supplementary figure 3.1). Coverage of reads was consistent across all samples, with few having noticeable lower or higher coverage. All samples were included in further analysis. FastQC results indicated that the read quality was very high throughout the entire read (Illumina quality scores above 30 throughout reads) (Supplementary figure 3.2). From the STACKS pipeline a total of 249,468 SNPs were identified among 672 samples with 8x or more sequence coverage; of which 20,311 were present in greater than 75% of individuals in the population and had a Minor Allele Frequency (MAF) greater than 0.05. The average coverage per SNP was 15.6x in the offspring (F_1) and 23.9x in the parents (F_2).

3.4.2 Pedigree reconstruction based on molecular relatedness values

Figure 3.2: The pedigree structure of the Australasian snapper (*Chrysophrys auratus*) population investigated in this thesis (A). Each individual is represented either as male (blue triangle), female (red circle), or as unknown gender (grey square). Individuals are placed into rows based on the year hatched, and are connected to their respective parents by lines. The generations F₀, F₁, and F₂ are displayed. The number of offspring produced by each parent in the F₁ generation divided into females (dark grey) and males (light grey) (B).

Both parents were identified for 64% and 94% of the individuals in the F_1 and F_2 generations, respectively. The number of parents that contributed offspring was 60% ($n = 15$ out of 25) for the F_0 generation and 71% ($n = 53$ out of 70) for the F_1 generation. The missing parents from the F_1 generation were mainly located in the year classes produced from the missing wild F_0 cohort. The 6% that were missing parents from the F_2 generation did not pass the 95% confidence limit for assignment. The top two potential parents assigned for each individual were shown to be consistently male-female. Visualization of the pedigree showed that a large number of individuals had contributed from the F_0 to F_1 generation and from the F_1 to F_2 generation, but that the contributions were highly skewed, with some individuals contributing many more offspring than others (Figure 3.2). A closer look at these contributions in the F_1 parents showed that the highest producing female and male produced 39% and 16% of the offspring, respectively. The F_1 and F_2 generations contained a complex mixture of full-siblings, half-siblings, and unrelated individuals.

3.4.3 Inbreeding in the pedigree

The method-of-moments F coefficient (F_H) was calculated for 611 individuals in the dataset, which were either parents or offspring with known parents. The values ranged from a minimum of -0.57 to a maximum of 0.55 with a median of 0.02. Variation in the inbreeding values was lowest in the wild-caught F_0 generation, but this group also contained the fewest individuals (Figure 3.3A). The variation in inbreeding values for the F_1 generation increased dramatically (-0.57 to 0.25) over those in the wild-caught F_0 individuals, but the median did not change significantly (p -value = 0.7410, Figure 3.3A, Supplementary table 3.1A). In the F_2 generation, the variation decreased from that observed in the F_1 generation, but was still higher than in the original wild-caught F_0 individuals (Figure 3.3A). In the F_2 generation there was also a skewed distribution towards higher inbreeding values (Figure 3.3A). Subdividing the F_2 offspring into those that were the product of the first wild broodstock (w1), second wild broodstock (w2), or a combination (w1-w2) it was found that all groups were significantly different from each other (p -value < 0.001, Figure 3.3B, Supplementary table 3.1B). Further subdividing F_2 individuals from the w2 group into offspring resulting from full-sibling, half-sibling, and unrelated crossing events showed a clear relationship between the degree of parental relatedness and the coefficient of inbreeding (p -value < 0.05, Figure 3.3C, Supplementary table 3.1C). Offspring from full sibling crosses were the most inbred, followed by half-sibling crosses, and the offspring of unrelated individuals had the lowest coefficient of inbreeding.

3.4.4 Trait values and phenotypic correlations

Phenotype data were recorded for 11 traits in the F_2 generation and a full list of means, standard deviations, and measurement counts for each trait is shown in Supplementary table 3.2. The number of measurements per individual and year differed depending on the availability of the individuals. A drop in sample size (577 to 417 individuals) occurred

Figure 3.3: Inbreeding scores for Australasian snapper (*Chrysophrys auratus*) grouped based on A) generation, B) grandparent types (F₂ only), or C) parent types (w2-w2 only). The grandparent types are: w1-w1 = F₂ from the first wild cohort, w2-w2 = F₂ from the second wild cohort, and w1-w2 = F₂ from a mixture of the first and second wild cohort. Visualized are the 1st, 2nd (median), and 3rd quartile and whiskers extending 1.5 times the interquartile range from the median (95% confidence interval). The coefficient of inbreeding used is Nei's FIS and significant differences between groups are shown in Supplementary Table 2. Significant differences (< 0.01) were found between the w2-w2 and the other two grandparental types. Significant differences (< 0.05) were found between all the different parent types.

between year one and year three as a result of natural mortality. High variation was observed in growth-specific traits including fork length (year one: $155.8 \text{ mm} \pm 18.0$, year three: $257.7 \text{ mm} \pm 21.0$), peduncle length (year one: $128.4 \text{ mm} \pm 15.0$, year three: $214.3 \text{ mm} \pm 17.6$), and weight (year one: $75.2 \text{ gm} \pm 25.3$, year three: $363.1 \text{ gm} \pm 85.0$). Relative to fish length, the weight and height increased disproportionately over the two years between measurements. Skin darkness of the fish also increased from year one to year three (year one: 89.6 ± 10.0 , year three: 111.3 ± 8.4). Sex ratios measured during year three identified a slightly skewed sex ratio based on the ability to strip milt or not (female: 245, male: 182), assuming those without milt were female. The number of spots was highly varied, but remained consistent between year one and year three (year one: 43 ± 9.6 , year three: 43.7 ± 8.5).

Based on Pearson's correlation coefficients (Table 3.1), strong phenotypic correlations were observed between all the growth traits ($r > 0.93$ between fork length, peduncle length, and weight) which were also strongly correlated with age ($r > 0.93$ with fork length, peduncle length, and weight). Age was further moderately correlated with skin darkness ($r = 0.46$) and survival ($r = -0.31$). No major correlations were found between sex and other traits. Height 0.25 was found to be weakly positively correlated with age ($r = 0.25$), while height 0.75 was found to be weakly negatively correlated with age ($r = -0.26$).

Table 3.1: Pearsons correlation coefficient (r) between all traits in both year one and year three. Sample sizes for each trait are shown in Supplementary table 3.1. FL = fork length, PL = peduncle length, Wgt = weight, H_0.25 = height 0.25, H_0.5 = height 0.5, H_0.75 = height 0.75, Skin = Skin darkness.

-	Age	FL	PL	Wgt	H_0.25	H_0.5	H_0.75	Skin	Spots	Nostrils	Sex
FL	0.93										
PL	0.93	1.00									
Wgt	0.93	0.98	0.98								
H_0.25	0.25	0.18	0.18	0.21							
H_0.5	-0.05	-0.04	-0.04	0.07	0.50						
H_0.75	-0.26	-0.21	-0.22	-0.05	0.36	0.71					
Skin	0.46	0.46	0.46	0.57	0.04	-0.11	-0.09				
Spots	0.08	0.06	0.05	0.16	0.04	0.07	0.02	-0.03			
Nostrils	0.03	-0.03	-0.03	-0.06	-0.01	-0.06	-0.09	0.01	-0.16		
Sex	-0.04	0.00	0.00	0.03	-0.01	0.04	0.08	0.03	0.09	-0.11	
Survival	-0.31	-0.38	-0.38	-0.40	-0.06	0.01	0.10	-0.15	-0.12	0.03	-

3.4.5 Trait heritability and selection potential

Trait heritability estimated with the PRM models varied widely depending on the trait and year it was measured (Table 3.2). The three growth traits all had similar heritability, averaging 0.27 in year one and 0.10 in year three. The heritability of height was relatively stable between year one and year three, and increased as the measurement of height moved from the nose of the fish towards the tail (year one = 0.15 ± 0.09 to 0.26 ± 0.11 , year three = 0.14 ± 0.09 to 0.30 ± 0.14). The heritability of skin darkness was greater in year three than in year one (year one = 0.03 ± 0.03 vs year three = 0.22 ± 0.18), but overall was either extremely low or had high standard error. Number of blue spots was the most heritable trait in the data set based on the PRM models (year one = 0.45 ± 0.13 , year three = 0.63 ± 0.18). Heritability for the three binomial traits was only recorded once. Survival had low heritability (0.08 ± 0.06), but number of nostrils and sex were both moderately heritable (0.34 ± 0.12 and 0.16 ± 0.09 , respectively). Percentage gains per year for a selective breeding programme based on PRM heritability and trait distribution were calculated in both year one and year three (Table 3.2). The percentage gains for growth rate traits (lengths and weight) ranged from 4.6 to 15.7 in year one and 1.4 to 4.9 in year three. Number of blue spots had the highest predicted percentage gains per year (year one: 18.8, year three: 23.8). The remaining traits had percentage gains ranging from 0.7 to 3.3 for years one and year three.

The heritability of traits based on the GRM models had a similar pattern to the PRM models, but with some differences (Table 3.3). The heritability of growth traits was higher based on the GRM models than the PRM models – average 0.42 in year one. No maternal effect was observed for this trait. The heritability of height traits in the GRM models (average = 0.12) was much lower than the PRM models, and had relatively large maternal effects. Skin darkness had minimal heritability (0.06 ± 0.05). Number of blue spots was still one of the most heritable traits (0.44 ± 0.07), was similar to the results from the PRM model, and had minimal maternal effects (0.07 ± 0.06). Similar to the pattern in the PRM models, survival had low heritability (0.1 ± 0.05), but nostrils and sex both had moderate heritability (0.31 ± 0.07 and 0.31 ± 0.07 , respectively). All three binomial traits had low or no maternal effects. Percentage gains per year for a selective breeding programme based on GRM+M heritability and trait distribution were calculated in year one (Table 3.3). The percentage gains for growth rate traits (lengths and weight) ranged from 6.3 to 24.2% and was much higher than that predicted with the PRM models. Number of blue spots had similar predicted gains to the PRM models (13.1%). The remaining traits had very low percentage gains ranging from 0.2 to 0.8%.

Table 3.2: This table shows the heritability and standard errors based on the pedigree relatedness matrix (PRM) for each trait in year one and year three of the current study. Also shown are the potential selection gains (%) based on a selective breeding programme in which only the top 10% of individuals were used. The three binomial traits were only measured once for each individual.

Trait		Heritability		Selection gains %	
Category	Name	Year one	Year three	Year one	Year three
Continuous	Fork length	0.25 ± 0.10	0.10 ± 0.08	4.6	1.5
Continuous	Peduncle length	0.25 ± 0.10	0.09 ± 0.08	4.6	1.4
Continuous	Weight	0.30 ± 0.14	0.11 ± 0.10	15.7	4.9
Continuous	Height 0.25	0.15 ± 0.09	0.14 ± 0.09	1.6	1.0
Continuous	Height 0.50	0.25 ± 0.11	0.20 ± 0.12	2.0	1.7
Continuous	Height 0.75	0.26 ± 0.11	0.30 ± 0.14	2.7	3.3
Continuous	Skin darkness	0.03 ± 0.03	0.22 ± 0.18	0.7	3.3
Continuous	Spots	0.45 ± 0.13	0.63 ± 0.18	18.8	23.8
Binomial	Nostrils	0.34 ± 0.12		-	
Binomial	Sex	0.16 ± 0.09		-	
Binomial	Survival	0.08 ± 0.06		-	

Table 3.3: This table shows the heritability and standard errors based on the genomic relatedness matrix (GRM) for trait in year one of the current study. The results for three different models are shown: GRM = GRM model without maternal effects, GRM+M = GRM model with maternal effects, and M-only = a model only including maternal effects. Also shown are the potential selection gains (%) based on the GRM+M model and a selective breeding programme in which only the top 10% of individuals were used.

Trait		Heritability			Selection gains %
Category	Name	GRM	GRM+M	M-only	GRM+M
Continuous	Fork length	0.35 ± 0.07	0.38 ± 0.08	0.00 ± 0.00	6.3
Continuous	Penduncle length	0.35 ± 0.07	0.38 ± 0.08	0.00 ± 0.00	6.3
Continuous	Weight	0.51 ± 0.12	0.52 ± 0.12	0.00 ± 0.00	24.2
Continuous	Height 25	0.09 ± 0.05	0.07 ± 0.05	0.03 ± 0.04	0.7
Continuous	Height 50	0.18 ± 0.07	0.11 ± 0.07	0.11 ± 0.07	0.8
Continuous	Height 75	0.17 ± 0.06	0.08 ± 0.06	0.07 ± 0.05	0.7
Continuous	Skin darkness	0.06 ± 0.05	0.01 ± 0.04	0.01 ± 0.02	0.2
Continuous	Spots	0.44 ± 0.07	0.35 ± 0.08	0.07 ± 0.06	13.1
Binomial	Nostrils	0.31 ± 0.07	0.27 ± 0.08	0.02 ± 0.04	-
Binomial	Sex	0.31 ± 0.07	0.29 ± 0.12	0.00 ± 0.00	-
Binomial	Survival	0.10 ± 0.05	0.11 ± 0.06	0.00 ± 0.00	-

3.5 Discussion

In this study a cultured snapper population was investigated using a GBS dataset. This dataset was used to reconstruct the pedigree, investigate the genetic diversity and inbreeding rates in each generation, and calculate the heritability of important phenotypic traits.

The results of this study will help the development of this new population by providing information about what has occurred during the initial generations and informing future breeding efforts.

Pedigree reconstruction based on molecular relatedness indicated that many of the individuals in the F_0 and F_1 generations contributed to offspring in the next generation (e.g. 58 out of 70 individuals in the F_1 generation), but the contributions were highly skewed. Skewed contributions have been observed in a wide range of captive fish populations including Asian seabass (*Lates calcarifer*) (Liu et al., 2012), gilthead seabream (*Sparus aurata*) (Chavanne et al., 2014), and flounder (*Paralichthys olivaceus*) (Sekino et al., 2003). Contribution distortion can be problematic in a long-term breeding programme because of the negative effects that it has on genetic diversity. Because this study was carried out on post-juvenile fish, two main explanations could account for the skewed distribution. Firstly, the parents may have contributed unevenly to breeding events. Secondly, the parents may have contributed evenly, but their progeny had different survival rates. Given the fact that the F_2 snapper were produced from five consecutive days of eggs, which were produced at the peak of the egg production for snapper, the second of these two explanations seems likely. However, further work is needed to clarify what is occurring during this stage of the snapper production cycle.

The inbreeding statistic used in the current study was a method-of-moments F coefficient (F_H), which is also equivalent to Nei's F statistic. With genome wide markers this was previously found to be an accurate measure of inbreeding (Kardos et al., 2015). Some interesting patterns of inbreeding were observed that appeared to begin with the originally from the two F_0 wild cohorts. The F_0 grandparents were sourced from the wild and as such represent a baseline for other inbreeding statistics. Unfortunately, genetic samples were only available for the second F_0 wild cohort. However, we can gain some idea about the absent F_0 cohort through their progeny in the F_1 and F_2 generations. The average F_H (also Nei's F statistic) for the second F_0 wild cohort (0.055) was within ranges observed for other marine fish from wild populations, including orange clown fish (*Amphiprion percula*, 0.018) (Salles et al., 2016), brook trout (*Salvelinus fontinalis*, 0.098) (Pilgrim et al., 2012), winter flounder (*Pseudopleuronectes americanus*, 0.169 - 0.283) (O'Leary et al., 2013), and gilthead seabream (*Sparus aurata*, 0.00 - 0.319) (Zeinab et al., 2014). It is worth noting that the above study in winter flounder was reporting severe inbreeding for a wild population (O'Leary et al., 2013). While the average inbreeding rates in the current study did not change significantly between generations, the distribution of inbreeding values was noticeable different. Some of the changes to that variation could be explained by differences in sample size between the generations; however, further analysis indicated that F_1 individuals with low values (outbred) were primarily located in the offspring of the first F_0 cohort (w1) and those with high values (inbred) were located in the offspring of the second F_0 cohort (w2). These results suggest that some structure may be present in the source population and that while the second F_0 cohort was sourced from a single population,

the first cohort was sourced from multiple populations, which have subsequently produced outbred offspring. These results are particularly interesting because most previous studies have suggested that minimal structure is present in the wild snapper population around New Zealand (Bernal-Ramírez et al., 2003; Paul and Tarring, 1980; Smith et al., 1978) and none would be expected over the range from which these two F_0 wild cohorts were sourced. The differences in inbreeding continued into the F_2 generation with individuals that were the product of solely the first F_0 cohort (w1-w1), being significantly (p -value < 0.001) less inbred than those produced solely from the second F_0 cohort (w2-w2). F_2 individuals that were the product of crossing between the two F_0 wild cohorts (w1-w2) had the lowest average inbreeding values for any group. As expected, further subdivision based on parent type indicated that F_2 individuals that were from unrelated individuals had significantly lower inbreeding values than those from half-sibling or full-sibling parents.

Trait values and heritability differed largely across the 11 traits investigated. Growth traits (e.g. weight and length) are some of the most commonly reviewed traits for aquaculture selective breeding programmes. This is because they directly affect production rates and are often moderately heritable; for example, heritability of growth traits (weight or length) was 0.2 to 0.4 in mirror carp (*Cyprinus carpio*) (Chen et al., 2017), 0.4 in gilthead seabream (*Sparus aurata*) (Fernandes et al., 2016), 0.21 to 0.362 in half-smooth tongue sole (*Cynoglossus semilaevis*) (Liu et al., 2016), and 0.31 to 0.34 in Atlantic cod (*Gadus morhua* (Kristjánsson and Arnason, 2016)). In the current study, growth traits had an average heritability of 0.26 in year one and 0.11 in year three based on the PRM and 0.41 in year one based on the GRM. Additionally, because the three growth traits (fork length, peduncle length, and weight) were all highly correlated (> 0.97), selection for one should also affect the others. This could prove useful in a breeding programme because length can often be more easily measured using high-throughput methodologies than weight. Body shape measurements are another commonly measured group of traits; for example, heritability for shape traits ranged from 0.18 to 0.289 in Nile tilapia (*Oreochromis niloticus*) (Oliveira et al., 2016), 0.24 to 0.58 in gilthead seabream (*Sparus aurata*) (Boulton et al., 2011), and 0.34 in common sole (*Solea solea*) (Blonk et al., 2010). By comparison heritability for height traits in the current study ranged from 0.14 to 0.30 for both years measured, but was lower based on the GRM and had large maternal effects. Skin darkness was found to only be weakly heritable in the current study. This would indicate an environmental basis for variation in this trait, which matches anecdotal observations at PFR. The number of blue spots was one of the most heritable traits in the current study. The number of nostrils was found to be weakly heritable (PRM model: 0.34 ± 0.12 , GRM model: 0.34 ± 0.07). This is interesting because previous research at Plant & Food Research (unpublished) has indicated that this trait may be affected by tank-specific conditions within the first few months of life. If so, this heritability may instead be an artefact of distorted parental contributions and multiple larval rearing tanks with variable conditions. Further work is needed to clarify this. Sex and survival both had low heritability based on the PRM model (sex: 0.16 ± 0.09 , sur-

vival: 0.08 ± 0.06) and low or moderate heritability based on the GRM model (sex: 0.31 ± 0.07 , survival: 0.10 ± 0.05). Survival against specific diseases is often highly heritable for fish species; for example, 0.39 for viral resistance in Rainbow trout (*Oncorhynchus mykiss*) (Flores-Mara et al., 2017), 0.38 to 0.52 for streptococcus resistance in Nile tilapia (*Oreochromis niloticus*) (Shoemaker et al., 2017), and 0.12 to 0.45 for pasteurellosis resistance in the gilthead seabream (*Sparus aurata*) (Antonello et al., 2009). However, in this study no single disease event was shown to result in the death of the fish, which could explain the low heritability for survival. None of the phenotypic traits were significantly correlated with sex, indicating a lack of clearly visible sexual dimorphism in snapper population from this thesis. However, this may be affected by the age of the study population, if the external sexual traits are yet to develop. It should also be noted that the sex determination in the current study may be affected by the transition of immature females to sexually mature males and females around 2-5 years of age (Parsons et al., 2014) – with more females expected in this study (female: 245, male: 182) as some may not yet have gone through the transition by year three.

The percentage gains per generation for a selective breeding programme can be estimated using the selection response formula ($R = h^2S$). The degree of change in this formula is affected by both the heritability of the trait and the difference in mean value between the selected and non-selected parents, which is in turn affected by the intensity of selection pressure applied and the distribution of the trait. A selective breeding programme in snapper would likely apply selective pressure at around 1 year of age. Results from this study suggested that a selective breeding programme for weight using the top 10% of individuals could expect around 12.2% gains per generation based on the PRM model and 24.2% based on the GRM model. This is comparable or higher than the suggested rate in sea bream (*Sparus aurata*, 15.6%) (Thorland et al., 2006) and near the average of 13% suggested across a wide range of fish species (Gjedrem and Rye, 2016). However, the gains at three years of age were much lower, at around 5.2% per generation. This drop in gains at year three is a result of diminishing heritability as the fish age, but the reason for this drop in heritability is unknown and needs further investigation. A drop in heritability indicates that non-genetic factor (e.g. size based tank effects) is increasingly becoming the driver of variation in growth as the fish age. One final note on these calculations is that although weight and length had similar heritability (year one: 0.26, year three: 0.11), gains for length were less than half of those expected for weight. This is because length had a much narrower trait distribution than weight (length = 257.7 ± 21.0 , weight = 363.1 ± 85.0). This higher level of variance in weight provides more range for selection to act upon. Gains for the remaining traits were all low (5.3 or less), except for number of spots, which had some of the highest gains of any trait in the study based on the PRM models (year one: 18.4%, year three: 22.8%).

3.5.1 Future directions

This study represents the most in-depth genetic investigation of a cultured snapper population to date. The resources developed will be useful for a range of future investigations (see Chapter 4 and 5) and be useful for supporting the snapper selective breeding program (discussed in Chapter 6). One future direction is to further investigate the genetic basis of traits with moderate heritability from the current chapter (e.g. growth traits and number of blue spots). This can be done using Quantitative Trait Loci (QTL) and candidate gene mapping to identify genome locations and potentially even specific genes, which are affecting traits of interest. Further investigation of growth traits (fork length, peduncle length, and weight) and blue spots is carried out in Chapter 4 and 5, respectively. Another future direction for this research, is the application to the snapper breeding programme. This is discussed further in the Chapter 6. Importantly, the results from this study indicated that snapper show high potential for genetic improvement of economically important growth traits. However, some findings from this chapter, such as the contribution distortion, require further consideration within the developing breeding programme. The genomic resources developed in this chapter could be used to facilitate further domestication of snapper in a way that enhances economically important traits while maintaining genetic variation and reducing any potentially negative effects of inbreeding.

Authors

David T. Ashton, Peter A. Ritchie, & Maren Wellenreuther

Author Contributions

Conceived and designed the experiments: DTA, PR, and MW; Data collection: DTA; Data analysis: DTA; Wrote the primary manuscript: DTA; Prepared final manuscript: DTA, PR, and MW

Publication status

This chapter is currently under review for publication in G3.

4.1 Abstract

Characterizing the genetic variation underlying phenotypic traits is a central objective in biological research. This research has been hampered in the past by the limited genomic resources available for most non-model species. However, recent advances in sequencing technology and related genotyping methods are rapidly changing this. In this chapter the genome-wide SNP dataset and phenotype data from Chapter 3 was used to 1) construct the first linkage map for Australasian snapper (*Chrysophrys auratus* also *Pagrus auratus*), 2) scan for growth QTLs, and 3) search for candidate genes in the surrounding QTL regions. The newly constructed linkage map contained 11K SNP markers. Comparisons with available genome scaffolds indicated that overall marker placement was strongly correlated between the scaffolds and linkage map ($r = 0.7$), but the positioning of markers was not precise at finer scales (< 5 cM). Of the 24 linkage groups, which reflect the 24 chromosomes of this species, three were found to contain QTLs with genome-wide significance for growth-related traits. A scan for 13 known candidate growth genes located the genes for growth hormone, myogenin, and parvalbumin within 5.3, 9.6, and 25 cM of these genome-wide significant QTLs, respectively. The linkage map and QTLs found in this study will advance the investigation of genome structure and selective breeding in snapper.

4.2 Introduction

Characterizing the genetic variation that affects phenotypic traits is a central goal in biology. Understanding this variation can inform selective breeding programmes (Dekkers, 2012), be used to predict disease risk in medicine (Lehner, 2013), and help researchers to understand evolution in natural populations (Savolainen et al., 2013). While genetic research has typically been pioneered in laboratory model species, the development of cheaper high-throughput genomic methods (e.g. high-throughput DNA sequencing) is now allowing this research to be extended to a wide range of non-model species (Braasch et al., 2015; Hilario, 2015).

Locating and characterizing Quantitative Trait Loci (QTLs) is one common approach used to investigate how genetic variation influences a specific phenotype (Barria et al., 2017; Chen et al., 2017; Pértille et al., 2017). QTL mapping methods locate specific loci that are influencing a phenotypic trait based on a significant correlation between allelic variation of those loci and variation of the trait (Lynch and Walsh, 1998). New QTLs can be informative as standalone observations, or used to identify candidate genes and causative variants in the surrounding genome, which may be influencing the trait (Bettembourg et al., 2017). Genotype-phenotype datasets can also be used to develop multi-marker (polygenic) models which explain variation of one or more phenotypic traits (e.g. genomic selection models, Cros et al. (2017)).

Having a road-map of the genome (e.g. linkage map or high-quality genome assembly) is an important prerequisite for QTL mapping, as it allows the relative positioning of dif-

ferent marker loci. High-quality genome assemblies are most effective because they allow genetic markers to be positioned at a base-pair level - while also providing sequence information for the surrounding genome. However, most non-model species do not currently have chromosome-level genome assemblies and instead rely on linkage maps to ascertain the relative position of markers in the genome. Although they are much less precise than high-quality genome assemblies, linkage maps can serve the dual purpose of bridging the resource gap before a genome is developed and providing useful information to improve the arrangement of scaffolds during the genome assembly process (Fierst, 2015). Data sets that are developed for constructing a linkage map can also be used in QTL mapping.

Research on marine teleost fish is one area that has potential to benefit from more affordable and higher-throughput genomic technologies. Teleost fish are the largest group of vertebrates, with approximately 33,000 species of fish, most which have limited geographic range and many are of commercial significance. The high diversity of production species means that scientific efforts are spread thinly, with few to no genomic resources available for many species. One such species, and the focus of this study, is the marine finfish Australasian snapper (*Chrysophrys auratus* also *Pagrus auratus*). Snapper supports a valuable recreational and commercial inshore fishery around the northern parts of New Zealand, southern Australia, and some of the Pacific Islands and is a strong candidate for development into an aquaculture species in both New Zealand and Australia. However, until recently there has been limited genetic and no genomic research carried out on this species

One of the most important traits in farmed species is growth rate, as it directly affects the efficiency of production systems. In many animal species growth is a complex trait that is influenced by a network of genes (De-Santis and Jerry, 2007) and many environmental factors, such as seasonal variation in temperature, food availability, and competition (Handeland et al., 2008). Moreover, growth in fish is also known to be correlated with variation in other life-history traits, such as gonad maturation processes and reproductive timing (Bhatta et al., 2012; Park et al., 2016). Despite the numerous factors influencing growth, most studies that investigate growth report moderate to high heritability (e.g. 0.1 - 0.5) throughout a wide range of taxa (Tsai et al., 2015; Wang, 2009; Ye et al., 2017). A number of genes associated with growth rate in fish have been identified by other means (reviewed in De-Santis and Jerry (2007)) including growth hormone, insulin-like growth factors, and a range of myogenic growth regulators.

This study follows on from the previous chapter which investigated the pedigree structure, inbreeding, and heritability of traits in a new aquaculture snapper population (see Chapter 3). Growth traits were found to have moderate heritability levels in this previous chapter. The primary goal of this chapter was to identify specific genetic variation underlying growth rate. To do this the genome-wide SNP and phenotyping data generated in Chapter 3 was used to 1) construct a high density linkage map, 2) conduct QTL mapping for three measures of growth rate (peduncle length, fork length, and weight), and 3) investigate the position of 13 candidate growth genes and their position relative to growth rate

QTLs.

4.3 Methods

4.3.1 Study population

A snapper breeding programme is under development at The New Zealand Institute for Plant & Food Research Limited (PFR) and currently includes a population with three generations held at the Nelson Research Centre (see Chapter 3). Data from the two most recent generations ($F_1 = 70$ individuals, $F_2 = 577$ individuals) were used in the current study. Uncontrolled tank-based spawning of the F_1 generation in a single tank was used to produce the offspring F_2 generation. This resulted in a complex pedigree, meaning that we obtained a combination of full-sibling, half-sibling, and unrelated individuals in the F_2 generation. The F_2 offspring were held in a single tank until they were approximately one year old and then split evenly among four tanks with comparable feeding, light, water flow, aeration, and tank design. All research carried out in this study was reviewed and approved by the animal ethics committee of Victoria University of Wellington (Application number 2014R19).

4.3.2 Phenotyping

Images were collected for each fish at 436-487 days (hereafter referred to as ‘year one’ fish) and 1045-1131 days (hereafter referred to as ‘year three’ fish). Fish were anaesthetized with Aqui-S™ to allow handling of fish when collecting data. The three measures of growth rate used in the current study were fork length, peduncle length, and weight. Fork length was the distance from the nose to the fork in the tail. Peduncle length was the distance from the nose to the narrowest cross-section across the tail. Length measurements were made by collecting images of each individual and then making measurements from those images. Images were collected with a Panasonic DMC-GH4 camera (16 megapixels) under natural lighting inside a large tunnel-house constructed of light diffusing material. The distance between the camera and fish varied between photos so a ruler was included in each image to provide a scale. Weight was collected manually by placing the fish on a Mettler-Toledo SB8001 scale with 0.1g accuracy. The number of individuals measured differed between year one and year three as a result of natural mortality during the study.

4.3.3 Genotyping by Sequencing

Samples of fin tissue were collected for all fish and DNA was extracted from these samples using a modified salt extraction protocol (see Chapter 3). DNA was genotyped using a modified Genotyping By Sequencing (GBS) approach (Hilario, 2015). A total of eight pooled libraries were sequenced on an Illumina HiSeq 2500. Each pool contained DNA from 96 individuals which were individually barcoded. Duplicate or triplicate samples

were prepared for each of the parents and grandparents, and single samples for each of the offspring. Sequencing data from the GBS libraries were processed using the STACKS pipeline (Catchen et al., 2013) and the genotyping results exported to a GENEPOP format txt file. When variant calling, sequence stacks were aligned to an initial snapper genome (version 1, number of scaffolds = 5,998, N50 = 1.5Mb, total bp = 738Mb, Wellenreuther et al. unpublished) in the STACKS pipeline. Single Nucleotide Polymorphisms (SNPs) from the STACKS pipeline were filtered for $> 7\times$ coverage, presence in $>75\%$ of the population, and a Minor Allele Frequency (MAF) > 0.05 .

4.3.4 Linkage map construction

Parents for each individual in the dataset were previously identified using Cervus v3.0.7 (Kalinowski et al., 2007) with a 95% confidence limit in Chapter 3. A linkage map was constructed based on the genotyping and pedigree data in LEPMAP v2.0 (Rastas et al., 2016). Data from the largest 14 F_2 families (full and half-sibling families with 10 or more offspring) were used, and included a total of 269 offspring and 14 parents. A total of 20,311 SNPs were located in the initial genotyping file. Markers were separated into chromosomes with the SeparateChromosomes module (logarithm of odds (LOD) limit = 14, minimum markers per linkage group = 50). The best order was then generated with the OrderMarkers module. Markers near the start and end of each linkage group (start and end 10% based on centiMorgan (cM) distance) were removed if they were more than 3 cM from the next closest marker. The linkage map was visualized using MapMaker v3.0. The accuracy of the final linkage map was reviewed by comparing the linkage map position (cM) with the position of markers on available genome scaffolds (base-pairs) from the snapper genome. This was possible by using the STACKS output files, which showed the scaffolds on which each marker was placed as well as their Base Pair (bp) position. The Pearson's correlation (r) between marker position in cM and bp position on the scaffolds was calculated for all scaffolds that contained >50 SNPs. A linear regression was also generated for each of these scaffolds and the mean and 95

4.3.5 QTL identification

Quantitative Trait Loci identification was carried out using the general model implemented in QTDT v2.6.1 (Abecasis et al., 2000) and the half-sibling regression method implemented in GRIDQTL v3.3.0 (Knott et al., 1996; Seaton et al., 2006). Both methods utilize parents as controls for population stratification (QTDT = parent-offspring-trios, GridQTL = mother-offspring or father-offspring duo's) and can use multiple offspring per family from a complex pedigree (Abecasis et al., 2000; Knott et al., 1996). Genotyping data from the F_2 generation ($n = 542$) and their F_1 parents, and phenotyping data from the F_2 generation were used in QTDT. This included a total of 10,716 SNPs in the genotyping dataset. A Bonferroni correction was used to calculate a 95% genome wide significance level for QTDT results (i.e. $0.05 / \text{number of markers}$). Phenotype variance explained by each QTL was estimated

as the difference between the R^2 value for the QTDT model with and without the within family genetic effects. A Bonferroni correction was used to calculate 95% significance limits (i.e. $0.05 / \text{number of markers}$) at a linkage group and genome-wide scale. Genotyping data from the largest F_2 half-sibling family ($n = 224$), their F_1 parents, and a subset of 1006 SNP loci (randomly filtered to a minimum distance of one SNP every 1 centiMorgan (cM)) were used in GridQTL along with phenotyping data from the F_2 offspring. A 1 cM filtering limit was selected because disequilibrium decay within the half-sibling family occurred over even larger distances (e.g. 5 cM Figure 4.4) and GridQTL struggled with higher marker numbers. Genome-wide and linkage group-wide 95% significance levels were calculated in GridtQTL using a permutation based procedure with 1000 permutations. The results from both programs were visualised using the ggplot2 library in the R statistical environment v3.2.3 (R Core Team, 2013). Before running all analysis, the genotype data were filtered for Mendelian errors by dropping loci for any individual that contained alleles not observed in either of the two parents. The phenotype measurements used for the analysis were standardized by tank and date collected to correct for temporal and tank effects.

4.3.6 Candidate genes and their location

The position of 13 candidate growth genes for fish (De-Santis and Jerry, 2007) and QTLs for growth identified in this study were compared using the available snapper genome (Wellenreuther et al., unpublished) and the newly constructed linkage map. To do this, the sequence for each candidate gene was located on the NCBI nucleotide database from the closest related teleost species - either the DNA or mRNA sequence. DNA sequences were aligned with the genome scaffolds by selecting the largest exon for the target gene and aligning with the “Map to Reference” alignment using the “Geneious mapper” in Geneious v10.0.9 (Kearse et al., 2012); alignment sensitivity was set to “High Sensitivity / Medium” with default settings. For mRNA sequences the sequences were aligned with the “Map to Reference” alignment using the “RNA Seq” mapper in Geneious v10.0.9; alignment sensitivity was set to “High Sensitivity / Medium” with the maximum gap size increased to 1000 bp. For each alignment the percentage of matching base pairs was reported for the largest exon. The linkage group and cM position of the scaffold containing specific candidate genes was then located using the STACKs output files and the newly constructed linkage map.

4.4 Results

4.4.1 Phenotyping

Fork length, peduncle length, and weight were recorded at the “year one” and “year three” time points for each fish. The distribution and relative sizes of fish in year one and year three are illustrated in Figure 4.1. In the first set of measurements the mean and standard deviation for fork length, peduncle length, and weight were 160.1 ± 15.0 mm, 132.1 ± 12.3 mm, and 89.8 ± 24.3 g, respectively (Table 4.1). In the second set of measurements the

Figure 4.1: Peduncle length measurements at year one ($n = 568$) and year three ($n = 314$). Images on the left side of the diagram are to scale relative to each other and show the smallest and largest individual fish from a tank of F_2 individuals at year one and year three, respectively.

same measures were 257.7 ± 21.0 , 214.3 ± 17.0 , and 363.2 ± 84.9 , respectively (Table 4.1). The three measures for growth were all found to be strongly positively correlated (Pearson's $r > 0.93$, Chapter 3). Strong positive correlation was also observed between year one and year three for each measure (Pearson's $r = 0.71 - 0.73$; Supplementary figure 4.1).

Table 4.1: Growth rate measurements for Australasian snapper (*Chrysophrys auratus*) at year one and year three.

	Year one			Year three		
	n	mean	stdev	n	mean	stdev
Fork length (mm)	568	160.1	15.0	314	257.7	20.2
Peduncle length (mm)	568	132.1	12.3	314	214.5	17.0
Weight (g)	530	75.5	20.3	247	361.9	82.3

Figure 4.2: Visualization of the linkage map including a total of 10,968 SNPs placed along the map in 24 linkage groups. The 24 linkage groups are equivalent size and number to represent the expected 24 chromosomes of Australasian snapper (*Chrysophrys auratus*).

4.4.2 Genotyping by Sequencing

A total of 1.6 billion reads were produced for all eight pooled GBS libraries with approximately 2, 4 or 6 million reads for each single, duplicate, or triplicate individual library respectively. Using the STACKs pipeline a total of 20,311 SNPs were found after filtering for $>7\times$ coverage, present in 75% of the individuals in the population, and > 0.05 MAF. The average coverage per SNP was 15.6x in the offspring (F_1) and 23.9x in the parents (F_2).

4.4.3 Pedigree structure and linkage map

Parents were previously identified for 93% of the individuals in the F_1 and F_2 generations (see Chapter 3). The remaining 7% with missing parents were mainly located in the F_1 generation, and were the result of F_0 wild-caught individuals that were not available for sampling at the time of this thesis research. A mixture of full sibling and half sibling families was present in the F_2 generation.

Figure 4.3: Comparison of linkage group (cM) and genome scaffold (Mb) position for loci placed on the four largest scaffolds available from the genome. The position was strongly correlated between the two approaches (scaffold_1 = 0.91, scaffold_2 = 0.53, scaffold_3 = 0.78, scaffold_4 = -0.95), but there is also noise around the precise placement on the linkage map.

Figure 4.4: Linkage disequilibrium (R^2 correlation) between markers in the dataset plotted against the distance (cM) between those markers on the linkage map. This statistic was calculated for the entire QTL mapping dataset (multiple families, families = 137, $n = 539$) and for the data from the largest full-sibling family in the dataset ($n = 48$). The results show a consistent decay of linkage disequilibrium across the length of the linkage maps and $> 5x$ higher linkage disequilibrium in the largest family than the entire dataset.

A total of 10,968 SNPs were positioned on the linkage map (Figure 4.2). The total length of the sex-averaged linkage map was 1,363.0 cM with an average marker spacing of 0.129 cM. The lengths of the male and female maps were 1401.5 cM and 1359.0 cM, respectively. The female and male recombination rates were 3.28 cM/million base-pairs (Mb) and 1.93 cM/Mb based on comparison with available scaffolds from the initial snapper genome (version 1, number of scaffolds = 5,998, N50 = 1.5 Mb, total base pairs = 738 Mb, Wellenreuther et al. unpublished). When aligning the genome scaffolds to the linkage map, individual scaffolds were placed exclusively onto one of the 24 linkage groups and a total of 701Mb (95% of the total base-pairs in the genome) of scaffolds were able to be placed on the linkage map. Visualization of the linkage map and scaffold position for the four largest scaffolds showed a clear relationship between the ordering of markers (Pearson's $r = 0.91, 0.53, 0.78, \text{ and } -0.95$), but that some noise was apparent around the exact placement (Figure 4.3). The correlation for all scaffolds with 50 or more markers was similar ($r = 0.74 \pm 0.20$ for 1723 markers on 26 scaffolds). The 95% confidence interval of the residuals ranged from -5.7 to 3.2 cM with a mean of -1.25; indicating that markers were placed within approximately 4.5 cM of their correct location. Investigation of the extent of linkage disequilibrium within the dataset showed a clear decay of linkage disequilibrium over the length of the linkage groups (Figure 4.4). When looking at a single F_2 family we see a high degree of linkage disequilibrium. However, when looking at the whole F_2 generation the decay of linkage disequilibrium is much greater, with minimal linkage disequilibrium observed even over small distances.

4.4.4 QTL mapping

Multiple QTLs were found for all three measures of growth rate (Figure 4.5, table 4.2, and supplementary table 4.1) and many of the QTLs were shared among the three measures of growth (fork length, peduncle length, and weight) and between the two software used (QTDT and GridQTL). Genome-wide significant QTLs were located on linkage groups 3, 11, 16 for fork length and peduncle length in QTDT and linkage groups 3, and 16 for all growth traits in GRIDQTL (Figure 4.5 and Table 4.2). The length trait QTLs on linkage groups 3, 11, and 16 from QTDT were also significant at a chromosome wide level for weight. The genome-wide significant markers for QTDT in year one had moderate effect sizes ranging from a minimum R^2 of 0.04 to a maximum of 0.05 for markers in QTDT (Table 4.2). No genome-wide significant QTLs were found for growth traits in year three using either QTDT or GridQTL.

Figure 4.5: QTL scans at year one in the F_2 generation of *Chrysophrys auratus* for fork length (A and D), peduncle length (B and E), and weight (C and F) using QTDT and GridQTL. The black horizontal line on each graph indicates the 95% genome-wide significance level. Genome-wide significant QTLs were found using both software on linkage groups 3 and 16. A third genome-wide significant QTL was found on linkage group 11 using QTDT, but not GridQTL. QTLs were shared between all traits within each software, except for weight in QTDT which had no genome-wide significant QTLs.

Table 4.2: Putative QTL markers (based on QTDT) in Australasian snapper (*Chrysophrys auratus*) that were significant for at least one trait at a 95% genome wide significance level of 5.33. Proportion phenotypic variance explained (R^2) was estimated in QTDT as the difference between the R squared values of the total model and the genotype model. Loci are reported including their linkage group number (LG) and position (cM). Genome wide (**) and linkage group wide (*) significance are indicated with asterix.

LG	cM	locus	Fork length		Peduncle length		Weight	
			-log10(p)	R^2	-log10(p)	R^2	-log10(p)	R^2
3	20.9	40730_33	6.4**	0.04	5.7**	0.04	4.3*	0.03
11	10.0	99093_35	5.7**	0.05	5.7**	0.04	5.2*	0.04
16	47.9	62074_47	5.7**	0.05	5.5**	0.04	4.5*	0.03
16	40.2	39092_27	5.5**	0.05	5.2*	0.05	4.5*	0.05

4.4.5 Candidate genes

The base-pair position on the genome scaffolds were found for all 13 candidate genes including growth hormone, growth hormone receptor, growth hormone receptor type 1, growth hormone receptor type 2, insulin like growth factor 1, insulin like growth factor 2, myogenic factor 1, myogenic factor 2, myogenic regulatory factor 4, myogenic regulatory factor 6, myogenin, myostatin, and parvalbumin (Supplementary table 4.2). Based on the largest exon, all genes exhibited high base-pair similarity with the target genome position (88.7 to 99.3%). Of the candidate genes investigated, growth hormone, myogenin, and parvalbumin were located on linkage groups containing genome-wide significant QTLs (Supplementary table 4.2). Growth hormone on linkage group 16 was 5.3 cM from the nearest QTL marker (QTDT) and 30.2 cM from the nearest QTL peak (GRIDQTL). Myogenin on linkage group 3 was 9.6 cM from the nearest QTL marker (QTDT) and 18.4 cM from the nearest QTL peak (GRIDQTL). Parvalbumin on linkage group 16 was 25 cM from the nearest QTL marker (QTDT) and 7.8 cM from the nearest QTL peak (GRIDQTL).

4.5 Discussion

This study assembled the first ever linkage map for Australian snapper *Chrysophrys auratus*. Proof checking against an initial snapper genome (Wellenreuther et al. unpublished) indicated that the linkage groups were of reasonable quality given the inherent data limitations (e.g. sample size). QTL mapping revealed eight markers on three linkage groups that were significantly associated with growth rate. Three candidate genes for growth rate were located on the same linkage groups as these QTL. These genomic resources will be used to advance the selective breeding programme currently underway on the study population and will form the basis of further genomic investigation in snapper.

Linkage maps have been used extensively in conjunction with QTL mapping studies to arrange markers into the order they appear in the target species genome (Boulton et al., 2011; Cnaani et al., 2004; Greenwood et al., 2011). The usefulness of a linkage map is

determined by marker density and overall accuracy and precision. Historically, most first generation linkage maps in fish have been constructed with just a few hundred markers and did not have genome sequences available to evaluate their accuracy (Castaño-Sánchez et al., 2010). However, a few recent examples have increased the number of markers (Castaño-Sánchez et al., 2010; Ninwichian et al., 2012) and some have also begun to utilize available genome data for checking the linkage map accuracy and/or even check genome assemblies (Tsai et al., 2016; Wang et al., 2017). The snapper linkage map had 11K markers for a 738 Mb genome, compared with 96K markers in Atlantic salmon (2.97 Gb genome), which is perhaps the most comprehensive fish linkage map to date (Tsai et al., 2016). The average Pearson’s correlation between the largest genome scaffolds (bp) and linkage map (cM) was 0.74. Comparatively, the same statistic for the recently constructed salmon linkage map were 0.81 for the male map and 0.92 for the female map. While the correlation between the linkage map and genome scaffolds was clear, it was also apparent that at a fine scale (< 5 cM intervals) there is some variation around the exact placement of SNPs. This variation is most likely the result of inherent limitations in the dataset (e.g. sample size, number of recombination events), but some of this variation could also be the result of differential recombination patterns across the genome (as observed in Roesti et al. (2013)).

Using the newly constructed linkage map and available genome scaffolds, we were able to calculate the sex-specific recombination rates for snapper (female = 3.28 cM/Mb, male = 1.93 cM/Mb). This reflects similar observations in other fish species, with females often having higher recombination rates than males (Castaño-Sánchez et al., 2010; Kucuktas et al., 2009; Tsai et al., 2016). Overall, the recombination rates in this study are similar for those found in several fish species including stickleback (*Gasterosteus aculeatus*, 3.11 cM/Mb, Roesti et al. (2013)), Asian seabass (*Lates calcarifer*, 2.4-2.8 cM/Mb, Wang et al. (2017)), and channel catfish (*Ictalurus punctatus*, 2.6 cM/Mb, Li et al. (2015)). Recent work has indicated that the different recombination rates for male and female fish may be related to the processes that occur during the formation of the egg or sperm (Theodosiou et al., 2016).

While dense, accurate, and precise linkage maps are useful for QTL mapping, very large pedigrees are needed to construct them – for example, studies can require many thousands of individuals and families to get precise marker placement (He et al., 2011). Specifically, linkage maps can lack precision because they require sufficient numbers of recombination events between markers to determine their ordering. With dense genotyping datasets and moderate to small sample sizes, such as the current study, there will be too few recombination events between markers that are close together on the genome. This limits their precision at finer scales, although the arrangement of markers at larger scales is accurate. Reference genomes provide an alternative way to position markers at their exact base pair location, but accuracy does depend on the stage and quality of the genome assembly (Berner et al., 2014). However, when a species does not have a fully assembled reference genome, linkage maps can prove valuable for determining the position of markers and aiding the arrangement of large genome scaffolds into new reference genomes (Fierst,

2015). Future work in snapper would benefit greatly from a finalized genome assembly. To do this, the linkage map could be used to position large genome scaffolds into chromosomes. From here, the arrangement of markers on the scaffold could be used to arrange the markers in the linkage map at a finer scales. The linkage map may also be useful as an independent check of the genome assembly into large scaffolds.

The target trait for this chapter was growth rate - measured using fork length, peduncle length, and weight. Growth rate is one of the primary targets for selective breeding programmes because it relates directly to production output. In fish, it typically has moderate heritability, which was also found in this thesis research for snapper (see Chapter 3). Other factors that can affect growth rate include feed amounts, fish density, and tank design (e.g. size, aeration, water flow). Steps were taken to control these factors in the current study by standardizing the conditions between tanks and by standardizing measures from each tank before using them in QTL analysis.

Genome-wide significant QTLs were found in the first year using both QTL mapping software. The genome-wide significance QTLs were largely shared among the three measures (peduncle length, fork length, and weight), which reflects that these are all measures of the same underlying trait (growth rate). The proportion of phenotypic variance explained by each QTL ranged from 0.03 to 0.05, which is low compared with that observed in these some other fish studies. For example results for other fish species include, Atlantic salmon (*Salmo salar*: 0.06 to 0.08, Besnier et al. (2015)), tilapia (*Oreochromis niloticus*: 0.06 to 0.19, Cnaani et al. (2004)), chinook salmon (*Oncorhynchus tshawytscha*: 0.14 to 0.33, Everett and Seeb (2014)), brill (*Scophthalmus rhombus*: 0.08 to 0.12, Houston et al. (2009)), and catfish (*Ictalurus furcatus*: 0.01 to 0.23, Hutson et al. (2014)). The identification of multiple QTLs affecting growth rate indicates a polygenic basis for growth rate. No genome-wide significant QTLs were found in year three. This is most likely the result of the smaller sample size available in year three, but may also reflect the lower heritability of growth rate in year three for the sample population (see Chapter 3). The lower sample size in year three was the result of natural mortality over the course of the study.

Determining precision of QTL placement is an important step in the QTL mapping process as it provides useful information about where variants responsible for an observed QTL signal (e.g. candidate genes or causative alleles) are likely to be located. If high rates of linkage or linkage disequilibrium are present between markers, a confidence interval for the QTL region can be estimated - as seen in R/QTL (Broman et al., 2003). However, in the current study, pairwise correlation between markers across the linkage map indicated very low linkage disequilibrium between markers over even relatively short distances (< 5 cM). This is most obvious when comparing the linkage disequilibrium within the single largest family to that observed in the entire dataset. However, it is worth noting that on linkage groups with genome-wide significant QTLs, there does appear to be a number of markers surrounding each QTL that are responding to the QTL signal. As such, it seems likely that there is some linkage disequilibrium between markers at a fine scale (< 5 cM),

but that this may be obscured by the low precision of marker placement on the linkage map. If true, the optimal approach to get more precise placement of QTL regions will be improved SNP positioning using either a second improved iteration of the linkage map, the genome assembly, or a combination of both these resources. Until this is done it is likely that causative genetic variations underlying the QTL signals will only be located to within a 5 cM scale of their actual position.

Previous studies have outlined a range of genes and molecular networks that are thought to be candidates for further investigation of growth rate in teleost species (De-Santis and Jerry, 2007). In this study the position of candidate genes surrounding the detected growth QTLs were located, however, the large distances between the genes and the QTL peaks indicate that one needs to be cautious about a definitive link between the QTLs and candidate genes. This was possible using the available genome sequence data to link gene positions back to their nearest markers. Central to growth rate in most species is the somatotrophic axis, which consists of the growth hormone releasing hormone (GHRH), growth hormone inhibiting hormone (GHIH), growth hormone (GH), and insulin-like growth factors (IGF-1 and -II) (De-Santis and Jerry, 2007). Of these, growth hormone and insulin-like growth factor I and II were able to be mapped to the linkage map in the current study. Growth hormone was located near a QTL of genome-wide significance (approximately 5.3 cM). In *Sparus aurata*, a close relative of snapper, a microsatellite repeat in the promoter region has previously been implicated for differences in growth rate (Almuly et al., 2000). This gene would be a good candidate in snapper because it is close to a QTL in the current study, and the causative microsatellite has been observed in a range of teleost species. Myogenic regulatory factors (myogenin, MyoD, myf-5, and myf-6) are another set of potential candidate genes (De-Santis and Jerry, 2007). These regulatory factors have been implicated in growth in terrestrial vertebrates, but not in fish species. In this study, the myogenin gene was located on linkage group 3 near a genome-wide significant QTL (approximately 9.6 cM). In pigs (*Sus domesticus*), a polymorphism in the promoter region of myogenin was found to account for up to 5.8% of differences in weight (te Pas et al., 1999), but no research has investigated its effect in teleost species. A final candidate gene (parvalbumin) was located on the same linkage group as a genome wide significant QTL, but was much further away from a putative growth QTL than the previous two genes (approximately 25 cM). A mutation in the promoter region of this gene was previously found to be involved in weight differences in the finfish species *Lates calcarifer* (Xu et al., 2006).

4.6 Future directions

The linkage map from this chapter can place markers with reasonable accuracy at broader scales (< 5 cM), but future work is needed to more precisely position the QTL regions and candidate genes. Finalizing the snapper genome assembly would be the most effective way to achieve this. Specifically, markers could be positioned accurately to a base-pair level using a finalized genome assembly rather than within relatively large regions generated by

the linkage map. This would allow the distance between QTL regions and candidate genes to be determined more precisely and thereby improve the certainty that specific candidate genes are associated with the observed QTL peaks. The snapper genome is currently being assembled at Plant & Food Research, and will use the linkage map from this thesis to help with arrangement of large scaffolds (discussed further in Chapter 6). Another future direction for the current research is to use the genome resources from this and the preceding chapter (Chapter 3) to investigate the genetic basis of other traits of interest within the snapper breeding programme. The chapter following this (Chapter 5) uses the linkage map and the pedigree structure from Chapter 3 and 4 to investigate the genetic basis of blue spots in same snapper.

Identifying the genetic basis of pattern and number of blue spots in the Australasian snapper (*Chrysophrys auratus*) 5

Authors

David T. Ashton, Peter A. Ritchie, & Maren Wellenreuther

Author Contributions

Conceived and designed the experiments: DTA, PR, and MW; Data collection: DTA; Data analysis: DTA; Wrote the primary manuscript: DTA; Prepared final manuscript: DTA, PR, and MW

Publication status

This chapter has been prepared for publication, but will be published after submitting my thesis.

5.1 Abstract

In this chapter, the genetic basis of blue spots in the marine teleost Australasian snapper (*Chrysophrys auratus* also *Pagrus auratus*) was investigated. This trait has recently become of interest within the snapper breeding programme for its potential for their use as visual identifiers of specific individuals (MPI, 2015), their potential use for creating unique colour morphs (e.g. a visually different colour morph for selectively breed snapper strain), and related research about their underlying biology and how to maintain their blue colouration after death (important for presentation of a fresh product in the whole fish market). Image-based phenotyping was used to extract the coordinates of blue spots for 501 fish in the cultured snapper population. Each fish was previously genotyped at 10,716 loci using Genotyping by Sequencing (GBS). The genotyping and phenotyping data was used to explore the similarity of the spot patterns between full-siblings, half-siblings and unrelated individuals and to conduct a scan for QTLs associated with number of blue spots. The pattern of blue spots was found to be highly variable, but not predicted by the level of relatedness. Quantitative Trait Loci (QTLs) were detected on 12 of the 24 linkage groups - one of which (on linkage group 23) was shared between the two QTL mapping software used. A blast search identify 73 genes within this QTL region. One of these, the gene for tyrosinase, forms a key part of the melanin pathway and has been previously associated with the number of iridophores, which is the main cell type that form blue colour patterns in other fish species.

5.2 Introduction

Colour patterns are one of the most recognizable phenotypes across the animal kingdom and are commonly involved in communicating and signalling. Some species have evolved specific mechanisms to produce spectacular colours (e.g. external patterns of cichlid fish, Maan and Sefc (2013)), while other species extract colours from their diet or naturally occurring compounds (e.g. blue feet of blue-footed booby, Velando et al. (2006)). Signalling using colour provides a wide range of advantages – from camouflage to escaping predators through to courtship displays to secure a mate. In the well-studied case of the peppered moth (*Biston betularia*), which underwent a rapid selective sweep during the Industrial Revolution, camouflage patterns aid individuals in avoiding detection by predatory birds (Clarke et al., 2008). In the case of the marine two-spotted goby (*Gobiussculus flavescens*), bright colour is the target of sexual selection during mating events, and bright individuals have higher reproductive success (Amundsen and Forsgren, 2001).

The study species for this thesis, Australasian snapper (*Chrysophrys auratus* also *Pagrus auratus*) exhibits at least three distinct colours including overall red colouration, a large numbers of blue spots across the length of their body, and variability in dark colouration in the PFR cultured population. Of these, snapper's blue spots are particularly interesting, because their genetic basis and biological function is not known, but the number of spots is

highly heritable (Chapter3). Greater knowledge about the genetic and molecular basis of these spots has a number of potential applications. Maintaining a fresh looking fish after death is important for the whole fish market, but the blue spots fade rapidly after death (< 5 days depending on storage conditions). A better understanding of their genetic and molecular basis may help with research to maintain these colour patterns post-harvest. This trait also has some quite unique applications within the breeding programme. Primarily, a specific colour morph (e.g. high levels of blue spots) could be used as a branding mechanism for selectively bred snapper (e.g. similar to colouration variation in apple strains).

Despite its wide ranging applications, all colouration (except bioluminescence) is based on the reflection or omission of light from a surface. Two ways that animals achieve this is via pigments or structural variation (e.g. reflective crystals). Pigments are compounds that selectively absorb a particular wavelength of light, which influences the colour of an organism (e.g. melanin). A pigment is typically contained within specialized cells called chromatophores, which are able to reflect light and in some cases can even rapidly change their reflectivity and translucency by muscle contractions (e.g. colour changes in cuttlefish, Suzuki et al. (2011)). The other mechanism for producing specific colours is structural variation (iridophores), which affect the reflection of light (e.g. iridescent iridophores in male guppies, Kottler et al. (2014)). Interestingly, blue colours in the natural world, such as those exhibited by snapper, are most often produced by iridophores rather than pigments. This is particularly true in the case of fish where the only documented blue pigment is found in the mandarin fish (*Synchiropus splendidus*) and psychedelic fish (*Synchiropus picturatus*) (Schaefer et al., 2014).

Although the molecular basis of blue spots in snapper is not known there is some research that may be informative from other species. Interestingly, while animal colour patterns are a complex phenotype produced through a network of molecular and developmental interactions, the genetic basis of pigment cell development and differentiation are often conserved between mammals and teleost fish (Braasch et al., 2007; Hubbard et al., 2010; Mundy, 2009). As an example, Sox10, Mitf, Kit, and Ednrb genes are all known to influence colour patterns in both mice (*Mus musculus*) and teleost fish (Braasch et al., 2007), as components of the melanin pathway (Braasch et al., 2007; Hoek et al., 2008). Although it is most often associated with melanin research in humans, this pathway is implicated in affecting both chromatophore and iridophore cell types (most likely the cell type responsible for blue spots in snapper). For example Sox10 and Ednrb, which are part of this pathway, affected both chromatophore driven pelt colour in mice (*Mus musculus*) (Braasch et al., 2007) and iridophore (blue spots) development in the teleost zebrafish (*Danio rerio*) (Lopes et al., 2008; Parichy et al., 2000). Another gene involved in this pathway, which affects iridophore numbers in a teleost (zebrafish, *Dani rerio*), is leukocyte tyrosine kinase (LTK) or tyrosinase (Lopes et al., 2008).

This study investigated the genetic basis of variation in the blue spots of the Australasian snapper (*Chrysophrys auratus* also *Pagrus auratus*). While snapper are predominantly rec-

ognized for their red colouration, they also have distinct blue spots (Figure 5.1). Similar blue patterns are observed intermittently throughout the Sparidae family, from solid blue stripes in *Polysteganus undulosus* to the smaller blue spots observed in the snapper. Genotyping by Sequencing (GBS) data was used in combination with blue spot trait information derived from images of 501 individuals. Specifically, 1) the similarity of blue spot patterns was compared between full-siblings, half-siblings and unrelated individuals, 2) Quantitative Trait Loci (QTL) associated with the number of blue spots were detected, and 3) a search was carried out for candidate genes located near putative QTL loci. The findings are discussed and compared with the broader literature and how they might apply to the selective breeding program.

5.3 Methods

5.3.1 Study population

A three generation snapper pedigree located at the Seafood Research Facility in Nelson, New Zealand, was used for this study (see Chapter 3). This pedigree is part of a breeding programme carried out by the New Zealand Institute for Plant and Food Research. Uncontrolled single tank spawning of a wild broodstock ($F_0 = 50$ individuals) was used to generate the F_1 generation ($F_1 = 70$ individuals), and upon reaching maturity the F_1 population was again used to produce the F_2 generation ($F_2 = 577$ individuals) in the same manner. This resulted in a complex pedigree, meaning that a combination of full-sibling, half-sibling, and unrelated individuals were produced in both generations. The F_2 offspring were held in a single tank until ~ 1 years old and then split evenly among four tanks - where the feeding, light, water flow, aeration, tank design was the same across these tanks. The research carried out in this study was conducted under VUW Animal Ethics approval 2014R19.

5.3.2 Phenotyping of iridophore pattern and number

Images were collected for each fish at 436-487 days (hereafter referred to as ‘year one’ fish) and 1045-1131 days (hereafter referred to as ‘year three’ fish). Fish were anaesthetized with Aqui-STM to allow handling of fish when collecting data. To capture the images, individual fish were placed alongside a ruler on a white background. All images were taken of the left side of the fish with a Panasonic DMC-GH4 camera (16 megapixels). Images of each fish were analysed with OpenCV v2.0 through a custom Python v3.6 script. The script allowed blue spot data to be extracted via two steps: 1) grey-scale thresholding was used to find the outline of the fish against the background, and 2) the blue spots were located on the surface of the fish by finding the top 0.5% pixels in the blue colour channel within the outlined area for each fish. Once identified, the XY coordinate for each blue spot was exported to a memory file along with the outline of the fish. This memory file was then used when investigating the genetic basis of blue spot patterns and used to count the number of blue spots for each individual for QTL mapping.

5.3.3 Genotyping by Sequencing

Samples of fin tissue were collected for all fish and DNA was extracted using a modified salt extraction protocol. DNA was genotyped using a modified Genotyping-by-Sequencing approach (Hilario, 2015). A total of eight pooled libraries were sequenced on an Illumina HiSeq 2500. Each library pool contained DNA from 96 individuals which were sequenced simultaneously using individual ‘barcode’ sequences to enable sample multiplexing. Duplicate or triplicate samples were prepared for each of the parent and grandparents and single samples for each of the offspring. Sequencing data from the GBS libraries was processed using the STACKs pipeline (Catchen et al., 2013) and the genotyping results exported to a GENEPOP format txt file. Variant calling sequence stacks were aligned to an initial snapper genome (version 1, number of scaffolds = 5,998, N50 = 1.5Mb, total bp = 738Mb, Wellenreuther et al. unpublished) in the STACKs pipeline. Single Nucleotide Polymorphisms (SNPs) from the STACKs pipeline were filtered and retained if they had $> 7x$ coverage, were present in $>75\%$ of the population, and had a minor allele frequency > 0.05 . Only loci that were able to be located on a linkage map were included in the final genotype file. The final genotyping file was also filtered for Mendelian errors, i.e. by dropping loci for any offspring that contained alleles not observed in either of the two parents.

5.3.4 Genetic basis of iridophore patterns

To investigate if there was a genetic basis to the blue spot patterns, the similarity of patterns was compared between individuals from three relatedness subgroups - namely full-siblings, half-siblings, and unrelated individuals. To compare the spot patterns (i.e. a set of XY coordinates for each blue spot on the left hand side of the fish) for two individuals, the XY coordinates were aligned to a common 2-dimensional position. This alignment was performed using an affine transformation and two reference points from the fish outline (nose and top of the narrowest point across the tail) (Supplementary figure 5.1). Once both fish had been aligned to the common position the average distance between the closest blue spots in each pattern were measured. For each blue spot in the first individual the closest blue spot in the second individual was located and the distance in pixels measured. For each pair of individuals in the dataset the average distance and relatedness subgroup was saved to a memory file. A violin plot was constructed in Python v3.6 using the matplotlib library which compared the average distance between blue spot patterns for the three subgroups. The regression coefficient of determination was generated to compare the number of spots for each individual in year one and year three.

5.3.5 QTL mapping for iridophore number

Quantitative Trait Loci identification was carried out using the general model implemented in QTDT v2.6.1 (Abecasis et al., 2000) and the half-sibling regression method implemented in GRIDQTL v3.3.0 (Seaton et al., 2006; Knott et al., 1996). Both methods were selected

for the current study because they utilize parents as controls for population stratification (QTDT = parent-offspring-trios, GridQTL = mother-offspring or father-offspring duo's) and can use multiple offspring per family from a complex pedigree (Abecasis et al., 2000; Knott et al., 1996). Phenotypes from year one were used for QTL mapping. Genotyping data from the F₂ generation and their F₁ parents, and phenotyping data from the F₂ generation were used in QTDT (n = 501). This included a total of 10,716 SNPs in the genotyping dataset. A Bonferroni correction was used to calculate a 95% genome wide significance level for QTDT results (i.e. 0.05 / number of markers). Phenotype variance explained by each QTL was estimated as the difference between the R² value for the QTDT model with and without the within family genetic effects. Genotyping data from the two largest F₂ half-sibling families (n = 224 and 64), their F₁ parents, and a subset of 1006 SNP loci (randomly filtered to a minimum distance of one SNP every 1 centiMorgan (cM)) were used in GridQTL along with phenotyping data from the F₂ offspring. A 1 cM filtering limit was selected because previous results from this population indicated limited linkage disequilibrium decay within the half-sibling family over even larger distances (e.g. 5 cM Chapter 4, Figure 4.4) and GridQTL struggled with higher marker numbers. Genome-wide and linkage group-wide 95% significance levels were calculated in GridtQTL using a permutation based procedure with 1000 permutations. The results from both software were visualised using the ggplot2 library in the R statistical environment v3.2.3 (R Core Team, 2013).

5.3.6 Searching for candidate genes in the QTL region

A search for candidate genes was carried out in the genome region around the QTL on linkage group 23 using available genome scaffolds. To do this, SNPs within the genome-wide significant QTL region were used to identify genome scaffolds that were represented within the QTL region. The scaffolds were then separated into 200 kb sections and blasted against the NCBI nucleotide database using Geneious v10.0.9 (Kearse et al., 2012) by applying the Discontiguous Megablast option with standard settings and a limit of 50 hits per 200 kb section. The hits for all 200 kb sections were pooled and filtered to remove multiple hits to the same location, sequences that were not informative about content (e.g. non-annotated genome sequences), and hits with sequence similarity scores lower than 80%. If the same gene was found for multiple species then the one with the longest matching sequence was retained. The names for the remaining gene sequences were investigated using a survey of published studies. A keyword search for the gene name in Google Scholar was used to find likely pathways and known effects for the target genes. Up to four keywords were recorded for each gene as a way of categorizing the different genes. Relevant information relating to the position of the candidate gene (e.g. scaffold location and nearest markers) was plotted using the ggplot2 library in the R statistical environment v3.2.3 (R Core Team, 2013).

5.4 Results

5.4.1 Iridophore pattern and number

Figure 5.1: A histogram of iridophore spot numbers for all photos collected in this study and two representative individuals with high and low numbers of iridophore spots (illustrated with blue dots).

Images were collected for each individual at 436-487 days (year one) and 1045-1131 days (year three). The blue spots were clearly visible in all images, although surface reflection did affect the ability to view some spots. The number of blue spots on the left side of the fish ranged from 11 to 75 and averaged 41.6 (Figure 5.1).

5.4.2 Genetic basis of iridophore patterns

A comparison of the iridophore patterns from each individual in the dataset with every other individual in the dataset indicated there was no association between relatedness (full-siblings, half-siblings, unrelated individuals) and the similarity of the patterns; based on average distance between closest spots (Figure 5.2). The distribution of results changed slightly between the different subgroups as a result of larger sample sizes for the less related pairs of individuals. However, a shorter average distance between patterns, which would indicate more similar patterns, was not observed for the more closely related subgroups.

5.4.3 QTL mapping of blue spot numbers

Genome-wide significant QTLs were found with both QTDT and GridQTL for the largest half-sibling family. No genome-wide significant QTLs were found with GridQTL and the second largest half-sibling family. The genome wide 95% significance limit was 5.33 ($-\log_{10}(\text{p-value})$) and 13.04 (F-statistic) for QTDT and GridQTL, respectively. A total of 24 of the 10,716 SNPs were found to be significantly associated with the number of blue

Figure 5.2: Comparison of iridophore pattern similarity between Australasian snapper (*Chrysophrys auratus*) with different levels of relatedness. Similarity was measured based on the average distance in pixels between the closest spots between the spot patterns from two individuals. The violin plots show the distribution of distances and the total number of pair-wise comparisons in each relatedness group.

spots using the complete dataset in QTDT (Figure 5.3). These SNPs were located across 12 of the 24 linkage groups. In contrast to the many linkage groups that were found to be associated with number of blue spots using QTDT, only one genome-wide significant QTL on linkage group 23 was detected when using GridQTL (Figure 5.3). Importantly, this QTL on linkage group 23 overlapped with one of the ones detected by QTDT. Based on the GridQTL results this QTL covered a 4 cM region at the start of the linkage group (1-4 cM, Figure 5.4).

5.4.4 Candidate gene search in the QTL region

A candidate gene search was carried out in the 4 cM QTL region on linkage group 23. The 47 SNPs in this QTL region were located on nine of the available genome scaffolds – ranging in size from 0.2 to 5.3 million base pairs (Mb) and covering a 13.2 Mb region based on available genome scaffolds. A BLAST search of these scaffolds (in 200 Kb sections) against the NCBI nucleotide database returned a total of 3428 matching sequences. A total of 73 unique genes were retained after filtering for overlapping hits, informative sequence descriptions, >80%

Figure 5.3: QTL scan results from QTDT (A) and GridQTL (B). The sample size for QTDT and GridQTL was 501 and 224, respectively. Black horizontal lines on both figures are the genome-wide 95% confidence level based on Bonferroni (QTDT) and permutation (GridQTL) correction. Grey horizontal lines on the GridQTL figure are the linkage group wide 95% confidence level based on permutation. A genome-wide significant QTL was located on linkage group 23 using both software applications. QTDT results are represented as dots and GridQTL as lines, because they were tested using association and linkage, respectively.

Table 5.1: Putative QTL markers (based on QTDT) in Australasian snapper (*Chrysophrys auratus*) for blue spot numbers that were significant at a 95% genome wide significance level. Proportion phenotypic variance explained (R^2) was estimated in QTDT as the difference between the R squared values of the total model and the genotype model. Loci are reported including their linkage group number (LG), position (cM), genome scaffold number (Scaffold), and base-pair position (BP).

SNP	LG	cM	Scaffold	BP	$-\log_{10}(\text{p-value})$	R^2
56107_36	1	1.542	47	2079501	20.3	0.19
52938_57	4	43.45	43	1203662	6.0	0.07
97872_78	4	44.013	43	1112657	8.7	0.08
73434_39	4	53.037	766	21889	5.4	0.06
60704_40	5	29.044	53	1862038	5.7	0.06
79023_18	5	36.869	8	2648741	6.4	0.07
10994_23	6	0.529	140	679334	6.5	0.08
66032_44	7	6.335	624	15618	6.2	0.05
62949_18	7	48.365	57	1228578	7.3	0.08
56774_10	10	7.322	48	141232	6.5	0.06
3540_44	10	12.473	10	2539646	6.3	0.07
74978_79	12	34.699	7	1668156	5.7	0.04
75043_80	12	36.906	7	2345150	6.1	0.07
3132_11	13	26.825	108	640587	6.3	0.06
56635_23	13	33.913	487	30445	7.2	0.09
12523_46	13	43.869	149	436197	8.2	0.08
57358_40	17	26.733	49	2021027	5.5	0.05
35161_42	20	48.816	277	353720	6.0	0.05
60918_13	20	53.462	541	57063	6.7	0.05
40102_24	21	37.442	310	74701	6.0	0.05
86127_59	23	1.205	147	75263	5.7	0.06
12149_37	23	1.543	147	359213	6.4	0.06
58740_46	23	4.476	50	1812804	5.4	0.04
64745_48	23	7.274	5	5335154	6.0	0.06

sequence similarity, and removing duplicates for a single gene (see Supplementary table 5.1). The matching sequences came from a total of 26 species – with the most represented species being *Lates calcarifer* followed by *Seriola dumerili* (Supplementary figure 5.2). A literature search for these 73 genes indicated that they were involved in a wide range of processes including signalling, regulation, development, metabolism, ubiquitination, disease, stress, morphology, growth, olfaction, and colour. One of the genes found, tyrosinase, has a direct link with colour in fish. Positioning the markers from this scaffold on linkage group 3 indicated that markers from this scaffold were overlapping the genome-wide QTL region and ranged further out into the surrounding linkage group wide QTL region (1 – 11 cM on linkage group 3, Figure 5.4). The two SNPs closest to the tyrosinase gene were positioned directly on top of the QTL peak and to the side of the main peak in the linkage group wide QTL region (Figure 5.4). Tyrosinase was centered at 405,342 base-pairs, and the closest and second closest SNP markers were centered at 1,432,593 and 1,579,267 base-pairs on the

largest scaffold (CauratusV1_scaffold_5), respectively.

Figure 5.4: The QTL scan of linkage group 23 from GridQTL. The horizontal black and grey lines shows genome-wide (GW) and linkage-group-wide (LGW) 95% significance levels. The grey dashed line indicates the location of the genome scaffold containing the tyrosinase gene. The asterisk * indicate the position of the two SNP markers closest to the tyrosinase gene.

5.5 Discussion

This is the first ever study to investigate the genetic basis of blue spots in the snapper. The first step in this study was to collect data about the pattern and number of blue spots in snapper. An automated image analysis script was used to extract the data for each individual. The blue spots in snapper are variable in number, size and placement and also found in a number of other species from the same genus. Visual comparison from the left and right side of individuals suggest that the blue spots on each side are independently positioned (i.e. are not mirror images) but appear similar in number. Unfortunately, data was not collected in the current thesis to precisely quantify this, but this could be done in a future study. Blue spots are observed in a number of species related to snapper. Phylogenetic studies indicate that around six teleost species (*Argyrops spinifer*, *Evynnis japonica*, *Cheimerius nufar*, *Pagrus auriga*, *Pagellus bellottii*, and *Pagrus pagrus*) are located in close phylogenetic proximity to snapper (Orrell and Carpenter, 2004) – plus an additional three species of *Pagrus* (*Pagrus africanus*, *Pagrus major*, and *Pagrus caeruleostictus*), which were not included in this previous phylogenetic analysis (Froese and Pauly, 2017). Based on the few images available in databases (e.g. FishBase, Froese and Pauly (2017)), similar blue spots are a characteristic for many of the species of the sparriid family. Particularly, in

the aptly named blue-spotted seabream (*Pagrus caeruleostictus*) (Froese and Pauly, 2017). Further work is needed to develop a more complete image dataset for Sparids, which would allow a more comprehensive phylogenetic comparison of these external phenotypic traits. This would be particularly useful for developing hypothesis about the possible biological function of the blue spots in this group of marine fishes.

The spot pattern comparison method used in this chapter (affine transformation followed by measurement of the average distance between closest spots) did not find greater similarity between related individuals than unrelated individuals. This could indicate that there is minimal or no genetic basis to the positioning of spots. This contrasts with the findings in zebrafish (*Danio rerio*) - where genetic factors and interactions between different cells types have been shown to be important for patterning (Frohnhofer et al., 2013). It should be noted that the method used in this chapter is just one approach that could be used to compare spot patterns between individuals. Further work could be undertaken to develop other mathematical descriptions of these patterns. It is also interesting that while no connection between relatedness and pattern was found in the current study, the number of spots was found to have relatively high heritability (see Chapter 3).

Two QTL mapping software (GridQTL and QTDT) were used to further investigate the genetic basis for the number of blue spots. Both software control for relatedness in the dataset using segregation distortion of alleles from the parents to the offspring to identify QTLs (Allison, 1997; Knott et al., 1996). Interestingly, somewhat different results were found with the two software. The main difference was that many more QTLs were found with QTDT than with GridQTL. One explanation for this difference in QTLs detected is that QTDT was run with a larger sample size than GridQTL (501 compared to 224 individuals) and therefore had higher detection power. Additionally, QTDT uses cumulative information content from multiple families when identifying significant QTLs, while GridQTL uses a single half-sibling family. If the QTL signal (i.e. the underlying genetic differences between individuals used to identify the QTL) was not present in the half-sibling family but was present in the remaining families, then QTDT might detect it but GridQTL would not. Overall the results of this work suggest a polygenic basis to the number of blue spots in snapper, but the exact variants involved may be dependent on the family being investigated. Future work could gain increased QTL detection power by focusing on a single family, in which the QTL is present. However, if possible, utilizing multiple families or unrelated individuals, in which the QTL was present, would improve the resolution of the QTL, because the greater number of recombination events would result in lower linkage disequilibrium between markers.

The candidate gene search for this chapter was focused on the single QTL region found by both QTDT and GridQTL. A total of 73 genes were found in this region. The functions of these genes included signalling, regulation, development, metabolism, ubiquitination, disease, stress, morphology, growth, olfaction, and colour. Of the 73 genes that were found in the target QTL region, tyrosinase, was the only clear candidate for influencing the number

of blue spots in snapper. Tyrosinase (or tyrosine kinase) is a key element of the melanin pathway, which has been extensively investigated for its role in controlling colouration (D'Mello et al., 2016). While much of the research on this pathway has focused on melanin synthesis and cancer research, this pathway has also been implicated in fish iridophore (blue spots) development. Specifically, in zebrafish (*Danio rerio*) different Sox-10 (a key element in the melanin pathway) mutant fish lack which would otherwise be present in all adults or have higher numbers iridophores (Lopes et al., 2008). Leukocyte tyrosine kinase (tyrosinase) mutants also lack iridophores and Sox-10 mutants have been shown to have higher numbers of Leukocyte tyrosine kinase expressing cells (Lopes et al., 2008). The combination of the current studies QTL region directly above the tyrosinase location and the previous research implicating this pathway and specifically tyrosinase in iridophore development provides robust support for this as a candidate gene for blue spot numbers in snapper. Additionally, because of the widespread persistence of this tyrosinase, the melanin pathway, and blue spots (iridophores) throughout the bony fish species a cross species investigation of this trait and gene could be particularly useful for understanding its evolution and possible function.

5.6 Future directions

The results from this work provide a starting point for future investigations of blue spots in snapper and other species which may share this trait. As discussed in the previous chapter the candidate gene research in this thesis could be advanced in the future by finalizing the snapper genome assembly. This would allow more precise placement of candidate genes and QTL peaks than using the linkage map and thereby improve the certainty of any linkages between candidate genes and the QTL peaks. The current chapter together with the referenced studies provides a robust starting point for further investigations of the link between and evolution of blue spot numbers (iridophores), tyrosinase, and the melanin pathway in fishes. Given the well-known and widespread effects of the melanin pathway and presence of blue spots throughout the Sparidae genera, further characterization of this trait will be useful for understanding the evolution of this important regulatory network. Further work on the image analysis software would also be advantageous including increasing the types of information collected (e.g. vibrancy, alternative pattern comparison methods, and additional colours) and extension of the software to all species in the Sparidae genera. Finally for application in the breeding programme, the QTLs identified in the current chapter could be included in breeding decisions if there was interest in developing a visually recognisable selectively breed snapper strain.

Authors

David T. Ashton, Peter A. Ritchie, & Maren Wellenreuther

Author Contributions

Wrote the primary manuscript: DTA; Prepared final manuscript: DTA, PR, and MW

Publication status

This discussion chapter was primarily written for this thesis, but some of the content was also included in a conference proceedings (Appendix 1) and a related publication in *Trends in Ecology & Evolution*. I helped co-author this publication following the invited session entitled *Genomics for improved fisheries management and conservation: have the promises been fulfilled?* at the 7th World Fisheries Congress in Korea in 2016.

Bernatchez, L., Wellenreuther M., Araneda C., Ashton D. T., Barth J. M. I., Beacham T. D., Maes G. E., Martinsohn J. T., Miller K. M., Naish K. A., Ovenden J. R., Primmer C. R., Young Suk H., Therkildsen N. O., and Withler, R. E. (2017) Harnessing the power of genomics to secure the future of seafood. *Trends in Ecology & Evolution*, 32(9), 665-680 - DOI:10.1016/j.tree.2017.06.010

6.1 Overview

This thesis generated a wide range of resources and findings that will be useful for future research and the Australasian snapper (*Chrysophrys auratus* also *Pagrus auratus*) selective breeding programme at Plant & Food Research (PFR). The sections in this chapter that discuss the integration of genomic information into breeding programme are particularly crucial, because despite the value of genomics to aquaculture breeding it has been observed that such techniques are often underutilized (as discussed in Bernatchez et al. (2017)). This chapter starts by reviewing the major findings of this thesis and then discusses the future research and applications, including the application to the snapper breeding programme.

6.2 Major findings

6.2.1 Reconstruction of a detailed pedigree

One of the first major analysis in this thesis was the reconstruction of the pedigree for the snapper population at PFR and the related findings (see Chapter 3). The founding broodstock for this population had been collected prior to the start of this thesis. These individuals were collected from the wild and then two subsequent generations were produced using tank-based spawning. During tank-based spawning all individuals had similar but uncontrolled opportunity to contribute gametes to the breeding event. Consequently, the generation that each offspring belonged to was known, but the identity of its parents was not.

Pedigree reconstruction using the genotyping data identified the parents for 93% of the individuals within the F_1 and F_2 generations of the population. Based on this information, it was discovered that only a small subset of the broodstock individuals ($F_0 = 15/25 +$ undocumented first wild cohort, $F_1 = 58/75$) had successfully contributed to the next generation. The number of breeding individuals in a population (i.e. effective population size) has a significant impact on the maintenance of genetic diversity and inbreeding depression over time in small populations (Jamieson and Allendorf, 2012; Frankham et al., 2014). An effective population size (N_E) of 50 individuals was previously proposed to prevent inbreeding depression in outbreeding diploid species (see review by Jamieson and Allendorf (2012)). However, more recently it has been suggested that a minimum effective population size >100 individuals is required to limit inbreeding depression to 10% over 5 generations in similar populations (Frankham et al., 2014). Although these numbers have been proposed for wild populations with a conservation focus, the snapper population (to date) could be considered similar – as it is an outbreeding diploid species with uncontrolled spawning. Doubling the number of snapper broodstock would be required to meet the more recent >100 rule. See the future research and applications section of this chapter for a more detailed discussion about ways forward for the snapper breeding programme.

An additional finding of the current study was that among those individuals that did produce offspring there was strong contribution distortion - some individuals contributed

significantly more offspring than others. For example, one F_1 female sired 39% of the F_2 offspring, but another F_1 female sired just one offspring. A total of 27 of the 35 females contributed offspring. The contribution distortion observed in this thesis is important for understanding this dynamic in the snapper breeding programme because contribution distortion will directly increase the rate at which genetic diversity is lost. Skewed contributions have been observed in a wide range of captive fish populations, including Asian seabass (*Lates calcarifer*) (Liu et al., 2012), gilthead seabream (*Sparus aurata*) (Chavanne et al., 2014), and flounder (*Paralichthys olivaceus*) (Sekino et al., 2003). The reason or reasons for this contribution distortion within the snapper population is not known. One possibility is that it is a chance effect generated by individual differences in contribution to breeding events (eggs were collected from 5 days out of a 3-4 month breeding season). Alternatively, it may be more directed in that it represents differential survival of offspring due to a genetic (e.g. greater fitness in an aquaculture environment) or potentially non-genetic (e.g. egg quality) factor. Further investigation is required to identify the exact cause. Whatever the underlying reason, reducing the severity of contribution distortion in any breeding programme with a small population, such as the snapper, will improve the long-term persistence of genetic variation.

6.2.2 Quantification of inbreeding

Following the pedigree reconstruction, the inbreeding level was estimated for each individual using the method-of-moments F coefficient (F_H) (see Box 1 in Keller and Waller (2002) for a discussion of this statistic). This statistics is calculated as (observed homozygotes – expected homozygotes) / (total observations – expected homozygotes) (Kardos et al., 2015) and is equal to Nei's F statistic, but calculated using a different formula. The observed number of homozygote loci is expected to be high for an individual with more closely related parents. In this thesis, the mean inbreeding levels did not change significantly between generations. However, a comparison between the F_2 progeny from the two founding cohorts detected significantly higher inbreeding levels in the second wild cohort than in the first wild cohort. Investigation of different parent combinations (e.g. full-sibling, half-sibling, and unrelated crosses) within the F_2 generation also found significantly different inbreeding for all three groups. On average offspring produced from full-sibling parents had higher inbreeding values than those from half-sibling parents, which in turn had higher values than unrelated parents.

The inbreeding results in this thesis could have important implications for the success of the snapper breeding programme over the next 5+ generations. The level of inbreeding in small populations is determined by the effective population size and the number of generations since it was isolated from a larger source population (i.e. inbreeding accumulates each generation, Hely et al. (2013)). Because it accumulates each generation, inbreeding, which may not have any obvious impacts in the current generations, could affect the long term viability of the snapper population. Specifically, inbreeding depression due to the effect of

deleterious recessive mutations could reduce fertility, survival, and fitness of future progeny (e.g. all of these effects were observed in zebrafish (*Danio rerio*) with inbreeding levels equivalent to snapper half and full-sibling crosses, Mrakovčić and Haley (1978)). The rate of inbreeding in the snapper population is being promoted by uncontrolled group spawning between full and half-sibling parents. The lower inbreeding coefficients for F_2 progeny produced by crosses between the two wild founding populations, could indicate that merging these two lineages has produced some outbred offspring. However if uncontrolled, the high contribution distortion identified through the pedigree analysis, coupled with the strong inbreeding effect from full-sibling and half-sibling crosses, will likely override this outbreeding in the next few generations.

Finding different inbreeding levels between the two wild founding cohorts may prove an interesting result for the wild snapper population. As reported in Chapter 3, the level of inbreeding was significantly higher in the F_2 progeny of the second cohort than the F_2 progeny of the first cohort. This finding suggests that some individuals within the second cohort are relatively highly related to each other (perhaps full or half-siblings). Furthermore, the significant difference between F_2 progeny when crossing the two founding cohorts and the F_2 progeny within the two founding cohorts indicates that there is also higher relatedness within the cohorts than between the cohorts. Although, some studies have reported evidence for genetic structure over long distances within the wild snapper population (Adcock et al., 2000; Bernal-Ramírez et al., 2003), no distinction has been observed previously between the two spatially close founding cohorts (from populations in Tasman Bay and the Marlborough Sounds). The most recent findings were based on variation at eight microsatellite loci and a small sample from the two populations (Tasman Bay and Marlborough Sounds), but no difference was found (Ashton, 2013). The current results support the idea that a more powerful genetic approach is needed to properly test whether there is genetic structuring within the wild population. More detailed information on this is something that is often needed when applying stock models to marine species with large population sizes.

6.2.3 Trait heritability and potential gains from selective breeding

Heritability is typically one of the first statistics generated when investigating the genetic basis of a phenotypic trait. Heritability can be defined as the proportion of variation in a trait which is attributable to genetic factors (Visscher et al., 2008). In this thesis research, heritability was estimated for 11 phenotypic traits in the snapper population (Chapter 3). These traits were selected based on their relevance to snapper selective breeding programme and the research goals for other projects being carried out within PFR. Growth rate, was the primary trait of interest for the breeding programme, because it is a known limitation on the aquaculture potential of the species. The growth rate of snapper is currently slower than its two closest relatives, which are already commercial aquaculture species (*Sparus aurata* and *Pagrus major*). Consequently, a selective breeding programme to increase growth rate will likely be central to its development as an aquaculture species. Importantly, growth

Figure 6.1: Gains per generation (%) for a selective breeding programme based on the breeder's equation ($R = h^2 S$) and a range of heritability and selective pressures. The selection differential (S) in the equation was calculated based on a normal distribution equivalent to fork length in snapper (mean = 155.8, standard deviation = 18). The dashed line indicates the position of snapper fork length based on its heritability in year one ($h^2 = 0.28$). Selection pressure (%) reaches 50% when the top 50% of the population are selected as parents and 90% when the top 10% of the population are selected as parents.

rate (measured as fork length, peduncle length, and weight) heritability was found to have moderate heritability in year one when the main selection of individuals for the breeding programme would occur (pedigree relatedness model average: 0.26, genomic relatedness model average: 0.41, see Chapter 3). The other main trait investigated in this study was the snapper's external blue spots (Chapter 5). This trait is of particular interest because its genetic basis and biological function is not known and there are a number of application within the breeding programme (including the development of specific colour morphs within the breeding programme), which rely on a greater understanding of these spots. The current thesis begins to address this knowledge gap. The results of the current thesis, which show that the number of blue spots in year one is relatively highly heritability (pedigree relatedness model: 0.45 ± 0.13 , genomic relatedness model: 0.44 ± 0.07 , see Chapter 3) and could therefore be rapidly modified through a selective breeding programme.

Heritability provides a broad view of the relative importance of genetic and environmental factors in producing a specific trait (Visscher et al., 2008). While heritability estimates are specific to the environment in which an individual is raised, the similarity of environments that a species is distributed through can often mean that any findings about

heritability can be extrapolated more broadly. Importantly, the heritability of a trait can also guide the degree to which future research should be focused on either genetic or environmental factors – with highly heritable traits more likely to gain insights from further investigation of genetic factors. Studies which investigate both genetic and environmental factors simultaneously are preferable, however studies with such a broad scope are often not feasible for financially and logistical reasons. For example, the current study investigated growth in one environment, because it would not have been possible to get sufficient genetic power while distributing samples across multiple environments and staying within the financial limits for this research. Growth rate heritability in the current study indicated that in year one a reasonable proportion of the variance is explained by genotype information (pedigree relatedness model average = 0.26), which is a promising finding for the success of a selective breeding programme. However, the drop in heritability in year three (pedigree relatedness model average = 0.10) suggests that as fish get older, the genetic potential of growth is modulated to a strong degree by an environmental influence. Further investigation of environmental factors is needed to understand why this drop in heritability occurred from year one to year three and how it might affect growth rate in the later part of a commercial breeding cycle (e.g. enhancing growth at 1 - 2 years old). The blue spot trait was the reverse – with heritability increasing from year one (pedigree relatedness model = 0.45 ± 0.13) to year three (pedigree relatedness model = 0.63 ± 0.18).

Heritability can be used to inform selective breeding programmes at a population level using the simple but effective selection response formula ($R = h^2S$) (Kelly, 2011). In this formula, the response to selection (R) is determined by the narrow-sense heritability of the trait (h^2) and the selection differential between the population and the subset of individuals selected as parents (S). The selection differential is calculated by subtracting the trait mean of the parents from the trait mean for the population. To better understand the dynamics of the selection response formula, a visualization of its distribution was presented in Figure 6.1, assuming a normally distributed trait equivalent to snapper fork length in year one (mean = 160.1, standard deviation = 15.0). From Figure 6.1, the interactions between heritability and selective pressure can be seen when predicting the gains of a selective breeding programme. Specifically, gains are directly proportional to the trait heritability (i.e. the doubling of heritability leads to a doubling in the selection response) and increasing the selective pressure also has a large impact on response (e.g. increasing the selective pressure from the top 20% of individuals to the top 1% one can double the expected gains). Based on the heritability estimates and trait distributions from the current study in year one and selection of the top 10% of individuals as broodstock, this formula indicated that gains for the selective breeding in snapper would be approximately 4.6% and 15.7% per generation for fork length and weight, respectively (see Chapter 3). The high heritability of blue spots means that if included as a target of selective breeding this trait would result selective gains of 18.8 to 23.8% in year one and year three, respectively.

6.2.4 Linkage map and recombination

The snapper linkage map was a new and important genetic resources produced by this thesis research. It is one of the densest linkage maps reported to date for the fish family Sparidae. A total of 11K Single Nucleotide Polymorphisms (SNPs) were arranged into a linkage map containing 24 linkage groups. These 24 linkage groups are of equivalent size and number to match the expected 24 snapper chromosomes. Comparison with the scaffolds from an initial snapper genome being developed at PFR (Wellenreuther et al. unpublished) (see Chapter 4, Figure 4.3) indicated that the linkage map accurately placed the markers at broader scales (> 5 cM), but that at finer scales there was some precision limitations. Correlation of marker positions using the 26 largest scaffolds (base-pairs) and the linkage map (centimorgans) produced an R value of 0.74, indicating a moderate to high correlation between the two methods.

The development of a fine-scale linkage map has important implications for several areas of future research. The primary use of linkage maps is to function as “road-maps” for the genome, prior to the assembly of a species-specific genome. These maps (linkage map or genome) allow the position of individual structures to be determined within the genome. Historically, linkage maps have been the primary genome mapping tool because they could be constructed with limited genetic data. However, high-quality genome assemblies are now becoming the new “gold standard” for Quantitative Trait Loci (QTL) and candidate gene applications (Braasch et al., 2015). However, for most non-model species (e.g. snapper) genomes have not yet been completely assembled and many have not even been sequenced. Consequently, linkage maps still have at least two functions; firstly, as an initial map of the genome before the genome is assembled (ideally to provide pseudo chromosome information as demonstrated in this thesis), and secondly, as a resource to inform the genome assembly procedure. In the current thesis, the linkage map was used to arrange markers in the QTL analysis, find relative candidate gene positions using integrated marker and genome scaffold information, and was an important component during further assembly of the genome. Prior to construction of this linkage map snapper genome sequencing data had been assembled into a total of 5,998 scaffolds. Using the linkage map as a scaffold a total of 701 Mb (95% of the total genome) were able to be placed somewhere along the linkage map. While the snapper linkage map will likely be surpassed in the next few years by the completed genome assembly, it was a crucial element for this immediate research.

6.2.5 Identification of QTLs and candidate genes

Quantitative trait loci were identified for both of the primary traits investigated in this thesis – growth rate and number of blue spots. Both traits were selected for QTL mapping because they exhibited moderate levels of heritability in Chapter 3 and they are of interest to the snapper breeding programme at PFR. Three measures were used to investigate growth included fork length, peduncle length, and weight. The results for all three measures were highly similar, including shared QTL peaks and markers, QTL phenotypic

variance explained (R^2), and pairwise correlation of the three measures (Pearsons $r = 0.98$). Genome-wide significant QTLs for growth were found on three linkage groups (3, 11, and 16). A search for 13 candidate genes located three growth related genes on the same linkage groups as these putative QTLs. The number of blue spots was selected as a second trait for QTL mapping. Interestingly, this trait had one of the highest heritability levels of any trait in the current thesis. Genome-wide significant QTLs were found on 12 out the 24 linkage groups (full dataset), of which one was verified with a second method (half-sibling family design). A blast search of the genome scaffolds within this QTL peak identified 73 unique genes. Of these, the gene for tyrosinase, was located directly on the peak of the QTL. Tyrosinase is a key component of the melanin pathway, and has been identified as a regulator of iridophores, which form blue spots in other fish species (Lopes et al., 2008; Parichy et al., 2000).

The data and results from QTL mapping will be important resources for the snapper selective breeding programme and future research. Two approaches which could be used to integrate this information into the breeding programme are Marker Assisted Selection (MAS) (Dekkers, 2012) and Genomic Selection (GS) (Arruda et al., 2016). With MAS, QTLs affecting a trait are collectively used to predict whether an individual should be included in the broodstock (i.e. predict its estimated breeding value). These genetically based estimated breeding values have higher accuracy compared to phenotypic data alone. Genomic selection is different to MAS in that it does not use previously identified QTLs, but uses the raw genotyping and phenotyping dataset directly (Dekkers, 2012). The estimated breeding value is calculated for GS based on all genetic markers simultaneously, regardless of the magnitude of their effect on the trait (Arruda et al., 2016). Importantly, studies have shown that by using all markers simultaneously GS can predict individual breeding values more accurately than MAS (Arruda et al., 2016). Considering both of these methods, the best progress in the snapper breeding programme will most likely be achieved if the raw dataset is incorporated directly into the breeding programme using a GS approach rather than through MAS and the QTL mapping results.

Although MAS based on the QTL mapping results may not be the most effective way forward for the snapper breeding programme, the QTL mapping and candidate gene results are informative about the genetic basis of traits. One limitation of QTL mapping results is that a single study is unlikely to be sufficient for describing the genotype-phenotype map for a trait. Consequently, the transfer of knowledge between genetic studies is fundamental to achieve a deep understanding of the genotype-phenotype process. Candidate gene identification is one mechanism, which can be used to achieve this. Importantly, candidate genes identified in one study or species can be informative for other studies, because many genes and molecular pathways are conserved between even distantly related species (for example the melanoma pathway is shared from teleosts to humans, Braasch et al. (2007) and Lopes et al. (2008)). Each study that provides a putative link between a gene and a specific trait can contribute to the growing body of knowledge about this trait and gene. Another method

for sharing information between QTL mapping studies was recently described by Sutherland et al. (2016), who developed software (MapComp) to link the QTL mapping datasets of two distinct species using an intermediate genome. The development of increasing numbers of reference genomes for a wide range of species (Braasch et al., 2015) is supporting this type of cross-study and cross-species investigation by providing intermediate genomes.

6.3 Future research and applications

6.3.1 Snapper selective breeding programme

One of the primary goals of this research was to generate genomic resources and information which would inform the snapper breeding programme at PFR. This research has provided a number of findings that will be highly informative for the breeding programme. However, incorporating the results into the decision making within the breeding programme is almost as complicated as collecting them. A number of decisions must be considered when incorporating genomic information into a breeding programme (see Box 1) - many of which have complex interactions affecting the outcome of the breeding programme. Some decisions (for example, “Should the programme use phenotypic or genomic selection?”) appear definitive based on reports in the literature (e.g. Yabe et al. (2018)), but may actually be more complicated to answer for a specific breeding programme. Below is a brief discussion about the decisions outlined in Box 1 and the findings of this thesis research. However, the ultimate answers to these decisions will require ongoing discussions with facility managers and staff. To be most effective, this thesis should be followed by a simulation study, which can incorporate the current results (e.g. information about skewed contribution distortions) into a breeding plan that considers and balances the interplay between different factors (e.g. genetic diversity vs selective pressure within both manual and tank-based spawning).

6.3.1.1 Breeding

The results of this research indicated that snapper have highly distorted contribution levels, which may be a consequence of uncontrolled tank-based spawning or subsequent family-specific mortality. The most effective method to control this problem would be to use manual crossing techniques, in which gametes are collected from individual fish and then fertilization is carried out *in-vitro*. These manual crosses would have to be held in separate tanks if future research was attempting to monitor family-specific mortality. Despite its potential benefits, inconsistent collection of unfertilized eggs from female snapper currently makes this approach unworkable for the selective breeding programme. An alternative approach is to continue the use of tank-based spawning, but use genotype data to select less related individuals when constructing the broodstock for the next generation (i.e. walk-back selection). This approach would benefit from any additional steps that reduce the contribution distortions. For example, holding the broodstock in multiple tanks and standardizing the volumes or fertilized eggs from each tank may also help with mitigating the

contribution distortions. Until a consistent manual crossing technique for snapper has been developed tank-based spawning will be the only option. If walk-back selection is to be used then further research is needed to determine if it would be effective for controlling inbreeding in a multi-generational breeding programme (e.g. a simulation study). This simulation should be constructed with the specific configuration of the current breeding programme (e.g. facility capacity, biological limitations of the species, target selective breeding gains, target inbreeding limits, and relevant results from the current thesis).

Box 1: Decisions relevant for integrating genomic results into the snapper breeding programme

1. Breeding: Should the programme use manual crosses or tank-based spawning?
2. Selection: Should the programme use phenotypic or genomic selection?
3. Genetic diversity: Does the current population contain sufficient genetic diversity for a selective breeding programme?
4. Genetic diversity: How should the programme balance genetic diversity versus selective pressure?

6.3.1.2 Selection

The decision to use either phenotypic or genomic selection is particularly interesting because it interacts with all other decisions in Box 1. Previous research indicates that selection based on phenotypic breeding values alone would produce lower gains (0 to 8% lower) than genomic breeding values assuming the same intensity of selective pressure was applied with both methods (Meuwissen and Goddard, 1996; Rutkoski, 2015; Yabe et al., 2018). However, if genotyping costs limit the number of individuals in the breeding programme then selective pressure could be lower for genomic selection methods. The effect of changes in selective pressure for a specific breeding programme can be broadly evaluated using the selection response formula ($R = h^2S$). This formula shows that the response to selection (adjusted to percent gains per generation in this thesis) can be improved by both increasing the intensity of selective pressure and selecting traits with higher heritability. Selective intensity is limited by the number of individuals that can be genotyped and any other requirements placed on the selection process to managing genetic diversity (e.g. 100 parents required to maintain genetic diversity). For example, selecting 100 parents from 1000 genotyped individuals would be selecting the top 10% of individuals (i.e. 90% selective pressure, Figure 6.1) and selecting 100 parents from 10,000 genotyped individuals would be selecting the top 1% (i.e. 99% selective pressure, Figure 6.1). Based on the model in Figure 6.1, increasing the intensity of selective pressure in this way from 90% to 99% would increase potential gains by 30% for fork length in the snapper breeding programme. For growth traits (the primary target for aquaculture) in the snapper breeding programme the number

Figure 6.2: A combined phenotypic and genomic selective breeding approach for snapper applied to a single population. A total of 100K fish would start in the hatchery and be selected until only the 100 fastest growing individuals were retained. Genomic selection would be implemented once genotyping has occurred (1K fish). The correlation between growth at different ages (e.g. $R = 0.8$) would determine how effectively early phenotypic selection (e.g. 6 months) represented growth at harvest size (e.g. 2-3 years). “Walk-back” selection would be used after genotyping to control for contribution distortion.

of individuals that could be genotyped each generation is currently much lower than the number of individuals that could be phenotyped (1,000 versus perhaps 100,000). If these individuals were produced from tank-based spawning, then controlling for relatedness (e.g. no full-sibling crosses) would further reduce this number to well below the 1000 genotyped. In a best case scenario this would limit genomic selection to approximately 90% selective pressure (100 broodstock from 1,000 genotyped individuals). Phenotypic selection on the other hand, could potentially have selective pressure as high as the top 99.9% (100 broodstock from 100,000 individuals) because it is not limited by genotyping costs. However, basing selection solely on phenotype would be unwise given the level of contribution distortions observed in the current thesis research. To balance the limitations and advantages of phenotypic and genomic selection for the snapper breeding programme, a combination of both could provide the best results. Phenotypic selection could be applied during the early life stages (< 2 years old). After which genotyping would be carried out and genomic selection together with steps to prevent inbreeding could be applied to the remaining individuals. With this approach phenotypic selection for the top 1% of individuals (1000 from

100,000) could be applied before genotyping is carried out. The 1000 individuals would then be genotyped and the top 10% of these selected as parents - while also controlling for contribution distortions and relatedness. This approach would increase selective pressure and gains beyond that possible using solely genomic selection (>99% vs 90% selective pressure or approximately 30). Finally, although growth rate is the primary target for the snapper breeding programme, the other main trait investigated in this study (blue spots) also has potential application within the breeding programme. The relatively high heritability of this trait (year one = 0.54 and year three = 0.67) indicates that selection for this trait would result in rapid gains (18 to 22.8% per generation). The potential use of a unique snapper colour morph could be discussed as part of the breeding programme. It should be noted that one of the primary limitations of genomic selection, which is perpetuating this suggestion for a combined (phenotypic and genomic) breeding approach, is the cost of genotyping. However, this cost is steadily decreasing, which could allow future iterations of the breeding cycle to move more heavily towards genomic selection. The cost of genotyping should also be balanced against the cost of maintaining snapper populations for longer. For example, genotyping at year one (5,000 fish) rather than year 2 (1,000 fish) would increase genotyping costs by 5x, but could reduce the population from 5,000 fish to perhaps 200 fish one year earlier.

6.3.1.3 Genetic diversity

The final two questions in Box 1 related to genetic diversity of snapper in the breeding population. The first was a question about whether the current population contains sufficient genetic diversity for a selective breeding program. The current study is not sufficient to completely answer this question. However, an important step forward was made by reconstructing the snapper pedigree and observing the level of contribution distortion. Based on the small number of founding individuals (< 50) and the high contribution distortions in the first few generations it would be highly advisable to supplement the genetic diversity of the current population before proceeding further with the breeding program. The simulation work discussed in the previous sections could be used to completely address the question about level of genetic diversity in the current population and how selective pressure and genetic diversity should be managed over the generations of the breeding programme (final question in Box 1). Simulations could be built around the design used by Hely et al., 2013, while also including relevant snapper specific parameters (e.g. contribution distortions, facility capacity).

6.3.2 Snapper genome assembly

High-quality genome assemblies are a foundational resources for genomic research. They provide the position of genetic components within the genome and can be used to integrate results from otherwise disparate studies (e.g. comparing the QTL regions with genes in transcriptome analysis for the same trait). They can also be used to compare genomic

information between species, for example, full genomes (Vij et al., 2016), genetic markers and QTL regions (e.g. MapComp, Sutherland et al. (2016), and candidate genes (current study). A high-quality snapper genome assembly is currently being developed at PFR. At the start of this thesis research the assembly consisted of 5,998 scaffolds ($N50 = 1.5$ Mb), which included most of the genome (701Mb or 95% of total unique base pairs). The next step for assembling this genome is to continue arranging and merging these scaffolds to form super-scaffolds. The linkage map constructed in this thesis can support this process by enabling sequence gaps to be bridged between the individual scaffolds (see for example Figure 6.3 where scaffolds have been linked to their relevant position on the linkage map). The large number of markers and accuracy at broad scales (> 5 cM) support its use for this application. Finalizing the snapper genome will likely involve merging a combination of resources including the existing scaffolds, additional genome sequencing, the linkage map, and genome comparisons to related species. Development of the snapper genome will be a major step forward for genomic investigations in this and closely related teleost species.

6.3.3 Refining the genetic architecture of traits and gene discovery

This was the first research project that investigated the genetic basis of growth and number of blue spots in snapper. Quantitative trait loci were found for both traits and some potential candidate genes were also located. Future investigations could build on the results of this work by improving the precision of QTL regions, finding additional candidate genes, and further improving support for candidate genes from this study. The development of a high-quality snapper genome assembly will be an essential step for this future work. A polished genome would allow markers from genotyping data sets to be arranged more precisely than was possible using the linkage map. This would in turn improve the precision of QTL regions. Candidate genes are expected to be within these regions so narrowing them will reduce the number of potential candidates and increase the level of statistical support for the remaining genes. It should be noted that future candidate gene research could also include investigation of non-genic elements within the QTL regions (e.g. regulatory sequences). Identification of candidate genes or regulatory sequences in these regions could also be supported by transcriptome analysis. Candidates which were shared between both QTL mapping and transcriptome analysis would have a high level of support.

6.3.4 Exploration of polygenic and phenomic methods

Polygenic analysis and phenomics are two topics which are currently of particular interest in the search for genetic variation underlying phenotypic traits (Boyle et al., 2017; Houle et al., 2010). Polygenic analysis addresses the fact that most traits have a complex genetic basis consisting of multiple causative genetic variants and attempting to investigate these variants individually may not provide a complete understanding of the genotype-phenotype map (Boyle et al., 2017). Phenomics is the full set of phenotypes for an individual and is seen as the counterpart to genomics (Houle et al., 2010). Despite the importance of

Figure 6.3: A visual representation of the genome scaffolds (outer circle) positioned along the linkage map (inner circle) using SNPs that are shared between the Genotyping by Sequencing (GBS) and version 1 of the snapper genome assembly (Chapter B).

these two topics many genetic studies (including this thesis) still use relatively simplistic methods; for example, a small number of simple phenotypes and predominantly monogenic QTL mapping software. While these methods can produce informative and relevant results, future research could make significant advances by developing and implementing polygenic and phenomic methods. One of the primary limiting factors for polygenic and phenomic methods are that they are often significantly more complicated than other approaches. For example, software has been developed to carry out polygenic QTL mapping (e.g. QTLBIM, Yandell et al. (2007)), but is much more complex to use than monogenic software (e.g. QTDT which was used in this thesis, Abecasis et al. (2000)). Multivariate approaches can also be used for QTL mapping with phenomic data (Galesloot et al., 2014), but again are

much more complex to use than software developed for a single phenotypic trait. Future investigations that identify the genetic basis of phenotypic variation could benefit from further software development (e.g. intuitive user interfaces and consistent data formats) and more user friendly documentation to improve access to these programs. An additional limiting factor for phenomics is that phenotypic data collection technology is often not capable of collecting sufficient volumes of data. Unlike genomics which has had a recent boost with high-throughput sequencing, phenotype data collection is often still manually collected. The development of high-throughput phenotyping technology will be central to more widespread incorporation of phenomic approaches into QTL mapping studies.

6.4 Conclusion

The research carried out in this thesis addressed a broad range of questions and has implications for an equally broad range of areas. Some of the results are directly informative for the snapper breeding programme (e.g. need for further development of controlled breeding). Others will likely require a future simulation study to fully integrate them into the breeding programme (e.g. balancing the consequences of the different decisions within the breeding programme). This thesis has also produced a number of genomic resources that will be useful for future work (e.g. linkage map). Finally, the integrated nature of genomic research means that the results and resources from this thesis will likely be applicable to other research programmes involving snapper, other teleost species, and perhaps even to more distantly related species.

A.2 Chapter 2

Supplementary figure 2.1: This figure shows the number of unique publications from the literature review (Chapter 2) that were found in the initial Web of Science search. Publications were separated by groups, nonqtl = not a QTL mapping study, included = included in the literature review, nonfish = not a fish related study, aprior = a QTL study based on prior work, language = no access to an english version of the publication, other = misc papers excluded from further analysis

Supplementary figure 2.2: This figure shows the different mapping populations used by QTL mapping studies in the literature review (Chapter 2). The proportion of studies which used specific family sizes is shown on the y-axis and colour coding indicates the specific crossing strategy used to produce the mapping population.

A.3 Chapter 3

Supplementary figure 3.1: This illustration shows the number of sequencing reads and SNPs produced by GBS at different stages of the filtering and for different individuals in the *Chrysophrys auratus* data set (Chapters 3, 4, 5, 6). F_0 and F_1 individuals had 2-3 times the amount of sequencing of F_2 individuals. The total number of reads dropped slightly after filtering with Fastq-mcf. The number of SNP loci started at 249,468 and then drops considerable with each of the filtering steps.

Supplementary figure 3.2: These graphs show the FastQC quality scores across the 100bp reads for each of the eight pooled sequencing plates. The quality was high across the full length of the reads for all plates, except slightly reduced quality towards the end of plate 6.

Supplementary table 3.1: This table shows the p-values for Welch two sample t-tests comparing the difference of the mean F statistic (F_H) between the three generations (A), between F^2 individuals grouped by grandparent type (B), and between F^2 individuals grouped by parent type (C). P-values have been adjusted for multiple comparisons using a Bonferroni correction.

	A			B			C	
	F_0	F_1		w1-w1	w1-w2		Full-sib	Half-sib
F_1	0.7410	-	w1-w2	4.1e-05		Half-sib	0.0148	
F_2	1.0000	0.0890	w2-w2	< 2e-16	< 2e-16	Unrelated	6.3e-12	0.0047

Supplementary table 3.2: This table shows the mean, standard deviation (SD), and number of measurements (n) for each continuous trait for year one and year three for the current thesis. Sample sizes were lower in year three than year one as a result of natural mortality.

Year	Trait	Mean	SD	n
one	Fork length (mm)	160.1	15.0	568
one	Peduncle length (mm)	132.1	12.3	568
one	Weight (g)	89.8	23.9	506
one	Height_0.25 (pixels)	377.2	20.4	568
one	Height_0.5 (pixels)	460.9	19.2	568
one	Height_0.75 (pixels)	388.4	20.2	568
one	Skin darkness	89.6	10.0	568
one	Spots	44.7	9.5	568
three	Fork length (mm)	257.8	20.1	314
three	Peduncle length (mm)	214.5	17.0	314
three	Weight (g)	361.9	82.3	247
three	Height_0.25 (pixels)	381.9	14.2	314
three	Height_0.5 (pixels)	460.9	19.6	314
three	Height_0.75 (pixels)	384.2	21.4	314
three	Skin darkness	111.3	8.4	314
three	Spots	43.5	8.4	314

A.4 Chapter 4

Supplementary figure 4.1: These figures show the correlation between fork length, peduncle length, and weight in year one and year three for the same individuals. Moderate correlation ($R = 0.710 - 0.730$) was found for all traits.

Supplementary table 4.1: This table shows the details for a complete list of linkage group-wide and genome-wide significant QTLs for growth related traits in the current study. Included are linkage group (LG), centimorgan position (cM), marker name (loci), significance level ($-\log_{10}(p)$), and proportion phenotypic variance explained (R^2). The genome-wide significance limit ($-\log_{10}(p)$) was 5.33.

LG	cM	loci	Fork length		Peduncle length		Weight	
			$-\log_{10}(p)$	R^2	$-\log_{10}(p)$	R^2	$-\log_{10}(p)$	R^2
3	12.2	29323_42	5.30	0.05	5.30	0.05	4.22	0.04
3	20.9	40730_33	5.52	0.04	5.70	0.04	4.30	0.03
3	21.6	56457_62	4.30	0.04	4.52	0.05	4.05	0.04
3	25.5	36800_49	4.00	0.03	4.10	0.03		
3	25.7	62747_60	4.15	0.04	4.40	0.04	4.22	0.04
3	25.9	36768_54	4.40	0.03	4.70	0.04		
3	28.8	93148_67	4.52	0.04	4.70	0.04	4.70	0.04
3	30.8	36929_14	4.52	0.03	4.70	0.04	4.15	0.03
3	36.7	44757_8	4.05	0.03	4.70	0.03		
5	32.1	100144_33	4.52	0.04	4.22	0.03	4.52	0.04
11	0.0	6283_25	4.22	0.03	4.10	0.03		
11	0.7	17818_78	4.00	0.03	4.00	0.03	4.15	0.03
11	1.7	17817_58	4.70	0.03	4.70	0.03	4.70	0.03
11	2.7	43276_58	4.40	0.03	4.40	0.04	4.22	0.04
11	4.0	6330_87	4.40	0.04	4.52	0.04	4.70	0.05
11	5.0	99820_51	4.00	0.03				
11	10.0	99093_35	5.70	0.05	5.70	0.04	5.22	0.04
12	10.8	32372_36	5.00	0.04	4.22	0.03		
12	15.8	7329_66	4.22	0.03	4.00	0.03		
13	0.3	80827_64	4.00	0.03				
13	5.7	42861_57	4.70	0.03	4.15	0.03		
13	5.7	42868_38	4.15	0.03				
13	5.7	112760_12	4.00	0.03				
16	20.2	19587_24	4.00	0.03				
16	20.2	19725_52	4.00	0.04				
16	21.2	17603_14	5.15	0.05	5.00	0.05		
16	22.7	47450_65	4.05	0.03	4.10	0.03		
16	24.6	103530_18	4.10	0.03			4.22	0.04
16	40.2	39092_27	5.52	0.05	5.22	0.05	4.52	0.05
16	41.0	39102_79	4.00	0.03				
16	42.7	45481_29	4.05	0.03			4.00	0.03
16	47.0	68334_77	4.00	0.04			4.00	0.04
16	47.9	62074_47	5.70	0.05	5.52	0.04	4.52	0.03
16	48.6	24347_75	4.52	0.04	4.10	0.03	4.22	0.03
16	49.8	27578_5	4.15	0.03			4.30	0.04
16	51.0	75441_82	4.00	0.03				

Supplementary table 4.2: This table shows the details for all the candidate growth genes investigated in this thesis. Included are the gene names (Gene), source species (Species), Accession # (Accession), sequence type (Type), linkage group number (LG), linkage group position (cM and bp), and % sequence match (Match). An asterix * indicates that the gene was located on a linkage group also containing a genome-wide significant QTL for growth rate.

Gene	Species	Accession	Type	LG	cM	bp	Match
growth-hormone*	<i>Pagrus major</i>	AB904715.1	DNA	16	53.2	147	99.3
myostatin	<i>Pagrus major</i>	AY965686.1	DNA	6	38.3	403	99.0
growth-hormone-receptor	<i>Epinephelus coioides</i>	KR269817.1	DNA	9	58.3	769	88.7
growth-hormone-receptor-type-I	<i>Sparus aurata</i>	AH014067.4	DNA	9	58.3	760	95.4
growth-hormone-receptor-type-II	<i>Sparus aurata</i>	AH014068.4	DNA	18	50.3	875	94.1
myogenin*	<i>Sparus aurata</i>	EF462192.1	DNA	3	9.6	534	96.8
myogenic-factor-MYOD1	<i>Sparus aurata</i>	AF478568.1	DNA	7	12.6	591	97.0
myogenic-factor-MYOD2	<i>Sparus aurata</i>	AF478569.1	DNA	14	39	546	96.5
myogenic-regulatory-factor-4	<i>Epinephelus coioides</i>	KR269828.1	DNA	15	25.2	510	93.8
myogenic-regulatory-factor-6	<i>Sparus aurata</i>	JN034421.1	mRNA	15	25.2	521	95.6
insulin-like-growth-factor-I	<i>Sparus aurata</i>	DQ118098.1	DNA	15	25.2	186	99.5
insulin-like-growth-factor-II	<i>Epinephelus coioides</i>	KR269813.1	DNA	7	23.4	240	96.7
parvalbumin*	<i>Sparus aurata</i>	GU060310.1	mRNA	16	15.2	311	97.7

A.5 Chapter 5

Supplementary figure 5.1: This figure illustrates the affine transformation process used to compare blue spot patterns for two individuals with different size and rotation. The two reference points (A and B) are used to rotate and resize both individuals and their spots to a common position. Once this is done the average distance between closest spots was used as a measure of pattern similarity.

Supplementary figure 5.2: This figure shows the distribution by species of blast hits for the blue spot QTL region on linkage group 23. The blast search included all scaffolds from v1 of the snapper genome that contained a SNP marker which was also found within the QTL region above the 95% genome-wide significance limit. Total number of unique genes found using the BLAST results was 73.

Supplementary table 5.1: This table shows the details for all the candidate blue spot genes investigated in this thesis. Included are the Accession # (Accession), sequence type (Type), source organism (Organism), scaffold number from v1 of the snapper genome (Scaffold), start and end base-pair position of the gene (Start and End), % match score for the alignment (Match), and the brief gene description from NCBI.

Accession	Type	Organism	Scaffold	Start	End	Match	Description
BN000676	mRNA	<i>Xenopus tropicalis</i>	571	53061	54659	82.3	Tc1-like transposon, Beagle
HQ677723	DNA	<i>Ctalarus punctatus</i>	571	100254	101881	90.8	TLR20-1 and TLR20-2
XM_008278968	mRNA	<i>Stegastes partitus</i>	571	175196	176998	80.7	nemo-like kinase (nlk)
XM_018684032	mRNA	<i>Lates calcarifer</i>	662	107892	109111	88	14-3-3 protein epsilon-like (LOC108888178)
XM_022767632	mRNA	<i>Seriola dumerili</i>	662	131186	131781	91.3	adapter molecule crk-like (LOC111238220)
XM_015389889	mRNA	<i>Yprinodon variegatus</i>	662	144107	144773	80.1	leucine-rich repeat-containing protein 75A-like (LOC107094335)
XM_018693281	mRNA	<i>Lates calcarifer</i>	147	1013811	1015518	83.6	archain 1 (arcn1)
XM_018684016	mRNA	<i>Lates calcarifer</i>	147	1308719	1310134	84.4	ubiquitin carboxyl-terminal hydrolase 2-like (LOC108888168)
XM_005804516	mRNA	<i>Xiphophorus maculatus</i>	147	1438221	1438752	82.7	transmembrane protein 199 (tmem199)
KC217618	mRNA	<i>Sparus aurata</i>	147	412911	413856	91.6	succinate dehydrogenase [ubiquinone] cytochrome b small subunit B (SDHD)
XM_022768088	mRNA	<i>Seriola dumerili</i>	147	580886	582031	82.1	ADAM metalloproteinase with thrombospondin type 1 motif 15 (adamts15)
XM_022218018	mRNA	<i>Acanthochromis polyacanthus</i>	147	640416	641400	93.6	zinc finger and BTB domain containing 44 (zbtb44)
XM_022764515	mRNA	<i>Seriola dumerili</i>	211	200257	201800	85.6	E3 ubiquitin-protein ligase TRIM11-like (LOC111235940)
XM_018701312	mRNA	<i>Lates calcarifer</i>	211	289411	291018	84.3	unc-51 like autophagy activating kinase 2 (ulk2)
AY965681	DNA	<i>Ptaretus stoutii</i>	211	320744	322431	88.2	nonfunctional variable lymphocyte receptor B (VLRB)
XM_016468617	mRNA	<i>Sinocyclocheilus anshuiensis</i>	211	396382	398128	80.6	zinc finger BED domain-containing protein 1-like (LOC107674596)
XM_018701241	mRNA	<i>Lates calcarifer</i>	211	438084	439775	85.8	amyloid beta precursor protein binding protein 2 (appbp2)
XM_018701230	mRNA	<i>Lates calcarifer</i>	211	457009	457736	85.3	protein phosphatase, Mg2+/Mn2+ dependent 1D (ppm1d)
XM_008425197	mRNA	<i>Oecilia reticulata</i>	211	620463	621887	82.2	T-box 2 (tbx2)
L48684	DNA	<i>Asbora heteromorpha</i>	211	901801	902477	87.1	Tc1-like transposon DNA, transposase gene
AF013120	DNA	<i>Sparus aurata</i>	286	118286	119076	89.8	apolipoprotein A-1
XM_022205801	mRNA	<i>Acanthochromis polyacanthus</i>	286	199444	200646	84.2	SIK family kinase 3 (sik3)
HM159991	DNA	<i>Sparus aurata</i>	286	608642	608917	86.3	Mx3 protein (Mx3)
AB904715	DNA	<i>Pagrus major</i>	286	688909	689106	94.4	gene for growth hormone
AB780555	DNA	<i>Haplochromis chilotes</i>	315	266405	268009	83.2	DNA containing V2R gene cluster region
XM_023830917	mRNA	<i>Aramormyrops kingsleyae</i>	315	494319	495417	88.8	kingsleyae REM2 and RAB like small GTPase 1 (rsg1)
XM_010782568	mRNA	<i>Notothenia coriiceps</i>	315	603054	604601	80.8	suppressor protein SRP40-like (LOC104955286)
XM_023282894	mRNA	<i>Amphiprion ocellaris</i>	5	1086989	1089572	88.5	frizzled class receptor 9 (fzd9)
XM_020618168	mRNA	<i>Monopterus albus</i>	5	1186453	1190877	82	zinc finger and BTB domain containing 38 (zbtb38)
XM_018660047	mRNA	<i>Lates calcarifer</i>	5	1224810	1226323	94.7	thyroid hormone receptor interactor 12 (trip12)
XM_018660056	mRNA	<i>Lates calcarifer</i>	5	1456680	1460050	88.6	SPHK1 interactor, AKAP domain containing (sphkap)
XM_020600385	mRNA	<i>Monopterus albus</i>	5	2170028	2175863	83.2	SLIT and NTRK like family member 3 (slitrk3)
XM_018681600	mRNA	<i>Lates calcarifer</i>	5	254666	256047	81.6	frizzled class receptor 4 (fzd4)
XM_023406624	mRNA	<i>Seriola lalandi</i>	5	326001	327484	84.1	dorsalis glutamate metabotropic receptor 5 (grm5)
AY965681	DNA	<i>Ptaretus stoutii</i>	5	1383057	1383496	92.3	nonfunctional variable lymphocyte receptor B (VLRB)
XM_022768915	mRNA	<i>Seriola dumerili</i>	5	2749319	2750528	81.1	golgi integral membrane protein 4 (golim4)
XM_022768907	mRNA	<i>Seriola dumerili</i>	5	2803144	2804048	82.2	pentraxin 3 (ptx3)

A. SUPPLEMENTARY MATERIAL

XM_018662119	mRNA	<i>Lates calcarifer</i>	5	3039789	3040707	92.3	TCDD inducible poly(ADP-ribose) polymerase (tiparp)
XM_022768858	mRNA	<i>Seriola dumerili</i>	5	3237800	3239181	80.3	ribosomal protein S6 kinase beta-1 (LOC111239086)
XM_008275957	mRNA	<i>Stegastes partitus</i>	5	3254872	3256164	83.2	vacuole membrane protein 1-like (LOC103353141)
XM_018662098	mRNA	<i>Lates calcarifer</i>	5	3273477	3274765	82.2	protein NPAT (LOC108873773)
XM_018662103	mRNA	<i>Lates calcarifer</i>	5	3470491	3471892	96	Down syndrome cell adhesion molecule homolog (LOC108873775)
XM_022746336	mRNA	<i>Seriola dumerili</i>	5	3694916	3707881	90.6	sacsin molecular chaperone (sacs)
XM_023401320	mRNA	<i>Seriola lalandi</i>	5	3848932	3850855	86.5	dorsalis dishevelled binding antagonist of beta catenin 3 (dact3)
XM_023282482	mRNA	<i>Amphiprion ocellaris</i>	5	3865546	3867518	82	GRAM domain containing 1B (gramd1b)
XM_005474590	mRNA	<i>Oreochromis niloticus</i>	5	3996160	3998462	81	CXADR like membrane protein (clmp)
XM_022748531	mRNA	<i>Seriola dumerili</i>	5	404916	405768	91.9	tyrosinase (tyr)
XM_023406646	mRNA	<i>Seriola lalandi</i>	5	482837	485690	80.8	dorsalis nuclear mitotic apparatus protein 1 (numa1)
AB219244	DNA	<i>Pagrus major</i>	5	4552964	4554025	99.3	somatolactin
XM_018698903	mRNA	<i>Lates calcarifer</i>	5	4616271	4619576	83.4	ETS proto-oncogene 1, transcription factor (ets1)
XM_022761181	mRNA	<i>Seriola dumerili</i>	5	4831669	4833033	80.5	SRP receptor alpha subunit (srpra)
AJ810410	mRNA	<i>Xenopus tropicalis</i>	5	5063390	5064259	82	Tc1-like transposon, Froggy
XM_023420455	mRNA	<i>Seriola lalandi</i>	5	5292748	5294828	84.5	dorsalis RAC-beta serine/threonine-protein kinase (LOC111665447)
XM_020641767	mRNA	<i>Labrus bergylta</i>	5	612228	612924	83.2	ClpB homolog, mitochondrial AAA ATPase chaperonin (clpb)
XM_023406634	mRNA	<i>Seriola lalandi</i>	5	876750	878196	84.4	dorsalis phosphodiesterase 2A (pde2a)
XM_008309821	mRNA	<i>Ynoglossus semi-laevis</i>	5	896294	897534	82.3	cullin 3 (cul3)
XM_018684101	mRNA	<i>Lates calcarifer</i>	50	1156781	1158385	85.4	EPH receptor A4 (epha4)
XM_020094887	mRNA	<i>Aralichthys olivaceus</i>	50	135691	136550	88	endothelin converting enzyme like 1 (ece1)
GU129139	mRNA	<i>Salmo salar</i>	50	1474417	1475735	95.8	IgH locus A
FJ705451	DNA	<i>Icentrarchus labrax</i>	50	1929245	1930237	96.1	forkhead box L2 (foxl2)
XM_023281314	mRNA	<i>Amphiprion ocellaris</i>	50	2051894	2054149	83.6	calcium binding protein 39 (cab39)
KF018667	mRNA	<i>Sparus aurata</i>	50	2121151	2122414	83.4	NAD-dependent protein deacetylase sirtuin-2 (SIRT2)
XM_023266390	mRNA	<i>Amphiprion ocellaris</i>	50	2168130	2171100	83	gap junction delta-2 protein-like (LOC111566028)
XM_022205745	mRNA	<i>Acanthochromis polyacanthus</i>	50	315612	316580	85.1	solute carrier family 12 member 9-like (LOC110958997)
XM_023405187	mRNA	<i>Seriola lalandi</i>	50	2284720	2286733	86.4	dorsalis kelch-like protein 17 (LOC111654248)
XM_022760523	mRNA	<i>Seriola dumerili</i>	50	2298785	2302422	90.9	F-box protein 46 (fbxo46)
XM_023405180	mRNA	<i>Seriola lalandi</i>	50	2386700	2389062	82.5	dorsalis numb-like protein (LOC111654240)
XM_018687626	mRNA	<i>Lates calcarifer</i>	50	2451583	2452153	87.1	double PHD fingers 1 (dpf1)
XM_022760500	mRNA	<i>Seriola dumerili</i>	50	309553	310048	87.8	endonuclease domain-containing 1 protein-like (LOC111232836)
XM_008301869	mRNA	<i>Stegastes partitus</i>	50	866594	868702	83.4	secretogranin II (scg2)
XM_022760496	mRNA	<i>Seriola dumerili</i>	50	989541	991059	88.4	E3 ubiquitin-protein ligase rnf152-B-like (LOC111232834)

Appendix 1: Genomic resources for the marine finfish Australasian snapper *Chrysophrys auratus*

B

Authors

David T. Ashton, Peter A. Ritchie, Ross Crowhurst, & Maren Wellenreuther

Author Contributions

Conceived and designed the experiments: DTA, PR, and MW; Data collection: DTA; Data analysis: DTA; Wrote the primary manuscript: DTA; Prepared final manuscript: DTA, PR, and MW; Assembled version one of the snapper genome; Ross Crowhurst

Publication status

This appendix provides content from my thesis and some related research at Plant & Food Research, which I presented at 7th World Fisheries Congress in Korea in 2016. This content was presented in a special session entitled *Genomics for improved fisheries management and conservation: have the promises been fulfilled?*. This special session was sponsored by the OECD Co-operative Research Programme on Biological Resource Management for Sustainable Agricultural Systems. Following the conference all presenters, including myself, co-authored a paper based on this content for Trends in Ecology and Evolution (Bernatchez et al., 2017).

Bernatchez, L., Wellenreuther M., Araneda C., Ashton D. T., Barth J. M. I., Beacham T. D., Maes G. E., Martinsohn J. T., Miller K. M., Naish K. A., Ovenden J. R., Primmer C. R., Young Suk H., Therkildsen N. O., and Withler, R. E. (2017) Harnessing the power of genomics to secure the future of seafood. *Trends in Ecology & Evolution*, 32(9), 665-680 - DOI:10.1016/j.tree.2017.06.010

B.1 Abstract

The increasing global demand for protein-rich food source is a significant challenge for sustainable production. In New Zealand, wild-capture fisheries have already peaked or are falling due to exploitation pressures and variable recruitment, leading to substantial fluctuations of natural stocks. Promising solutions to alleviate these pressures and protect the future of these stocks include implementation of fisheries management that reflects the biology of the target species, stock enhancement through re-seeding methods, and alternative seafood production through domestication of new aquaculture species. All of these solutions can benefit from a good understanding of the genomics of the target species (e.g. structure of wild genetic variation). However, for many fish species genomic resources (e.g. fully assembled genomes, genetic markers) are either severely limited or not available. The native marine finfish Australasian Snapper (*Chrysophrys auratus*) is one such species, which although it supports an important New Zealand fishery has until recently had no genomic data available. Work is now being undertaken to develop genomic resources for this species, which can be used to improve the domestication process for cultured populations and help inform management of the wild fisheries.

B.2 Introduction

Fisheries is a major source of revenue for New Zealand (\$1.42 billion), which has the fourth largest exclusive economic zone in the world (Figure B.1). Wild-catch is the primary source for most fisheries, but with many species at or above maximum production capacity there is an increased focus on the potential for development of aquaculture species and supplementation of natural populations. Fisheries management in New Zealand is carried out under the Quota Management System (QMS), which was introduced in the 1980's following concern about the rapid depletion of previously healthy stocks. This together with other legislation has led to NZ being recognized as having one of the top fisheries management systems in the world (Alder et al., 2010). However, with increasing pressure to produce more and higher quality seafood and the development of new technologies (e.g. genomics) an ongoing focus needs to be on how we can continue to improve management of this precious resource.

The Australasian snapper (*Chrysophrys auratus*) is one of New Zealand's most valuable inshore finfish species (\$36.8m pa in exports) (MPI, 2013). In addition to the commercial fishery, snapper populations also support a major recreational fishery. While it is not a current aquaculture species in New Zealand, the close sister species (*C. major*) is a highly cultured fish in Japan and a three generation experimental cultured population has been established at Plant and Food Research in Nelson (Figure B.2). Genomic data has not been used when developing management strategies for New Zealand snapper populations. One of the main reasons for this has been the limited availability of genomic resources for this species. Until recently, the only genetic resources available for this species were a few

Figure B.1: Despite its relatively small size, New Zealand has the 4th largest exclusive economic zone in the world.

microsatellites transferred from *C. major* and short mitochondrial DNA sequences (Adcock et al., 2000; Hauser et al., 2002; Bernal-Ramírez et al., 2003). Over the last two years, work has been undertaken at Plant and Food Research to “genomically enable” snapper by developing a range of genomic resources, including a fully assembled and annotated genome, development of a large dataset of genetic markers, and development of linkage map. A brief overview of this work is given below.

B.3 Material and methods

B.3.1 Snapper populations

Research on snapper at Plant & Food Research has been carried out using samples collected from a three generation cultured snapper pedigree held in Nelson. The F_0 generation (grandparents) is comprised of individuals ($n = 50$) collected from Tasman Bay in the South Island of New Zealand. These were held in two separate populations and used to produce five F_1 year classes using tank based spawning. In 2013 a selection of individuals

Figure B.2: Cultured Australasian snapper (*Chrysophrys auratus*) at the Nelson hatchery of Plant & Food Research Limited.

($n = 70$) from the five F_1 year classes were combined into a single population and used to produce a F_2 generation ($n = 577$) by tank based spawning.

B.3.2 Genotyping by Sequencing

A modified Genotyping by Sequencing (GBS) approach was used to genotype all individuals in the three generation pedigree (Hilario, 2015). A double digest (restriction enzymes: PstI and MspI) GBS approach was used to generate the libraries. A total of 672 individuals were split into 96 individuals per sequencing lane, with 2-3 replicates used for each of the F_0 and F_1 individuals. A total of 8x lanes of sequencing were carried out on a HiSeq 2000 at the Australian Genome Research Facility (AGRF) in Australia. Single Nucleotide Polymorphisms (SNPs) were extracted from the raw data using the STACKs pipeline (Catchen et al., 2013) after checking and trimming reads with fastqc and fastq-mcf, respectively. Relatedness estimates were calculated for individuals using an R package developed by Dodds et al., 2015 and checked with a second software package Cervus (Kalinowski et al., 2007). A linkage map was constructed using Lep-MAP2 (Rastas et al., 2016), which was based on the 10 largest F_1 and F_2 families in the data set.

Figure B.3: Historical snapper biomass (Percent of original biomass) in the Hauraki Gulf over the last century (MPI, 2013)

B.3.3 Genome sequencing

A genome assembly for snapper was being constructed using one F_1 individual. Sequencing data included a 160bp (PE151), 3kbp (PE101), 8kbp (PE101), 10kbp (PE50), and 20kbp (PE50) mate pair libraries. The initial genome assembly was carried out using ALLPATHS (Butler et al. 2008). The scaffolds from the initial assembly and the linkage map from the GBS data were used to check the match between ordering on the linkage map and scaffolds. Comparison between the linkage map and scaffolds was then plotted using the CIRCOS software (Krzyszowski et al. 2009). Optical mapping was also carried out to aid the genome assembly process.

B.3.4 Phenotyping and QTL mapping

Phenotyping and Quantitative Trait Loci (QTL) mapping were carried out using the GBS data and three generation pedigree. All traits currently being measured are non-lethal and include growth rate information, disease information, time series images, and other observations recorded in pedigreed fish while being held at Plant & Food Research. Automated phenotyping software is being developed, which can extract a wide range of phenotypic shape and growth information from the images collected. Initial QTL scans were carried out on 1 year old fish using GridQTL (Seaton et al., 2006) and 10 families from the pedigree with mixed levels of relatedness.

B.4 Results

A total of 1.6 GB of data was produced from the 8 lanes of sequencing with each individual having between 2 – 6 Million reads. FastQC results indicated that the data quality was very high, with 1.5 GB remaining after filtering out low quality reads. Running the data through STACKs generated a total of 249,468 SNPs with greater than 7x coverage. 20,311 SNPs remained after filtering for those with a Minor Allele Frequency (MAF) above 0.05 and that were present in 75% of the individuals. This data was used to determine the pedigree structure of the cultured population, generate a dense linkage map, and identify several QTLs for growth and colour traits. The linkage map is of high quality and consists of 10,968 markers in 24 linkage groups. Data from a range of sequencing libraries is being used to construct a full genome sequence. The linkage map and optical mapping have also proved suitable to help link together genome scaffolds. Before linkage mapping the genome assembly consisted of 5,998 scaffolds. Of those, 1,479 were able to be placed somewhere on the linkage map. Those 1,479 scaffolds comprised a total of 701Mb (95% of the total bp in the genome). The 1,479 scaffolds were combined into 24 linkage groups which match our expected 24 chromosomes, and together had a N50 scaffold value of 1.5 Mb. Following integration of the optical mapping data to improve the genome assembly the N50 scaffold value increased to 4.4 Mb.

B.5 Discussion

The development of high throughput techniques (e.g. next-generation sequencing) has led to the transition from relatively basic genetic methods into a new area of research known as genomics. While prior genetic methods used relatively low numbers of neutral genetic markers, genomic methods can use large numbers of neutral and adaptive genetic markers, full genome sequencing data, and more effectively address the complex array of interactions that occur across the genome (see Wenne et al. (2007) and Valenzuela-Quinonez (2016)). Due the significant increases in power over previous genetic methods there is much interest in using genomics within fisheries research. Potential applications for genomics in fisheries include informing researchers and managers about the structure of populations, how wild populations may respond to different selective pressures (e.g. climate change, fishing), how the genetics of escapees or released fish from cultured populations might affect wild populations, and how wild source populations can best be used when constructing new cultured populations (see Wenne et al. (2007) and Valenzuela-Quinonez (2016)).

Despite significant benefits for fisheries management and research, genomics is still underutilized or not used at all in many New Zealand fish species. One of the main reasons for this is the high resource requirements of genomics based approaches. Additionally, unlike land based animal culture, which is dominated by relatively few species (e.g. sheep, cattle, pigs), fish, which are still largely wild-caught, have the highest species diversity for any vertebrate group. This results in most genomic research being spread thinly across the dif-

ferent species and limited or no genomic resources being available for many key fisheries or aquaculture species. However, the reduced cost of sequencing and the development of new genotyping method for species with no prior information (e.g. genotyping by sequencing) are beginning to overcome this limitation (Elshire et al., 2011; Hilario, 2015).

Snapper is a good example of a fish species reflecting the opportunity for, but limited application of genomic resources. It is an important recreational and commercial teleost fish species in New Zealand and worth \$35M pa in exports and supports an important local fishery. Some population structure may be present although it is poorly understood (due to being investigated with a few weak genetic markers) (Bernal-Ramírez et al., 2003). Currently, snapper is only harvested from the wild, but there is interest to develop this species for aquaculture and potentially restocking of the wild populations. Wild snapper populations have recently undergone a significant reduction in total biomass (for example, 20% of original biomass in the Hauraki Gulf, see Figure B.3) (MPI, 2013) and restocking could present on method for recovering this stock. Despite a wide range of potential applications, no genomic approaches have been used in the research or management of this species. Current management has made use of available genetic data (MPI, 2013), but this data has limited capacity for informing managers because it is based on relatively few genetic markers. Initial projects are underway at Plant & Food Research to start addressing this shortage of genomic data, including development of a genome assembly, a linkage map, and the identification QTLs affecting phenotypic traits of interest.

With a large amount of genomic data now produced for the cultured snapper populations held at Plant & Food Research, the focus for future work is likely to centre on investigation of natural populations, annotation of the assembled genome, and further QTL mapping. Investigation of the natural population will likely include resequencing and genotyping of individuals to determine the structure of the wild populations, direction and strength of gene flow, and the amount of adaptive genetic variation in different regions of New Zealand.

B.6 Conclusions

Despite the potential benefits of genomic research to fisheries management, the lack of genomic resources for many species is an ongoing limitation to its use in fisheries management. Australasian snapper is one such species. Until recently, no genomic data was available for this species. Work has now been undertaken to “genomically enable” snapper at Plant & Food Research including development of a genome assembly, a linkage map, and the identification QTLs affecting phenotypic traits of interest. Further work is needed to better characterize the structure and diversity of the wild population.

B.6.1 Country specific: New Zealand

Although fisheries is an important area of research in New Zealand, there is currently no government funding directly related to support fisheries genomics. However, as is the case

with snapper genomics at Plant & Food Research, funding that is not specifically targeted at fisheries genomics may be used for this if the genomics research presents the best way of addressing a specific question. Some sources of funding for fisheries and genomics in New Zealand include, 1) the Ministry of Business, Innovation, and Employment (MBIE), 2) Seafood Innovation Limited (SIL), and 4) Crown Research Institute (e.g. Plant & Food Research) core funding. The mission of SIL is primarily focused on quality of harvested fisheries products and is therefore less likely to fund fisheries genomics. Crown Research Institute core funding may be the predominant source of funding for fisheries genomics, which is also how the snapper genomics at Plant and Food Research has been funded until now.

Approximately 130 species are targeted commercially in New Zealand's exclusive economic zone. Limited genetic data and no genomic data is available for most species. Some major species such as snapper and hoki (*Macruronus novaezelandiae*) have undergone some basic analysis for structure, but this has been carried out with relatively weak neutral genetic markers or less powerful non-genetic methods (Bernal-Ramírez et al., 2003; Ballara and O'Driscoll, 2014). Because these cases have mostly confirmed a lack of structure or that existing management structure is sufficiently complex they have not required changes to existing management strategies. As far as genetics informing catch levels goes, this has typically been superseded by recorded catch and population modelling methods (for example, MPI (2013)).

B.7 Acknowledgements

This paper summarizes the content of a talk delivered at the Session: 'Genomics for improved fisheries management and conservation: have the promises been fulfilled?' at the 7th World Fisheries Congress in South Korea, 2016. We would like to thank the OECD Co-operative Research Programme on Biological Resource Management for Sustainable Agricultural Systems for providing support to all invited speakers for attending this symposium.

Bibliography

- Abecasis, G. R., Cardon, L. R., and Cookson, W. O. (2000). A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics*, 66(1):279–292.
- Adcock, G. J., Bernal Ramírez, J. H., Hauser, L., Smith, P., and Carvalho, G. R. (2000). Screening of DNA polymorphisms in samples of archived scales from New Zealand snapper. *Journal of Fish Biology*, 56(5):1283–1287.
- Alder, J., Cullis-Suzuki, S., Karpouzi, V., Kaschner, K., Mondoux, S., Swartz, W., Trujillo, P., Watson, R., and Pauly, D. (2010). Aggregate performance in managing marine ecosystems of 53 maritime countries. *Marine Policy*, 34(3):468–476.
- Aljanabi, S. M. and Martinez, I. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research*, 25(22):4692–4693.
- Allison, D. B. (1997). Transmission-disequilibrium tests for quantitative traits. *American Journal of Human Genetics*, 60(3):676–690.
- Almuly, R., Cavari, B., Ferstman, H., Kolodny, O., and Funkenstein, B. (2000). Genomic structure and sequence of the gilthead seabream (*Sparus aurata*) growth hormone-encoding gene: identification of minisatellite polymorphism in intron I. *Genome*, 43(5):836–845.
- Amundsen, T. and Forsgren, E. (2001). Male mate choice selects for female coloration in a fish. *Proceedings of the National Academy of Sciences*, 98(23):13155–13160.
- Antonello, J., Massault, C., Franch, R., Haley, C., Pellizzari, C., Bovo, G., Patarnello, T., de Koning, D.-J., and Bargelloni, L. (2009). Estimates of heritability and genetic correlation for body length and resistance to fish pasteurellosis in the gilthead sea bream (*Sparus aurata* L.). *Aquaculture*, 298(1):29–35.
- Arcade, A., Labourdette, A., Falque, M., Mangin, B., Chardon, F., Charcosset, A., and Joets, J. (2004). BioMercator: integrating genetic maps and QTL towards discovery of candidate genes. *Bioinformatics*, 20(14):2324–2326.
- Aronesty, E. (2011). *ea-utils: Command-line tools for processing biological sequencing data*.

- Arruda, M. P., Lipka, A. E., Brown, P. J., Krill, A. M., Thurber, C., Brown-Guedira, G., Dong, Y., Foresman, B. J., and Kolb, F. L. (2016). Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (*Triticum aestivum* L.). *Molecular Breeding*, 36(7):84.
- Ashbrook, D. G., Williams, R. W., Lu, L., and Hager, R. (2015). A cross-species genetic analysis identifies candidate genes for mouse anxiety and human bipolar disorder. *Frontiers in Behavioral Neuroscience*, 9.
- Ashton, D. T. (2013). Population genetics of New Zealand *Pagrus auratus* and genetic variation of an aquaculture broodstock.
- Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of Experimental Medicine*, 79(2):137–158.
- Ayllon, F., Kjærner-Semb, E., Furmanek, T., Wennevik, V., Solberg, M. F., Dahle, G., Taranger, G. L., Glover, K. A., Almén, M. S., Rubin, C. J., Edvardsen, R. B., and Wargelius, A. (2015). The *vgl3* locus controls age at maturity in wild and domesticated Atlantic salmon (*Salmo salar* L.) males. *PLOS Genetics*, 11(11):e1005628.
- Baerwald, M. R., Meek, M. H., Stephens, M. R., Nagarajan, R. P., Goodbla, A. M., Tomalty, K. M. H., Thorgaard, G. H., May, B., and Nichols, K. M. (2016). Migration-related phenotypic divergence is associated with epigenetic modifications in rainbow trout. *Molecular Ecology*, 25(8):1785–1800.
- Ballara, S. and O’Driscoll, R. (2014). Catches, size, and age structure of the 2011-12 hoki fishery, and a summary of input data used for the 2013 stock assessment. *New Zealand Fisheries Assessment Report 2014/05*, pages 1–117.
- Barria, A., Christensen, K. A., Correa, K., Jedlicki, A., Lhorente, J. P., Davidson, W., and Yáñez, J. M. (2017). Genome-wide association study and genomic predictions for resistance against *Piscirickettsia salmonis* in Coho salmon (*Oncorhynchus kisutch*) using ddRAD sequencing. *bioRxiv*, page 124099.
- Battaglione, S. C. and Talbot, R. B. (1992). Induced spawning and larval rearing of snapper, *Pagrus auratus* (Pisces: Sparidae), from Australian waters. *New Zealand Journal of Marine and Freshwater Research*, 26(2):179–185.
- Benfey, P. N. and Mitchell-Olds, T. (2008). From genotype to phenotype: systems biology meets natural variation. *Science*, 320(5875):495.
- Bernal-Ramírez, J. H., Adcock, G. J., Hauser, L., Carvalho, G. R., and Smith, P. J. (2003). Temporal stability of genetic population structure in the New Zealand snapper, *Pagrus auratus*, and relationship to coastal currents. *Marine Biology*, 142(3):567–574.

- Bernatchez, L., Wellenreuther, M., Araneda, C., Ashton, D. T., Barth, J. M. I., Beacham, T. D., Maes, G. E., Martinsohn, J. T., Miller, K. M., Naish, K. A., Ovenden, J. R., Primmer, C. R., Young Suk, H., Therkildsen, N. O., and Withler, R. E. (2017). Harnessing the power of genomics to secure the future of seafood. *Trends in Ecology & Evolution*, 32(9):665–680.
- Berner, D., Moser, D., Roesti, M., Buescher, H., and Salzburger, W. (2014). Genetic architecture of skeletal evolution in European lake and stream stickleback. *Evolution*, 68(6):1792–1805.
- Besnier, F., Glover, K. A., Lien, S., Kent, M., Hansen, M. M., Shen, X., and Skaala, . (2015). Identification of quantitative genetic components of fitness variation in farmed, hybrid and native salmon in the wild. *Heredity*, 115:47.
- Bettembourg, M., Dardou, A., Audebert, A., Thomas, E., Frouin, J., Guiderdoni, E., Ahmadi, N., Perin, C., Dievart, A., and Courtois, B. (2017). Genome-wide association mapping for root cone angle in rice. *Rice*, 10:45.
- Bhatta, S., Iwai, T., Miura, C., Higuchi, M., Shimizu-Yamaguchi, S., Fukada, H., and Miura, T. (2012). Gonads directly regulate growth in teleosts. *Proceedings of the National Academy of Sciences*, 109(28):11408–11412.
- Blonk, R. J. W., Komen, J., Tenghe, A., Kamstra, A., and van Arendonk, J. A. M. (2010). Heritability of shape in common sole, *Solea solea*, estimated from image analysis data. *Aquaculture*, 307(1):6–11.
- Bostock, J., McAndrew, B., Richards, R., Jauncey, K., Telfer, T., Lorenzen, K., Little, D., Ross, L., Handisyde, N., Gatward, I., and Corner, R. (2010). Aquaculture: global status and trends. *Philosophical Transactions of the Royal Society Biological Sciences*, 365(1554):2897–2912.
- Boulton, K., Massault, C., Houston, R. D., de Koning, D. J., Haley, C. S., Bovenhuis, H., Batargias, C., Canario, A. V. M., Kotoulas, G., and Tsigenopoulos, C. S. (2011). QTL affecting morphometric traits and stress response in the gilthead seabream (*Sparus aurata*). *Aquaculture*, 319(1):58–66.
- Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186.
- Braasch, I., Peterson, S. M., Desvignes, T., McCluskey, B. M., Batzel, P., and Postlethwait, J. H. (2015). A new model army: Emerging fish models to study the genomics of vertebrate Evo-Devo. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 324(4):316–341.
- Braasch, I., Schartl, M., and Volff, J.-N. (2007). Evolution of pigment synthesis pathways by gene and genome duplication in fish. *BMC Evolutionary Biology*, 7:74.

- Broman, K. W., Wu, H., Sen, S., and Churchill, G. A. (2003). R/qt1: QTL mapping in experimental crosses. *Bioinformatics (Oxford, England)*, 19(7):889–890.
- Camara, M. D. and Symonds, J. E. (2014). Genetic improvement of New Zealand aquaculture species: programmes, progress and prospects. *New Zealand Journal of Marine and Freshwater Research*, 48(3):466–491.
- Castaño-Sánchez, C., Fuji, K., Ozaki, A., Hasegawa, O., Sakamoto, T., Morishima, K., Nakayama, I., Fujiwara, A., Masaoka, T., Okamoto, H., Hayashida, K., Tagami, M., Kawai, J., Hayashizaki, Y., and Okamoto, N. (2010). A second generation genetic linkage map of Japanese flounder (*Paralichthys olivaceus*). *BMC Genomics*, 11:554.
- Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22(11):3124–3140.
- Chavanne, H., Parati, K., Cambuli, C., Capoferri, R., Jiménez, C. A., and Galli, A. (2014). Microsatellites markers to depict the reproductive and genetic patterns of farmed gilthead seabream (*Sparus aurata*): illustration by a case study on mass spawning. *Aquaculture Research*, 45(4):577–590.
- Chen, H., Semagn, K., Iqbal, M., Moakhar, N. P., Haile, T., N'Diaye, A., Yang, R.-C., Hucl, P., Pozniak, C., and Spaner, D. (2017). Genome-wide association mapping of genomic regions associated with phenotypic traits in Canadian western spring wheat. *Molecular Breeding*, 37(11):141.
- Clarke, C. A., Mani, G. S., and Wynne, G. (2008). Evolution in reverse: clean air and the peppered moth. *Biological Journal of the Linnean Society*, 26(2):189–199.
- Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibé, B., Bouix, J., Caiment, F., Elsen, J.-M., Eychenne, F., Larzul, C., Laville, E., Meish, F., Milenkovic, D., Tobin, J., Charlier, C., and Georges, M. (2006). A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nature Genetics*, 38(7):813–818.
- Cnaani, A., Zilberman, N., Tinman, S., Hulata, G., and Ron, M. (2004). Genome-scan analysis for quantitative trait loci in an F2 tilapia hybrid. *Molecular Genetics and Genomics*, 272(2):162–172.
- Cohen-Zinder, M., Seroussi, E., Larkin, D. M., Loor, J. J., Wind, A. E.-v. d., Lee, J.-H., Drackley, J. K., Band, M. R., Hernandez, A. G., Shani, M., Lewin, H. A., Weller, J. I., and Ron, M. (2005). Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Research*, 15(7):936–944.
- Cortijo, S., Wardenaar, R., Colomé-Tatché, M., Gilly, A., Etcheverry, M., Labadie, K., Caillieux, E., Hospital, F., Aury, J.-M., Wincker, P., Roudier, F., Jansen, R. C., Colot,

- V., and Johannes, F. (2014). Mapping the epigenetic basis of complex traits. *Science*, 343(6175):1145–1148.
- Crick, F. H. C., Barnett, L., Brenner, S., and Watts-tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, 192:1227.
- Cros, D., Bocs, S., Riou, V., Ortega-Abboud, E., Tisné, S., Argout, X., Pomiès, V., Nodichao, L., Lubis, Z., Cochard, B., and Durand-Gasselin, T. (2017). Genomic pre-selection with genotyping-by-sequencing increases performance of commercial oil palm hybrid crosses. *BMC Genomics*, 18:839.
- Dahle, G., Jørstad, K. E., Rusaas, H. E., and Otterå, H. (2006). Genetic characteristics of broodstock collected from four Norwegian coastal cod (*Gadus morhua*) populations. *ICES Journal of Marine Science*, 63(2):209–215.
- Dahm, R. (2005). Friedrich Miescher and the discovery of DNA. *Developmental Biology*, 278(2):274–288.
- Darwin, C. and Wallace, A. (1858). On the tendency of species to form varieties; and on the perpetuation of Varieties and species by natural means of selection. *Journal of the Proceedings of the Linnean Society of London. Zoology*, 3(9):45–62.
- De-Santis, C. and Jerry, D. R. (2007). Candidate growth genes in finfish — Where should we be looking? *Aquaculture*, 272(1):22–38.
- Dekkers, J. C. (2012). Application of genomics tools to animal breeding. *Current Genomics*, 13(3):207–212.
- Desta, Z. A. and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, 19(9):592–601.
- D’Mello, S. A. N., Finlay, G. J., Baguley, B. C., and Askarian-Amiri, M. E. (2016). Signaling pathways in melanogenesis. *International Journal of Molecular Sciences*, 17(7).
- Dodds, K. G., McEwan, J. C., Brauning, R., Anderson, R. M., van Stijn, T. C., Kristjánsson, T., and Clarke, S. M. (2015). Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics*, 16:1047.
- Doolan, B. J., Allan, G. L., Booth, M. A., and Jones, P. L. (2008a). Effect of carotenoids and background colour on the skin pigmentation of Australian snapper *Pagrus auratus* (Bloch & Schneider, 1801). *Aquaculture Research*, 39(13):1423–1433.
- Doolan, B. J., Allan, G. L., Booth, M. A., and Jones, P. L. (2008b). Effect of carotenoids and background colour on the skin pigmentation of australian snapper *pagrus auratus* (bloch & schneider, 1801). *Aquaculture Research*, 39(13):1423–1433.

- Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450.
- Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, 29(1):51–63.
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. *PLOS ONE*, 6(5):e19379.
- Everett, M. V. and Seeb, J. E. (2014). Detection and mapping of QTL for temperature tolerance and body size in Chinook salmon (*Oncorhynchus tshawytscha*) using genotyping by sequencing. *Evolutionary Applications*, 7(4):480–492.
- Falconer, D. and Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. Longman, Burnt Mill, England, 4th edition.
- Fang, Z., Polacco, M., Chen, S., Schroeder, S., Hancock, D., Sanchez, H., and Coe, E. (2003). cMap: the comparative genetic map viewer. *Bioinformatics (Oxford, England)*, 19(3):416–417.
- FAO, editor (2016). *Contributing to food security and nutrition for all*. Number 2016 in The state of world fisheries and aquaculture. Rome. OCLC: 957090850.
- Fernandes, T., Herlin, M., Dolores López Belluga, M., Ballón, G., Martinez, P., Toro, M., and Fernández, J. (2016). Estimation of genetic parameters for growth traits in a hatchery population of gilthead sea bream (*Sparus aurata* L.). *Aquaculture International*, 25.
- Fierst, J. L. (2015). Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Frontiers in Genetics*, 6.
- Flores-Mara, R., Rodríguez, F. H., Banger, R., Lhorente, J. P., Neira, R., Newman, S., and Yáñez, J. M. (2017). Resistance against infectious pancreatic necrosis exhibits significant genetic variation and is not genetically correlated with harvest weight in rainbow trout (*Oncorhynchus mykiss*). *Aquaculture*, 479:155–160.
- Foscarini, R. (1988). A review: Intensive farming procedure for red sea bream (*Pagrus major*) in Japan. *Aquaculture*, 72(3):191–246.
- Frankham, R., Bradshaw, C. J., and Brook, B. W. (2014). Genetics in conservation management: Revised recommendations for the 50/500 rules, red list criteria and population viability analyses. *Biological Conservation*, 170:56 – 63.
- FRDC (2018). Snapper 2016.
- Froese, R. and Pauly, D. (2017). FishBase. *FishBase*.

- Frohnhöfer, H. G., Krauss, J., Maischein, H.-M., and Nüsslein-Volhard, C. (2013). Iridophores and their interactions with other chromatophores are required for stripe formation in zebrafish. *Development*, 140(14):2997–3007.
- Gagnaire, P.-A., Normandeau, E., Pavey, S. A., and Bernatchez, L. (2013). Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (*Coregonus clupeaformis*). *Molecular Ecology*, 22(11):3036–3048.
- Galesloot, T. E., van Steen, K., Kiemeny, L. A. L. M., Janss, L. L., and Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. *PLOS ONE*, 9(4):e95923.
- Gentry, R. R., Froehlich, H. E., Grimm, D., Kareiva, P., Parke, M., Rust, M., Gaines, S. D., and Halpern, B. S. (2017). Mapping the global potential for marine aquaculture. *Nature Ecology & Evolution*, 1(9):1317–1324.
- Gibson, G. (2011). Rare and common variants: twenty arguments. *Nature reviews. Genetics*, 13(2):135–145.
- Gilmour, A., Gogel, B., Cullis, B., and Thompson, R. (2009). *ASReml User Guide*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK, 3 edition.
- Gjedrem, T., Robinson, N., and Rye, M. (2012). The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. *Aquaculture*, 350-353:117–129.
- Gjedrem, T. and Rye, M. (2016). Selection response in fish and shellfish: a review. *Reviews in Aquaculture*, pages n/a–n/a.
- Goddard, M. E. and Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, 124(6):323–330.
- Goddard, M. E. and Whitelaw, E. (2014). The use of epigenetic phenomena for the improvement of sheep and cattle. *Frontiers in Genetics*, 5.
- Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11:R86.
- Gonzalez, E. B., Aritaki, M., Knutsen, H., and Taniguchi, N. (2015). Effects of large-scale releases on the genetic structure of red sea bream (*Pagrus major*, Temminck et Schlegel) populations in Japan. *PLOS ONE*, 10(5):e0125743.
- Greenwood, A. K., Jones, F. C., Chan, Y. F., Brady, S. D., Absher, D. M., Grimwood, J., Schmutz, J., Myers, R. M., Kingsley, D. M., and Peichel, C. L. (2011). The genetic basis of divergent pigment patterns in juvenile threespine sticklebacks. *Heredity*, 107(2):155.

- Grigorenko, E. L. (2005). The inherent complexities of gene–environment interactions. *The Journals of Gerontology: Series B*, 60(Special_Issue_1):53–64.
- Grover, A. and Sharma, P. C. (2016). Development and use of molecular markers: past and present. *Critical Reviews in Biotechnology*, 36(2):290–302.
- Gutierrez, A. P., Yáñez, J. M., Fukui, S., Swift, B., and Davidson, W. S. (2015). Genome-Wide Association Study (GWAS) for growth rate and age at sexual maturation in Atlantic salmon (*Salmo salar*). *PLOS ONE*, 10(3):e0119730.
- Handeland, S. O., Imsland, A. K., and Stefansson, S. O. (2008). The effect of temperature and fish size on growth, feed intake, food conversion efficiency and stomach evacuation rate of Atlantic salmon post-smolts. *Aquaculture*, 283(1):36–42.
- Hauser, L., Adcock, G. J., Smith, P. J., Bernal Ramírez, J. H., and Carvalho, G. R. (2002). Loss of microsatellite diversity and low effective population size in an overexploited population of New Zealand snapper (*Pagrus auratus*). *Proceedings of the National Academy of Sciences of the United States of America*, 99(18):11742–11747.
- He, C., Weeks, D. E., Buyske, S., Abecasis, G. R., Stewart, W. C., and Matise, T. C. (2011). Enhanced genetic maps from family-based disease studies: population-specific comparisons. *BMC Medical Genetics*, 12:15.
- Hely, F. S., Amer, P. R., Walker, S. P., and Symonds, J. E. (2013). Optimised parent selection and minimum inbreeding mating in small aquaculture breeding schemes: a simulation study. *animal*, 7(1):1–10.
- Helyar, S. J., Lloyd, H. a. D., Bruyn, M. d., Leake, J., Bennett, N., and Carvalho, G. R. (2014). Fish product mislabelling: failings of traceability in the production chain and implications for illegal, unreported and unregulated (IUU) fishing. *PLOS ONE*, 9(6):e98691.
- Hemmer-Hansen, J., Therkildsen, N. O., and Pujolar, J. M. (2014). Population genomics of marine fishes: next-generation prospects and challenges. *The Biological Bulletin*, 227(2):117–132.
- Henderson, C. (1984). Applications of linear models in animal breeding. *University of Guelph, Ontario*.
- Hilario, E. (2015). The restriction enzyme target approach to Genotyping by Sequencing (GBS). In *Plant Genotyping, Methods in Molecular Biology*, pages 271–279. Humana Press, New York, NY.
- Hoek, K. S., Schlegel, N. C., Eichhoff, O. M., Widmer, D. S., Praetorius, C., Einarsson, S. O., Valgeirsdottir, S., Bergsteinsdottir, K., Schepsky, A., Dummer, R., and Steingrimsdottir, E. (2008). Novel MITF targets identified using a two-step DNA microarray strategy. *Pigment Cell & Melanoma Research*, 21(6):665–676.

- Hollenbeck, C. M., Portnoy, D. S., and Gold, J. R. (2016). A method for detecting recent changes in contemporary effective population size from linkage disequilibrium at linked and unlinked loci. *Heredity*, 117(4):207–216.
- Hong, E. P. and Park, J. W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics & Informatics*, 10(2):117–122.
- Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the next challenge. *Nature Reviews Genetics*, 11:855.
- Houle, D., Mezey, J., Galpern, P., and Carter, A. (2003). Automated measurement of *Drosophila* wings. *BMC Evolutionary Biology*, 3(1):25.
- Houston, R. D., Bishop, S. C., Hamilton, A., Guy, D. R., Tinch, A. E., Taggart, J. B., Derayat, A., McAndrew, B. J., and Haley, C. S. (2009). Detection of QTL affecting harvest traits in a commercial Atlantic salmon population. *Animal Genetics*, 40(5):753–755.
- Hu, Z. and Xu, S. (2008). A simple method for calculating the statistical power for detecting a QTL located in a marker interval. *Heredity*, 101(1):48–52.
- Hubbard, J. K., Uy, J. A. C., Hauber, M. E., Hoekstra, H. E., and Safran, R. J. (2010). Vertebrate pigmentation: from underlying genes to adaptive function. *Trends in Genetics*, 26(5):231–239.
- Hulata, G. (2001). Genetic manipulations in aquaculture: a review of stock improvement by classical and modern technologies. *Genetica*, 111(1-3):155–173.
- Hutson, A. M., Liu, Z., Kucuktas, H., Umali-Maceina, G., Su, B., and Dunham, R. A. (2014). Quantitative trait loci map for growth and morphometric traits using a channel catfish $\times$ blue catfish interspecific hybrid system. *Journal of Animal Science*, 92(5):1850–1865.
- Huxley, J. (1943). *Evolution: the modern synthesis*, volume 1. Allen & Unwin, London, 1 edition.
- Ioannidis, J. P. A. and Kavvoura, F. K. (2006). Concordance of functional in vitro data and epidemiological associations in complex disease genetics. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 8(9):583–593.
- Jamieson, I. G. and Allendorf, F. W. (2012). How does the 50/500 rule apply to MVPs? *Trends in Ecology & Evolution*, 27(10):578–584.
- Jin, S., Zhang, X., Jia, Z., Fu, H., Zheng, X., and Sun, X. (2012). Genetic linkage mapping and genetic analysis of QTL related to eye cross and eye diameter in common carp (*Cyprinus carpio* L.) using microsatellites and SNPs. *Aquaculture*, 358-359:176–182.

- Johnson, S. L., Nguyen, A. N., and Lister, J. A. (2011). *mitfa* is required at multiple stages of melanocyte differentiation but not to establish the melanocyte stem cell. *Developmental Biology*, 350(2):405–413.
- Kahraman, A., Avramov, A., Nashev, L. G., Popov, D., Ternes, R., Pohlenz, H.-D., and Weiss, B. (2005). PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. *Bioinformatics*, 21(3):418–420.
- Kalinowski, S. T., Taper, M. L., and Marshall, T. C. (2007). Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, 16(5):1099–1106.
- Kardos, M., Luikart, G., and Allendorf, F. W. (2015). Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity*, 115(1):63–72.
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649.
- Keller, L. F. and Waller, D. M. (2002). Inbreeding effects in wild populations. *Trends in Ecology & Evolution*, 17(5):230–241.
- Kelly, K., J. (2011). The breeder’s equation. 4(5):5.
- Kirschner, J., Weber, D., Neuschl, C., Franke, A., Böttger, M., Zielke, L., Powalsky, E., Groth, M., Shagin, D., Petzold, A., Hartmann, N., Englert, C., Brockmann, G. A., Platzer, M., Cellerino, A., and Reichwald, K. (2012). Mapping of quantitative trait loci controlling lifespan in the short-lived fish *Nothobranchius furzeri* – a new vertebrate model for age research. *Aging Cell*, 11(2):252–261.
- Knott, S. A., Elsen, J. M., and Haley, C. S. (1996). Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theoretical and Applied Genetics*, 93(1-2):71–80.
- Ko, W.-Y., David, R. M., and Akashi, H. (2003). Molecular phylogeny of the *Drosophila melanogaster* species subgroup. *Journal of Molecular Evolution*, 57(5):562–573.
- Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, 44(9):1066–1071.
- Kottler, V. A., Koch, I., Flötenmeyer, M., Hashimoto, H., Weigel, D., and Dreyer, C. (2014). Multiple pigment cell types contribute to the black, blue, and orange ornaments of male Guppies (*Poecilia reticulata*). *PLoS ONE*, 9(1).

- Kristjánsson, T. and Arnason, T. (2016). Heritability of economically important traits in the Atlantic cod *Gadus morhua* L. *Aquaculture Research*, 47(2):349–356.
- Kucuktas, H., Wang, S., Li, P., He, C., Xu, P., Sha, Z., Liu, H., Jiang, Y., Baoprasertkul, P., Somridhivej, B., Wang, Y., Abernathy, J., Guo, X., Liu, L., Muir, W., and Liu, Z. (2009). Construction of genetic linkage maps and comparative genome analysis of Catfish using gene-associated markers. *Genetics*, 181(4):1649–1660.
- Laikre, L., Palm, S., and Ryman, N. (2005). Genetic population structure of fishes: implications for coastal zone management. *Ambio*, 34(2):111–119.
- Le Bras, Y., Dechamp, N., Krieg, F., Filangi, O., Guyomard, R., Boussaha, M., Bovenhuis, H., Pottinger, T. G., Prunet, P., Le Roy, P., and Quillet, E. (2011). Detection of QTL with effects on osmoregulation capacities in the rainbow trout (*Oncorhynchus mykiss*). *BMC Genetics*, 12:46.
- Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nature Reviews Genetics*, 14(3):168.
- Li, Y., Liu, S., Qin, Z., Waldbieser, G., Wang, R., Sun, L., Bao, L., Danzmann, R. G., Dunham, R., and Liu, Z. (2015). Construction of a high-density, high-resolution genetic map and its integration with BAC-based physical map in channel catfish. *DNA Research*, 22(1):39–52.
- Li, Y. and Tollefsbol, T. O. (2011). DNA methylation detection: bisulfite genomic sequencing analysis. In *Epigenetics Protocols*, Methods in Molecular Biology, pages 11–21. Humana Press.
- Liu, F., Yang, Y., Li, Y., Guo, H., Dai, H., Gao, J., Zhang, Y., Ma, J., Liu, Y., Wei, Z., Shao, C., and Chen, S. (2016). Phenotypic and genetic parameter estimation of juvenile growth and bottom color traits in half-smooth tongue sole, *Cynoglossus semilaevis*. *Acta Oceanologica Sinica*, 35(10):83–87.
- Liu, P., Xia, J. H., Lin, G., Sun, F., Liu, F., Lim, H. S., Pang, H. Y., and Yue, G. H. (2012). Molecular parentage analysis is essential in breeding Asian seabass. *PLoS ONE*, 7(12).
- Lopes, S. S., Yang, X., Müller, J., Carney, T. J., McAdow, A. R., Rauch, G.-J., Jacoby, A. S., Hurst, L. D., Delfino-Machín, M., Haffter, P., Geisler, R., Johnson, S. L., Ward, A., and Kelsh, R. N. (2008). Leukocyte tyrosine kinase functions in pigment cell development. *PLOS Genetics*, 4(3):e1000026.
- Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. *PLOS Computational Biology*, 13(5):e1005457.
- Lynch, M. and Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sinauer Sunderland, MA.

- Maan, M. E. and Sefc, K. M. (2013). Colour variation in cichlid fish: developmental mechanisms, selective pressures and evolutionary consequences. *Seminars in Cell & Developmental Biology*, 24(6):516–528.
- Mackay, T. F. C. (2001). The genetic architecture of quantitative traits. *Annual Review of Genetics*, 35(1):303–339.
- Mackay, T. F. C., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10(8):565–577.
- Manichaikul, A., Dupuis, J., Sen, a., and Broman, K. W. (2006). Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. *Genetics*, 174(1):481–489.
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.
- Massault, C., Bovenhuis, H., Haley, C., and de Koning, D.-J. (2008). QTL mapping designs for aquaculture. *Aquaculture*, 285(1):23–29.
- McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369.
- McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in science conference*, pages 51–56.
- Mendel, G. (1866). Versuche über Pflanzen-Hybriden.[Experiments on Plant Hybrids] Verhandlungen des naturforschenden Vereines in Brunn, der naturfoschung Vereins. 4: 3–47.
- Meuwissen, T. and Goddard, M. (1996). The use of marker haplotypes in animal breeding schemes. *Genetics, Selection, Evolution : GSE*, 28(2):161–176.
- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.
- MPI (2013). Review of sustainability and other management controls for snapper 1 (SNA 1). *MPI Discussion Paper*.
- MPI (2015). Evaluation of tagging programme designs for SNA 1 and SNA 8. *New Zealand Fisheries Assessment Report 2015/35*.
- MPI (2016). Fisheries assessment plenary may 2016: stock assessments and stock status. *MPI Fisheries Assessment Plenary*.

- Mrakovčić, M. and Haley, L. E. (1978). Inbreeding depression in the zebra fish brachydanio rerio (hamilton buchanan). *Journal of Fish Biology*, 15(3):323–327.
- Mueller, A. J., Canty-Laird, E. G., Clegg, P. D., and Tew, S. R. (2017). Cross-species gene modules emerge from a systems biology approach to osteoarthritis. *npj Systems Biology and Applications*, 3(1):13.
- Mundy, N. I. (2009). Conservation and convergence of colour genetics: MC1r mutations in brown cavefish. *PLoS Genetics*, 5(2):e1000388.
- Murata, O., Harada, T., Miyashita, S., Izumi, K.-I., Maeda, S., Kato, K., and Kumai, H. (1996). Selective breeding for growth in red sea bream. *Fisheries Science*, 62:845–849.
- Navarro, A., Lee-Montero, I., Santana, D., Henríquez, P., Ferrer, M. A., Morales, A., Soula, M., Badilla, R., Negrín-Báez, D., Zamorano, M. J., and Afonso, J. M. (2016). IMAFISH_ml: A fully-automated image analysis software for assessing fish morphometric traits on gilthead seabream (*Sparus aurata* L.), meagre (*Argyrosomus regius*) and red porgy (*Pagrus pagrus*). *Computers and Electronics in Agriculture*, 121:66–73.
- NCBI (2018). GenBank and WGS Statistics.
- Nekrutenko, A. and Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, 13(9):667–672.
- Ninwichian, P., Peatman, E., Liu, H., Kucuktas, H., Somridhivej, B., Liu, S., Li, P., Jiang, Y., Sha, Z., Kaltenboeck, L., Abernathy, J. W., Wang, W., Chen, F., Lee, Y., Wong, L., Wang, S., Lu, J., and Liu, Z. (2012). Second-generation genetic linkage map of catfish and its integration with the BAC-based physical map. *G3: Genes/Genomes/Genetics*, 2(10):1233–1241.
- O’Leary, S. J., Hice, L. A., Feldheim, K. A., Frisk, M. G., McElroy, A. E., Fast, M. D., and Chapman, D. D. (2013). Severe inbreeding and small effective number of breeders in a formerly abundant marine fish. *PLOS ONE*, 8(6):e66126.
- Omics (2018). Omic tools database.
- Orrell, T. M. and Carpenter, K. E. (2004). A phylogeny of the fish family Sparidae (porgies) inferred from mitochondrial sequence data. *Molecular Phylogenetics and Evolution*, 32(2):425–434.
- Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nature Reviews Genetics*, 12(7):465–474.
- Parichy, D. M., Mellgren, E. M., Rawls, J. F., Lopes, S. S., Kelsh, R. N., and Johnson, S. L. (2000). Mutational analysis of endothelin receptor b1 (rose) during neural crest and pigment pattern development in the zebrafish *Danio rerio*. *Developmental Biology*, 227(2):294–306.

- Park, I.-S., Gil, H. W., Lee, T. H., Nam, Y. K., and Kim, D. S. (2016). Comparative study of growth and gonad maturation in diploid and triploid marine medaka, *Oryzias dancena*. *Development & Reproduction*, 20(4):305–314.
- Parsons, D. M., Sim-Smith, C. J., Cryer, M., Francis, M. P., Hartill, B., Jones, E. G., Port, A. L., Lowe, M., McKenzie, J., Morrison, M., Paul, L. J., Radford, C., Ross, P. M., Spong, K. T., Trnski, T., Usmar, N., Walsh, C., and Zeldis, J. (2014). Snapper (*Chrysophrys auratus*): a review of life history and key vulnerabilities in New Zealand. *New Zealand Journal of Marine and Freshwater Research*, 48(2):256–283.
- Paul, L. J. and Tarring, S. C. (1980). Growth rate and population structure of snapper, *Chrysophrys auratus*, in the East Cape region, New Zealand. *New Zealand Journal of Marine and Freshwater Research*, 14(3):237–247.
- Paulin, C. D. (1990). *Pagrus auratus*, a new combination for the species known as “snapper” in Australasian waters (Pisces: Sparidae). *New Zealand Journal of Marine and Freshwater Research*, 24(2):259–265.
- Pilgrim, B. L., Perry, R. C., Keefe, D. G., Perry, E. A., and Dawn Marshall, H. (2012). Microsatellite variation and genetic structure of brook trout (*Salvelinus fontinalis*) populations in Labrador and neighboring Atlantic Canada: evidence for ongoing gene flow and dual routes of post-Wisconsinan colonization. *Ecology and Evolution*, 2(5):885–898.
- Pértille, F., Moreira, G. C. M., Zanella, R., Nunes, J. d. R. d. S., Boschiero, C., Rovadoski, G. A., Mourão, G. B., Ledur, M. C., and Coutinho, L. L. (2017). Genome-wide association study for performance traits in chickens using genotype by sequencing approach. *Scientific Reports*, 7.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575.
- R Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.
- Rastas, P., Calboli, F. C. F., Guo, B., Shikano, T., and Merilä, J. (2016). Construction of ultradense linkage maps with Lep-MAP2: stickleback F2 recombinant crosses as an example. *Genome Biology and Evolution*, 8(1):78–93.
- Reiter, L. T., Potocki, L., Chien, S., Gribskov, M., and Bier, E. (2001). A systematic analysis of human disease-associated gene sequences in *Drosophila melanogaster*. *Genome Research*, 11(6):1114–1125.
- Rezende, F. M., Ferraz, J. B. S., Eler, J. P., Silva, R. C. G., Mattos, E. C., and Ibáñez-Escriche, N. (2012). Study of using marker assisted selection on a beef cattle breeding program by model comparison. *Livestock Science*, 147(1):40–48.

- Riddell, N. and Crewther, S. G. (2017). Integrated comparison of GWAS, transcriptome, and proteomics studies highlights similarities in the biological basis of animal and human myopia. *Investigative Ophthalmology & Visual Science*, 58(1):660–669.
- Risch, N. and Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science*, 268(5217):1584–1589.
- Rockman, M. V. (2011). The qtn program and the alleles that matter for evolution: all that’s gold does not glitter. *Evolution*, 66(1):1–17.
- Rockman, M. V. and Kruglyak, L. (2008). Breeding designs for recombinant inbred advanced intercross lines. *Genetics*, 179(2):1069–1078.
- Roesti, M., Moser, D., and Berner, D. (2013). Recombination in the threespine stickleback genome—patterns and consequences. *Molecular Ecology*, 22(11):3014–3027.
- Rutkoski, J. (2015). Genetic gain from phenotypic and genomic selection for quantitative resistance to stem rust of wheat. *The plant genome*, 8(2).
- Salles, O. C., Pujol, B., Maynard, J. A., Almany, G. R., Berumen, M. L., Jones, G. P., Saenz-Agudelo, P., Srinivasan, M., Thorrold, S. R., and Planes, S. (2016). First genealogy for a wild marine fish population reveals multigenerational philopatry. *Proceedings of the National Academy of Sciences of the United States of America*, 113(46):13245–13250.
- Samorè, A. B. and Fontanesi, L. (2016). Genomic selection in pigs: state of the art and perspectives. *Italian Journal of Animal Science*, 15(2):211–232.
- Sandoval-Castillo Jonathan, Robinson Nick A., Hart Anthony M., Strain Lachlan W. S., and Beheregaray Luciano B. (2018). Seascape genomics reveals adaptive divergence in a connected and commercially important mollusc, the greenlip abalone (*Haliotis laevis*), along a longitudinal environmental gradient. *Molecular Ecology*, 0(0).
- Sarropoulou, E., Nousdili, D., Magoulas, A., and Kotoulas, G. (2008). Linking the genomes of nonmodel teleosts through comparative genomics. *Marine Biotechnology*, 10(3):227.
- Sauvage, C., Vagner, M., Derôme, N., Audet, C., and Bernatchez, L. (2012). Coding gene SNP mapping reveals QTL linked to growth and stress response in brook charr (*Salvelinus fontinalis*). *G3: Genes/Genomes/Genetics*, 2(6):707–720.
- Savolainen, O., Lascoux, M., and Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews Genetics*, 14(11):807–820.
- Schaefer, W. F., Schmitz, M. H., Blazer, V. S., Ehlinger, T. J., and Berges, J. A. (2014). Localization and seasonal variation of blue pigment (sander cyanin) in walleye (*Sander vitreus*). *Canadian Journal of Fisheries and Aquatic Sciences*, 72(2):281–289.

- Seaton, G., Hernandez-Sanchez, J., Grunchev, J.-A., White, I., Allen, J., Koning, D. D., Wei, W., Berry, D., Haley, C., and Knott, S. (2006). GridQTL: A grid portal for QTL mapping of compute intensive datasets. In *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production (Belo horizonte, Brazil)*.
- Sekino, M., Saitoh, K., Yamada, T., Kumagai, A., Hara, M., and Yamashita, Y. (2003). Microsatellite-based pedigree tracing in a Japanese flounder *Paralichthys olivaceus* hatchery strain: implications for hatchery management related to stock enhancement program. *Aquaculture*, 221(1):255–263.
- Shakhova, O., Cheng, P., Mishra, P. J., Zingg, D., Schaefer, S. M., Debbache, J., Häusel, J., Matter, C., Guo, T., Davis, S., Meltzer, P., Mihic-Probst, D., Moch, H., Wegner, M., Merlino, G., Levesque, M. P., Dummer, R., Santoro, R., Cinelli, P., and Sommer, L. (2015). Antagonistic cross-regulation between Sox9 and Sox10 controls an anti-tumorigenic program in melanoma. *PLOS Genetics*, 11(1):e1004877.
- Sham, P. C. and Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346.
- Shoemaker, C. A., Lozano, C. A., LaFrentz, B. R., García, J. C., Soto, E., Xu, D.-H., Beck, B. H., and Rye, M. (2017). Additive genetic variation in resistance of Nile tilapia (*Oreochromis niloticus*) to *Streptococcus iniae* and *S. agalactiae* capsular type Ib: Is genetic resistance correlated? *Aquaculture*, 468:193–198.
- Silva, P. F., McGurk, C., Thompson, K. D., Jayasuriya, N. S., and Bron, J. E. (2015). Development of a quantitative semi-automated system for intestinal morphology assessment in Atlantic salmon, using image analysis. *Aquaculture*, 442:100–111.
- Slate, J. (2005). Invited review: quantitative trait locus mapping in natural populations: progress, caveats and future directions. *Molecular Ecology*, 14(2):363–379.
- Slate, J. (2013). From beavis to beak color: A simulation study to examine how much QTL mapping can reveal about the genetic architecture of quantitative traits. *Evolution*, 67(5):1251–1262.
- Smith, P. J., Francis, R. I. C. C., and Paul, L. J. (1978). Genetic variation and population structure in the New Zealand snapper. *New Zealand Journal of Marine and Freshwater Research*, 12(4):343–350.
- Song, K., Li, L., and Zhang, G. (2016). Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. *Scientific Reports*, 6:35736.
- Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., and Robinson, G. E. (2015). Big data: astronomical or genetical? *PLOS Biology*, 13(7):e1002195.

- Stranger, B. E., Stahl, E. A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383.
- Sutherland, B. J. G., Gosselin, T., Normandeau, E., Lamothe, M., Isabel, N., Audet, C., and Bernatchez, L. (2016). Salmonid chromosome evolution as revealed by a novel method for comparing RADseq linkage maps. *Genome Biology and Evolution*, 8(12):3600–3617.
- Suzuki, M., Kimura, T., Ogawa, H., Hotta, K., and Oka, K. (2011). Chromatophore activity during natural pattern expression by the squid *Sepioteuthis lessoniana*: contributions of miniature oscillation. *PLOS ONE*, 6(4):e18244.
- te Pas, M. F., Soumilion, A., Harders, F. L., Verburg, F. J., van den Bosch, T. J., Galesloot, P., and Meuwissen, T. H. (1999). Influences of myogenin genotypes on birth weight, growth rate, carcass weight, backfat thickness, and lean weight of pigs. *Journal of Animal Science*, 77(9):2352–2356.
- Theodosiou, L., McMillan, W. O., and Puebla, O. (2016). Recombination in the eggs and sperm in a simultaneously hermaphroditic vertebrate. *Proceedings of the Royal Society B: Biological Sciences*, 283(1844).
- Thorland, I., Papaioannou, N., and Kottaras, L. (2006). The Kego breeding programs for sea bream (*Sparus aurata*) and sea bass (*Dicentrarchus labrax*) in Greece.
- Tong, J. and Sun, X. (2015). Genetic and genomic analyses for economically important traits and their applications in molecular breeding of cultured fish. *Science China Life Sciences*, 58(2):178–186.
- Tsai, H. Y., Hamilton, A., Guy, D. R., Tinch, A. E., Bishop, S. C., and Houston, R. D. (2015). The genetic architecture of growth and fillet traits in farmed Atlantic salmon (*Salmo salar*). *BMC Genetics*, 16:51.
- Tsai, H. Y., Robledo, D., Lowe, N. R., Bekaert, M., Taggart, J. B., Bron, J. E., and Houston, R. D. (2016). Construction and annotation of a high density SNP linkage map of the Atlantic salmon (*Salmo salar*) genome. *G3: Genes/Genomes/Genetics*, 6(7):2173–2179.
- Valenzuela-Quinonez, F. (2016). How fisheries management can benefit from genomics? *Briefings in Functional Genomics*, 15(5):352–357.
- Vallejo, R. L., Leeds, T. D., Gao, G., Parsons, J. E., Martin, K. E., Evenhuis, J. P., Fragomeni, B. O., Wiens, G. D., and Palti, Y. (2017). Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree-based model in rainbow trout aquaculture. *Genetics Selection Evolution*, 49:17.
- van Gurp, T. (2011). Deena Bioinformatics.

- Velando, A., Beamonte-Barrientos, R., and Torres, R. (2006). Pigment-based skin colour in the blue-footed booby: an honest signal of current condition used by females to adjust reproductive investment. *Oecologia*, 149(3):535–542.
- Viazzi, S., Van Hoestenbergh, S., Goddeeris, B. M., and Berckmans, D. (2015). Automatic mass estimation of Jade perch *Scortum barcoo* by computer vision. *Aquacultural Engineering*, 64:42–48.
- Vij, S., Kuhl, H., Kuznetsova, I. S., Komissarov, A., Yurchenko, A. A., Van Heusden, P., Singh, S., Thevasagayam, N. M., Prakki, S. R. S., Purushothaman, K., Saju, J. M., Jiang, J., Mbandi, S. K., Jonas, M., Hin Yan Tong, A., Mwangi, S., Lau, D., Ngoh, S. Y., Liew, W. C., Shen, X., Hon, L. S., Drake, J. P., Boitano, M., Hall, R., Chin, C.-S., Lachumanan, R., Korlach, J., Trifonov, V., Kabilov, M., Tupikin, A., Green, D., Moxon, S., Garvin, T., Sedlazeck, F. J., Vulture, G. W., Gopalapillai, G., Kumar Katneni, V., Noble, T. H., Scaria, V., Sivasubbu, S., Jerry, D. R., O'Brien, S. J., Schatz, M. C., Dalmay, T., Turner, S. W., Lok, S., Christoffels, A., and Orbán, L. (2016). Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. *PLOS Genetics*, 12(4):e1005954.
- Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, 90(1):7–24.
- Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Reviews. Genetics*, 9(4):255–266.
- Visscher, P. M., Thompson, R., and Haley, C. S. (1996). Confidence intervals in QTL mapping by bootstrapping. *Genetics*, 143(2):1013–1020.
- Wan, S., Li, Q., Liu, T., Yu, H., and Kong, L. (2017). Heritability estimates for shell color-related traits in the golden shell strain of Pacific oyster (*Crassostrea gigas*) using a molecular pedigree. *Aquaculture*, 476:65–71.
- Wang, C. (2009). Quantitative genetic estimates of growth-related traits in the common carp (*Cyprinus carpio* L.): A review. *Frontiers of Biology in China*, 4(3):298–304.
- Wang, L., Bai, B., Liu, P., Huang, S. Q., Wan, Z. Y., Chua, E., Ye, B., and Yue, G. H. (2017). Construction of high-resolution recombination maps in Asian seabass. *BMC Genomics*, 18.
- Wang, S., Hard, J. J., and Utter, F. (2002). Salmonid inbreeding: a review. *Reviews in Fish Biology and Fisheries*, 11(4):301–319.
- Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356):737–738.

- Weight, C., Parnham, D., and Waites, R. (2007). Technical advance: LeafAnalyser: a computational method for rapid and large-scale analyses of leaf shape variation. *The Plant Journal*, 53(3):578–586.
- Wellenreuther, M. and Hansson, B. (2016). Detecting polygenic evolution: problems, pitfalls, and promises. *Trends in Genetics*, 32(3):155–164.
- Wenne, R., Boudry, P., Hemmer-Hansen, J., Lubieniecki, K. P., Was, A., and Kause, A. (2007). What role for genomics in fisheries management and aquaculture? *Aquatic Living Resources*, 20(3):241–255.
- Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Use R! Springer-Verlag, New York.
- Wikipedia (2018). Australasian snapper. Page Version ID: 826047433.
- Wray, N. and Visscher, P. M. (2008). Estimating trait heritability. *Nature Education*, 1:1–1.
- Xu, Y. X., Zhu, Z. Y., Lo, L. C., Wang, C. M., Lin, G., Feng, F., and Yue, G. H. (2006). Characterization of two parvalbumin genes and their association with growth traits in asian seabass (*Lates calcarifer*). *Animal Genetics*, 37(3):266–268.
- Yabe, S., Hara, T., Ueno, M., Enoki, H., Kimura, T., Nishimura, S., Yasui, Y., Ohsawa, R., and Iwata, H. (2018). Potential of genomic selection in mass selection breeding of an allogamous crop: an empirical study to increase yield of common buckwheat. *Frontiers in Plant Science*, 9:276.
- Yandell, B. S., Mehta, T., Banerjee, S., Shriner, D., Venkataraman, R., Moon, J. Y., Neely, W. W., Wu, H., von Smith, R., and Yi, N. (2007). R/qtlbim: QTL with bayesian interval mapping in experimental crosses. *Bioinformatics (Oxford, England)*, 23(5):641–643.
- Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., Robinson, M. R., Perry, J. R. B., Nolte, I. M., van Vliet-Ostaptchouk, J. V., Snieder, H., The LifeLines Cohort Study, Esko, T., Milani, L., Mägi, R., Metspalu, A., Hamsten, A., Magnusson, P. K. E., Pedersen, N. L., Ingelsson, E., Soranzo, N., Keller, M. C., Wray, N. R., Goddard, M. E., and Visscher, P. M. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, 47:1114.
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of heritability for human height. *Nature genetics*, 42(7):565–569.
- Ye, B., Wan, Z., Wang, L., Pang, H., Wen, Y., Liu, H., Liang, B., Lim, H. S., Jiang, J., and Yue, G. (2017). Heritability of growth traits in the Asian seabass (*Lates calcarifer*). *Aquaculture and Fisheries*, 2(3):112–118.

- Zeinab, Z., Shabany, A., and Kolangi-Miandare, H. (2014). Comparison of genetic variation of wild and farmed Bream (*Abramis brama orientalis*; berg, 1905) using microsatellite markers. *Molecular Biology Research Communications*, 3(3):187–195.
- Zhang, C. and Hsieh, a. T.-F. (2013). Heritable epigenetic variation and its potential applications for crop improvement. *Plant Breeding and Biotechnology*, 1(4):307–319.
- Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824.