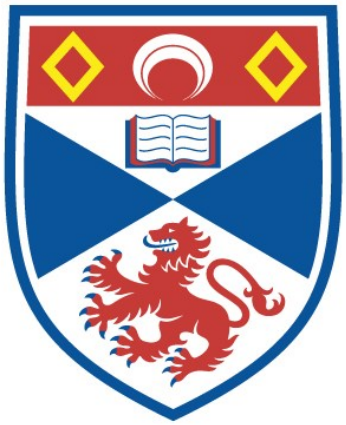
THE DEVELOPMENT OF MOLECULAR TOOLS AND RESOURCES FOR SELECTIVE BREEDING IN AQUACULTURE

Luke Earl Holman

A Thesis Submitted for the Degree of MPhil
at the
University of St Andrews



2017

The development of molecular tools and resources for selective breeding in aquaculture

Luke Earl Holman

University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of MPhil
at the
University of St Andrews

Submission: 13 January 2017

1. Candidate's declarations:

I, Luke E. Holman, hereby certify that this thesis, which is approximately 23,200 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student and candidate for the degree of Master of Philosophy in March 2015; the higher study for which this is a record was carried out in the University of St Andrews between 2015 and 2016.

Date 09.12.2016 Luke E. Holman

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Master of Philosophy in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date 09.12.2016 Ian A. Johnston

3. Permission for publication:

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

PRINTED COPY
Embargo on all or part of print copy for a period of 1 year on the following ground: Publication would preclude future publication.

ELECTRONIC COPY
Embargo on all or part of electronic copy for a period of 1 year on the following ground: Publication would preclude future publication

ABSTRACT AND TITLE EMBARGOES

I agree to the title and abstract being published

Date 09.12.2016
Luke E. Holman …… Ian A. Johnston ……

Acknowledgements

This work is the product of the most hectic part of my life to date. I have split my time between my MPhil and Xelect, something I would have found impossible if not for the supportive and encouraging people I have had the privilege of working with these last two years.
I would first like to thank my supervisor Ian Johnston whose original suggestion of this MPhil was a very pleasant surprise. Ian has been a consistent and encouraging supervisor and has always found time to offer his insight on whatever problem I have brought to him. Ian's positive influence on my work will be obvious to those who have shared his guidance, and I wish him all the best in his retirement from Professor to CEO.
Throughout my time at Xelect and the progress of this thesis I have been greatly influenced by Tom Ashton. I will miss his infectious enthusiasm and curiosity. Tom has endured my long lunch break snorkels, far flung holidays and expensive lab mistakes, for which I am immensely grateful.
Aubrie Onoufriou has been a constant friend, she is thoughtful and caring. She has kept the wheels turning long after I would have abandoned my post and can *always* find the thing I have lost in the lab.
Throughout my MPhil research Dani Garcia and Clara Coll have provided support and insight on challenging stats, SNPs and buffers. I will miss them both.
This journey would not be possible if not for the constant and steady support of my parents Cathy and Chris Holman.
Finally, my partner Charlotte who has shared so much of my time with this MPhil has my immeasurable gratitude for her patience and understanding. Weekends in wild places at her side permeate my memories of this work.

Thank you all.

Table of Contents

Abstract

Human population growth is predicted to continue well into the 21st century, and beyond. The provision of selectively bred organisms will be an essential part of global food security. While the selective breeding of terrestrial animals has been essential to the human success story, the breeding of aquatic organisms has only recently received serious attention. Aquaculture research urgently needs both specific genetic resources for existing aquatic species, and generalised workflows and pipelines for the generation of resources for newly cultivated species. This study presents a stochastic simulation of a selective program for the gilthead seabream (*Sparus aurata* L.). The simulation models the change of a non-selective breeding program to a scheme improving growth rate by mass selection. The effect of selection on growth rate, inbreeding and projected profits are modelled explicitly. The simulation predicts a profitable and sound breeding scheme for gilthead seabream and can also be easily adapted for new traits and species. A workflow for the filtration of an optimal number genetic variants for molecular parentage assignment was also developed and validated in Atlantic salmon (*Salmo salar* L.). A discovery dataset of 102 Atlantic salmon from three distinct aquaculture strains were subject to restriction site associated DNA marker sequencing. The resultant single nucleotide polymorphisms were filtered according to quality, property and suitability for probe-based high-throughput genotyping technology. The final SNP panel consisted of 94 mass genotyping assays that gave 100% accurate parentage in independent samples of known pedigree. Finally, a set of standardised trait descriptors were designed for bivalve molluscs to accompany next generation sequencing submissions. These standards are needed to provide consistent trait measurements between investigators for quality control and to enable interoperability of phenotypic and genotypic data in future meta analyses.

Chapter 1. General Introduction

1.1 The Birth of Aquaculture

The earliest records of aquaculture date back to over four thousand years ago. Around 2000BC the banks of the Yangzhe River in China thronged with merchants trading fish 'seed'. As fresh food, ornament and status symbol, the social elite stocked their ponds with common carp; known as *lee*. During the reign of Emperor Li in 618AD the culture, killing and eating of the emperor's homonym was outlawed (Nash, 2011). Around this time the polyculture of the bighead, silver, mud and grass carp began, with each of these fish occupying a distinct niche in a pond ecosystem (Hao-Ren, 1982).

There are also records of aquaculture practices from other ancient civilizations. The Ancient Egyptians relied heavily on fish from the Nile River as a source of protein (Beveridge and Little, 2007). Figure 1.1 reproduces a bas relief from a tomb dated to around 2000 BC. In this image a Nobleman is depicted fishing from an enclosed pond. However, there is no indication if this practice is true aquatic species culture or a form of proto-aquaculture.

Figure 1.1: Bas relief detailing an Ancient Egyptian Nobleman fishing an enclosed pond, from the Tomb of Thebaine, reproduced from Beveridge and Little (2007).

Additionally, there is considerable evidence that the Ancient Romans not only stocked ponds (*piscinae*) with various aquatic life (Balon, 1995), but employed advanced technologies such as artificial feeding and aeration (Kron, 2008). Clearly technologically advanced forms of aquaculture have existed for thousands of years, but never on the scale or complexity found in modern aquaculture production. Global human population increase and the demand for food security has resulted in technological advances such as feed enhancement, advanced enclosures and, most relevant to this work, selective breeding.

1.2 Selective Breeding in Aquatic Organisms

The selection and breeding of favourable aquatic organisms also has ancient heritage. In his book *Treatise on Fish Culture*, published in 475 BC, the Chinese politician turned fish culturist Fan Li details 'aquahusbandry' practices and, in particular, notes favourable traits for the selection of breeding carp. Despite this early start only around 8.2% of the contemporary world's aquaculture production currently uses genetic selection (Gjedrem and Rye, 2016). Nevertheless, aquatic species are promising candidates for artificial selection because their unique life history traits provide an excellent response to selection. The response to selection, or genetic gain, is determined by several parameters for a given population. For a trait that is normally distributed this relationship can be expressed as Equation 1.1, from Falconer and MacKay (1997).

(1.1)

$$Phenotypic\ Gain = h^2 \times i \times \sigma_P$$

The first parameter, selection intensity (i), is a value that corresponds with the proportion of the total population which is selected. The second term is hereditability (h^2) for the given trait, which is explored in detail in Section 1.3. The last is the phenotypic standard deviation of the trait within the population (σ_P). Through the synthesis of this equation, and the life history traits of aquatic species, it is clear why aquatic organisms are well suited to artificial selection programs compared to terrestrial organisms.

Aquatic organisms have much greater fecundity than terrestrial farm animals, for whom, in most cases, the total reproductive output of the generation is limited by the female contribution. This can vary from hundreds to millions of eggs per individual in wild species (Duarte and Alcaraz, 1989). This range in fecundity may reflect both the variation in size of aquatic species and also the trade-off between investing many high quality offspring or few of lower quality (Duarte and Alcaraz, 1989). In an aquaculture environment competition for food and space is minimised and many more offspring can survive than in natural conditions. As a result of this high fecundity, high selection intensity can be applied to a population because the required number of brood stock for a given generation represents a smaller proportion of the population. Additionally, increased fecundity per individual allows for mating designs to be adopted that provide decreased levels of inbreeding or the testing of many mating combinations. Factorial mating (Busack and Knudsen, 2007), for example, allows for the number of effective breeders to be greatly increased in comparison to a random breeding scheme.

The second parameter that allows for strong selection in aquatic species is a larger standard deviation for commercially relevant traits in comparison to terrestrial animals. Example values for bodyweight are shown in Table 1.1, reproduced from Gjedrem and Baranski (2009). A larger standard

deviation (SD) in a trait allows for greater selection as there exists much more variation upon which to act.

Table 1.1: Example average body weight and coefficient of variation (CV) in different farmed species, reproduced from Gjedrem and Baranski (2009).

Species	**Body Weight (kg)**	**CV**	**Reference**
Atlantic Salmon	6.61	19	Rye and Refstie (1995)
Rainbow Trout	3.41	21	Gjerde and Schaffer (1989)
Rohu carp	0.3	31	Gjerde (pers. comm.)
Whiteleg Shrimp	20.3	20	Gitterle et al. (2005)
Broiler	1.51	8	Rensmoen (pers. comm.)
Pig	151	10	Sehested (pers. comm.)
Cattle, bulls	440	7	Steine (pers. comm.)

Increased fecundity and trait variation are key life history traits that quantitative geneticists often cite as advantages in the selection of marine organisms, but practical implementation and commercial viability are often the primary consideration when designing selective breeding programs. Two economic advantages of aquatic organisms are; their much reduced feed conversion ratio, relative to most terrestrial animals, allowing them to more efficiently convert feed into biomass; secondly the low value per animal allows for large-scale informant trials with destructive measurements such as flesh quality or disease challenge trials to be implemented cost-effectively. However, the low value of single animal prohibits high investment in individuals in a mating program and may prohibit the use of technologies such as SNP genotyping arrays, as used intensively in bovine agriculture (Wiggans et al., 2011).

1.3 Heritability of Phenotypic Traits

The observed phenotype of a given individual can be partitioned into two unobserved effects; the genotypic and environmental (Equation 1.2). The variation in a given phenotypic trait is a function of the sum of the variance of the two unobserved traits, as shown in Equation 1.3.

(1.2)

$$Phenotype\ (P) = Genotype(G) + Environment(E)$$

(1.3)

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

The goal of selective breeding is to maximise the phenotype through the manipulation of the genotypic contribution by selection. Since the genotypic effect cannot be directly measured, it cannot be directly selected upon. Selection must therefore act on the phenotype of a trait. The response to this selection depends strongly on the hereditability of the trait. Heritability is the proportion of the trait phenotype that corresponds with genetic effect in a population. Heritability can be expressed as narrow or broad sense heritability. Broad sense heritability (H^2) is the ratio of the total genetic variance to the phenotypic variance (Equation 1.4). The total genetic variance includes additive, epistatic and dominance effects, shown in Equation 1.5.

(1.4)

$$H^2 = {\sigma_G^2}\Big/{\sigma_P^2}$$

(1.5)

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$$

Narrow sense heritability (h^2) is the ratio of the additive genetic variance to the phenotypic variance (Equation 1.6) and is most commonly used in quantitative genetics.

(1.6)

$$h^2 = {\sigma_A^2}\Big/{\sigma_P^2}$$

Selective breeding focusses on the selection of additive genetic variance and its transmission over generations, epistatic and dominance effects are to be avoided and ignored by maintaining low inbreeding in selected populations. Therefore, following references to heritability are under the narrow sense definition.

Under narrow sense heritability the remainder of the contribution to the phenotypic value of the trait not corresponding with the genetic effect is the environmental contribution. Environmental contribution, in this case, may refer to factors related to the environmental conditions such as temperature or pH, it may also refer to additional sources of variance such as measurement error.

Heritability is commonly measured in an index that ranges from 0 to 1. A heritability of 0 indicates that none of the variance in the trait is produced from genetic effects while a value of 1 indicates that the entire variance of trait is determined by additive genetic variance. The calculation of heritability index allows practitioners to make predictions of the genetic gain for the trait of interest. However, heritability for a given trait can vary across space and time for a population of genetically identical individuals. It is therefore recommended to calculate heritability for the trait of focus in the population of interest. The calculation of the heritability is often the first step in a breeding program. The index is frequently then incorporated into various predictive and descriptive statistics. Heritability values for traits in aquaculture species are shown in Table 1.2 adapted from Gjedrem (2000).

Table 1.2: Example heritability for a range of traits in aquaculture species with standard errors (±). Adapted from Gjedrem (2000)

Trait	**Species**	**h^2**	**Reference**
Body Weight	Rainbow Trout	0.21	Gjerde and Schaeffer (1989)
	Atlantic Salmon	0.35±0.10	Rye and Refstie (1995)
	Coho Salmon	0.30±0.10	Hershberger et al. (1990)
	Chinook Salmon	0.20±0.10	Winkelman and Peterson (1994)
	Arctic Char	0.40±0.19	Nilsson (1992)
Fat Percentage	Rainbow Trout	0.47	Gjerde and Schaeffer (1989)
	Atlantic Salmon	0.30±0.09	Rye and Gjerde (1996)
	Arctic Char	0.06±0.08	Elvingson and Nilsson (1994)
Flesh Colour	Rainbow Trout	0.27	Gjerde and Schaeffer (1989)
	Atlantic Salmon	0.09±0.05	Rye and Gjerde (1996)
Disease	Rainbow Trout	0.16±0.03	Rye et al. (1990)
	Atlantic Salmon	0.0±0.02	Rye et al. (1990)
	Arctic Char	0.34±0.14	Nilsson (1992)

The values in Table 1.2 range from a heritability of 0 – signifying no genetic effect on the phenotype in question, to a value of 0.47 – where 47% of the population phenotypic variation is explained by additive genetic variance. The success of a selective breeding program for the improvement of a trait depends on examination and quantification of the heritability of the trait in question.

1.4 Selection Methods

The criteria for selection of organisms and subsequent improvement on traits can proceed by many methods. The choice of selection method depends on the parameters of the trait, the life history of the organism and, frequently, the resources available. In all cases the aim is to select individuals in the population who have the highest breeding value (BV) while avoiding inbreeding. BV is the value, compared to the mean trait value, that an organism passes onto its offspring. For example, a value of +50g of growth above the mean at harvest being passed by a fish to its offspring would give the fish a BV of +50. Since this value cannot be directly measured it has to be estimated. The estimation of this value depends on the method of selection and is called the predicted (or estimated) breeding value (PBV or EBV). The accuracy of any selection method is the correlation of the PBV/EBV to the BV.

Mass selection (also known as individual selection) is the simplest method of selection and proceeds by population truncation based on the trait values of individuals. It allows for large gains in traits with high heritability due to the phenotypic variance in these traits reflecting a large proportion of the additive genetic contribution. The accuracy of mass selection is a direct function of the heritability of the trait. A trait with a heritability of 0.4 will have a correlation of 0.4 between the PBV and BV. Mass selection is not suitable, however, when heritability for a trait is low: the accuracy of selection becomes very poor and gains are marginal.

There are also some practical issues with the implementation of mass selection. The effects of uncontrolled environmental factors have a large effect of the success of selection programs. An example commonly found in aquaculture breeding programs is age. A small difference in age when stocked in a shared environment can have a disproportionate effect on final weight. In organisms where it is impossible to group together individuals of the same age, mass selection is unlikely to provide the expected gains. Additional problems are found when trying to combine multiple populations under mass selection in different enclosures. Here there may be additional unmeasured factors that cause the overall mean of the enclosures to differ. Some progress may be made using the values in comparison to the mean or selecting by standard deviations. Selective breeding programs using mass selection in catfish (Moav and Wohlfarth, 1976) and tilapia (Hulata et al., 1986; Teichert-Coddington and Smiterman, 1988) have shown poor results and demonstrate the difficulties in achieving a large realised response to selection. These studies also highlight an additional concern with mass selection: unchecked inbreeding. The large fecundity of aquatic species often leads to the presence of a high number of siblings in a given enclosure. Variation will exist both within and between families and for a given trait and it is likely that selection with act unevenly between families; some families will be over-represented in the selected population resulting in higher inbreeding than expected by chance. Inbreeding can also be increased by unequal contribution of parents to each generation of selective breeding program. Highly skewed contributions (a small number of parents contributing a large number of offspring) are common in mass spawning events.

Empirical evidence has shown this effect is significant in both mass spawning fish (Loughnan et al., 2013) and shellfish species (Li et al., 2009b).

Family selection is an alternative method to mass selection where the unit of selection is not individual fish, but family trait means (Lush, 1947). The term family could mean full or half-sibling families or a more broadly related group of related fish. In this scheme families are ranked according to their mean trait value and selection proceeds using either entire families or a sub batch. This method is more appropriate when the heritability for the trait of interest is low, as the number of individuals used to calculate the family mean is a function of the accuracy of the BV for the family. The strength of this approach depends on sufficient numbers of individuals per family and accuracy is decreased when small numbers of individuals are used to estimate the family mean. Family selection allows for selection upon traits that require destructive measurement. Large families can be separated and the breeding value of the family can be calculated using only a proportion of the total offspring, leaving the unmeasured individuals available for breeding. A common trait that uses destructive breeding estimation is disease resistance which is calculated in disease challenge trials. In these trials survivability to an introduced pathogen is tested in a large mixed family group. This scheme has the advantage that families with high survivability can be selected without having to directly expose all selection candidates to disease (Fjalestad et al., 1993).

Inbreeding is of concern in this mating scheme: selection upon families inherently involves the crossing of a narrow genetic base to form the next generation. Additionally, the maintenance of many separate families may not be economically viable with as many as 200 separate families being maintained in family selection programs of Atlantic Salmon (Gjedrem, 2010). Finally, the overall selection applied is weaker under family selection in comparison to other methods, as only between family variation is utilised.

Within family selection is a method performed within large family groupings. As with between family selection the family grouping can be any permutation of relationships, but the method relies on a high proportion of shared genetic background within a family. Selection proceeds under this scheme by treating the family trait mean as zero across many families and selecting upon individual variation from the family mean. This method is particularly effective when there is a large environmental component that is common to families. It is also practically advantageous, as it economises breeding space and can be performed with minimal manual handling of selection candidates. Trait gains in body weight of around 12.4% per generation have been achieved in Nile Tilapia using within family selection. However, the within family selection method has been shown to provide the poorest gains in aquaculture fish (Gall and Huang, 1988). Therefore, within family selection may be considered an economical way to begin a breeding program with minimal cost and practical consideration being played off against poor predicted gains (Bolivar and Newkirk, 2002).

Practitioners often use a mixture of different selection methods in what has become known as combined selection. Here breeding values from both individual and family selection are combined to

form a combined value of the individual. Usually combined selection programs use two or more measures of breeding value in the derivation of an individual's comparative merit. A combined within and between family selection program has been successful in channel catfish (Bondari, 1983; Bondari, 1986) providing gains in a variety of economically relevant traits. However, by contrast most practitioners of combined selection do not select directly on the phenotype. It has become commonplace to use a variety of phenotypic data such as family, parent or progeny trait means to derive a selection index. A selection index is an overall assessment of an organisms breeding value for a selective breeding scheme with many target traits. However, selection indices can also be used for a single trait in order to produce greater gain per generation in comparison to multi trait selection.

The most common selection index currently found in aquaculture is Best Linear Unbiased Prediction or BLUP, originally described in Henderson (1975). This approach used in conjunction with mixed model methods has become commonplace in selective breeding; using all available trait information, correcting for known cofactors and incorporating pedigree structure. The BLUP analysis produces estimated breeding values (BLUP-EBV) on which practitioners can select the next generation of breeders. This method of selection is particularly useful when heritability is low as the accuracy can exceed the heritability, unlike in mass selection. BLUP has been used successfully in tilapia (Gall and Bakar, 2002), Coho Salmon (Neira et al., 2006) and rainbow trout (Kause et al., 2005) selective breeding programs. However, the commercial nature of detailed information on selective breeding programs may mean that many more species and traits are under BLUP selection than indicated in academic literature.

Whist BLUP selection carries a range of advantages over direct phenotypic selection it may also increase inbreeding at a greater rate as individuals with high BLUP-EBV scores often originate from the same family grouping, as seen in Gall and Bakar (2002). The utility and flexibility of the method is one of its key strengths and method have been developed, such as optimal contribution selection (Meuwissen, 1997), to balance gains and inbreeding utilising BLUP in mixed model methods.

All the above selection methods are independent of underlying genetic variation and are only concerned with the average amount of shared genomic ancestry as a function of the inheritance of entire chromosomes through mating. However, as research into the underlying genetic causes of variation in traits continues, information is produced that may allow the exploitation of genetic markers to assist breeding goals. Often these methods are implemented through the addition of genetic markers into BLUP models. In the case of specific markers, favourable and unfavourable alleles are determined through a quantitative trait locus (QTL) association experiment. The favourable allele is then used as an associated marker to improve the phenotype. These methods are known as marker-assisted selection (MAS). The underlying assumption is that a trait is controlled by an unknown number of genes and favourable variations within the genes will be linked according to linkage disequilibrium (LD) found in the genome. There have been documented cases of

commercially valuable traits dependant on a small number of loci with large effect (Barson et al., 2015; Houston et al., 2008). However, studies in wild fish populations support the theory that many (if not most) traits are dependant on a large number of loci with small effect (Bernatchez, 2016; Hoban et al., 2016).

In traits where a single locus has a large effect, the success in detecting the QTL depends highly on the density of the marker used, and the length of LD. In the case of markers such as microsatellites or amplified fragment length polymorphisms (AFLP) it would be a significant economic investment to genotype enough markers across the genome to have a significant chance of detecting an association between a trait and an allele. However, in the case of microsatellites, the number of alleles per marker is very high allowing for large power to detect QTLs in aquaculture populations that have less than a couple of hundred breeders. Microsatellite QTLs have been identified for upper thermal tolerance in rainbow trout (Perry et al., 2001) and for lymphocystis disease resistance in Japanese flounder (Fuji et al., 2007) indicating this approach can provide information of use to selective breeding programs with MAS.

The advent of Next Generation Sequencing (NGS) has made it economically possible to discover many thousands of SNP markers across each chromosome. The dramatic increase in marker coverage across a genome has resulted in a significant number of SNP QTL discoveries in the last decade. Relevant examples include Infectious Pancreatic Necrosis (IPN) (Houston et al., 2008) and Pancreatic Disease (PD) (Gonen et al., 2015) resistance in Atlantic Salmon, bacterial cold water disease (BCWD) in Rainbow Trout (Vallejo et al., 2014) and White spot syndrome virus disease in the tiger shrimp (Robinson et al., 2014).

An alternative approach to MAS for polygenic traits is genomic prediction. Genomic prediction (or genomic selection) uses markers distributed across the entire genome so that every QTL affecting the trait of interest is in linkage with a marker. This enables the genomic breeding value of the individual to be accurately estimated. The original implementation of genomic prediction (Meuwissen et al., 2001) proposed three different methods with different QTL/marker relationship effects assumed in each case. One method (genome-wide BLUP) used mixed model methods while two others (BayesA and BayesB) used a Bayesian method. Work in this area continues, and current efforts are well contrasted in Gianola et al. (2009). The effect of genomic selection in dairy cattle, with increased BV estimation accuracy, has been greater per generation trait gains internationally (Hayes et al., 2009). This early success has catalysed genomic selection programs in other livestock species (Hayes et al., 2013), crops (Heffner et al., 2009) and forest trees (Grattapaglia and Resende, 2010).

The utility of genomic selection in aquaculture has been explored under stochastic simulations. Early work has assessed its accuracy against BLUP selection (Nielsen et al., 2009), exploring different strategies for implementation (Sonesson and Meuwissen, 2009) and examining the relationship between power and number of markers (Lillehammer et al., 2013). Some empirical work

has begun to compare the accuracy of genomic prediction to other methods in Atlantic salmon (Ødegård et al., 2014; Tsai et al., 2015; Tsai et al., 2016) and large yellow croacker (Dong et al., 2016). However, studies demonstrating increased trait gains over time using genomic prediction are lacking in aquaculture species. This may reflect the deficit in genomic resources for aquaculture species or the low per animal value preventing genomic selection becoming commonplace. Alternatively, it may be simply that no programs, using genomic selection, have shown results. Public knowledge on commercial selection programs is severely lacking so inference on worldwide uptake is difficult.

1.5 Inbreeding in Selective Breeding Programs

The mating of full-sib (consanguineous) individuals in selective breeding programs is of major concern because it leads to inbreeding depression: a group of phenotypic changes (usually undesired). In selection programs the aim is to maximise genetic gains while minimising inbreeding as to avoid any negative effect on phenotype.

At a genetic level there are two major hypotheses for the causes of inbreeding depression (Kristensen et al., 2010). The partial dominance hypothesis is concerned with the effects of deleterious recessive alleles, occurring more frequently as homozygotes in inbred lines. In contrast the overdominance hypothesis is concerned with the genome-wide heterozygosity, positing that inbred lines have overall increased levels of homozygote sites which decrease overall fitness. Evidence in support of each hypothesis can be supplied through the crossing of inbred lines (Roff, 2002). These experiments proceed by inbreeding many lines from a single source population, deleterious mutations are then fixed or lost through genetic drift over many generations. The crossing of these lines under the partial dominance hypothesis should result in the fitness of the inbred cross being greater than the mean of the inbred lines, due to the purging of deleterious alleles. Under overdominance the fitness of the cross is the mean of the two inbred lines. There is empirical evidence for partial dominance over overdominance in most cases (Charlesworth and Charlesworth, 1999; Roff, 2002). Practically in selective breeding programs the causative mechanisms of inbreeding are of utility when predicting recovery from an inbred state, but in most cases the aim is measurement and management.

Inbreeding is measured using the coefficient of inbreeding (F) which is calculated per individual and is defined in relation to a base population of unrelated individuals in which all alleles are unique. However, the definition of terms such as ‘related’ or ‘unrelated’ is highly subjective and dependant upon the time-scale at which mating is considered significant. For example, the most recent common ancestor (MRCA) of all humans is thought to have lived just a few thousand years ago (Rohde et al., 2004), if the base population for calculating the coefficient of inbreeding of all modern day humans included the MRCA the statistic would not consider any modern day humans completely unrelated. In practice, a base population from which to calculate the inbreeding coefficient is often

limited by available information. The coefficient of inbreeding is a value between 0 and 1, with 0 being completely outbred and 1 being completely inbred. The value is often referred to as a percentage, an increase in inbreeding (ΔF) of 1% per generation refers to a mean increase in F (or ΔF) by 0.01 between a generation.

Reporting of the effect of inbreeding in monitored populations is generally reported as a linear regression of the trait value on the inbreeding coefficient expressed as a percentage per individual. For example, Gjerde et al. (1983) found per 10% increase in inbreeding 4.5-6.1% reduction adult growth. A summary of empirical work on the effects of inbreeding on a range of traits in (mainly) salmonid aquaculture fish can be found in Kincaid (1983). An exhaustive overview on the effects of inbreeding on characteristics in aquaculture species has not been collated, but Leroy (2014) provides a good analysis in terrestrial animals over many traits and seven organisms found an average decrease of 0.137% of the mean of the trait per 1% inbreeding. This value includes many traits (342 analysed out of 1218) where inbreeding had no effect on the mean in the tested parameters.

1.6 Next-generation Sequencing in Aquaculture Research

Genomic and transcriptonomic experiments in aquaculture species use NGS technologies to sequence many billions of nucleotides. There are a range of different commercial technologies operating in this space and the technical and methodological differences are well covered in Goodwin et al. (2016). Generally, these platforms provide short (<500bp in most cases) reads of DNA fragments that are then analyzed using PCR incorporated 'barcode' regions to separate out individuals, or runs, for downstream analysis. The 'next' in next generation sequencing is in reference to the significant increase in output from the classic Sanger (chain-termination) sequencing, now frequently called first generation sequencing. The increase in raw number of bases has allowed more complex experiments that require computer automation for analysis. However, despite the large decrease in the cost per nucleotide, sequencing is still expensive. Researchers may not have available funding to run the number of samples they require if the entire genome of each individual is to be sequenced, additionally the experimental design may not require such a high density of information. The solution to these problems has been to use reduced representation libraries, where a common subsample of the genome is sequenced in all individuals.

The most commonly used reduced representation technique is restriction site associated DNA marker sequencing or RAD-Seq. First presented in Miller et al. (2007), and followed by a more generalized protocol in Baird et al. (2008), the technique relies on the digestion of genomic DNA with restriction endonucleases and sequencing of the adjacent fragments. Over time the protocol has been adapted and similar methods have been developed under the RAD umbrella. The original RAD protocol was followed by genotyping-by-sequencing (GBS) (Elshire et al., 2011), ddRAD-seq (Peterson et al., 2012), 2b-RAD (Wang et al., 2012) and ezRAD (Toonen et al., 2013). These all use

the same principle of restriction enzyme digestion followed by sequencing. Details of the advantages and disadvantages of specific methods are found in Andrews et al. (2016). Common to all methods is the discovery of many thousands of SNPs located across the genome. This technique is used in aquaculture genomics projects for the discovery of SNPs for a wide range of purposes. Aquaculture applications of this technology include the association of discovered markers with QTLs; genomic selection; estimation of kinship parameters such as inbreeding or relatedness; generation of linkage maps; and strain and species identification. Some of the RAD-seq protocols require specialist equipment and all require some form of complex bioinformatic analysis. Overall RAD-seq is highly specialist, but commonplace in experiments aiming to profile a large number of markers for aquaculture genomics.

Another frequently used technique is whole transcriptome shotgun sequencing or RNA-seq. Before NGS the profiling of RNA was performed using Sanger sequencing of short (500-800bp) complimentary DNA (cDNA) fragments. These were called expressed sequence tags or ESTs. Usually only a very small proportion of the transcriptome was sampled using ESTs and they have been functionally replaced by RNA-seq (Wang et al., 2009). A modern RNA-seq experiment consists of the isolation of total RNA from tissue followed by cDNA synthesis, library preparation and sequencing. The technique allows for the rapid and vast expansion of sequenced functional genome regions, used for the characterization of genes with no reference sequence. The use of De Bruijn graphs has allowed for the de novo reconstruction of the expressed transcriptome from short reads (Grabherr et al., 2011). Very frequently existing databases such as The Gene Ontology database (Ashburner et al., 2000) are used to provide functional or gene pathway annotations of the assembly. Alternativly the technique can be used to test hypotheses about differential gene expression under trialed conditions (Li et al., 2011; Wang et al., 2009). RNA-seq is frequently used in aquaculture for the assembly of a draft transcriptome. This provides a wealth of functional sequence data and, if a number of representative individuals are used, it is possible to use the transciptomic data for SNP discovery. Experiments using comparative gene expression analyses provide information about feed conversion, disease resistance, stress and a variety of other commercially relevant parameters.
However, the annotation of genes is reliant on current database with a bias towards vertebrate linages, making RNA-seq experiments in invertebrate aquaculture species more difficult. Even when annotated correctly RNA-seq datasets are extremely complex and most feature splice variants, bioinformatics chimeras and other difficult features that require specialist attention.

As the per base cost of sequencing decreases further it is expected that it will become economically possible to sequence entire genomes of many individuals for aquaculture research. While it is possible to perform a *de novo* genome assembly using short read technologies, there are many methodological difficulties such as complex or repetitive regions (Chaisson et al., 2015). Whole genome resequencing is now possible for aquaculture species with a high quality reference, such as Atlantic salmon. In this kind of experiment, reads from individuals are mapped to the reference and

novel sequencing and structural variants can be discovered. Unlike RAD-seq or RNA-seq, the entire genome is surveyed for variation, therefore the likelihood of detecting causative variants in association analyses is much higher. For most species this option will be uneconomical and reduced representation techniques will remain dominant for SNP discovery in non-model species.

1.7 The Decay of Genomic Tools

The development of genomic tools represents the greatest hurdle for the implementation of these technologies in selective breeding programs. Here we define a genomic tool as a sequence, or the characteristics of a sequence, related to genomic information. It could be a series of SNPs known to associate with a trait, or a known gene containing many neutral SNPs. The tool may exist as a genotyping method, or as a generalised reference. The prohibitive costs of tool development mean that there is no commercial case for a single provider to engage in the research, and, in practice, many providers and researchers form a cooperative and pool their respective resources. This model has been a success in the sequencing of the Atlantic salmon genome (Davidson et al., 2010; Lien et al., 2016) and in the discovery of various QTL markers (Gonen et al., 2015; Houston et al., 2008). However, there has been little discussion of longevity of the marker panels, and little work has explored the potential for the decrease in accuracy, here called decay, over the progress of selection programs. The validation of genomic tools is, for the most part, concerned with the immediate gains possible in light of novel discoveries.

In the investigation of genomic decay, we are mainly concerned with SNP loci as these are currently the marker of choice for selective breeding. However, the question applies to all markers in varying degree. In the use of SNPs in QTL mapping, breeding value prediction, parentage or sex assignment there lies an assumption that the current mechanism of action will maintain effectiveness over time. In fact, there are many mechanisms of genomic decay through which these tools might produce less accurate results over time including recombination, fixation and mutation. In the case of markers being used to exploit regions of linkage disequilibrium across the genome, recombination events will reduce the degree of linkage as generations proceed. Genetic linkage distance is measured in centimorgans (cM). A cM corresponds with a 1% chance that a marker on a chromosome will be displaced onto the second chromosome (in diploid organisms) due to recombination. According to this definition over time all linkage disequilibrium will decay, but at varying rates across the genome. Allelic fixation (or fixation) can also cause the decay of genomic tools.

Fixation is a process by which, across a generation, all variation within a population at a locus is lost. Through this process loci can go from being polymorphic to monomorphic. Once fixed, a polymorphism ceases to become relevant unless alternative alleles are found in another population. Fixation can proceed through genetic drift where a breeding program may fix an allele by chance: this effect is proportional to the population size and in many cases selective breeding programs have a

small number of families. Another pathway to fixation in selection programs is through artificial selection, particularly in the case of QTL markers where the selection upon the favoured allele may have unintended consequences for whole genome diversity.

Mutation may result in the decay of effectiveness of genomic tools through a change or disruption of the molecular mechanism of action, in the case of traits with few loci of strong effect. Alternatively, mutation may cause the decay of genomic tools through the disruption of genotyping methods, resulting in null alleles. These are a major concern in parentage assignment with molecular markers, and are reviewed in Dakin and Avise (2004). Examples of the decay of genomic tools are not well documented, the continual improvement of current tools has prevented any long-term studies examining the potential effects of genomic tool decay. However, an example that provides some evidence of variable linkage can be found in markers for sex in the sdY gene in salmonids. Originally characterised in Rainbow Trout (Yano et al., 2012), markers targeting the region in sockeye salmon show imperfect linkage with sex (Larson et al., 2016). A similar experiment in Atlantic salmon produced markers with perfect sex linkage in 63 individuals (Houston et al., 2014). While this provides no evidence of the process over time, it does demonstrate that markers may be in variable linkage to a desirable trait.

There has been increased theoretical investigation into the decay of QTL-marker linkage in genomic selection and its effect on BV estimation accuracy, probably because of the high cost and delicate economics of this method. Through stochastic simulations of genomic selection Jannick (2010) found a rapid decrease in trait gains over time, corresponding with a decrease in genomic-EBV due to linkage disequilibrium decay. Using gene level simulations, Muir (2007) found the same effect was more significant under selection across generations, as commonly found in selective breeding programs.

1.8 Data Standards in Omics

Primary data quality in genomics and transcriptomics has been subject to little or no requirements or standardization beyond those of scientific journals for publications. Recently, however, there has been increasing work on producing standards for different data types and their associated metadata. A data standard refers to any number of requirements that the data should meet to be considered valid. There may be methodological standards (for example, a description of the given study design), data quality standards (for example, a certain proportion of data must not be missing) or standards in the presentation or notation of data (the use of specific ontological terms to avoid confusion of nomenclature between researchers). Metadata in this case refers to data associated with the primary data and the organisms from which they were obtained (such as sex, sample type, age, location) that may effect the experimental conclusions. Ideally, this metadata would also be subject to some form of standardization.

1.8.1 Problems with Poor Metadata or Primary Data

A lack of standards in data can result in false inference. Unreported parameters may result in insufficient power, methodological issues or simply poor experimental design. These problems are especially important in experiments with complex methodologies. Publication associated data presents a record of the research performed allowing other researchers to validate the findings. Data should be presented in a format that facilitates easy, unambiguous understanding. Meta-analyses and review papers often make links between different studies and synthesize findings, allowing for unique or expanded conclusions. Poor metadata and/or primary data standards prevent easy synthesis of data from multiple sources. The notation and curation of good, associated metadata allow for potentially confounding variables to be taken into account. Equally, standardization means shared variables can be combined with confidence in resultant findings.

1.8.2 Existing Relevant Standards

Standardization is a relatively new concept, but it has gained momentum with its debut into various fields. A good example of just such a standard, that sees continued use, is the minimum requirement of information for a genome sequence specification (MIGS) (Field et al., 2008). In this case researchers provide a range of supplementary and descriptive data alongside the publication of a genome assembly. The data is split into ‘minimum’ or ‘extra’ requirement. Researchers must supply all of the ‘minimum’ requirements and the ‘extra’ items are non-compulsory. The authors of the MIGS have since expanded the standard to include ‘any sequence’ (MIxS) (Yilmaz et al., 2011). Together these standards cover the bulk of genetic information uploaded to public databases and have seen sustained citation since their inception (Figure 1.2).

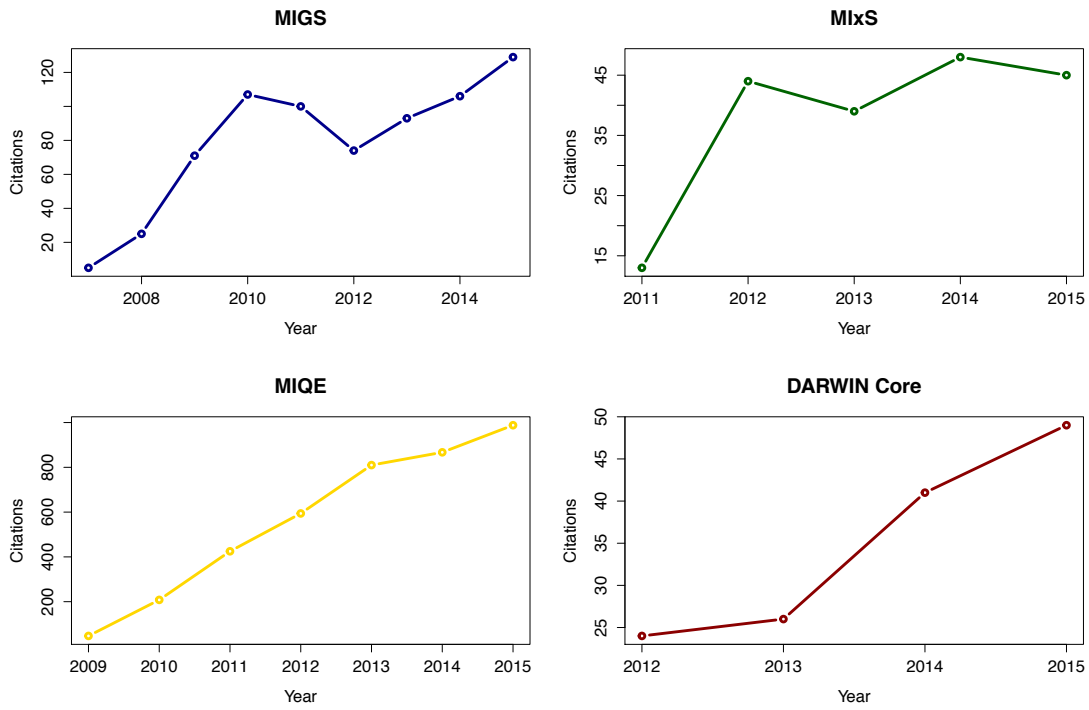


Figure 1.2 – Line charts showing Google Scholar citation number per year for MIGS (Minimum information for genome sequence specification), MIxS (minimum information for any(x) sequence specification), MIQE (minimum information for publication of real-time PCR experiments) and DARWIN Core – (a biodiversity data standard).

The MIGS and MIxS standards recommendations have been adopted by database services such as The Genomes Online Database (GOLD) to enrich the metadata attached to their genomic data. There has been limited adoption by Journals and only 'Standards in Genomic Science' Journal has made MIGS and MIxS compliance mandatory for submission of genome or sequence data.

The best example of a set of standards adopted by a community comes in the form of the minimum information for publication of quantitative real-time PCR experiment standards (MIQE) (Bustin et al., 2009). While quantitative PCR equipment has become commonplace in most modern molecular biology laboratories, there is a huge range of methodological concerns when planning and executing a qPCR experiment. The initial mRNA extraction and quality control is a difficult and important step to ensure template quality is sufficient for cDNA synthesis. Once high-quality mRNA has been extracted, a qPCR user has a huge range of cDNA synthesis kits to choose from and little consensus on optimal practice. The user must select appropriate control samples and sufficient replication to keep costs low, while applying sufficient power to detect the effect of interest. Key variables such as oligoneucleotide design and the selection of reference genes have a large impact on experimental results. The need for standards in qPCR was brought to the fore in a high profile publication (Huang, 2005) and subsequent retraction due to incorrect interpretation of qPCR results. The qPCR community then formed a set of standards ensuring mandatory reporting of the aforementioned variables. Since their publication of MIQE the set of standards have seen increasing

yearly citation. MIQE citation is commonly used at publication submission to assure reviewers of high quality qPCR experiments and allow practitioners to evaluate the effectiveness of different experimental parameters.

The DARWIN core standards (DwC) (Wieczorek et al., 2012) are an example of standards used in the description of non-genetic data. The aim is to present biodiversity data, using common terminology and formatting, to allow for interoperability between studies. These standards have seen modest citation (Figure 1.2) and have been incorporated into various international biodiversity initiatives such as the Global Biodiversity Information Facility whose aim is to supply worldwide biodiversity data via an online portal. Briefly, the DwC provides a set of terms that describe a particular record of an organism in a specific place and time. The terms have specific semantic definitions to permit precise human and machine interpretation.

1.8.3 Data Reporting in Aquaculture Genomics

Data reporting in aquaculture genomics is highly heterogeneous. Some authors choose to include a range of metadata associated with high-quality, well-annotated genomic datasets, others the minimum acceptable to pass peer-review. The process of publishing peer-reviewed journal articles has two opportunities to exert specific standards on data reporting. Before submission, the majority of authors choose to submit their genomic data to an online repository. At this stage each of the three major online nucleotide repositories, European Bioinformatics Institute (EMBL-EBI), National Centre for Biotechnology Information (NCBI) and DNA Data Bank of Japan (DDBJ), require identical minimum reporting in their archive for raw short-read sequence data. In all submissions data must be uploaded corresponding with a Project (or Study), a name and description for the scientific project, and a Sample, a description of source materials for the data object. A set of raw, short nucleotide reads, grouped into 'Runs', must then be uploaded along with an Experiment identifier. The Project, Sample, Experiment and Run identifiers represent the minimum data required to upload genomic data. Upon the upload of raw data, an accession number is issued that acts as a permanent link to the data for inclusion in publications. In this way, any author wishing to have a permanent link to their raw data is only required to report a Project Name to which the data is connected, a written sample description, an experiment identifier and to group the data into distinct runs.

Journals have the opportunity to apply certain conditions to articles, including specific data reporting standards. A search for 'Aquaculture' in the Thomsen Reuters Web of Science journal listings gives a total of eleven journals of which only two have specific data standards for nucleotide information (Appendix A). In both cases, an accession number from the aforementioned databases is specifically required, but no additional information is explicitly mandated in any journal documentation. Despite this requirement there are a growing number of examples of published work were the raw data has not been uploaded (Shi et al., 2014; Ren et al., 2016). Even when raw data is

included, there are cases were further information is required. One example can be found in transcriptome studies where authors circumvent the provision of data by uploading raw reads and omitting the final assembly. In these cases there may be some motivation to avoid publishing the assembled data for personal or private gain (cf. Nguyen et al., 2016).

Furthermore, no specific metadata is required to accompany genomic data. In many cases, the unenclosed metadata may include information key to experimental design (such as sex, in a sex-linked marker discovery) limiting reproducibility. Downstream analyses using the raw data may produce important intermediate files. There is little mandate for keeping or presenting these files, however they may be of critical use. A lack of standards can also result in unexpected consequences for non-specialists who may use unsuitable data formats, as shown in a recent example where gene abbreviations were incorrectly altered by a commonly used spreadsheet software with an autocorrect function (Ziemann et al., 2016).

1.9 Thesis Aims

This thesis aims to develop tools, workflows and resources for aquaculture practitioners working in diverse species worldwide. Chapter 2 details a stochastic simulation and bio-economic analysis of a proposed breeding programme for a mass spawning fish: the gilthead seabream. The simulations are written in R and can be easily adapted for any species and trait, so have wide applicability across aquaculture. Chapter 3 details the development and validation of a low-density SNP panel for parentage in Atlantic salmon. The panel is publically available and recommendations are made for the general workflow of developing panels in other species. Finally, Chapter 4 details a set of shellfish reporting standards that are currently implemented via the EMBl-EBI submission system. These are expected to be developed over the coming years in the expectation that practitioners will incorporate the standards into publications and reports.

Chapter 2. Bio-Economic Simulations of Breeding Programmes

2.1 Introduction

The initiation and continued success of a breeding programme relies on a huge number of variables. The specific aims and facilities of individual practitioners of selective breeding in aquaculture do not readily lend themselves to a one-size-fits all approach. It is, therefore, important to model the uncertainty, and predict the result, of a particular design of selective breeding programme. These models have become known as Bio-Economic simulations. They include biological and economic parameters often modelled using an explicit stochastic approach to predict the outcomes of complex selective breeding programmes; predicted gains in trait values, inbreeding and economic output. A comprehensive review summarising bio-economic modelling in aquaculture, since its inception in the 1980s, is provided by Llorente et al. (2016). This review surveys the large heterogeneity of approaches and aims; some studies use models that express biological and economic factors as a function, the solution for which represents an ideal optimum (Araneda et al., 2011; Bjørndal 1988); others model individual animals as agents in a computer program environment, generating summary statistics from the simulated population (Melià and Gatto, 2005; Robinson and Hayes, 2008). In this study, the latter approach has been used to model genetic gain and economic output using the latest statistical methods as implemented in Robinson et al. (2010a; 2010b).

The method first details the selective breeding design with regard to the tank layout, number of individuals per tank, mortality at various stages of development, mechanism of selection and mating design. Genetic gain simulations begin with estimates of genetic and environmental effects according to published or estimated heritability of the trait. Values for these effect are then randomly allocated to all individuals in the system. Selection proceeds according to the selective breeding design being tested, and offspring receive a genetic component of the phenotype equal to the sum of half of each parents' genetic additive value. Once phenotypic values have been simulated the results are recorded and further generations are simulated. The entire process is replicated many times and summary statistics calculated for the varying result of each replicate run. The variation between replicates in the simulation is due to the random allocation of phenotypic values and also according to the level of stochastic events in the simulated selection and mating design.

For each year in the simulation, profit and loss are projected according to cash flow as a result of production sales, and outgoings such as energy, labour and feed in the hatchery and growout. The added profit derived from a trait depends on the trait in question. Gain in growth rate for example, allows for more weight of fish to be produced from fixed energy and labour costs, but offset against increased feed cost, while a gain in disease resistance increases survival and produces profit from decreased mortality. In all cases, the cost of the selective breeding programme is integrated into the calculations. Stochastic economic factors such as feed cost or market prices can be simulated to give

estimates of profits under different scenarios. The combination of all these factors is then used to form a cost-benefit ratio detailing the various projected profits against the costs on a year-by-year basis.

The gilthead seabream (*Sparus aurata L.*) is a fish from the *Sparidae* family with a native distribution in the Mediterranean Sea and eastern coastal sections of the Atlantic Sea. Aquaculture production of the gilthead seabream was over 160,000 tonnes worldwide in 2014 (FAO, 2016) with major production in Greece, Turkey, Italy and Spain. The gilthead seabream is a sequential protandrous hermaphrodite with a complex mating system. In captivity, the bisexual gonad undergoes differentiation and spermatogenesis between 1-2 years before undergoing sex reversal with oocyte development beginning at 2-3 years (Zohar et al., 1978). In artificial populations, around 95% of individuals in year one of life develop male reproductive capacity (the remaining percent being immature), in year two a maximum male reproductively of 75% was observed compared to a maximum of 45% female maturity (Kissel et al., 2001). The species undergoes mass spawning in the wild. Attempts to produce individual crosses by artificial stripping had a success rate of around 10% due to difficulty in identifying ripe individuals (Gorshkov et al., 1997). In the wild, mating is usually between October and January, which presents issues for aquaculture production where eggs are required year round. Some progress has been made by using photoperiod manipulation to alter both growth (Vardar and Yıldırım, 2012) and spawning (Kissel et al., 2001) dynamics.

An additional concern with mass spawning species is unequal parental contribution and increased inbreeding as a result. Studies using molecular markers to determine parentage of mass spawned offspring in gilthead seabream have found an effective population size (N_e) values between 13-28 in batches of 40 fish (Borrell et al., 2011) and 14-18 in batches of 50-60 fish (Brown et al., 2005). N_e, most simply, is the number of breeding individuals, however more complex definitions are presented in Wang et al. (2016). The N_e values presented above correspond with inbreeding values of 1.7% - 7.7% in a randomly breeding population (Appendix B for calculation). However, the differential mating success found in the gilthead seabream will inflate these values well above the maximum of 1% suggested for selective breeding programmes in Gjedrem et al. (2005). The complex spawning dynamics of the gilthead seabream limit the methods of selection that can be applied. Selection between and within families is complicated. Difficulties in setting up crosses between individuals limit the mating designs that can be adopted, and large spawning events carry a highly skewed parental contribution which can vastly decrease the proportion of desired crosses. Additionally, the hermaphroditic nature of the fish make it difficult to synchronize the timing of selection generations and the intended crosses may not be made in spawning events due to sex change over time. It is hard to test the merit of families, as the unequal distribution prevents a large enough number of individuals being produced per family for accurate trait average calculations. Mass selection is, therefore, the method with the least technical difficulty for the improvement of most traits in gilthead seabream.

A survey on selective breeding programmes in Europe revealed that individual growth rate was most frequently included in gilthead seabream broodstock improvement programmes in comparison to all other traits (Chavanne et al., 2016). Growth rate is trait of significant economic importance; for a given space and production costs (feed withstanding) more weight of fish can be produced increasing the overall profitability of the operation. Additionally, estimates of heritability for growth rate in gilthead seabream are very high allowing for rapid response to selection. Studies calculating heritability in farmed conditions found an estimate of 0.38 ± 0.07 for body length (Antonello et al., 2009), and an estimate of 0.40 ± 0.03 for body weight (Fernandes et al., 2016).

A selective breeding programme aiming to improve growth rate is therefore a good example of a programme that requires explicit modelling. The trait is known to have high heritability but difficulties lie in preventing inbreeding and getting good response to selection due to unusual spawning dynamics. Understanding risk and uncertainty will provide the confidence necessary for breeding programme staff, producers and investors to support the selective breeding programme. This study aims to simulate the predicted gain in growth rate of gilthead seabream in response to a selective breeding programme. The rate of inbreeding and entire cost-benefit ratio of the scheme will also be examined. Finally, the parentage assignment accuracy will be assessed across the entire proposed breeding programme.

2.2 Methods

2.2.1 Selective Breeding Programme Design

The design of selective breeding programmes rarely begins with the capture of wild organisms. In this case, I simulated the initiation of a breeding programme from aquaculture bred stock not under any intended artificial selection. The organisation of the existing stock is a simplified version of that found in the breeding nucleus of a major commercial provider based in Greece. The basic schematic for the existing programme is shown in Figure 2.1.

In the ‘current’ system individuals from a wild population were randomly captured, and a population of 50 individuals were selected based on survivability to the artificial conditions, to form the F0 generation. The F0 were then mass spawned to produce five lines of 50 individuals per line. Each generation was bred to form the proceeding generation until F3, all breeding fish were spawned in groups of 50.

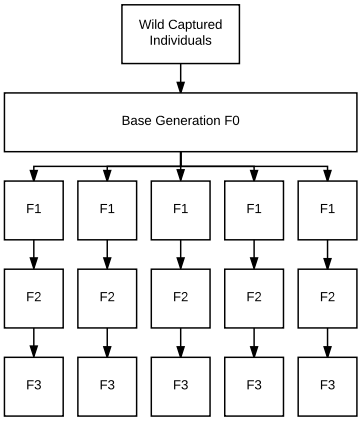


Figure 2.1: Flowchart detailing breeding over four generations in the 'current' breeding programme.

This 'current' scheme was adapted into the breeding programme as shown below in Figure 2.2. The scheme starts with the equal contribution of the five existing lines to a growout of 10,000 surviving fish in production conditions. A pre-selection of 750 from the 10,000 fish follows where the 750 fish, with the most favourable phenotype, are selected for genotyping. This step is performed to minimise costs and technical difficulties in genotyping and parentage analysis on the entire growout. Following molecular parentage assignment 200 fish are selected based on their estimated breeding value and these fish are used to form 4 tanks of 50 individuals. These 200 individuals are used to supply production and are also bred to form the growout from which the next generation is selected.

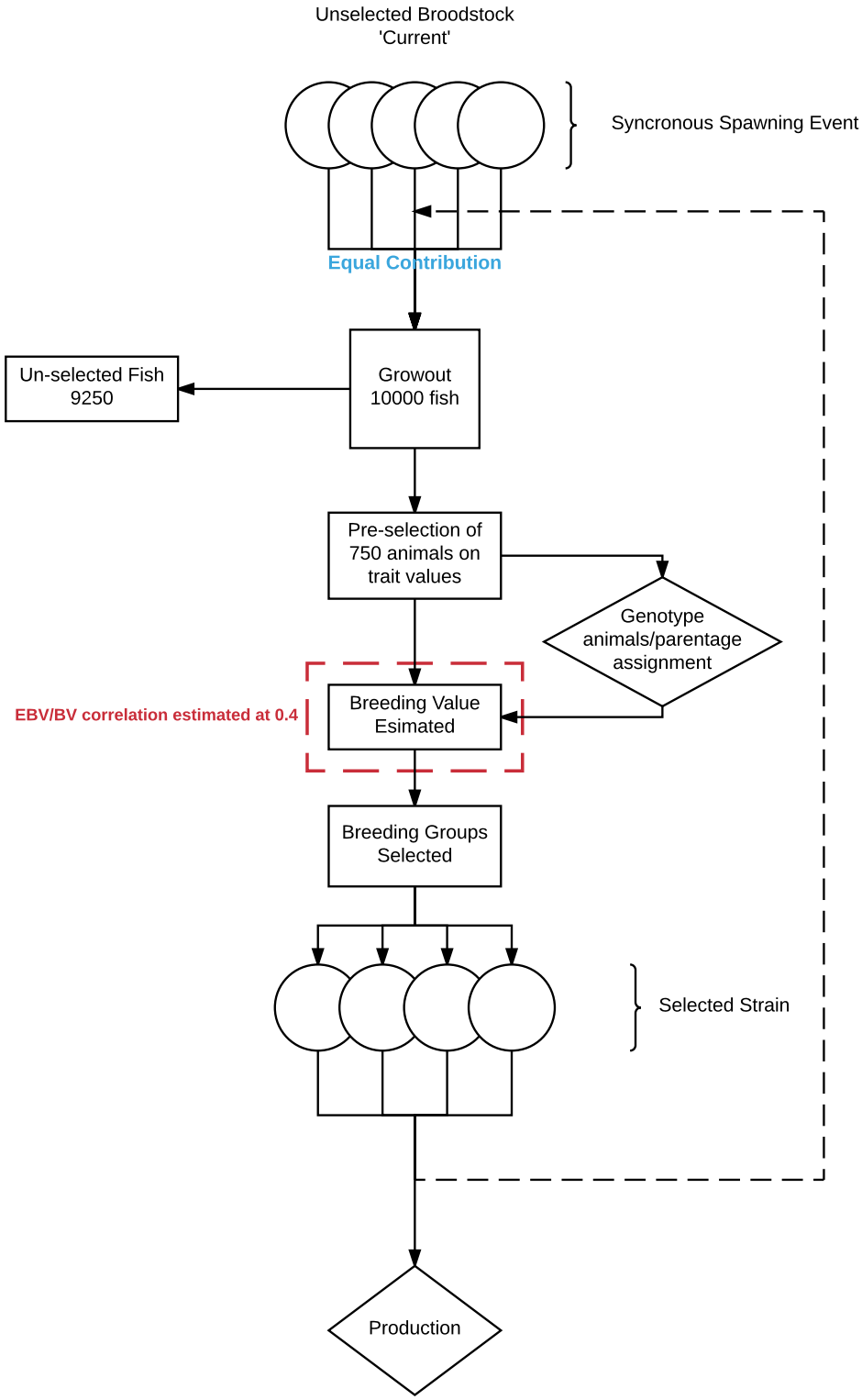


Figure 2.2: Flowchart detailing selective breeding programme from currently available breeding individuals to improved selective strain and export to production.

2.2.2. Simulation Implementation and Parameter Estimation

All stochastic simulations were programmed using R version 3.2.2. (R Core Team, 2016), and scripts are provided in Supplementary Data. Table 2.1 details the list of parameters used in simulations. This section outlines the estimation of these parameters.

Table 2.1: Table detailing simulation parameters used for all simulated scenarios.

Group	**Parameter**	**Value**	**Unit**
Burn-in Generation Parameters	generations simulated	5	generations
	generation size	100	individuals
	sex ratio	0.5	proportion of males
	mating proportion	1	proportion of total population mating
Current Scheme	breeding tanks simulated	5	number of tanks
	breeding generations	3	generations
	individuals per tank	50	number of fish
	tank sex ratio	0.4	proportion of males
Proposed Scheme	breeding tanks simulated	4	number of tanks
	breeding generations	10	generations
	individuals per tank	50	number of fish
	tank sex ratio	0.4	proportion of males
	correlation of EBV/BV	0.4	correlation coefficient
	number of fish per experimental growout	10000	number of fish
	number of fish pre-selected	750	number of fish
	number of fish selected	200	number of fish
Biological Parameters	heritability for growth	0.4	heritability
	mean weight at slaughter	400	grams
	SD of weight at slaughter	100	grams
Bio-economic Parameters	years of selection programme simulated	20	years
	time for fish to reach market size	18	months
	time for fish to reach sexual maturity	36	months
	total cost per kilo for fish	3.5	€/kg
	price per kilo at 'farm gate' for fish	4.4	€/kg
	profit per kilo fish sold	0.9	€/kg
	number of fish produced per year	100	millions of fish
	price per fish genotyped	7	€
	monthly selective breeding management costs	7000	€
	labour costs per experimental growout	1200	€/growout

All simulations begun with initial burn-in generations in which simulated individuals performed random mating. This was performed to simulate a background level of inbreeding and SNP segregation. Simulation work estimating genomic selection parameters (Sonesson and Meuwissen,

2009) used 4000 generations of burn-in breeding to establish linkage disequilibrium. Conversely, simulations of genetic gain (Robinson et al., 2010a) did not use any burn-in generations. In this study, initial work indicated that inbreeding values under simulations with no burn-in generations gave erroneously high values due to the shallow pedigree. Large numbers of burn-in generations resulted in unacceptable deviation of allele frequencies of simulated markers due to drift. Sex ratio in the burn-in generations was assumed to be 50:50 at time of wild capture. Sex ratio for each of the selected generations at mating was assumed to be 60:40 females to males, in line with estimates from commercial providers (personal communication).

The variation in size of gilthead seabream at commercial weight (~400g) varies among conditions and strains, Dupont-Nivet et al. (2008) reported a SD between 71.7 - 139.4g for four trailed stains. In this work for simplicity the SD at mean weight of 400g was estimated at 100g.

A correlation between the breeding value and the estimated breeding value was set at R=0.4 correlation coefficient. This value represents a minimum accuracy possible for the breeding programme, performing mass selection on phenotype alone with a heritability of 0.4 will give an accuracy of 0.4.

Economic parameters such as costs of growout and genotyping, selective breeding management and number of fish grown per year were provided by Xelect Ltd. (St Andrews, Scotland). The distribution of parental contribution for every mass spawning event was derived using a smoothed spline. This spline was created as a function of empirical data; the parental contribution found from 300 offspring assigned to 50 gilthead seabream individuals in a mass spawning event. Parental contribution data in the form of number of offspring allocated to each parent from a mass spawning event was provided by Xelect Ltd. The relative contribution of the parents is shown in Figure 2.3.

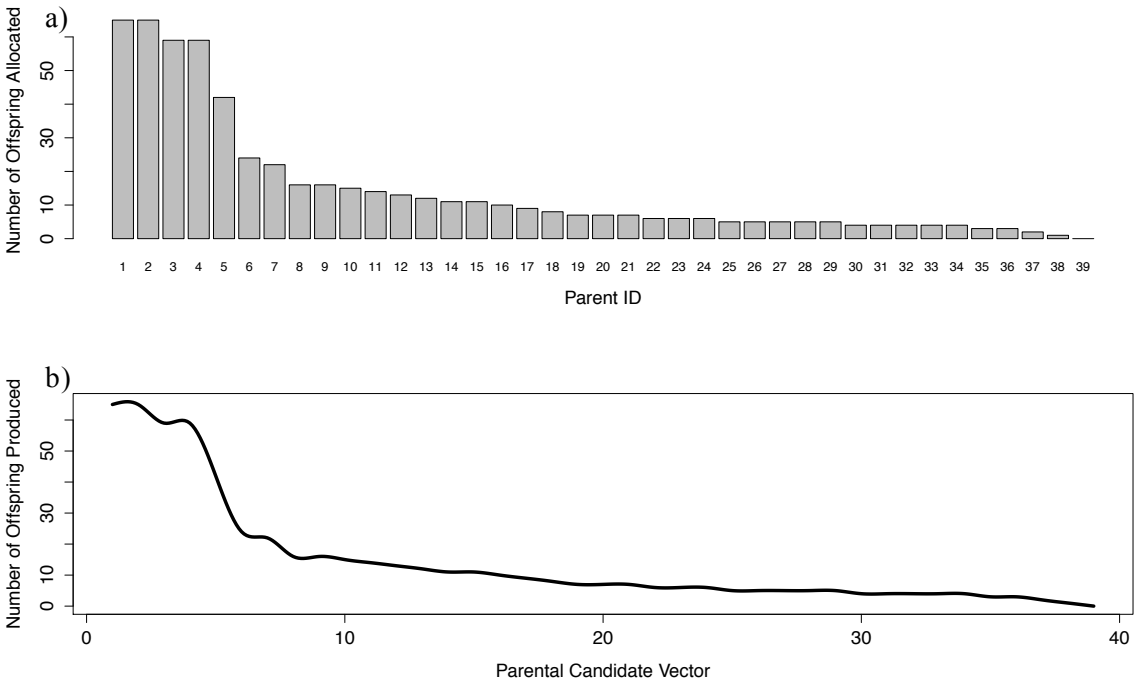


Figure 2.3: Boxplot (a) detailing per parent offspring contribution for a batch of 50 parents and 300 allocated offspring. Line chart (b) showing spline function derived from boxplot data in a).

For every occasion mating was simulated the likelihood of a parent being allocated an offspring was determined by the height of the spline at the position of the parent in a shuffled list of parents. For example, a parent randomly allocated a low ranking number below five has a very high likelihood of being the parent of a given offspring in comparison to a parent allocated a high number above 30. Parents were equally spaced in the parental candidate vector.

2.2.3. Breeding Programme Simulation

A simulation was written to simultaneously estimate expected genetic gain, inbreeding, bio-economic feasibility and parentage assignment accuracy under the proposed breeding programme. The whole simulation was replicated 100 times. Genetic gain proceeded as follows, distributions of the genetic (G) and environmental (E) components of the phenotype were simulated with mean (μ) values as shown in the Equations 2.1, 2.2 and 2.3.

(2.1)

$$Phenotype\ Distribution\ Mean = \mu_P = \mu_G + \mu_E$$

(2.2)

$$\mu_G = \mu_p \times h^2$$

(2.3)

$$\mu_E = \mu_p \times (1 - h^2)$$

The SD used in the generation of distribution of genetic effects changed according to the change in variance over selective breeding generations. The variance for complex traits such as growth is a product of variation at a very large number of loci. This model, known as the infinitesimal model, satisfies observed normality found in quantitative traits as a result of bi-allelic loci. In simulations the environmental variance remained constant across the breeding programme, while the genetic variance changed in accordance to the Bulmer Effect (Bulmer, 1971). The Bulmer effect mathematically predicts the effect of selection on variance in a trait, due to linkage disequilibrium. Equations 2.4 and 2.5 below from Falconer and MacKay (1997) detail first the derivation of k from selection intensity (i) and truncation point (x) and then the predicted Variance (V_A*) after selection through the derivation of heritability (h^2), k and selected population variance (V_A).

(2.4)

$$k = i(i - x)$$

(2.5)

$$V_A^* = (1 - h^2 k)V_A$$

Values for the truncation point and selection intensity, according to the proportion selected, were obtained from tables in Falconer and MacKay (1997). The genetic variance was calculated per generation according to this equation. At the end of each generation simulated, the heritability was recalculated, as the reduction in genetic variance causes a comparative increase in the environmental variance. The SD of the distribution of genetic effects was calculated as the square root of the variance after selection (Equation 2.6).

(2.6)

$$\sigma_G = \sqrt{V_A^*}$$

The SD of the distribution of environmental effects remained constant throughout the simulations, this value was calculated as a function of the starting phenotypic SD (σ_P) and the starting heritability (h^2) as Equation 2.7.

(2.7)

$$\sigma_E = \sqrt{(\sigma_P)^2 \times (1 - h^2)}$$

For the first generation of selective breeding (S0) the population mean was 400g with a SD of 100g. Individuals were randomly allocated phenotypic and environmental components from the distributions, and the individual phenotype was the sum of the two effects. A total of 10,000 offspring were simulated per generation of selection, parental contribution within tanks was determined using the spline as in section 2.2.2., contribution was equal between tanks to form the total growout. Offspring received a genetic value for the phenotype equal to the mean of the two parent's genetic values. Pre-selection was performed by truncating the highest 750 individuals from the growout. The final selection truncated the 200 individuals with the largest EBV. The values for these individuals were then recorded and the mean value calculated.

The 200 selected fish were allocated to one of four breeding tanks using an R script that attempted to minimise siblings sharing the same tank. Briefly, the script randomly allocated the fish to tanks, counted the number of individuals with a sibling relationship in the tank, recorded the value and then repeated the process a total of 1000 times. The tank layout with the minimum number of animals with siblings in the tank, was selected as the breeding design.

The mean phenotypic value of the selected fish was then used to generate new genetic and environmental contributions for the selected fish according to a normal distribution. This randomisation step was performed to avoid unrealistic selection on a small set of families over many generations. The genetic merit of a family will not remain constant over many generations, especially as the phenotype changes. Therefore, these simulations assumed that genetic merit for the trait was only inherited over a single generation. This mechanism underestimates gains and inbreeding in the system, but initial trials indicated that hereditable genetic merit across many generations results in biologically unrealistic parameter values.

Following allocation of breeding tanks and parameter randomisation, the growout was produced from the new parents for the next generation of selection. This cycle proceeded across 10 generations of selective breeding. Mean values for genetic gain at each generation were calculated.
Inbreeding was simulated as follows; for each replicate of the entire breeding programme a record of the pedigree for all fish was recorded; the first burn-in generation was designated the base generation and the coefficient of inbreeding was calculated for all individuals in the simulation using the R package 'Pedigree' (Coster, 2013); the mean inbreeding value was then calculated per generation.

Bio-economic simulations were performed on a monthly basis. For each month of 20 total simulated years the costs and sales were recorded. Only incomings and outgoings associated with the selective breeding programme were simulated to avoid complexities in simulating a large commercial production operation. The costs associated with the selective breeding programme were the growout

trial and the cost of selective breeding management and genotyping. Profit was calculated by subtracting the profit of a null distribution of unselected fish, against the profit from fish from the selective breeding programme. Under this derivation, zero profit would mean zero profit from undertaking the selective breeding programme. A cost-benefit ratio was calculated by taking the sum of yearly costs and income due to the selective breeding programme and dividing the income by the costs. This process was performed across all replicate independently.

Under the model 100 million fish per annum are produced under the selective breeding programme. The fish are harvested after 18 months of growth and only offspring from the most advanced generation of the selective breeding programme were used for production. The value of the harvest fish for each generation is according to a random distribution generated with mean and SD values according to the genetic gain simulation.

There are many underlying assumptions of the model. Firstly, it was assumed there were no practical constraints to the programme. For example, it was assumed that current facilities had sufficient space for selection and that offspring could be transported to production facilities. It was also assumed that the selective breeding programme did not require any new equipment purchases, and that the goals could be achieved using current facilities. The output of production was kept level across the breeding programme, it was assumed that the costs and profits scaled linearly, and that no discounts were applied for the purchase of extra feed or sales of produce. Finally, it was assumed that the broodstock fish could provide sufficient offspring for production.

Parentage assignment was performed on the simulated data to evaluate the utility of a low density (100 SNP) panel of genetic markers. The parentage assignment software COLONY v2.0.6.2. (Jones and Wang, 2010) was used in all cases. The pedigree file for each run of the simulation was subset into a file containing only parents (n=200) and pre-selected fish (n=750) for each of the selected generations. SIMPED (Leal et al., 2005) was used to simulate 100 bi-allelic markers with a mean MAF of 0.49 and a SD of 0.2. COLONY input files were created using the parameters as Appendix C with the number of runs altered to 1 to decrease computation time. A total of 10 COLONY files were run per entire run of the simulation, for a total of 1000 COLONY runs.

2.2.4. Statistical Analyses

All statistics were performed in R v3.2.2. A Shapiro-Wilk test was performed on all variables if statistical tests assumed normality. All data was statistically normally distributed unless reported otherwise. Linear regression was performed on genetic gain and inbreeding data. In both cases, generation was transformed into a numerical variable, as a proxy for time. In all cases Q-Q and residual plots were examined for violations of test assumptions. All assumptions were met unless reported otherwise.

2.3. Results

2.3.1. Predicted Phenotypic Gain for Proposed Selective Breeding Programme

Over all 100 replicates the mean phenotypic size at harvest increased as the selective breeding programme proceeded. Figure 2.4 shows the results for a single replicate across the progress of the breeding programme.

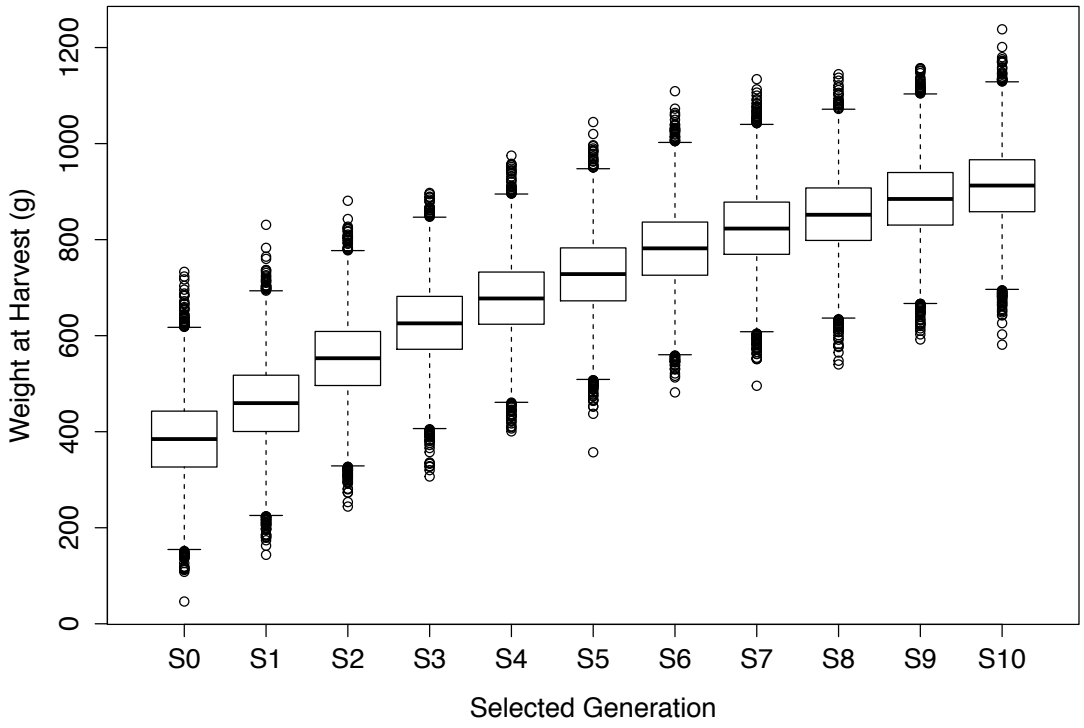


Figure 2.4: Boxplot detailing weight at harvest of 10,000 production fish for each generation of the selective breeding programme.

Figure 2.5 shows the mean weight at slaughter and 95% confidence intervals for all 100 replicate runs. Table 2.2 details the difference between mean harvest weights of adjacent generations. There was a significant (R^2=0.923, p > 0.001) positive relationship between generation and mean weight at harvest under the proposed breeding programme. There is a predicted increase of between 23.1g – 105.2g in final weight, corresponding with a doubling of average weight at slaughter in between 6 and 7 generations of selection. The increase in mean weight at harvest peak between S0-S1 and is lowest in the final generation simulated. Overall the increase in gains corresponds with an average per generation gain of 11.9 ± 6.5%.

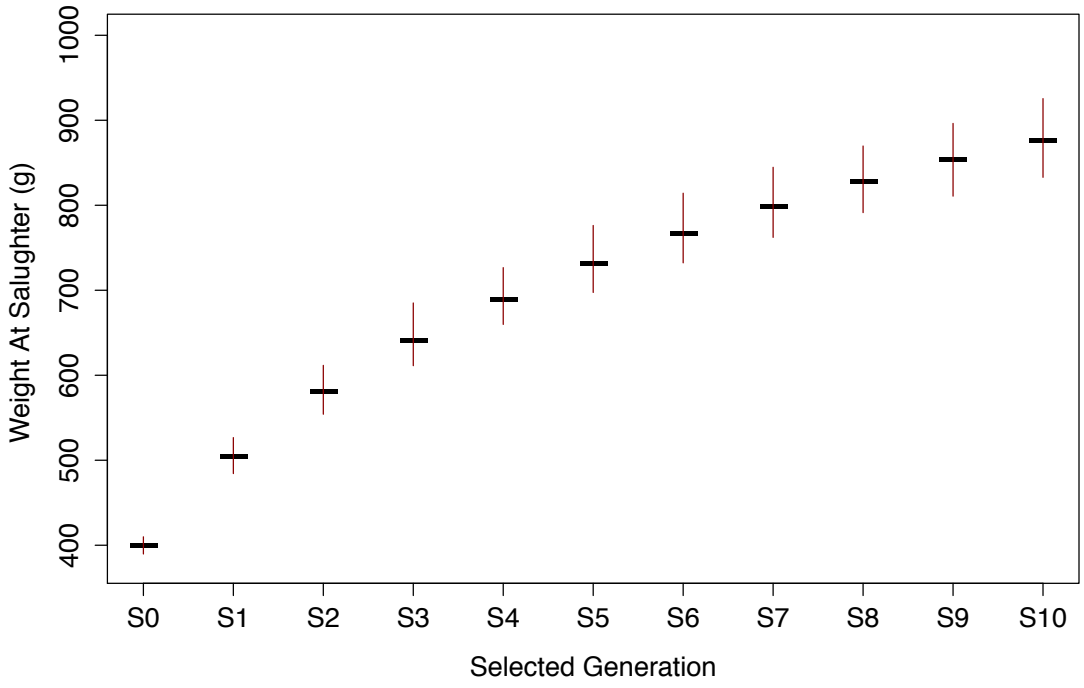


Figure 2.5: Plot detailing mean weight at harvest of 10,000 production fish over 100 replicate stochastic simulations of the selective breeding programme. The red lines correspond with 95% confidence intervals for the mean at each selective breeding generation.

Table 2.2: Table detailing mean increase of weight at slaughter and percentage increase against initial weight over 10 generations across 100 replicate stochastic runs of a selective breeding programme.

Generation	**Mean Increase in Weight at Slaughter (g)**	**Percentage Gain Per Generation relative to Initial Weight (%)**
S0_S1	105.19	26.30
S1_S2	75.65	18.91
S2_S3	60.79	15.20
S3_S4	48.18	12.05
S4_S5	41.39	10.35
S5_S6	36.24	9.06
S6_S7	31.8	7.95
S7_S8	28.55	7.14
S8_S9	25.75	6.44
S9_S10	23.06	5.77
Mean	47.66 ± 26.11	11.92 ± 6.53

2.3.2. Predicted Inbreeding for Proposed Selective Breeding Programme

The mean inbreeding coefficient across the progress of the entire selective breeding programme is detailed in Figure 2.6. Inbreeding is also shown for the wild and 'current' breeding schemes. There is a significant positive correlation between generation time and average inbreeding in both the 'current' (R^2=0.968, p > 0.001) and proposed (R^2=0.859, p > 0.001) schemes. The 'current' scheme has an average inbreeding increase of 2.55% per generation compared to the average increase in inbreeding of 1.24% found in the proposed breeding programme.

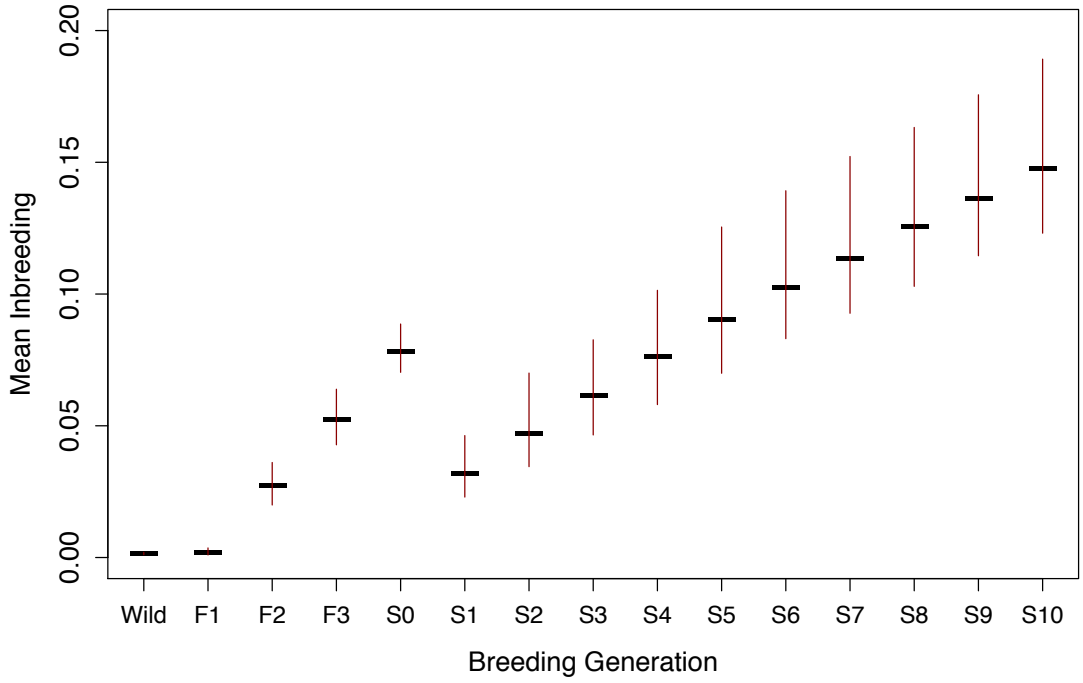


Figure 2.6: Plot detailing increase in inbreeding over the progress of the selective breeding programme. The plot starts with the inbreeding of the final wild caught generation, follows inbreeding of the 'current' system and then examines the inbreeding under the new proposed system. The black horizontal lines indicate mean inbreeding over the 100 replicates. The red vertical lines indicate 95% confidence intervals.

2.3.3. Predicted Profitability for Proposed Selective Breeding Programme

The bio-economic simulation predicted that the selective breeding programme would produce an overall profit in the fifth year. This corresponds with the first harvest of fish from selective breeding. However, the scheme loses a mean total of €360,117± €386 in the first 4 years of the selective breeding programme. At the 10th year of the programme, selective breeding is predicted to add a mean of €16,154,702 ± €1,509,956 of additional profit compared to the same scheme with no selective breeding. Cumulatively over the first 10 years of the scheme selective breeding is predicted

to add a mean profit of €76,087,456 ± €1,039,155 against the same scheme with no selective breeding.

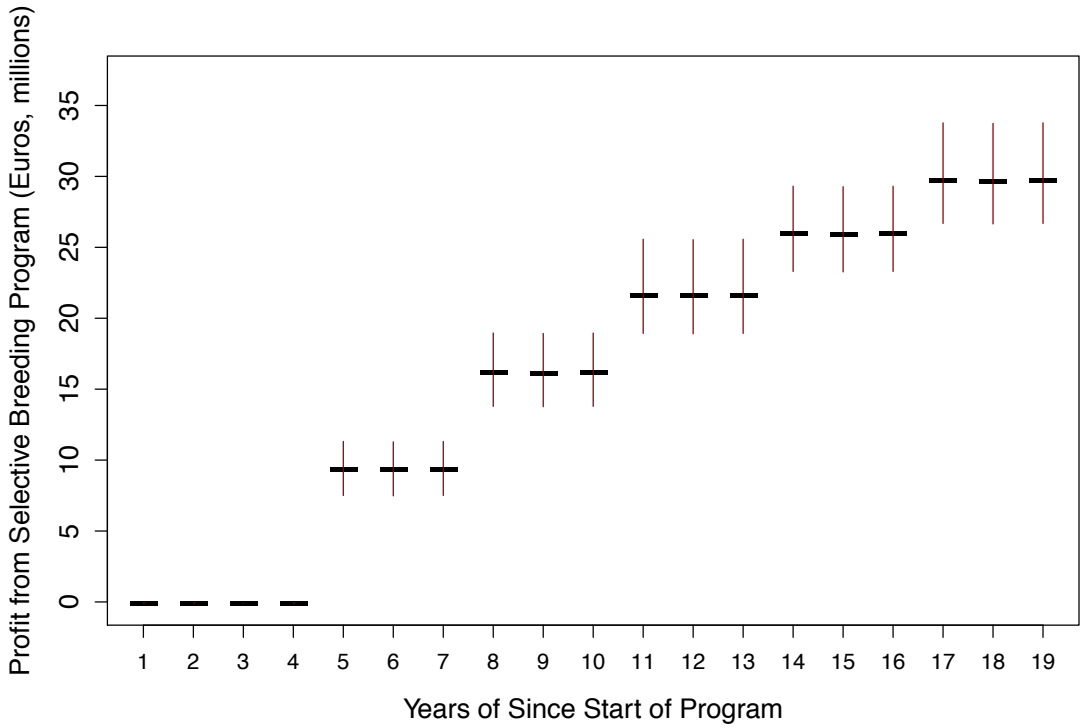


Figure 2.7: Plot detailing profit from selective breeding programme over time across 100 replicate runs of stochastic simulation. The black horizontal lines correspond with mean profit across the year and the red vertical lines correspond with 95% confidence intervals.

The mean cost-benefit ratio after the selective breeding programme was profitable was between 1:83 (year 7) and 1:355 (year 20). The years in which the cost of growout is paid show a decrease in cost-benefit ratio. See years 6,9 and 12 in Figure 2.8 for an example.

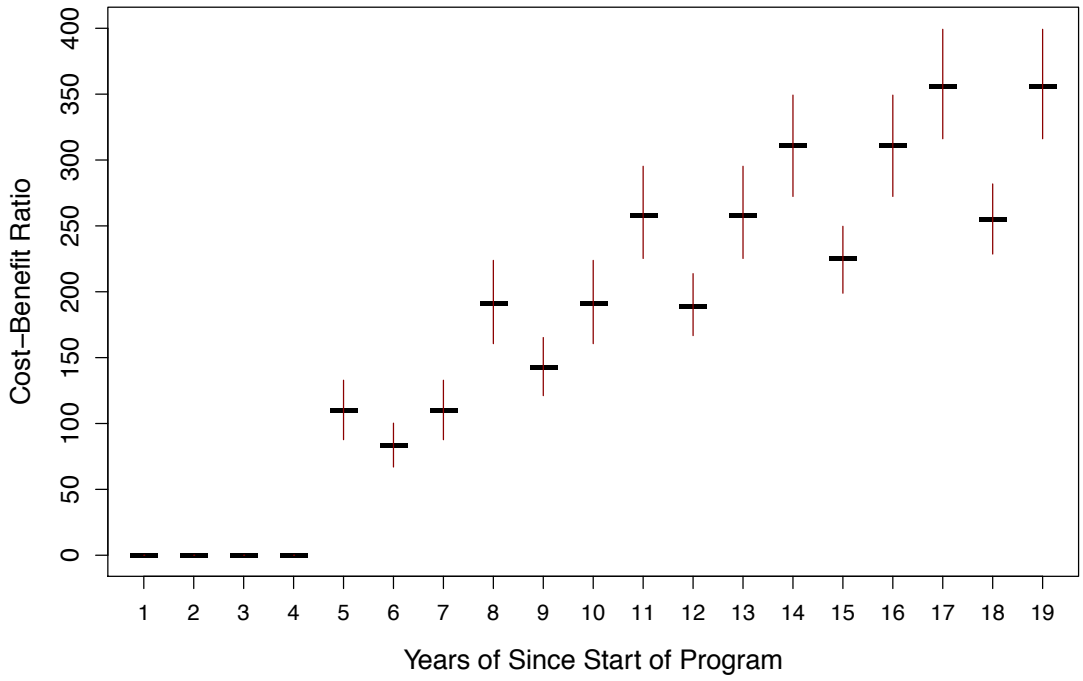


Figure 2.8: Plot detailing cost-benefit ratio from selective breeding programme over time across 100 replicate runs of stochastic simulation. The black horizontal lines correspond with mean cost-benefit ratio for the year and the red vertical lines correspond with 95% confidence intervals.

2.3.4. Parentage Assignment Accuracy for Proposed Selective Breeding Programme

A total of 100 parentage assignment runs were performed out of 1000 (10 per replicate, 100 replicates) due to computational constraints. The mean accuracy of parentage assignment across all runs was 99.995% There was no statistically significant ($F=0.094$, $p = 0.760$) difference between group accuracy means when data was grouped by generation. Overall accuracy was 100% for the majority of runs with 4 runs incorrectly assigning 1 parent (99.933%) and 2 runs incorrectly assigning 2 parents (99.867%).

2.4. Discussion

2.4.1. How Realistic are the Predicted Gains?

Overall the projected gains of around 12% per generation under the proposed selective breeding programme are in line with realised gains in growth in other aquaculture species. Kause et al. (2005) demonstrated a mean increase in growth of between 4.8-12.5% per generation in 5 generations of selective breeding in Rainbow Trout. Neira et al. (2006) found a mean increase of 13.9% weight per generation in 5 generations of selective breeding in Coho Salmon. Finally, Boliver and Newkirk (2002) found a mean increase of 12.4% body weight per generation could be maintained across a total of 12 generations in Nile Tilapia.

While the simulation results broadly agree with empirical expectations, some aspects of the predicted gains are unrealistic. In these simulations the BLUP breeding values for individuals were simulated, rather than explicitly calculated, due to technical and computational constraints. The estimation of BLUP requires many parental records and a large pedigree for accurate results. The accuracy of the BLUP-EBV over the breeding programme would initially be very low and then increase as pedigree records grew, in turn providing better prediction power. Simulations of BLUP-EBV accuracy indicate that accuracies can be as high as 0.53 in typical sib based aquaculture systems (Nielsen et al., 2009). The simulations will therefore underestimate the EBV accuracy in latter generations as under the simulated scheme the accuracy can never exceed the heritability. In turn this will underestimate expected gains.

An additional inaccuracy in using a fixed EBV accuracy is found in late generations due to the cumulative decrease in heritability. This phenomenon is a result of the Bulmer Effect and would cause an overall decrease in the accuracy of EBV, something not found in simulations. The overall effect of this is an overestimation of genetic gain in latter generations. An idealised selective breeding programme would balance these two variables switching to BLUP-EBV selection from mass selection once BLUP accuracy became high enough. The simulation of this trade-off is beyond the scope of this work. However, the similarity of the predicted and published realised gains indicate that the simulation of EBV estimation has utility in predicting phenotypic gains, possibly due to the two previously noted inaccuracies having opposing effects.

Identifying a single trait upon which to focus in selective breeding allows for rapid gains to be made. In reality it may be preferable to make smaller gains in many traits. Indeed, the use of selection indices for multi-trait selection has produced successful phenotypic gains across multiple traits in Atlantic Salmon selective breeding programmes (Friars et al., 1995; O'Flynn, 1999). However, there may be unforeseen consequences of selection upon one trait, and advantages to focussing on a single trait. For example, experiments trialling selection simultaneously on growth and disease resistance in Pacific White Shrimp (Argue et al., 2002) found that multi-trait selection indices (70% growth, 30% disease resistance) produced gains in survivability, but an overall decrease in growth.

These kind of experiments demonstrate the difficulty in predicting the outcome of multi-trait selection experiments, in all but the most intensively studied organisms, where trait co-variation can be quantified. Simulations of gains in multi-trait selective breeding programmes are uncommon and are unlikely to be a useful tool in predicting response to multi-trait selection response.

2.4.2. What Level of Inbreeding is Acceptable?

The mean inbreeding between the 'current' and proposed breeding programmes decreases by approximately half. This reduction is due to the effect of the equal contribution to the growout by the breeding tanks and reduction of inbreeding by avoidance of sib-sib mating. The simulation predicts that the selective breeding design will have an increase in inbreeding of 1.23% per generation.

There is limited public data on the levels of inbreeding present in current selective breeding programmes, studies instead focus on the effects of different mean values on phenotypic traits of interest. Despite this impediment pedigree data in Neira et al. (2006) found a mean estimate of 1.75% per generation in two strains of selected Coho Salmon. Additionally, a mean value of 1.3% per generation was found across 3 strains of Rainbow Trout in Pante et al. (2001). No impact of phenotype was reported in either example at this level of inbreeding.

A common textbook value provided for an acceptable level of inbreeding in aquaculture selective breeding programmes is around 1% per generation (Gjedrem, 2005). This estimate originates from theoretical work of Meuwissen and Woolliams (1994), here it is estimated that an N_e of between 25 and 250 animals per generation will not cause any loss of fitness. Under the equation presented in Appendix B these values of N_e result in inbreeding values of 2% and 0.2% respectively. However, the assumption of random mating is often violated in selective breeding programmes. Additionally, inbreeding values are calculated as a mean of all individuals, two populations with identical mean inbreeding values for the population may constitute two very different situations in terms of individual level inbreeding (a generalised format of this problem is demonstrated in Anscombe (1973)). Therefore, care must be taken when interpreting values of ΔF calculated from N_e, especially in species with particularly unbalance reproductive output. Despite these theoretical concerns the effect of inbreeding on the proposed selective breeding programme is predicted to be negligible.

2.4.3. Profitability of Selective Breeding Programmes

Predicted yearly profits from the selective breeding programme increase as the breeding programme continues, with a diminishing increase in profits per generation of selectively bred individuals due to a decrease in phenotypic gains according to the Bulmer Effect. The cost-benefit ratios predicted in the proposed selective breeding programme are similar to those found in other stochastic simulations. Robinson et al. 2010a found ratios of 1:50 - 1:30 depending on the selective breeding design in a programme for Asian sea bass in a simulation of 20 years. Much larger ratios

were found in Robinson et al. 2010b, ratios of 1:150 – 1:350 depending on breeding design for Greenlip and Blacklip Abaone over a simulation of 30 years. There are very few details of empirical cost-benefit ratios, however an estimate of 1:15 for Atlantic Salmon in the National Breeding Programme of Norway has been provided in Gjedrem (2000). All the provided examples incorporate some form of capital expenditure for the establishment of the breeding programme. In this work no capital expenditure was simulated, due to the low number of functional broodstock fish. It is very likely that a commercial provider with sufficient facilities to produce the number of fish simulated in the commercial growout, will have existing broodstock facilities and that the selective breeding programme will represent a change in practice from unselected to selected fish. Overall the proposed selective breeding programme represents an economically sound and profitable endeavour for a commercial provider.

Chapter 3. The Discovery and Validation of a Low Density SNP Panel for Parentage Assignment in Atlantic Salmon.

3.1 Introduction

An explicit understanding of the relationship between individuals in a breeding programme is essential to achieving the improvement of selected phenotypic traits. Pedigree data is often used in the estimation of genetic parameters such as heritability or genotype-environment interactions. They are used routinely in the estimation of BLUP-EBV.

A pedigree is the 'gold standard' of relationship data and is a central piece of animal breeding programmes. In terrestrial animal breeding this data is tracked by physical tags. However, aquatic species have numerous and small offspring making tagging difficult. Additionally, over the course of their life-cycle many aquatic organisms have growth trajectories that scale many factors of ten, this makes the marking of individual juveniles difficult or impossible.

One solution is to raise offspring in separate enclosures until they reach a size suitable or tagging. This approach has been successful in Atlantic salmon (Gjedrem, 1991), blue tilapia (Zak et al., 2014), Progift red Tilapia (Thodesen et al., 2013) and giant tiger prawn (Krishna et al., 2011). However, raising whole family groups in separate enclosures is costly and requires as many individual enclosures as families in the breeding programme. Many breeding programmes feature hundreds of families and so this approach may not be appropriate. Additionally, the confounding effect of tank may be significant between individuals as small differences in early life can dramatically shift the growth trajectory of an individual in aquaculture facilities. This may in turn bias the results of selective breeding if they cannot be controlled for.

A better solution is to use molecular markers to reconstruct the pedigree of a mixed, unknown set of individuals. The technique uses the Mendelian inheritance of molecular markers to determine parentage of a group of unknown offspring from a set of candidate parents. Parentage assignment using molecular data in aquaculture can be broadly broken into exclusion and likelihood methods (cf. Jones et al., 2010). Exclusion methods rely on the exclusion of parental relationships based on Mendelian incompatibility to determine the parents of a given individual. However, exclusion methods depend on data with low error and in practice use a method that matches parents under a set number of mismatches. In comparison, likelihood methods incorporate population level allele frequencies to determine the likelihood of parental contribution based on Mendelian inheritance. While no one method is universally applicable in all systems, likelihood methods generally give more accurate answers in cases of low marker data (Herlin et al., 2007; Vandeputte and Haffray, 2014). Commonly used likelihood parentage assignment programs include PAPA (Duchesne et al., 2002), CERVUS (Kalinowski et al., 2007) and COLONY (Jones and Wang, 2010). Most empirical

comparisons between likelihood methods indicate that COLONY gives the most accurate parentage assignment in most scenarios (Hauser et al., 2011; Liu et al., 2015; Weinman et al., 2015).

The first documented use of molecular markers in parentage is found in Brody et al. (1981) where allozyme markers were used to reconstruct a pedigree of a small number of common carp families in Israel. The use of molecular pedigree reconstruction became commonplace in the 1990s with the advent of microsatellite genotyping (Herbringer et al., 1995). In their review of microsatellite applications in fish genetics, Chistikov et al. (2006) identified that high levels of polymorphism, small size and ease of use are the main drivers of uptake across applied life sciences. However, theoretical work suggested that SNPs may have utility in high throughput parentage assignment due to low error rates and easy transferability between laboratories (Anderson and Garza, 2005). Empirical comparisons between SNPs and microsatellites for parentage assignment proceeded in Sockeye salmon (Hauser et al., 2011), Chinese Rhesus Macaques (Ross et al., 2014), Black Tiger Shrimp (Sellars et al., 2014), African Penguin (Labuschagne et al., 2015) and, most recently, black-throated blue warbler (Kaiser et al., 2016). This growing body of work indicates that around 80-100 SNPs give accurate parentage assignment and provide more accurate parentage assignment in comparison to microsatellites in most examples. SNP markers are now routinely used for parentage assignment in commercial selective breeding programmes, and the publication of validated panels is becoming more frequent (Nguyen et al., 2014; Liu et al., 2015). Some limited work has explored the effect of a diminishing number of SNPs on assignment accuracy (Liu et al., 2015; Weinman et al., 2014), but these cases have been limited to a small number of tested panels.

The Atlantic Salmon (*Salmo salar L.*) is a teleost fish of socio-economic importance. Historically, wild populations were distributed across the North Atlantic Ocean from the North East coasts of North America to Western Europe and around the south coast of Greenland. However, habitat destruction means that most contemporary populations are a fraction of their historical size. Atlantic salmon are an important human food source and wild populations are unable to support the global demand. In response to diminishing wild capture, experiments in the establishment of a selective breeding programme began in Norway in the 1970s (Gjedrem, 1991). A family-selection programme was implemented over five generations resulting in a strain with improved growth, time until sexual maturity, disease resistance and superior product quality (Gjedrem, 2010). The success of this initial breeding programme boosted worldwide production and Atlantic Salmon is now farmed in Tazmania, Chile, Scotland, Ireland, Canada and The United States (FAO, 2014). Atlantic salmon aquaculture production has increased 7% per annum over recent decades with current production at over 2 million metric tonnes/annum (FAO, 2014). This vast production has been supported by advanced selective breeding biotechnology. Genomic resources include high-density linkage maps (Gonen et al., 2014; Lien et al., 2011), validated 6K (Lien et al., 2011), 132K (Housten et al., 2014) and 151K (Yáñez et al., 2016) SNP genotyping arrays, and a recently sequenced genome draft (Lien et al., 2016). However, public validated tools for parentage assignment in Atlantic salmon are

currently limited to microsatellite markers (O'Reilly et al., 1998; Norris et al., 2000). This work aims to develop and validate a panel of SNP marker for parentage assignment in Atlantic Salmon, evaluating the effect of a diminishing number of markers on assignment accuracy. A generalised workflow informed by the Atlantic salmon panel was developed for use in other strains and species.

3.2. Methods

3.2.1. Genetically Diverse Discovery Samples

To ensure utility across the Atlantic salmon industry, a range of samples were chosen for the initial SNP discovery. A total of 102 Atlantic Salmon samples were selected for analysis from three distinct strains. The AG strain is the product of a breeding programme dating back to 1970. The founders of this strain were sampled from 41 Norwegian rivers as detailed in Gjedrem et al (1991). The SB strain is a product of the Norwegain Bolaks and Jakta strains and was founded in 2000. The NU strain was founded in 2007 from wild fish sampled from a single river on North Uist, Scotland. All strains have been subject to artificial selection as a part of a commercial breeding nucleus. The AG, SB and NU strains contributed 40, 41 and 21 samples respectively to the experiment.

3.2.2. Known Pedigree Training Samples

To ensure accurate parentage assignment and achieve further filtering of the SNP panel a training set of samples were used. These consisted of known crosses of Atlantic salmon individuals from the SalmoBreed AS breeding programme (Bergen, Norway), and broadly unrelated to the discovery samples. A total of 95 individuals were used, this consisted of 8 sires, 10 dams and 77 offspring. Each family contained between 7-8 offspring. There were a total of 10 families with 2 sets of 2 families sharing a single father. Adipose tissue was sampled from the live parents and stored in 70% (v/v) ethanol. Entire fry offspring, between 15-25mm, were supplied in 70% (v/v) ethanol.

3.2.3. DNA Extraction

A sample of ~0.5g of fast skeletal muscle tissue was taken from each individual from the discovery populations, a maximum of 48 hours after slaughter. Tissue was stored at -20 until DNA extraction. 20-40mg of tissue was homogenized in SSTNE buffer (Pardo et al., 2005) with 0.1% SDS (m/v) and 50µg of proteinase K for a total of 3 hours at 55 ^{o}C. Proteinase K was denatured with a 15-minute incubation at 70 ^{o}C. RNase was added and the solution was incubated at 37 ^{o}C for 1 hour. Protein was precipitated by adding 5M NaCl. DNA was then recovered from the supernatant and precipitated with isopropanol. The pellet was then washed four times in cold 75% ethanol. Pellets were dissolved in 100µl of nuclease free H_2O and all samples were stored at -80 ^{o}C.

3.2.4. Restriction Site Associated DNA Marker Sequencing (RAD-Seq)

DNA from the discovery population was used to perform a single digestion RAD-seq experiment by Floragenex Ltd. (Portland, USA) as described by Baird et al. (2008). Briefly, DNA was digested using Sbf1 restriction endonuclease, then individual samples were barcoded using custom Floragenex adapters followed by PCR amplification of the fragments. A library was constructed through equimolar pooling of all post-amplification fragments, and sequencing proceeded across two lanes of Illumina HiSeq 2000 platform. Sequencing data was de-multiplexed and quality trimmed to 90 base-pairs using custom Floragenex scripts. The surviving 90 base fragments were mapped to the Atlantic salmon genome (Assembly: ICSASG v_1, Accession: AGKD00000000.3) using BOWTIE v.0.12.8 allowing up to three mismatches (Langmeid et al., 2012). SAMTOOLS (Li et al., 2009a) and custom Floragenex scripts were used for SNP calling and variants were output as a Variant Call format (VCF) file.

3.2.5. SNP Selection

The first applied set of filters aimed to retain only high-quality SNP variants. Error introduced during sequencing and bioinformatic SNP calling can result in false positive variants being identified. SNPs were selected using the following quality filters using VCFtools v.0.1.12b (Danecek et al., 2011): minimum 15x minimum sequence depth, Phred scaled genotype quality per sample of 20+, minimum of 90% of samples genotyped.

The second set of applied filters aimed to retain SNPs with favourable properties for parentage assignment. Theoretical (Anderson and Garza, 2005) and empirical (Weinman et al., 2014) work supports the use of highly polymorphic neutral SNPs, with minor allele frequencies of above 0.15. SNPs with low levels of polymorphism in the focus population will have low power to distinguish relationships. Additionally, SNPs that are in linkage disequilibrium are unsuitable for parentage assignment as they provide overlapping information on genetic relationships. SNPrelate (Zheng et al., 2012) implemented in R v3.2.2 was used to apply the following property filters. SNPs were discarded if they deviated from Hardy-Weinburg equilibrium (HWE) at a significance level of $p < 0.10$ within populations. This inflated significance level was used to avoid discarding SNPs in HWE that failed to meet $P < 0.05$ to due to sampling error, as the number of samples from each population was low. SNPs were then discarded if they had a minor allele frequency (MAF) of below 0.15 or above 0.85 across the entire dataset. In this case MAF was calculated against a reference so a MAF of above 0.5 was possible as minor refers to non-reference allele as opposed to less frequent allele. Finally, in pairs of SNPs that had a pairwise linkage disequilibrium correlation coefficient of $R > 0.46$ one SNP was randomly discarded from the pair. This correlation coefficient was used as it resulted in a dataset roughly a tenth of the size of the post-quality filter dataset.

The final set of filters in the discovery population aimed to retain SNPs with characteristics suitable for probe based SNP genotyping. Here only bi-allelic SNPs were retained, as tri-allelic markers are not suitable for probe based genotyping platforms. An important criterion in genotyping assay success rate is nearby variants causing failure of PCR. To minimise this error SNPs with another variant within ±50bp were discarded using the vcf-annotate tool in VCFtools.

The mapping reference used for the RAD-seq did not contain chromosome level data, after property filters were applied the SNPs were mapped to chromosomes. This was performed by extracting 1kb up and downstream of each of the remaining SNP from the ICSASG v_1 Atlantic salmon assembly (Accession: AGKD00000000.3) using the faidx tool in SAMtools v1.2 (Li et al., 2009a). The 2kb fragment was then aligned to the v2 Atlantic salmon genome (Assembly: ICSASG v_2, Accession: AGKD00000000.4) (Lien et al., 2016) using the Basic Local Alignment Search Tool (BLAST v 2.2.30+) (Camacho et al., 2009). The single best hit was retained for each SNP, and the chromosome for the hit was recorded. SNPs with no hits were discarded. Surviving SNPs were then formatted for assay design and ordered via the Fluidigm D3 portal for the design and ordering of SNPtype assays. Assays were ordered in three batches of 96, 45 and 40. Each order aimed to achieve a balance of SNPs in the final panel across all 29 Atlantic salmon chromosomes.

3.2.6. SNP Genotyping

The training samples were subject to DNA extraction and SNP genotyping as follows. 20-40mg of tissue per sample was lysed using 200µl 10% (m/v) Chelex 100 (Sigma-Aldrich, St Louis, USA) and 50µg of proteinase K. The lysis proceeded at 55 ^{o}C for 1 hour followed by 15 minutes at 70 ^{o}C. PCR template consisted of 1:100 dilution of lysis in distilled H_2O.

SNP genotyping proceeded using Fluidigm SNPtype assays on the Fluidigm EP1 platform according to manufacturer recommended protocol. The protocol begins with a multiplex PCR that increases the number of target region copies for all trialled SNPs, generating an enhanced template for each individual. This template is then diluted, and microfluidics are used to load a reaction well with enhanced template and a second set of fluorescent allele specific primers for a second PCR reaction. In all cases, the Fludigim 96.96 Dynamic Array was used which allows for 96 assays and 96 samples for a total of 9,216 reaction chambers.

The initial multiplex PCR consisted of 0.5µM of each forward and reverse primer for each SNP region (a total of 96 regions), 1x Qiagen (Hilden, Germany) Multiplex Mastermix, 1.25µl diluted Chelex digestion and H_2O for a total reaction volume of 5µl. Thermal cycling consisted of an initial denaturation at 95 ^{o}C for 15 minutes, followed by 14 cycles of 95 ^{o}C for 15 seconds, followed by 60 ^{o}C for 4 minutes.

The second PCR proceeded in microfluidic chambers within the Fludigim 96.96 dynamic array, exact reaction concentrations are proprietary and unknown. For each assay and sample a

reaction mixture was loaded into the dynamic array. The assay mixture consisted of 1.5μM of each fluorescently labelled allele specific primer, 4μM of reverse primer, and 1X Fludigim Assay Loading reagent in a total of 4μl loaded solution. The sample mixture consisted of 2.083μl of 1:100 diluted multiplex PCR product form the first step, 1x Agilent (Santa Clara, USA) Brilliant III Probe Mastermix, 1x Fluidigm SNPtype reagent, 1x Fludigm Sample Loading reagent and 1x ROX reference dye in a total of 5μl loaded solution. Thermal cycling proceeded using the SNPtype 96.96 v1 program on the Fludigim FC1 cycler. Following genotyping the Fludigim SNP Genotyping Analysis software was used to automatically call SNPs, using a K-means clustering algorithm at a threshold of 85. All SNP calls were confirmed manually.

3.2.7. Microsatellite Genotyping and Analysis

During the training phase it was necessary to incorporate microsatellite markers, to validate the relationships in a subset of samples. Microsatellite genotyping proceeded as follows. PCR consisted of 0.5U of Agilent Paq5000 DNA polymerase, 0.2mM of each dNTP and 0.2μm of each forward and reverse primer and 2μl 1:100 diluted chelex lysis template. The total reaction volume was 10μl. Thermal cycling proceeded with an initial denaturation at 95 ^{o}C, followed by 35 cycles of 95 ^{o}C for 20 seconds, 58 ^{o}C temperature for 20 seconds, followed by 72 ^{o}C for 30 seconds. A final elongation step was performed at 72 ^{o}C for 5 minutes. Each microsatellite PCR was performed in simplex before pooling and fragment analysis using an Applied Biosystems (ABI) (Foster City, USA) 3730XL DNA Analyser by The University of Dundee's DNA Sequencing and Services department (Dundee, UK). Pools contained 5 microsatellite regions each labelled with distinct ABI dyes. In cases where dyes overlapped, the fragment length was used to determine the loci.
Microsatellite regions genotyped are shown in Table 3.1.

Table *3.1: Flowchart detailing microsatellite regions used in genotyping including length of repeat, fluorescent dye used in multiplex, multiplex group and source literature.*

Name	**Repeat Length**	**Dye**	**Analysis Group**	**Source**
SsuD190	4	6FAM	G1	King et al 2005
SSsp2213	4	VIC	G1	Paterson et al 2004
SsspG7	4	NED	G1	Paterson et al 2004
SSsp1605	4	NED	G1	Paterson et al 2004
Ssa197	2	PET	G1	O'Reilly et al 1996
Ssa85	2	6FAM	G2	O'Reilly et al 1996
Ssa171	2	6FAM	G2	O'Reilly et al 1996
Sssp2210	4	VIC	G2	Paterson et al 2004
SSsp2216	4	NED	G2	Paterson et al 2004
SSsp2215	4	PET	G2	Paterson et al 2004

Microsatellite traces were analysed using Geneious microsatellite analyser plugin (Biomatters Ltd, Auckland, New Zealand). Results were output in comma separated values format.

3.2.8. Parentage Assignment

All parentage assignment runs were performed in COLONY. The method implemented in COLONY uses a simulated annealing algorithm wherein the likelihood of many millions of potential pedigree arrangements are compared. Over a predetermined number of iterations, the method heuristically finds an optimum assignment. COLONY is suitable for this work not only because of it's relative accuracy, but also because it is simple to perform many thousands of runs simultaneously, and the preparation of input files can be automated using R. The main COLONY parameters are detailed here, additional parameters can be found in Appendix C. The computationally intensive full-likelihood method was implemented in all cases, with three separate replicates per COLONY run. Medium precision and run length was chosen and the allele frequency was updated to reflect assignment during the progress of the run. These parameters were chosen as a balance between computational time and reliability. COLONY runs were parallelised using GNU Parallel to speed up computation (Tange, 2011).

3.2.9. Panel Training and Further Filtering of SNPs

The training samples were genotyped using all 181 ordered assays. Genotyping assays may give erroneous results due to sequence artefacts, such as null alleles or sequence homology (as reviewed in Pompanon et al. (2005)). Therefore, it is recommended to use a training phase in the

development of SNP panels for parentage. PLINK v1.07 (Purcell et al., 2007) was used to confirm Mendelian inheritance of the SNP alleles across the ten training families. SNPs that exhibited more than one Mendelian error were discarded.

A second approach was used to further train the panel, based on early observations that simply removing SNPs with Mendelian errors does not always increase the panel's accuracy. The objective was to use the known pedigree of the training samples to rank the SNPs according to their ability to contribute to an accurate parentage assignment. The method used is illustrated in Figure 3.1 and detailed briefly here. The surviving SNPs were randomly sampled into 1000 separate COLONY runs each contain genotypes for 45 SNPs of the training population. The results of these runs were tabulated using an R script, and the SNPs in each run noted. For each SNP a ratio was calculated between COLONY runs that contained the SNP reconstructing the pedigree with 100% accuracy, and number of runs with <100% accuracy. SNPs were then ordered according to the ratio. Finally, COLONY runs were created sequentially omitting the SNPs with the lowest ratios until the panel gave 100% accuracy. All COLONY runs created in this algorithm had parameters as Appendix C.

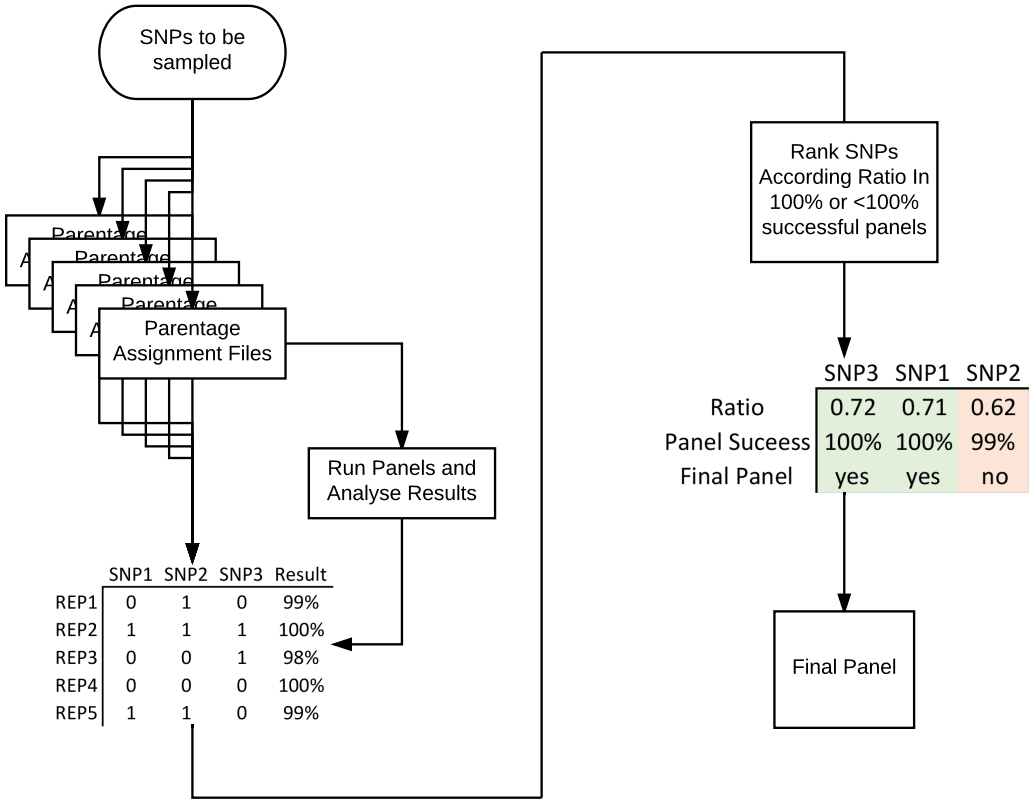


Figure 3.1: Flowchart detailing generalised SNP panel training method used to select an optimal subsample of variants. Shown example details an illustrative 5 replicate runs for 3 SNPs.

3.2.10. Number of SNPs Per Panel

In order to examine the effect of number of SNPs on panel success an R script was written to randomly subsample 100 panels of variable size from a dataset consisting of the training population genotyped at the final selected parentage panel. SNP panels were considered between 10 SNPs and 90 SNPs at intervals of 5 SNPs. As selective breeding programmes in aquaculture may have more than a hundred potential parents (Gjedrem, 2010), a second scenario was designed to test the ability of the SNP panel to assign parents in a situation with many candidate parents. A second dataset was created containing the training population and the discovery population. The R script used to subsample the first dataset was adapted to create COLONY runs that included the discovery population as both potential sires and dams.

3.3. Results

3.3.1. Sequencing and Mapping

Sequencing resulted in a total of 452.9 million reads, with a mean of 4.4 million reads per individual. Mapping resulted in a total of 56.7% of reads being mapped unambiguously to the reference sequence. All sequence data has been made public in the EMBL-EBI Short Read Archive (SRA) under the study accession PRJEB17687.

3.3.2. SNP Filtration

A total of 86,485 SNPs were identified before quality filtration. Following quality filters 17,283 SNPs were of sufficient quality to be considered as candidates. The property filters produced a total of 1517 SNPs that had suitable properties for use in parentage assignment. Following the property filters, SNPs were submitted for assay design and batches were selected from surviving SNPs to balance contribution across chromosomes.

SNP genotyping using the initial batch of 96 assays produced clear clustering of alleles for 54 assays (56.3%). In order to improve success for the following batches, an *in silico* validation step was performed where the primers, designed by the Fludigim software, were aligned to the ICSASG v_1 salmon genome assembly and hits reported. Only assays where all primers gave a single hit to the genome were considered as candidates in subsequent orders. This approach gave a greater success rate in the 2nd (33/45 -73.3%) and 3rd (27/40 - 67.5%) orders. The total assay success rate for the project was 111/181 (61.3%), with 23/181 (12.7%) of assays exhibiting poor clustering or unclear genotypes, and 47/181 (26.0%) exhibiting no clustering or discernible genotypes.

3.3.3. Mendelian Errors

The 111 assays that produced clear genotypes where then analysed for Mendelian errors. 16 Assays exhibited more than 1 Mendelian error in the training samples. Further examination revealed that the errors were unevenly distributed among families, with a single dam exhibiting 31 incompatible genotypes against her offspring. This dam and all putative offspring were subject to genotyping at 10 microsatellite markers, and Mendelian errors were detected in 7 microsatellites. The entire family was omitted from all further analyses. After omission of the erroneous family 16 assays still exhibited more than 1 Mendelian error.

3.3.4. Panel Training

A total of 95 assays remained after Mendelian error filters were applied. These were subsampled into 1000 COLONY runs which gave a mean parentage assignment accuracy of 99.03%, with a max of 100% and a minimum of 94.12%. After the calculation of ratios and sorting of runs, the omission of the single worst ranking SNP gave 100% run accuracy.

3.3.5. Final Panel

The final SNP panel consists of 94 SNPs distributed across 28 of 29 Atlantic salmon chromosomes and gives 100% accurate parentage assignment in the training population. Figure 3.2 details the minor allele frequencies for the final panel in the study populations. Primer information is provided in Supplementary Data.

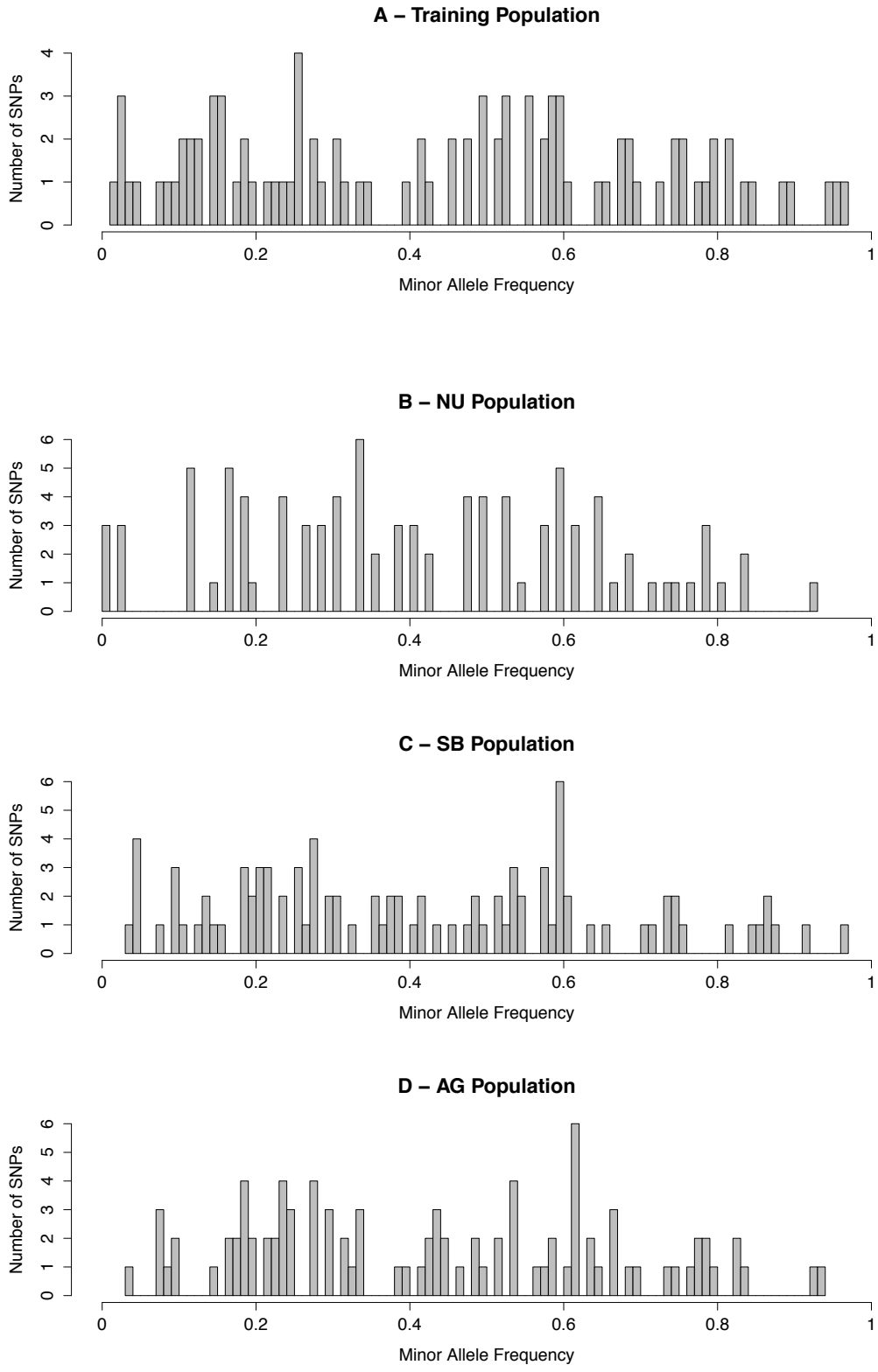


Figure 3.2: Histograms detailing minor allele frequency for the 94 SNP parentage panel across 4 study populations. A – The training population of 102 individuals, B-The NU discovery population of 20 individuals, C- The SB population of 41, D- The AG population of 40 individuals.

3.3.6. Variable SNP Panel Parentage Assignment

A total of 3,400 COLONY models were completed, comparing 17 different sizes of SNP panels, with 100 replicates per group in two different sets of samples. The distribution of assignment accuracy for the two sets of samples are shown below is shown in Figure 3.3 below

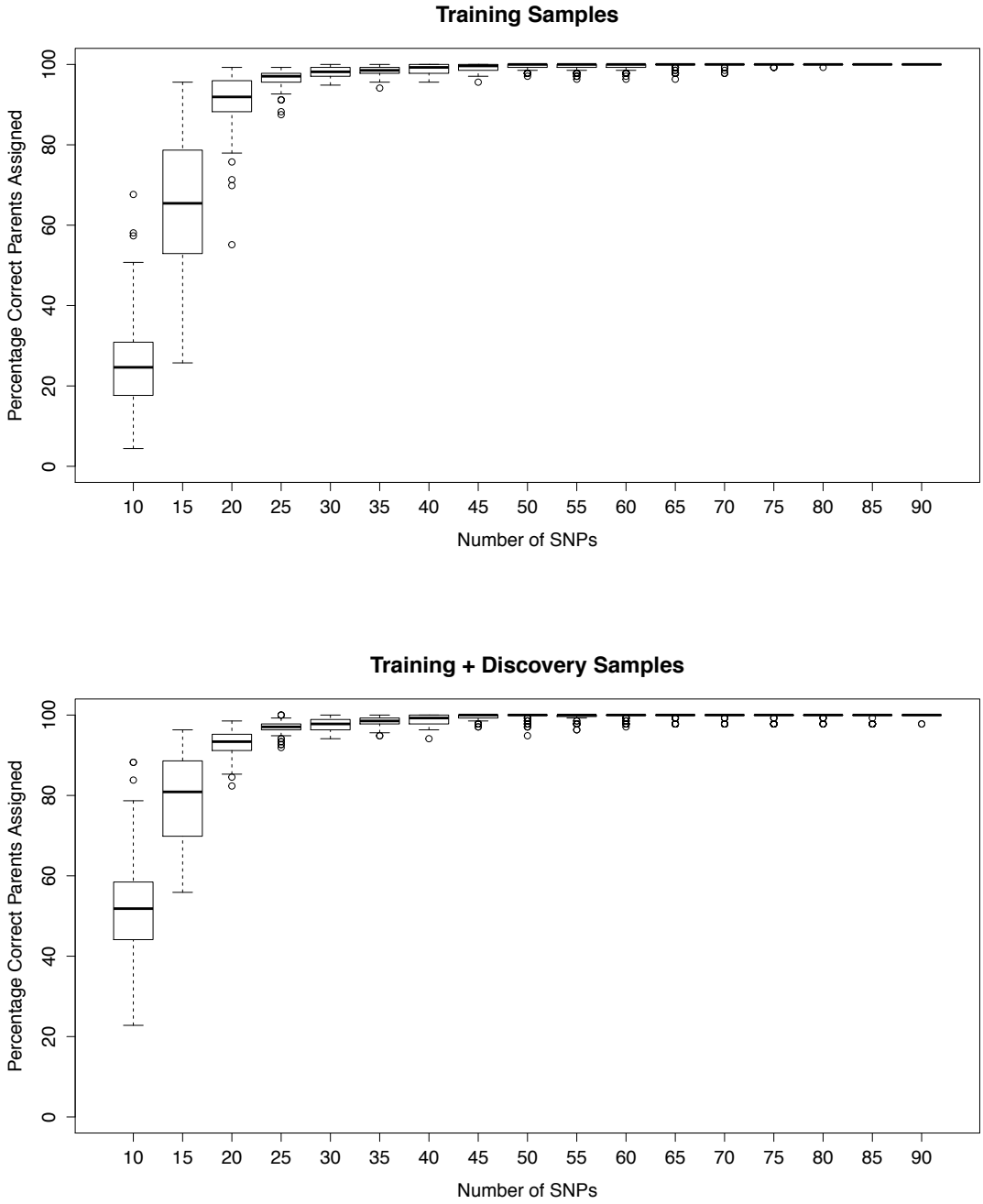


Figure 3.3: Boxplots detailing distribution of parentage assignment accuracy across randomly selected SNP panels of varying size, sampled from the final panel of 94 SNPs. The upper chart 'Training Samples' details the distribution for the results from the set of 85 training samples, made up of 17 parents and 68 offspring. The lower chart 'Training and Discovery Samples' details the distribution from the 85 training samples and the 102 Discovery samples provided as both potential Sires and Dams for a total of 289 individuals, made up of 221 parents and 68 offspring.

The trend in both sample groups is a wide range of parental assignment accuracy between panels sized 10-30 SNPs, followed by an asymptotic region from 30-90 SNPs. In the case of parental assignment, we are interested in how many SNPs are required to get 100% accuracy. The 95% confidence intervals derived from the 100 replicates provide an indication of the number of SNPs required to reliably get 100% accuracy. In the Training samples the lower 95% confidence interval reaches 100% at 90 SNPs, in the Training and Discovery samples the lower 95% confidence interval reaches 100% at 75 SNPs.

3.3.7. Generalised Workflow for the Selection of SNP Parentage Panels

A workflow for the filtration and selection of SNPs for parentage assignment is shown in Figure 3.4 below. The workflow uses a set of genetically diverse discovery individuals for initial filtering and then validates SNP genotyping assays in a second independent set of samples. The workflow begins with quality control filters to minimise false positive variants in the discovery dataset. Following quality filtration, SNPs are filtered using a property filter to retain SNPs with desirable qualities for parentage assignment. Assays are designed and only SNPs that meet assay design criteria are retained. The second set of samples are genotyped at the surviving SNPs and only assays that produce clear, unambiguous calls are retained. The pedigree data is then used to detect Mendelian errors and SNPs with greater than 1 error are discarded. Finally, the SNPs are ranked according to their ability to provide good parentage assignment, and the panel is truncated until the pedigree from the training samples is 100% accurately reconstructed.

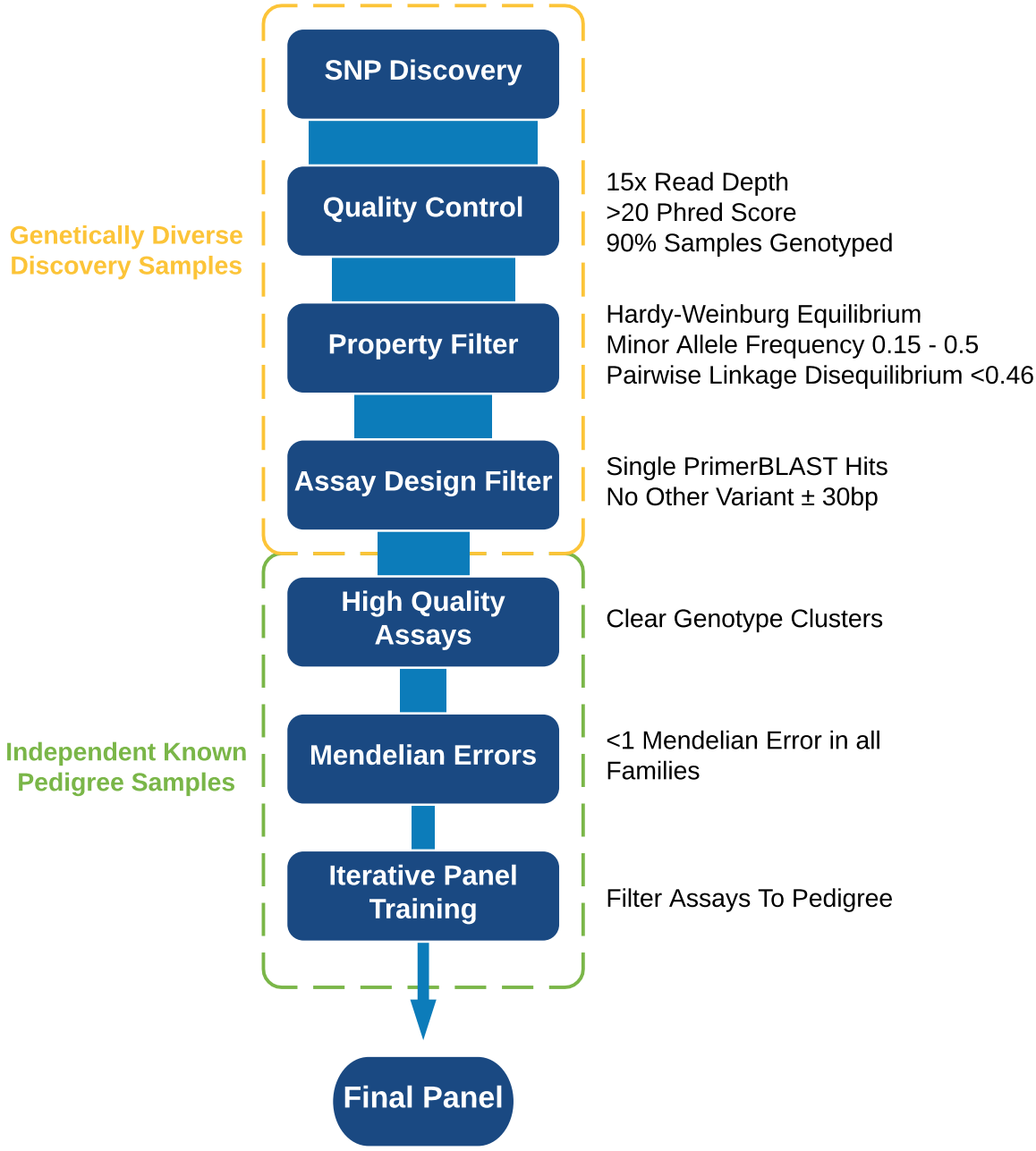


Figure 3.4: Flowchart of the workflow for the development of a SNP panel for parentage assignment.

3.4. Discussion

3.4.1. Sequencing

The number of reads and proportion of reads mapping is in line with published estimates in other species (Glazer et al., 2015; Hohenlohe et al., 2010). Furthermore, the number of recovered SNPs is approximately similar to other work using *Sbf1* restriction enzyme in Atlantic salmon RAD-seq (Houston et al., 2012).

3.4.2. SNP Assay Conversion Success

The success in the conversion of *in silico* SNP variants into genotyping assays is highly variable, with studies giving values as high as >99.9% (Weinman et al., 2014) and as low as 48% (Sánchez et al., 2009). The total conversion rate found here of 61.3% is in line with published estimates, but highlights the importance of a number of variables contributing to the proportion of validated SNPs. The source of the SNPs is a parameter that contributes to SNP conversion success. Discovery methods that produce high rates of false positive will give high SNP conversion failure. In this work, the use of RAD-seq for SNP discovery allowed for the discovery of many novel common variants among the diverse discovery population. However, the conversion of variants from SNP array technology to Fluidgim SNPtype assays demonstrated a much higher SNP conversion success rate (79.7%) in Rainbow trout (Liu et al., 2015) compared to the success rate found here. The development of the 57k SNP array used in Liu et al. (2015) resulted in the conversion failure of around 14% of the source SNPs (Palti et al., 2015). It is likely that this contributed to the higher success rate of the conversion of SNP array derived variants to functional SNPtype assays.

Another key variable in the validation of assays is the study species. All salmonid species underwent a whole genome duplication (WGD) around 88 million years ago (Macqueen and Johnston, 2014), which has resulted in a large number of retained paralogous regions (Lien et al., 2016). This can complicate SNP discovery and give rise to false positive variants in SNP datasets (Etter et al., 2011). It is likely that the SNP conversion success rate in this work was affected by complexities in the Atlantic salmon genome. Finally, the choice of genotyping assay may also have had a strong effect on the SNP conversion success. This effect may be due to differences in propriety primer design software, PCR cycling conditions, choice of fluorescent marker, or other methodological attributes that may have an influence on efficiency. For example, the SNP to assay conversion success, via the Agena Bioscience (Hamburg, Germany) MassARRAY system, ranges from 80.9% in Atlantic salmon (Freamo et al., 2011) to >99.9% in superb starlings (Weinman et al., 2015).

Very little work has attempted to discuss the relative contribution of these factors in the role of SNP assay validation. Humble et al., (2016) makes some attempt to assess the factors influencing

SNP validation quantitatively, focussing on practical methods to increase SNP validation. Notably the authors suggest the use of sequence alignment of primer and probe sequences to improve success rates, a method empirically supported in this work. There is an urgent need for studies to explore the effect of the aforementioned factors on SNP conversion success rate in a targeted intentional study.

3.4.3. Mendelian Errors in Samples of Known Pedigree

The exhibition of Mendelian errors in genotyping results may simply represent error in genotyping, the causes of which are well covered in Pompanon et al. (2005). However, in the case of the dam with 31 errors against her offspring, it is more likely to correspond with human error. The unintentional crossing of mass spawning broodstock animals in aquaculture facilities has been documented in Morvesen et al. (2013). However, this kind of error is unlikely to occur in Atlantic salmon breeding programmes where crosses do not produce mixed family batches. The most likely cause of this error is a mislabelled sample. Attempts to track and recover the true parent of the offspring were unsuccessful. Whist SNPs carry many advantages over microsatellite markers, in this case the use of microsatellite makers provided clear and definitive evidence of error in the provided samples.

3.4.4. Parentage Assignment Success

The assignment accuracy found across different sized SNP panels broadly agrees with other published results. Liu et al. (2015) found decreasing accuracy with 95 (100%), 68 (100%), 48 (99.2%) and 36 (92.5%) SNP sized parentage panels in Rainbow trout. Weinman et al. (2015) found similar results with decreasing accuracy with 10 different sized panels, seeing 100% accuracy with 102 SNPs and below 80% accuracy with ~35 SNPs. Directly comparing these values to this work is problematic as, especially at a low number of SNPs, there is high variance between panel success. There is still no consensus on the number of SNP markers that provide sufficient power in most cases. Vandeputte and Haffrey (2014) suggest a number between 100-450, while recent empirical work suggests 95 (Liu et al., 2015) or 97 (Kaiser et al., 2016) SNPs give 100% accuracy. These values may reflect a number that satisfied requirements for genotyping platform (such as 96 assays for a Fluidigm 96.96 chip) or a number that provides good results in the number of trailed samples. Overall the entire SNP panel of 94 variants provided here is expected to give accurate parentage assignment in most situations.

An unexpected finding is that including the discovery population in parentage assignment runs increased the accuracy of the assignments. The difference in parentage assignment accuracy between the training dataset, and the joint training and discovery dataset, is due to the effect of adding the discovery parents on the allele frequency estimation in COLONY (Jinliang Wang, Personal Communication, 4th December 2016). The likelihood of a given parentage assignment in COLONY depends on the calculated allele frequency of both offspring, and all parental candidates. Adding a

large number of individuals from populations unrelated to the training population, means the calculated allele frequencies deviate significantly from the true frequencies of the training population. This effect means that the individuals from the training population appear more related, in comparison to the same calculations of allele frequency for a run containing only the training population. The increase in calculated relatedness means that more correct assignments are made in the training population when discovery individuals are included. In this case, this effect results in a strong positive effect on the parentage assignment accuracy. However, in cases where there are many more parental candidates from the same population, the addition of individuals from other populations as parental candidates will result in decreased accuracy, in comparison to a run containing only the individuals from the same population. Further work is required to examine the effect of a large number of closely related candidates on parentage assignment accuracy. The evaluation of parentage panels should ideally replicate pedigrees found in the selective breeding programme as closely as possible.

3.4.5. The Formalisation of a Workflow for Parentage Panel Design

The generalised workflow presented here is the first example of any formalisation in the methods in the production of SNP panels for parentage. There are a growing number of validated SNP panels for parentage with disparate selection criteria and methods. While this method will not be applicable to all users, it can easily be adapted to suit different needs. In this case, the discovery and training samples were separate, but the two sample sets may be combined (cf. Liu et al., 2015). However, it may be difficult to acquire sufficiently large sample sets, with pedigree data, from many different populations. Additionally, there are a growing amount of publically available resources that provide sufficient data for the discovery phase, allowing practitioners to perform the initial workflow steps *in silico* before trialling their own populations.

The update of the genome assembly in the progress of this work allowed further investigation into the physical location of SNPs on the genome. In a case where a chromosome anchored assembly is available at SNP discovery, the pairwise linkage disequilibrium filter should be exchanged with a tool considering only physical distance. Since the length of LD can be estimated it would be simple to truncate the SNPs according to a minimum distance under which LD would be negligible.

Finally, the iterative training part of the workflow was designed with a number of SNPs per subpanel that gave intentionally imperfect results. It may be necessary to change this value as more or less individuals are included in the training panel. The principle is to ensure imperfect panel accuracy to allow for the ranking of SNPs based on variable success.

3.4.6. Genomic Decay in SNP Parentage Panels

The economic investment to develop a validated SNP parentage panel is significant. It is therefore important to have an estimate of how genomic decay might affect the ability for the panel to

remain accurate over time. In order to assess the decay of parentage assignment accuracy over time, a poor-quality parentage assignment was designed as a 'worst-case' scenario. Methods are detailed in Appendix D. As shown in Figure 3.5 in a 'worst-case' scenario the decay of low density SNP panels over the course of a selective breeding programme results in a significant ($F(8,891)=16.86$, $p > 0.001$) loss of accuracy. This loss is equivalent to 2.05% between the first and last generation.

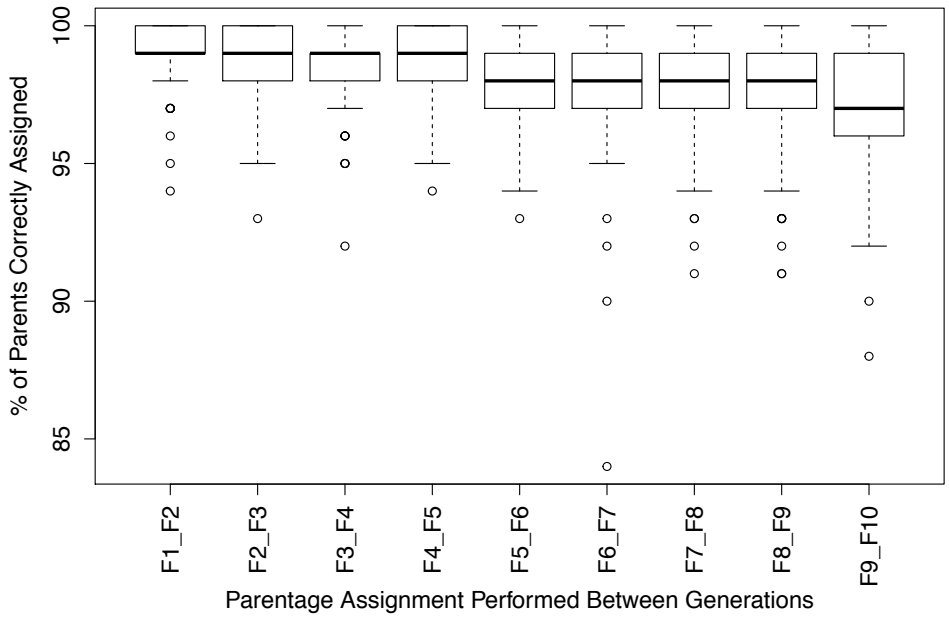


Figure 3.5: Boxplot detailing accuracy of parentage assignment across 100 replicate runs per between-generation parentage assignment in a simulated 'worst-case' scenario.

There is some information about the effect of decreasing parentage assignment accuracy on the goals of selective breeding programmes. Israel and Weller (2000) found that 10% incorrect parentage assignment gives 4.3% less genetic gain in a trial with simulated data. Meanwhile, Banos et al. (2001) simulated 11% parentage assignment error in empirical data showing a decrease of between 11%-18% of gains in a variety of milk related traits. However, as seen in Section 2.3.5., a more realistic SNP panel, with optimally selected variants used in mating designs found commonly in selective breeding programmes, exhibits no sign of genomic decay. Therefore, it is important to ensure the SNP panel is of high quality before its implementation into breeding programmes. This step should prevent any of the effects of genomic decay shown above.

Chapter 4. Shellfish Trait Standards

4.1 Introduction

The culture of molluscan species accounted for 23.6% of the worlds aquaculture output by weight in 2014 (FAO, 2014). Over 90% of this impressive production is based in Asia, and the main groups produced are clams, oysters, scallops, abalones and mussels (FAO, 2014). Despite major production, invertebrate genomes are generally less well understood in comparison to those of vertebrate species. Recent genome drafts of the Pearl (Takeuchi et al., 2012) and Pacific oyster (Zhang et al., 2012) have highlighted the complexities in molluscan genomes. A review (Astorga, 2014) examining the current state of mollusc knowledge in aquaculture noted a clear disparity between the reported total production, and available genomic resources. Since this publication, the amount of mollusc sequence data submitted to public databases has vastly increased (see Figure 4.1). However, the total data available is still small in comparison to the fish species.

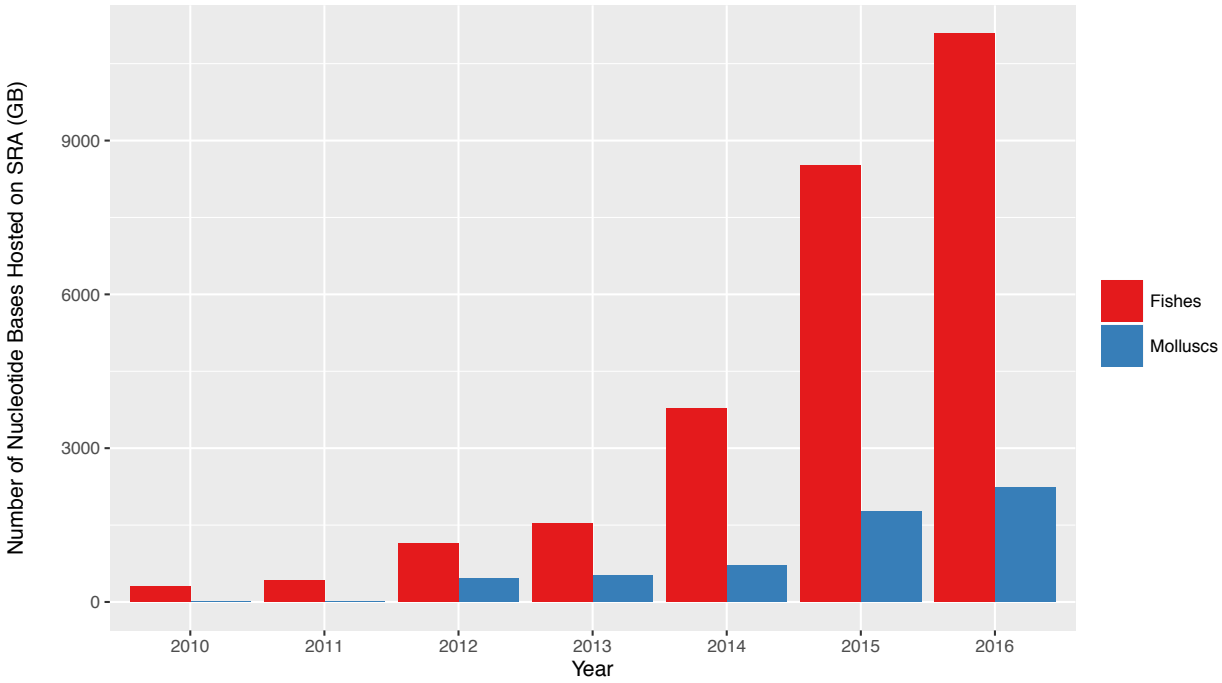


Figure 4.1: Bar chart detailing number of sequenced nucleotides hosted on the NCBI short read archive (SRA) over 2010-2016. Data is shown for the top 10 Fish and Mollusc species by aquaculture production according to FAO (2014).

A great deal of the nucleotide data produced has been in connection to the discovery of quantitative trait loci for desirable traits in mollusc species (cf. Yue, 2014). Additionally, cutting edge genomic

technologies, such as genomic selection, have begun to be adapted into aquaculture research of marine invertebrates (Dou et al., 2016).

The mapping of commercially relevant quantitative trait loci (QTL) to regions in the genome is the first step in the implementation of marker assisted selection in aquaculture breeding. The association of markers to the trait of interest relies on accurate and consistent trait measurements and careful control of confounding variables. Therefore, the measurement and reporting of traits, both desired and confounding, are good candidates for standardization. Additionally, once a marker or set of markers have been identified it is necessary to independently verify the association in separate stocks (Collard and Mackill, 2008). This may increase the total cost of the experiment and a sufficient number of independent individuals may not be available. One solution may be to verify the QTL marker using an existing trait database linked to genomic data. The current public nucleotide databases have minimal trait data reporting which prevents the use of public data for validation of new QTL markers.

This study aims to make the first steps in establishing a MAS database by taking advantage of the gap in molluscan information. The provision of a set of trait measurement and reporting standards for bivalve molluscs will hopefully result in greater data interoperability as aquaculture researchers increase their focus on the understudied group.

4.2. Methods

A sample dataset was created as a starting point in developing trait standards. Two hundred King Scallop (*Pecten maximus L.)* individuals were collected from a shallow (<30m) loch site (long/lat = 57.862237,-5.274639). A range of phenotypic data was collected as shown in Table 4.1.

Table 4.1: Table detailing collected phenotypic data for sample dataset of 200 King Scallops.

Trait	**Unit**	**Description**
scallop id	integer	unique integer identifying individual
total weight	grams	weight of entire organism
width	millimeters	greatest length parallel to the hinge
height	millimeters	greatest length perpendicular to the hinge
depth	millimeters	greatest length in z axis
soft tissue weight	grams	weight of organism without shell
skeletal muscle weight	grams	weight of skeletal muscle
smooth muscle weight	grams	weight of smooth muscle
wight of gonad	grams	weight of entire gonad
total weight of muscle	grams	weight of skeletal and smooth muscle
demoic acid	mg/kg	concentration of demoic acid in tissue homogenate
age class	winter rings	age class according to number of winter rings

The trait data was collected alongside tissue samples for DNA extraction. The aim was to emulate an association analysis with the aim of identifying QTLs for traits of economic importance. The data was combined with the M2B3 (Marine Microbial Biodiversity, Bioinformatics and Biotechnology) data reporting standards as detailed in Ten Hoopen et al. (2015). The M2B3 standards are a set of descriptors created as a minimum reporting standard for a marine microbial samples collected in the epipelagic zone. Each descriptor details a particular parameter related to the sample, such as collection location or sampling platform. These standards were compared in order to use existing infrastructure, and allow the shellfish standards to be implemented in the European Bioinformatics Institute (EBI) Web-In system, alongside nucleotide sequence submission. The M2B3 is implemented in the EBI Web-In system as a checklist of fields that are either optional, recommended or mandatory.

The standards were then filtered. An initial step filtered out the descriptors that applied only to microbial samples, such as size fraction of filter. The descriptors were then filtered for potentially redundant information, for example both the M2B3 and the trait standards contain descriptors on weight of sample. Finally all descriptors were designated optional, recommended or mandatory.

4.3. Results

4.3.1. Trait Descriptors

The final descriptors are detailed in Table 4.2 below.

Table 4.2: Information about a shellfish biological sample provided in conjunction with molecular data. A – Mandatory data to be provided with all samples, B –Recommended highly relevant data , C – Optional data relevant information to be provided if possible.

A

Descriptor Name	Descriptor Definition	Descriptor Requirement level	Descriptor Format	Example
*sample ID**	unique identifier for the sample	mandatory	Single-line text	lab barcode XY
*sample title**	a brief human readable description of the sample	mandatory	Single-line text	Sample obtained from the 9A progeny strain of parent strains 88 and 75. This sample has a biological replica XZ.
*organism scientific name**	scientific name of the organism	mandatory	NCBI Taxonomy ID	*Pecten maximus (taxid:6579)*
*sampling campaign**	refers to a finite or indefinite activity aiming at collecting data/samples, e.g. a cruise, a time series, a mesocosm experiment.	mandatory	Single-line text	MAS_EXPERIMENT_42.
*sampling station**	refers to the site/station where data/sample collection is performed.	mandatory	Single-line text	Loch Broom
*sampling platform**	Refers to the unique stage from which the sampling device has been deployed.	mandatory	Single-line text	Research Vessel Tara
*event date/time**	date and time in UTC when the sampling event started and ended, e.g. each CTD cast, net tow, or bucket collection is a distinct event. Format: yyyy-mm- ddThh:mm:ssZ	mandatory	Single-line text	2013-06-21T14:05:00Z/2013-06-21T14:46:00Z
*latitude start**	latitude of the location where the sampling event started, e.g. each CTD cast, net tow, or bucket collection is a distinct event. Format: ##.####, Decimal degrees; North= +, South= -; Use WGS 84 for GPS data	mandatory	Single-line text	-24.6666
*longitude start**	longitude of the location where the sampling event started, e.g. each CTD cast, net tow, or bucket collection is a distinct event. Format: ##.####, Decimal degrees; East= +, West= -; Use WGS 84 for GPS data	mandatory	Single-line text	-096.1012
*depth**	the distance below the surface of the water at which a measurement was made or a sample was collected. Format: ####.##, Positive below the sea surface. SDN:P06:46:ULAA for m.	mandatory	Single-line text	14.71
*protocol label**	identifies the protocol used to produce the sample, e.g. filtration and preservation	mandatory	Single-line text	BACT_NUC_W0.22-1.6
*environment biome**	biomes are defined based on factors such as plant structures, leaf types, plant spacing, and other factors like climate. Biome should be treated as the descriptor of the broad ecological context of a sample. Examples include: desert, taiga, deciduous woodland, or coral reef. EnvO (v 2013-06-14) terms can be found via the link: www.environmentontology.org/Browse-EnvO	mandatory	Single-line text	marine biome (ENVO: 00000447)
*environment feature**	environmental feature level includes geographic environmental features. Compared	mandatory	Single-line text	sea grass bed (ENVO: 01000059)

	to biome, feature is a descriptor of the more local environment. Examples include: harbor, cliff, or lake. EnvO (v 2013-06-14) terms can be found via the link: www.environmentontology.org/Browse-EnvO			
*environment material**	the environmental material level refers to the material that was displaced by the sample, or material in which a sample was embedded, prior to the sampling event. Environmental material terms are generally mass nouns. Examples include: air, soil, or water. EnvO (v 2013-06-14) terms can be found via the link: www.environmentontology.org/Browse-EnvO	mandatory	Single-line text	cobble sediment (ENVO: 01000115)
seabed habitat	classification of the seabed where the organism has been found; for European seabed habitats please use terms from http://eunis.eea.europa.eu/habitats-code-browser.jsp;	mandatory	Single-line text	B3.4 : Soft sea-cliffs, often vegetated
age	age of the organism the sample was derived from	mandatory	Single-line text	2 months
aquaculture origin	origin of stock and raised conditions, AO – Aquaculture origin WO – Wild origin AR – Aquaculture raised WR – Wild raised	mandatory	Single-line text controlled by a list of allowed values: AOAR,AOWR, WOAR, WOWR	WOAR
shellfish total weight	total weight of shellfish including shell at the time of sampling. Epifauna and epiphytes to be removed	mandatory	Single-line text	223g
shellfish soft tissue weight	total weight of all soft tissue, i.e. weight of entire organism without shell, at the time of sampling	mandatory	Single-line text	83g
shell length	length of shell (perpendicular to the hinge)	mandatory	Single-line text	123mm
shell width	width of shell (perpendicular angle to length)	mandatory	Single-line text	110mm

B

Descriptor Name	**Descriptor Definition**	**Descriptor Requirement Level**	**Descriptor Format**	**Example**
adductor weight	total weight of striated muscle and smooth muscle	recommended	Single-line text	33.2g
gonad weight	total weight of entire gonad tissue	recommended	Single-line text	6.7g
shell markings	visible markings on outer shell	recommended	Single-line text	Dark striations
toxin burden	concentration of toxins in the organism at the time of sampling	recommended	Single-line text	502mg/kg
marine region	the geographical origin of the sample as defined by the marine region name chosen from the Marine Regions vocabulary at http://www.marineregions.org/.	recommended	Single-line text	Adriatic Sea (MRGID:3314)

C

Descriptor Name	Descriptor Definition	Descriptor Requirement Level	Descriptor Format	example
sample collection device	the sampling device(s) used for the Event.	optional	Single-line text	Chain trawl
storage conditions (fresh/frozen/other)	explain how and for how long the sample was stored before DNA extraction	optional	Single-line text	-80 degree Celsius, 1month
sample health state	health status of the subject at the time of sample collection	optional	Single-line text controlled by a list of allowed values: healthy, diseased	diseased
sample disease status	list of diseases with which the subject has been diagnosed at the time of sample collection; can include multiple diagnoses; the value of the field depends on subject;	optional	Single-line text	Vibrio spp.
treatment agent	the name of the treatment agent used	optional	Single-line text	antibiotics
chemical compound	a drug, solvent, chemical, etc., with a property that can be measured such as concentration (http://purl.obolibrary.org/obo/CHEBI_37577).	optional	Single-line text	oxytetracycline (CHEBI:27701)

3.4.2. Implementation of Shellfish Standards

The standards are currently implemented via the EMBL-EBI Web-In system. An idealized schematic, detailing the process from sampling to data submission, is shown in Figure 4.2.

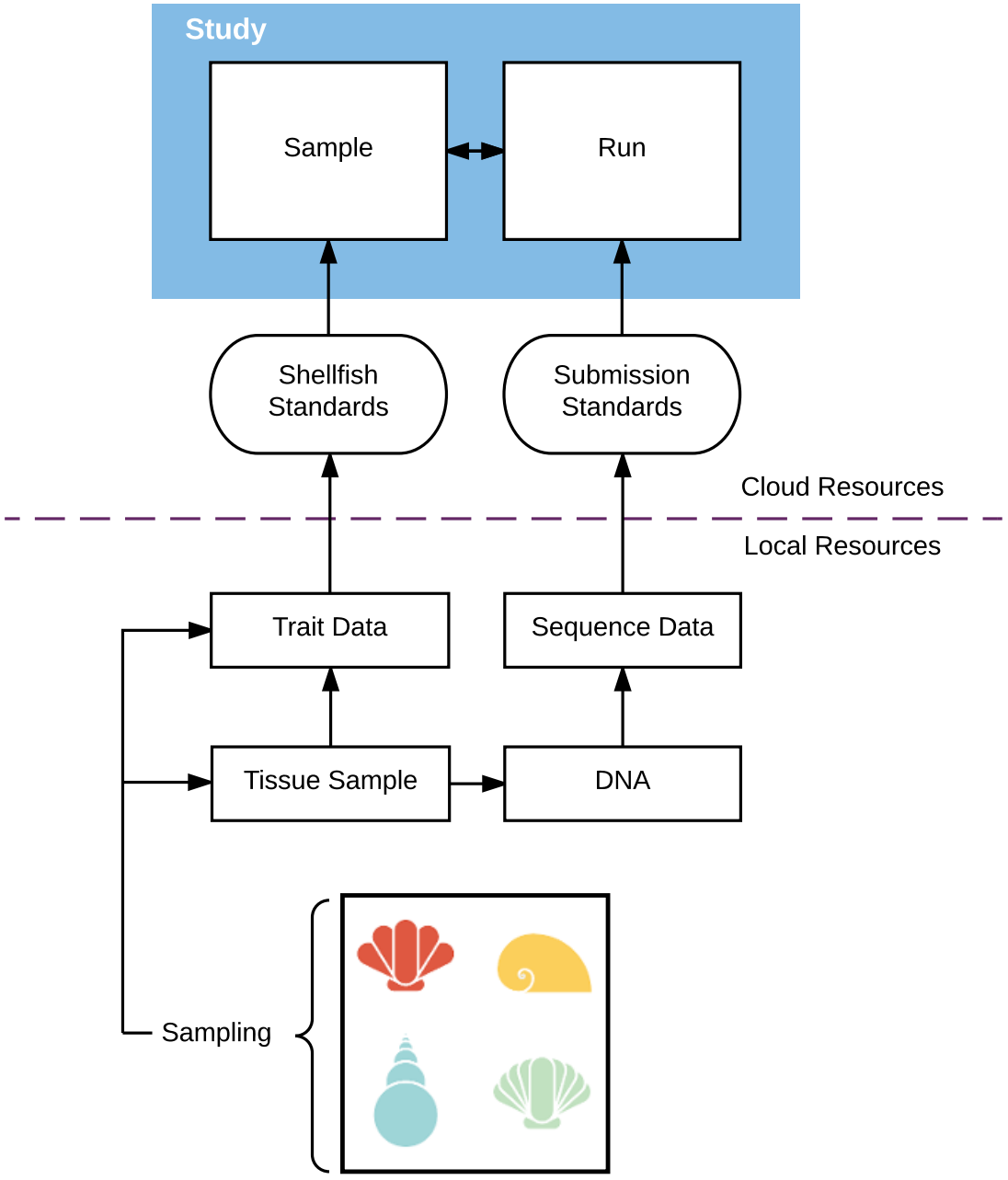


Figure 4.2: Flowchart detailing submission and standardisation of shellfish trait data and molecular data via the European Bioinformatics Institute.

4.4. Discussion

The standards outlined here represent a starting point in the implementation of phenotypic reporting in shellfish and more broadly in aquaculture genomics. However, the success of data standards depends heavily on uptake in their target community. In this case provision has been made only to optimize implementation. The adoption of these standards represents the largest hurdle in their success.

Collaboration between commercial operators has been demonstrated in dairy cattle BV prediction. The increase in precision, as a result of the increased sample size, incentivised the providers to share their records in a central database (Wiggans et al., 2009). However, the success of this model is based on very large investment in single individuals and little monopoly on specific strains between providers. The number of species found in aquaculture systems may mean this model is not viable. Despite the unsure future in commercial operators providing standardized (or any) data to public databases, the standards may see uptake in academic communities. Reproducibility is increasingly important in the genomics era (Begley and Ioannisis, 2015), and primary data provision is mandatory at point of publication in many cases (McNutt, 2014). Academic researchers currently operate in a way highly amenable to adopting these standards.

A provider wishing to upload data conforming to the shellfish standards presented here is met with an additional burden of data provision, this may hinder adoption of the standards. The mandatory data provision consists of 21 fields, of which 14 at most could be shared among samples in an experiment. Many researchers may already be recording the required data for their experimental design, but many of the required traits may be additional. For example, a researcher performing a genome wide association study (GWAS) on disease resistance would have to record a value for 14 extra traits, common to all individuals, and 7 traits individually, for all samples in the experiment. In an experiment with 500 samples, this would be a total of 3,514 additional records; large number of records required to meet standardisation. However, in this study the collection of the records detailed in Table 4.1 took approximately 0.3 person-hours per sample, equal to 150 person-hours for all samples. This cost is negligible in the scope of a large GWAS study, where the main cost is DNA sequencing. Additionally, the collection of further data allows researchers to expand the scope of their original experiment, opening up the possibility of supplementary discoveries.

Chapter 5. General Discussion

5.1 What has been achieved?

The worldwide population is predicted to reach 8.5 billion by 2030 (UN, 2015). Achieving food security for this many people will require extensive and advanced farming techniques in terrestrial, freshwater and marine biomes. Animal breeding research is in a state of constant evolution; technical advances in genome sequencing have opened up opportunities to increase the response to selection and computational innovation has enabled complex predictive and descriptive analyses to provide better understanding of selective breeding programmes. However, research is needed to address selective breeding for specific, important, organisms, while also developing pipelines, workflows and general tools that can be applied to new species.

This study aims to provide both specific and general solutions to several key problems in aquaculture research. Specific solutions are provided in Chapter 2 in the form of a simulated selective breeding programme for the gilthead seabream, a fish with unusual mating dynamics. The results indicate that the mating design schema provided here provides a starting point for a profitable selective breeding programme. In Chapter 3, a validated, low density SNP panel for parentage assignment is presented for Atlantic salmon, a species with active, selective breeding programmes across the world. The publicly available SNP parentage panel presented here allows producers to implement molecular parentage assignment with minimal development costs. Finally, Chapter 4 presents a set of trait standards for shellfish; a formalised way to ensure high quality data and metadata is submitted to public databases. This will, hopefully, further precipitate a culture of collaboration and reproducibility in QTL marker discovery between researchers and commercial providers.

The simulations detailed in Chapter 2 are written in the programming language R. The trait mean, SD and heritability can be substituted with values for other normally distributed traits of interest, to evaluate the predicted genetic gain in the gilthead seabream. Additionally, the mating design, reproductive output and generation gap can be adapted to reflect the values found in other species. With an intermediate understanding of the R programming language, the simulations can be further adapted to reflect traits with an effect on mortality, such as disease resistance. Overall, the simulation structure is highly amenable to adaption to new species of interest in aquaculture selective breeding programmes. The workflow for the selection of SNPs for parentage assignment, presented in Chapter 3, provides much needed clarity in molecular parentage assignment. The effect of various parameters on parentage assignment accuracy are clear to those working in the area, but no study to date has provided any formalisation of the important variables. Hopefully, the workflow will be adapted into the development of low density SNP panels, promoting similar methodology to enable comparison between assay type, organism or panel size.

5.2 Further Work on Bio-Economic Simulations

The simulation presented here provides a tool suitable for those who are familiar with the R programming language. However, not all academic and commercial researchers have a working fluency in the language. Future work should aim to increase the reach and applicability of the simulations presented in Chapter 2, through the implementation of a generalised, user-friendly software distribution. *EVA* is an example of an existing program that meets these criterea (Sørensen et al., 2008). This software has advanced features that are specific to terrestrial livestock breeding practices. No published work has used it for aquaculture breeding programmes. Alternatively, the web application framework *Shiney* would allow for the current R code the be incorporated into a user interface with minimal web development. The application could then be hosted in the public sphere and users could describe the trait of interest, mating dynamics and other breeding programme design elements. These inputs could then be used to generate the summary statistics and charts in Chapter 2. Users could interact with the application via their web browser and computation could be provided by 3rd parties to provide a cloud service. The entire simulation in Chapter 2 takes approximately 150 minutes to run on a single thread of an Intel 2.0 GHz (i7-4750HQ) processor. The hosting of the simulation via a web application would require the implementation of parallelised code in order to complete the computation in reasonable time. Since the bulk of the processing is running replicates, to estimate confidence, the current structure of the simulation is highly amenable to parallelisation. The R packages *foreach* and *doParallel* would make these changes to the R code very simple.

5.3 The Future of Low-Density SNP Panels for Parentage

Microsatellites and low density SNP panels will likely remain useful for accurate parentage assignment for many years. However, there are two potential technologies that may become dominant in the near future. The first is the use of multiplex PCR reactions to amplify regions of interest containing informative SNPs for parentage analysis, followed by sequencing via the next-generation sequencing instruments. This method is known as Genotyping-in-Thousands by Sequencing or GT-Seq and allows for the genotyping of 50-500 SNPs in thousands of individuals economically with high accuracy (Campbell et al., 2015). This technology is currently used by the Columbia River Inter-Tribal Fish Commission (Portland, Oregon, USA) in hatchery and fishery management operations, for several species of conservation concern. A second method that is likely to become more common is the use of several, tightly linked markers within an PCR amplicon, to give significantly more power to small regions of DNA. This concept is first found in Jones et al. (2010) where ‘the linked SNPs become a sort of ‘super-locus’, potentially with many alleles, provided the rate of recombination is low enough that haplotypes are stably inherited’. Since the publication of Jones et al. (2010) the

widespread adoption of RAD-Seq has made the discovery of regions with many polymorphic loci simple, and the publication of GT-Seq has demonstrated the utility of NGS for SNP genotyping. These haplotype 'super-loci' would have the polymorphism and power of microsatellites and the cost and accuracy equal or greater to today's current SNP genotyping platforms.

5.4 Maintaining Standards

The shellfish standards presented in Chapter 4 require further feedback and updates from the broader aquaculture community to improve their utility over time. The next step is to approach user groups and request that they begin uploading standard compliant data. Once the checklist of standards reaches a stage with a consistent user base, funding bodies could be approached. The aim would be to ensure all funded projects upload standard compliant data. After a sufficient number of records is reached the database will become increasingly useful for users, as a source of validation trait and nucleotide data.

6. General Appendix

6.1 Appendix A: Table of Aquaculture Journals

Journal Name	Publisher	2015 Impact Factor	Standards
Aquaculture	Elsevier	1.893	Genbank Accession numbers mentioned but not mandated
Aquaculture Economics and Management	Taylor and Francis Inc	1.175	None detailed
Aquaculture International	Springer	0.96	None detailed
Aquaculture Research	Wiley-Blackwell	1.606	None detailed
Aquaculutre Environment Interactions	Inter-Research	1.985	None detailed
Aquauclture Nutrition	Wiley-Blackwell	1.511	None detailed
Israeli Journal of Aquaculture	n/a	0.252	None detailed
Journal of the World Aquaculture Society	Wiley-Blackwell	0.665	None detailed
North Amercian Journal of Aquaculture	Taylor and Francis Inc	0.76	None detailed
Reviews in Aquaculture	Wiley-Blackwell	4.769	Genbank Accession number required

6.2 Appendix B: Calculations for ΔF from N_e

Equation 6.1 below is from Gjedrem et al. (2005) Chapter 6, page 75.

(6.1)

$$N_e = 1/(2\Delta F)$$

Re-arranging gives Equation 6.2.

(6.2)

$$\Delta F = \frac{1}{N_e}/2$$

A N_e of 13 gives a ΔF of 0.0769, whist N_e of 28 gives a ΔF of 0.0179.

6.3 Appendix C: COLONY Parameters

1234 - Seed for random number generator
Updating allele frequency
Dioecious species
Inbreeding present
Diploid species
Polygamy for both males and females
No clone inference
Full sibship size scaling selected
No sibship prior
Unknown population allele frequency
3 Runs
Medium runs selected
Full-Likelihood(FL) Method
Medium precision
Runs where performed under the command line version of COLONY v 2.0.6.2.

6.4 Appendix D: Methods for Panel Decay

The scenario followed the 'current' scheme shown in Figure 6.1 and continued for 10 generations from F0-F10, each linage was simulated separately with no mixing. Breeding was according to probabilities as described in Section 2.2.2.

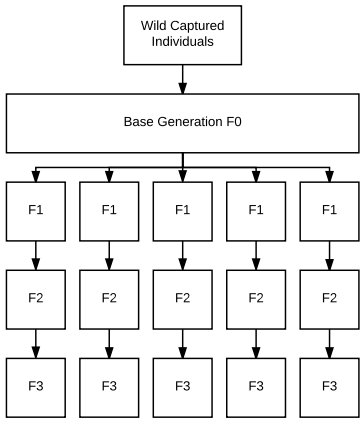


Figure 6.1: Flowchart detailing breeding over four generations in the 'current' breeding programme.

A pedigree for the 10 generations for a single linage was then produced. The program SIMPED was used to simulate bi-allelic SNP marker data for all individuals in the pedigree according to Mendelian inheritance. A non-optimal SNP panel was simulated for this data to ensure some error in parentage assignment occurred in order to examine how the error changed over the progress of the breeding programme. A total of 50 SNPs were simulated per individual, the SNPs had a mean minor allele frequency of 0.48 and a SD of 0.26 at the beginning of the simulation. Additionally, 1% error was simulated using a custom R function that randomly selected 2% of total data and selected a new

allele from random. This function results in approximately 1% error as alleles may be replaced with the same allele due to chance.

The simulated data was then subset into nine parentage assignment runs, each run contained 50 offspring and 50 parents from adjacent generations. COLONY v 2.0.6.2 was then used to allocate parents to each of the 50 offspring. COLONY run parameters followed Appendix C. The entire simulation was replicated 100 times. Results were compared against the known pedigree for each parentage assignment run using a R script.

The decay percentage data was arcsine transformed to meet the assumptions of the two-way ANOVA. A Tukey's Honest Significant Difference test was performed on the ANOVA result as shown in Table 6.1.

The ANOVA showed a significant ($F(8,891)=16.86$, $p > 0.001$) difference between group means.

*Table 6.1: Table detailing results of a Tukey's HSD test on a two-way ANOVA testing for a significant difference between generation group parentage assignment accuracy. Group difference is untransformed mean per group difference in percentage accuracy. The groups compared are shown in the left-most column. The right-most column details significance of the p value. NS = > 0.05, * = < 0.05, ** = < 0.01, ***= < 0.001.*

Group Comparison	**Group Differences (%)**	**p adj**	**sig**
F2_F3-F1_F2	-0.36	0.832	NS
F3_F4-F1_F2	-0.61	0.025	*
F4_F5-F1_F2	-0.59	0.048	*
F5_F6-F1_F2	-1.05	< 0.0001	***
F6_F7-F1_F2	-1.46	< 0.0001	***
F7_F8-F1_F2	-1.62	< 0.0001	***
F8_F9-F1_F2	-1.55	< 0.0001	***
F9_F10-F1_F2	-2.05	< 0.0001	***
F3_F4-F2_F3	-0.25	0.689	NS
F4_F5-F2_F3	-0.23	0.820	NS
F5_F6-F2_F3	-0.69	0.005	**
F6_F7-F2_F3	-1.1	< 0.0001	***
F7_F8-F2_F3	-1.26	< 0.0001	***
F8_F9-F2_F3	-1.19	< 0.0001	***
F9_F10-F2_F3	-1.69	< 0.0001	***
F4_F5-F3_F4	0.02	1.000	NS
F5_F6-F3_F4	-0.44	0.553	NS
F6_F7-F3_F4	-0.85	0.079	NS
F7_F8-F3_F4	-1.01	0.002	**
F8_F9-F3_F4	-0.94	0.016	*
F9_F10-F3_F4	-1.44	< 0.0001	***
F5_F6-F4_F5	-0.46	0.403	NS
F6_F7-F4_F5	-0.87	0.042	*
F7_F8-F4_F5	-1.03	0.001	**
F8_F9-F4_F5	-0.96	0.008	**
F9_F10-F4_F5	-1.46	< 0.0001	***
F6_F7-F5_F6	-0.41	0.989	NS
F7_F8-F5_F6	-0.57	0.517	NS
F8_F9-F5_F6	-0.5	0.868	NS
F9_F10-F5_F6	-1	0.017	*
F7_F8-F6_F7	-0.16	0.978	NS
F8_F9-F6_F7	-0.09	1.000	NS
F9_F10-F6_F7	-0.59	0.236	NS
F8_F9-F7_F8	0.07	1.000	NS
F9_F10-F7_F8	-0.43	0.891	NS
F9_F10-F8_F9	-0.5	0.555	NS

7. References

Anderson, E.C., Garza, J.C., 2005. The Power of Single-Nucleotide Polymorphisms for Large-Scale Parentage Inference. Genetics 172, 2567–2582. doi:10.1534/genetics.105.048074

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., Hohenlohe, P.A., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17, 81–92. doi:10.1038/nrg.2015.28

Angiuoli, S. V, Gussman, A., Klimke, W., Cochrane, G., Field, D., Garrity, G.M., Kodira, C.D., Kyrpides, N., Madupu, R., Markowitz, V., Tatusova, T., Thomson, N., White, O., 2008. Toward an Online Repository of Standard Operating Procedures (SOPs) for (Meta)genomic Annotation. Omi. A J. Integr. Biol. 12, 137–141. doi:10.1089/omi.2008.0017

Anscombe, F.J., 1973. Graphs in Statistical Analysis. Am. Stat. 27, 17–21. doi:10.1080/00031305.1973.10478966

Antezana, E., Kuiper, M., Mironov, V., 2009. Biological knowledge management: the emerging role of the Semantic Web technologies. Brief. Bioinform. 10, 392–407. doi:10.1093/bib/bbp024

Antonello, J., Massault, C., Franch, R., Haley, C., Pellizzari, C., Bovo, G., Patarnello, T., de Koning, D.-J., 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, 29–35. doi:10.1016/j.aquaculture.2009.10.022

Araneda, M.E., Hernández, J.M., Gasca-Leyva, E., 2011. Optimal harvesting time of farmed aquatic populations with nonlinear size-heterogeneous growth. Nat. Resour. Model. 24, 477–513. doi:10.1111/j.1939-7445.2011.00099.x

Argue, B.J., Arce, S.M., Lotz, J.M., Moss, S.M., 2002. Selective breeding of Pacific white shrimp (*Litopenaeus vannamei*) for growth and resistance to Taura Syndrome Virus. Aquaculture 204, 447–460. doi:10.1016/S0044-8486(01)00830-4

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29. doi:10.1038/75556

Astorga, M.P., 2014. Genetic considerations for mollusk production in aquaculture: current state of knowledge. Front. Genet. 5, 1–6. doi:10.3389/fgene.2014.00435

Baird, N. a., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z. a., Selker, E.U., Cresko, W. a., Johnson, E. a., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3, 1–7. doi:10.1371/journal.pone.0003376

Balon, E.K., 1995. Origin and domestication of the wild carp, *Cyprinus carpio*: from Roman gourmets to the swimming flowers. Aquaculture 129, 3–48. doi:10.1016/0044-8486(94)00227-F

Banos, G., Wiggans, G.R., Powell, R.L., 2001. Impact of Paternity Errors in Cow Identification on Genetic Evaluations and International Comparisons. J. Dairy Sci. 84, 2523–2529. doi:10.3168/jds.S0022-0302(01)74703-0

Barson, N.J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G.H., Fiske, P., Jacq, C., Jensen, A.J., Johnston, S.E., Karlsson, S., Kent, M., Moen, T., Niemelä, E., Nome, T., Næsje, T.F., Orell, P., Romakkaniemi, A., Sægrov, H., Urdal, K., Erkinaro, J., Lien, S., Primmer, C.R., 2015. Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. Nature 0, 1–4. doi:10.1038/nature16062

Begley, C.G., Ioannidis, J.P.A., 2015. Reproducibility in Science: Improving the Standard for Basic and Preclinical Research. Circ. Res. 116, 116–126. doi:10.1161/CIRCRESAHA.114.303819

Bernatchez, L., 2016. On the maintenance of genetic variation and adaptation to environmental change: considerations from population genomics in fishes. J. Fish Biol. 1–38. doi:10.1111/jfb.13145

Beveridge, M.C.M., Little, D.C., 2007. The History of Aquaculture in Traditional Societies, in: Ecological Aquaculture. Blackwell Science Ltd, Oxford, UK, pp. 1–29. doi:10.1002/9780470995051.ch1

Bik, E.M., Casadevall, A., Fang, F.C., 2016. The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications. MBio 7, e00809-16. doi:10.1128/mBio.00809-16

Bjørndal, T., 1988. Optimal Harvesting of Farmed Fish. Mar. Resour. Econ. 5, 139–159. doi:10.1086/mre.5.2.42628926

Bolivar, R.B., Newkirk, G.F., 2002. Response to within family selection for body weight in Nile tilapia (*Oreochromis niloticus*) using a single-trait animal model. Aquaculture 204, 371–381. doi:10.1016/S0044-8486(01)00824-9

Bondari, K., 1983. Response to bidirectional selection for body weight in channel catfish. Aquaculture 33, 73–81. doi:10.1016/0044-8486(83)90387-3

Bondari, K., 1986. Response of channel catfish to multi-factor and divergent selection of economic traits. Aquaculture 57, 163–170. doi:10.1016/0044-8486(86)90193-6

Borrell, Y.J., Gallego, V., García-Fernández, C., Mazzeo, I., Pérez, L., Asturiano, J.F., Carleos, C.E., Vázquez, E., Sánchez, J.A., Blanco, G., 2011. Assessment of parental contributions to fast- and slow-growing progenies in the sea bream *Sparus aurata* L. using a new multiplex PCR. Aquaculture 314, 58–65. doi:10.1016/j.aquaculture.2011.01.028

Boudry, P., Collet, B., Cornette, F., Hervouet, V., Bonhomme, F., 2002. High variance in reproductive success of the Pacific oyster (*Crassostrea gigas*, Thunberg) revealed by microsatellite-based parentage analysis of multifactorial crosses. Aquaculture 204, 283–296. doi:10.1016/S0044-8486(01)00841-9

Brody, T., Wohlfarth, G., Hulata, G., Moav, R., 1981. Application of electrophoretic genetic markers to fish breeding. IV. Assessment of breeding value of full-sib families. Aquaculture 24, 175–186. doi:10.1016/0044-8486(81)90054-5

Bulmer, M.G., 1971. The Effect of Selection on Genetic Variability. Am. Nat. 105, 201–211. doi:10.1086/282718

Busack, C., Knudsen, C.M., 2007. Using factorial mating designs to increase the effective number of breeders in fish hatcheries. Aquaculture 273, 24–32. doi:10.1016/j.aquaculture.2007.09.010

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., Vandesompele, J., Wittwer, C.T., 2009. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. Clin. Chem. 55, 611–622. doi:10.1373/clinchem.2008.112797

Bustin, S. a, Beaulieu, J.-F., Huggett, J., Jaggi, R., Kibenge, F.S., Olsvik, P. a, Penning, L.C., Toegel, S., 2010. MIQE précis: Practical implementation of minimum standard guidelines for fluorescence-based quantitative real-time PCR experiments. BMC Mol. Biol. 11, 74. doi:10.1186/1471-2199-11-74

Bustin, S.A., 2010. Why the need for qPCR publication guidelines?-The case for MIQE. Methods 50, 217–226. doi:10.1016/j.ymeth.2009.12.006

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421. doi:10.1186/1471-2105-10-421

Cameron Brown, R., Woolliams, J.A., McAndrew, B.J., 2005. Factors influencing effective population size in commercial populations of gilthead seabream, *Sparus aurata*. Aquaculture 247, 219–225. doi:10.1016/j.aquaculture.2005.02.002

Campbell, N.R., Harmon, S.A., Narum, S.R., 2015. Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Mol. Ecol. Resour. 15, 855–867. doi:10.1111/1755-0998.12357

Chaisson, M.J.P., Wilson, R.K., Eichler, E.E., 2015. Genetic variation and the *de novo* assembly of human genomes. Nat. Rev. Genet. 16, 627–640. doi:10.1038/nrg3933

Charlesworth, B., Charlesworth, D., 1999. The genetic basis of inbreeding depression. Genet. Res. 74, S0016672399004152. doi:10.1017/S0016672399004152

Chavanne, H., Janssen, K., Hofherr, J., Contini, F., Haffray, P., Komen, H., Nielsen, E.E., Bargelloni, L., 2016. A comprehensive survey on selective breeding programs and seed market in the European aquaculture fish industry. Aquac. Int. 24, 1287–1307. doi:10.1007/s10499-016-9985-0

Chistiakov, D.A., Hellemans, B., Volckaert, F.A.M., 2006. Microsatellites and their genomic distribution, evolution, function and applications: A review with special reference to fish genetics. Aquaculture 255, 1–29. doi:10.1016/j.aquaculture.2005.11.031

Collard, B.C., Mackill, D.J., 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos. Trans. R. Soc. B Biol. Sci. 363, 557–572. doi:10.1098/rstb.2007.2170

Coster, A., 2013. Package "Pedigree."

Dakin, E.E., Avise, J.C., 2004. Microsatellite null alleles in parentage analysis. Heredity (Edinb). 93, 504–509. doi:10.1038/sj.hdy.6800545

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. doi:10.1093/bioinformatics/btr330

Davidson, W.S., Koop, B.F., Jones, S.J.M., Iturra, P., Vidal, R., Maass, A., Jonassen, I., Lien, S., Omholt, S.W., 2010. Sequencing the genome of the Atlantic salmon (*Salmo salar*). Genome Biol. 11, 403. doi:10.1186/gb-2010-11-9-403

Dong, L., Xiao, S., Wang, Q., Wang, Z., 2016. Comparative analysis of the GBLUP, emBayesB, and GWAS algorithms to predict genetic values in large yellow croaker (*Larimichthys crocea*). BMC Genomics 17, 460. doi:10.1186/s12864-016-2756-5

Dou, J., Li, X., Fu, Q., Jiao, W., Li, Y., Li, T., Wang, Y., Hu, X., Wang, S., Bao, Z., 2016. Evaluation of the 2b-RAD method for genomic selection in scallop breeding. Sci. Rep. 6, 19244. doi:10.1038/srep19244

Duarte, C.M., Alcaraz, M., 1989. To produce many small or few large eggs: a size-independent reproductive tactic of fish. Oecologia 80, 401–404. doi:10.1007/BF00379043

Duchesne, P., Godbout, M.H., Bernatchez, L., 2002. PAPA (package for the analysis of parental allocation): A computer program for simulated and real parental allocation. Mol. Ecol. Notes 2, 191–193. doi:10.1046/j.1471-8286.2002.00164.x

Dupont-Nivet, M., Vandeputte, M., Vergnet, A., Merdy, O., Haffray, P., Chavanne, H., Chatain, B., 2008. Heritabilities and GxE interactions for growth in the European sea bass (*Dicentrarchus labrax* L.) using a marker-based pedigree. Aquaculture 275, 81–87. doi:10.1016/j.aquaculture.2007.12.032

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6, 1–10. doi:10.1371/journal.pone.0019379

Elvingson, P., Nilsson, J., 1994. Phenotypic and genetic parameters of body and compositional traits in Arctic charr, *Salvelinus alpinus* (L.). Aquac. Res. 25, 677–685. doi:10.1111/j.1365-2109.1994.tb00732.x

Etter, P.D., Bassham, S., Hohenlohe, P. a, Johnson, E. a, Cresko, W. a, 2011. Molecular Methods for Evolutionary Genetics, Molecular Methods for Evolutionary Genetics, Methods in Molecular Biology, vol. 772, Methods in Molecular Biology. Humana Press, Totowa, NJ. doi:10.1007/978-1-61779-228-1

Falconer, D.S., Mackay, T.F.C., 1997. Introduction to quantitative genetics (4th edn), Trends in Genetics. [Amsterdam, The Netherlands: Elsevier Science Publishers (Biomedical Division)], c1985-.

FAO, 2014. FAO Yearbook. Fishery and Aquaculutre Statistics. FAO yearbook.Fishery and aquaculture statistics.

FAO, 2016. FishStatJ. Universal software for fishery statistical time series., Food and Agricultural Organization, Fisheries Department, fishery Information, Data and Statistics Unit ROME:FAO

Fernandes, T., Herlin, M., Belluga, M.D.L., Ballón, G., Martinez, P., Toro, M.A., Fernández, J., 2016. Estimation of genetic parameters for growth traits in a hatchery population of gilthead sea bream (*Sparus aurata* L.). Aquac. Int. doi:10.1007/s10499-016-0046-5

Field, D., Amaral-Zettler, L., Cochrane, G., Cole, J.R., Dawyndt, P., Garrity, G.M., Gilbert, J., Glöckner, F.O., Hirschman, L., Karsch-Mizrachi, I., Klenk, H.P., Knight, R., Kottmann, R., Kyrpides, N., Meyer, F., Gil, I.S., Sansone, S.A., Schriml, L.M., Sterk, P., Tatusova, T., Ussery, D.W., White, O., Wooley, J., 2011. The Genomic Standards Consortium. PLoS Biol. 9, 8–10. doi:10.1371/journal.pbio.1001088

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M.J., Angiuoli, S. V, Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De Vos, P., DePamphilis, C., Edwards, R., Faruque, N., Feldman, R., Gilbert, J., Gilna, P., Glöckner, F.O., Goldstein, P., Guralnick, R., Haft, D., Hancock, D., Hermjakob, H., Hertz-Fowler, C., Hugenholtz, P., Joint, I., Kagan, L., Kane, M., Kennedy, J., Kowalchuk, G., Kottmann, R., Kolker, E., Kravitz, S., Kyrpides, N., Leebens-Mack, J., Lewis, S.E., Li, K., Lister, A.L., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S.-A., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T., San Gil, I., Wilson, G., Wipat, A., 2008. The minimum information about a genome sequence (MIGS) specification. Nat. Biotechnol. 26, 541–547. doi:10.1038/nbt1360

Fjalestad, K.T., Gjedrem, T., Gjerde, B., 1993. Genetic improvement of disease resistance in fish: an overview. Aquaculture 111, 65–74. doi:10.1016/0044-8486(93)90025-T

Freamo, H., O'Reilly, P., Berg, P.R., Lien, S., Boulding, E.G., 2011. Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. Mol. Ecol. Resour. 11, 254–267. doi:10.1111/j.1755-0998.2010.02952.x

Friars, G.W., Bailey, J.K., Flynn, F.M.O., 1995. Applications of selection for multiple traits in cage- reared Atlantic salmon (*Salmo salar*) 137, 213–217.

Fuji, K., Hasegawa, O., Honda, K., Kumasaka, K., Sakamoto, T., Okamoto, N., 2007. Marker-assisted breeding of a lymphocystis disease-resistant Japanese flounder (*Paralichthys olivaceus*). Aquaculture 272, 291–295. doi:10.1016/j.aquaculture.2007.07.210

Gall, G.A.E., Huang, N., 1988. Heritability and selection schemes for rainbow trout: body weight. Aquaculture 73, 43–56. doi:10.1016/0044-8486(88)90040-3

Gall, G.A.., Bakar, Y., 2002. Application of mixed-model techniques to fish breed improvement: analysis of breeding-value selection to increase 98-day body weight in tilapia. Aquaculture 212, 93–113. doi:10.1016/S0044-8486(02)00024-8

Gallardo, J.A., García, X., Lhorente, J.P., Neira, R., 2004. Inbreeding and inbreeding depression of female reproductive traits in two populations of Coho salmon selected using BLUP predictors of breeding values. Aquaculture 234, 111–122. doi:10.1016/j.aquaculture.2004.01.009

Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., Fernando, R., 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics 183, 347–363. doi:10.1534/genetics.109.103952

Gitterle, T., Rye, M., Salte, R., Cock, J., Johansen, H., Lozano, C., Arturo Suárez, J., Gjerde, B., 2005. Genetic (co)variation in harvest body weight and survival in Penaeus (*Litopenaeus*) vannamei under standard commercial conditions. Aquaculture 243, 83–92. doi:10.1016/j.aquaculture.2004.10.015

Gjedrem, T., 2000. Genetic improvement of cold-water fish species. Aquac. Res. 31, 25–33. doi:10.1046/j.1365-2109.2000.00389.x

Gjedrem, T., 2010. The first family-based breeding program in aquaculture. Rev. Aquac. 2, 2–15. doi:10.1111/j.1753-5131.2010.01011.x

Gjedrem, T., AKVAFORSK, Å., 2005. Selection and breeding programs in aquaculture. Springer.

Gjedrem, T., Baranski, M., 2009. Selective Breeding in Aquaculture: An Introduction, Reviews: Methods and Technologies in Fish Biology and Fisheries. Springer Netherlands, Dordrecht. doi:10.1007/978-90-481-2773-3

Gjedrem, T., Gjøen, H.M., Gjerde, B., 1991. Genetic origin of Norwegian farmed Atlantic salmon. Aquaculture 98, 41–50. doi:10.1016/0044-8486(91)90369-I

Gjedrem, T., Rye, M., 2016. Selection response in fish and shellfish: a review. Rev. Aquac. 1–12. doi:10.1111/raq.12154

Gjerde, B., Schaeffer, L.R., 1989. Body traits in rainbow trout. II. Estimates of heritabilities and of phenotypic and genetic correlations. Aquaculture 80, 25–44. doi:10.1016/0044-8486(89)90271-8

Gjerde, B., Gunnes, K., Gjedrem, T., 1983. Effect of inbreeding on survival and growth in rainbow trout. Aquaculture 34, 327–332. doi:10.1016/0044-8486(83)90212-0

Glazer, A.M., Killingbeck, E.E., Mitros, T., Rokhsar, D.S., Miller, C.T., 2015. Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-sequencing. G3 5, 1463–72. doi:10.1534/g3.115.017905

Gonen, S., Baranski, M., Thorland, I., Norris, A., Grove, H., Arnesen, P., Bakke, H., Lien, S., Bishop, S.C., Houston, R.D., 2015. Mapping and validation of a major QTL affecting resistance to pancreas disease (salmonid alphavirus) in Atlantic salmon (*Salmo salar*). Heredity (Edinb). 115, 405–414. doi:10.1038/hdy.2015.37

Gonen, S., Lowe, N.R., Cezard, T., Gharbi, K., Bishop, S.C., Houston, R.D., 2014. Linkage maps of the Atlantic salmon (*Salmo salar*) genome derived from RAD sequencing. BMC Genomics 15, 166. doi:10.1186/1471-2164-15-166

Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351. doi:10.1038/nrg.2016.49

Gorshkov, S., Gordin, H., Gorshkova, G., Knibb, W., 1997. Reproductive constraints for family selection of the gilthead seabream (*Sparus aurata* L.). Isr. J. Aquac. - Bamidgeh 49, 124–134.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. doi:10.1038/nbt.1883

Grattapaglia, D., Resende, M.D. V, 2011. Genomic selection in forest tree breeding. Tree Genet. Genomes 7, 241–255. doi:10.1007/s11295-010-0328-4

Guo, X., 2009. Use and exchange of genetic resources in molluscan aquaculture. Rev. Aquac. 1, 251–259. doi:10.1111/j.1753-5131.2009.01014.x

Hao-Ren, L., 1982. Polycultural System of Freshwater Fish in China. Can. J. Fish. Aquat. Sci. 39, 143–150. doi:10.1139/f82-015

Hartl, D., Clark, A., 2007. Principles of Population Genetics. {Sinauer Associates}.

Hauser, L., Baird, M., Hilborn, R., Seeb, L.W., Seeb, J.E., 2011. An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. Mol. Ecol. Resour. 11, 150–161. doi:10.1111/j.1755-0998.2010.02961.x

Hayes, B.J., Bowman, P.J., Chamberlain, A.J., Goddard, M.E., 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92, 433–443. doi:10.3168/jds.2008-1646

Hayes, B.J., Lewin, H.A., Goddard, M.E., 2013. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. Trends Genet. 29, 206–214. doi:10.1016/j.tig.2012.11.009

Hedgecock, D., Launey, S., Pudovkin, A.I., Naciri, Y., Lapègue, S., Bonhomme, F., 2007. Small effective number of parents (Nb) inferred for a naturally spawned cohort of juvenile European flat oysters *Ostrea edulis*. Mar. Biol. 150, 1173–1182. doi:10.1007/s00227-006-0441-y

Heffner, E.L., Sorrells, M.E., Jannink, J., 2009. Genomic Selection for Crop Improvement. Crop Sci. 49, 1. doi:10.2135/cropsci2008.08.0512

Henderson, C.R., 1975. Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 31, 423. doi:10.2307/2529430

Herbinger, C.M., Doyle, R.W., Pitman, E.R., Paquet, D., Mesa, K.A., Morris, D.B., Wright, J.M., Cook, D., 1995. DNA fingerprint based analysis of paternal and maternal effects on offspring growth and survival in communally reared rainbow trout. Aquaculture 137, 245–256. doi:10.1016/0044-8486(95)01109-9

Herlin, M., Taggart, J.B., McAndrew, B.J., Penman, D.J., 2007. Parentage allocation in a complex situation: A large commercial Atlantic cod (*Gadus morhua*) mass spawning tank. Aquaculture 272, S195–S203. doi:10.1016/j.aquaculture.2007.08.018

Hershberger, W.K., Myers, J.M., Iwamoto, R.N., Mcauley, W.C., Saxton, A.M., 1990. Genetic changes in the growth of coho salmon (*Oncorhynchus kisutch*) in marine net-pens, produced by ten years of selection. Aquaculture 85, 187–197. doi:10.1016/0044-8486(90)90018-I

Hoban, S., Kelley, J.L., Lotterhos, K.E., Antolin, M.F., Bradburd, G., Lowry, D.B., Poss, M.L., Reed, L.K., Storfer, A., Whitlock, M.C., 2016. Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. Am. Nat. 188, 000–000. doi:10.1086/688018

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A., Cresko, W.A., 2010. Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. PLoS Genet. 6, e1000862. doi:10.1371/journal.pgen.1000862

Houston, R.D., Taggart, J.B., Cézard, T., Bekaert, M., Lowe, N.R., Downing, A., Talbot, R., Bishop, S.C., Archibald, A.L., Bron, J.E., Penman, D.J., Davassi, A., Brew, F., Tinch, A.E., Gharbi, K., Hamilton, A., 2014. Development and validation of a high density SNP genotyping array for Atlantic salmon (*Salmo salar*). BMC Genomics 15, 90. doi:10.1186/1471-2164-15-90

Houston, R.D., Haley, C.S., Hamilton, A., Guy, D.R., Tinch, A.E., Taggart, J.B., McAndrew, B.J., Bishop, S.C., 2008. Major Quantitative Trait Loci Affect Resistance to Infectious Pancreatic Necrosis in Atlantic Salmon (*Salmo salar*). Genetics 178, 1109–1115. doi:10.1534/genetics.107.082974

Houston, R., Davey, J., Bishop, S., Lowe, N., Mota-Velasco, J., Hamilton, A., Guy, D., Tinch, A., Thomson, M., Blaxter, M., Gharbi, K., Bron, J., Taggart, J., 2012. Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. BMC Genomics 13, 244. doi:10.1186/1471-2164-13-244

Huang, T., 2005. The mRNA of the Arabidopsis Gene FT Moves from Leaf to Shoot Apex and Induces Flowering. Science (80-.). 309, 1694–1696. doi:10.1126/science.1117768

Hulata, G., Wohlfarth, G.W., Halevy, A., 1986. Mass Selection for Growth Rate in the Nile Tilapia (*Oreochromis niloticus*) l. Aquaculture 57, 177–184. doi:10.1016/0044-8486(86)90195-X

Humble, E., Martinez-Barrio, A., Forcada, J., Trathan, P.N., Thorne, M.A.S., Hoffmann, M., Wolf, J.B.W., Hoffman, J.I., 2016. A draft fur seal genome provides insights into factors affecting SNP validation and how to mitigate them. Mol. Ecol. Resour. 16, 909–921. doi:10.1111/1755-0998.12502

Israel, C., Weller, J.I., 2000. Effect of misidentification on genetic gain and estimation of breeding value in dairy cattle populations. J. Dairy Sci. 83, 181–187. doi:10.3168/jds.S0022-0302(00)74869-7

Jannink, J.-L., 2010. Dynamics of long-term genomic selection. Genet. Sel. Evol. 42, 35. doi:10.1186/1297-9686-42-35

Jones, A.G., Small, C.M., Paczolt, K.A., Ratterman, N.L., 2010. A practical guide to methods of parentage analysis. Mol. Ecol. Resour. 10, 6–30. doi:10.1111/j.1755-0998.2009.02778.x

Jones, O.R., Wang, J., 2010. COLONY: a program for parentage and sibship inference from multilocus genotype data. Mol. Ecol. Resour. 10, 551–555. doi:10.1111/j.1755-0998.2009.02787.x

Kaiser, S.A., Taylor, S.A., Chen, N., Sillett, T.S., Bondra, E.R., Webster, M.S., 2016. A comparative assessment of SNP and microsatellite markers for assigning parentage in a socially monogamous bird. Mol. Ecol. Resour. n/a-n/a. doi:10.1111/1755-0998.12589

Kalinowski, S.T., Taper, M.L., Marshall, T.C., 2007. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol. Ecol. 16, 1099–1106. doi:10.1111/j.1365-294X.2007.03089.x

Kause, A., Ritola, O., Paananen, T., Wahlroos, H., Mäntysaari, E.A., 2005. Genetic trends in growth, sexual maturity and skeletal deformations, and rate of inbreeding in a breeding programme for rainbow trout (*Oncorhynchus mykiss*). Aquaculture 247, 177–187. doi:10.1016/j.aquaculture.2005.02.023

Kincaid, H.L., 1983. Inbreeding in fish populations used for aquaculture. Aquaculture 33, 215–227. doi:10.1016/0044-8486(83)90402-7

King, T.L., Eackles, M.S., Letcher, B.H., 2005. Microsatellite DNA markers for the study of Atlantic salmon (*Salmo salar*) kinship, population structure, and mixed-fishery analyses. Mol. Ecol. Notes 5, 130–132. doi:10.1111/j.1471-8286.2005.00860.x

Kissil, G.W., Lupatsch, I., Elizur, A., Zohar, Y., 2001. Long photoperiod delayed spawning and increased somatic growth in gilthead seabream (*Sparus aurata*). Aquaculture 200, 363–379. doi:10.1016/S0044-8486(01)00527-0

Krishna, G., Gopikrishna, G., Gopal, C., Jahageerdar, S., Ravichandran, P., Kannappan, S., Pillai, S.M., Paulpandi, S., Kiran, R.P., Saraswati, R., Venugopal, G., Kumar, D., Gitterle, T., Lozano, C., Rye, M., Hayes, B., 2011. Genetic parameters for growth and survival in *Penaeus monodon* cultured in India. Aquaculture 318, 74–78. doi:10.1016/j.aquaculture.2011.04.028

Kristensen, T.N., Pedersen, K.S., Vermeulen, C.J., Loeschcke, V., 2010. Research on inbreeding in the "omic" era. Trends Ecol. Evol. 25, 44–52. doi:10.1016/j.tree.2009.06.014

Kron, G., 2008. Recontructing the techniques and potential productivity of Roman aquaculture in light of recent research and preactices, in: Hermon, E. (Ed.), Vers Une Gestion Integree de L'eau Dans L'empire Romain. Actes du Colloque International Universita Laval, Rome, pp. 175–185.

Labuschagne, C., Nupen, L., Kotzé, A., Grobler, P.J., Dalton, D.L., 2015. Assessment of microsatellite and SNP markers for parentage assignment in ex situ African Penguin (*Spheniscus demersus*) populations. Ecol. Evol. 5, 4389–4399. doi:10.1002/ece3.1600

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. doi:10.1038/nmeth.1923

Larson, W.A., McKinney, G.J., Seeb, J.E., Seeb, L.W., 2016. Identification and Characterization of Sex-Associated Loci in Sockeye Salmon Using Genotyping-by-Sequencing and Comparison with a Sex-Determining Assay Based on the *sdY* Gene. J. Hered. 107, 559–566. doi:10.1093/jhered/esw043

Leal, S.M., Yan, K., Müller-Myhsok, B., 2005. SimPed: A simulation program to generate haplotype and genotype data for pedigree structures. Hum. Hered. 60, 119–122. doi:10.1159/000088914

Leroy, G., 2014. Inbreeding depression in livestock species: review and meta-analysis. Anim. Genet. 45, 618–628. doi:10.1111/age.12178

Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323. doi:10.1186/1471-2105-12-323

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009a. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, R., L, Q., Yu, R., 2009b. Parentage Determination and Effective Population Size Estimation in Mass Spawning Pacific Oyster, *Crassostrea gigas* , Based on Microsatellite Analysis. J. World Aquac. Soc. 40, 667–677. doi:10.1111/j.1749-7345.2009.00286.x

Lien, S., Gidskehaug, L., Moen, T., Hayes, B.J., Berg, P.R., Davidson, W.S., Omholt, S.W., Kent, M.P., 2011. A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome

homeologies and striking differences in sex-specific recombination patterns. BMC Genomics 12, 615. doi:10.1186/1471-2164-12-615

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Matthew, P., Leong, J.S., Minkley, D.R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R.A., Schalburg, K. Von, Rondeau, E.B., Genova, A. Di, Samy, J.K.A., Vik, J.O., 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533, 200–205. doi:10.1038/nature17164

Lillehammer, M., Meuwissen, T.H.E., Sonesson, A.K., 2013. A low-marker density implementation of genomic selection in aquaculture using within-family genomic breeding values. Genet. Sel. Evol. 45, 39. doi:10.1186/1297-9686-45-39

Liu, S., Palti, Y., Gao, G., Rexroad, C.E., 2015. Development and validation of a SNP panel for parentage assignment in rainbow trout. Aquaculture 452, 178–182. doi:10.1016/j.aquaculture.2015.11.001

Llorente, I., Luna, L., 2016. Bioeconomic modelling in aquaculture: an overview of the literature. Aquac. Int. 24, 931–948. doi:10.1007/s10499-015-9962-z

Loughnan, S.R., Domingos, J.A., Smith-Keune, C., Forrester, J.P., Jerry, D.R., Beheregaray, L.B., Robinson, N.A., 2013. Broodstock contribution after mass spawning and size grading in barramundi (*Lates calcarifer*, Bloch). Aquaculture 404–405, 139–149. doi:10.1016/j.aquaculture.2013.04.014

Lush, J.L., 1947. Family Merit and Individual Merit as Bases for Selection. Part I. Am. Nat. 81, 241–261. doi:10.1086/281520

Macqueen, D.J., Johnston, I.A., 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc. R. Soc. B Biol. Sci. 281, 20132881–20132881. doi:10.1098/rspb.2013.2881

McNutt, M., 2014. Journals unite for reproducibility. Science (80-.). 346, 679–679. doi:10.1126/science.aaa1724

Melià, P., Gatto, M., 2005. A stochastic bioeconomic model for the management of clam farming. Ecol. Modell. 184, 163–174. doi:10.1016/j.ecolmodel.2004.11.011

Meuwissen, T.H.E, 1997. Maximizing the response of selection with a predefined rate of inbreeding. J. Anim. Sci. 75, 934. doi:10.2527/1997.754934x

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829. doi:11290733

Meuwissen, T.H.E., Woolliams, J.A., 1994. Effective sizes of livestock populations to prevent a decline in fitness. Theor. Appl. Genet. 89–89, 1019–1026. doi:10.1007/BF00224533

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., Johnson, E.A., 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17, 240–248. doi:10.1101/gr.5681207

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., Johnson, E.A., 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17, 240–248. doi:10.1101/gr.5681207

Moav, R., Wohlfarth, G., 1976. Two-way selection for growth rate in the common carp (*Cyprinus carpio* L.). Genetics 82, 83–101.

Morvezen, R., Cornette, F., Charrier, G., Guinand, B., Lapègue, S., Boudry, P., Laroche, J., 2013. Multiplex PCR sets of novel microsatellite loci for the great scallop *Pecten maximus* and their application in parentage assignment. Aquat. Living Resour. 26, 207–213. doi:10.1051/alr/2013052

Muir, W.M., 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124, 342–355. doi:10.1111/j.1439-0388.2007.00700.x

Mylonas, C.C., Zohar, Y., Pankhurst, N., Kagawa, H., 2011. Reproduction and Broodstock Management, in: Sparidae. Wiley-Blackwell, pp. 95–131. doi:10.1002/9781444392210.ch4

Nash, C.E., 2011. Seeds in Antiquity (2000 BC to AD 500), in: The History of Aquaculture. Wiley-Blackwell, Oxford, UK, pp. 11–23. doi:10.1002/9780470958971.ch2

Navarro, A., Zamorano, M.J., Hildebrandt, S., Ginés, R., Aguilera, C., Afonso, J.M., 2009. Estimates of heritabilities and genetic correlations for growth and carcass traits in gilthead seabream (*Sparus auratus* L.), under industrial conditions. Aquaculture 289, 225–230. doi:10.1016/j.aquaculture.2008.12.024

Neira, R., Díaz, N.F., Gall, G.A.E., Gallardo, J.A., Lhorente, J.P., Manterola, R., 2006. Genetic improvement in Coho salmon (*Oncorhynchus kisutch*). I: Selection response and inbreeding depression on harvest weight. Aquaculture 257, 9–17. doi:10.1016/j.aquaculture.2006.03.002

Nguyen, C., Nguyen, T.G., Nguyen, L. Van, Pham, H.Q., Nguyen, T.H., Pham, H.T., Nguyen, H.T., Ha, T.T., Dau, T.H., Vu, H.T., Nguyen, D.D., Nguyen, N.T.T., Nguyen, N.H., Van Quyen, D., Chu, H.H., Dinh, K.D., 2016. De novo assembly and transcriptome characterization of major growth-related genes in various tissues of *Penaeus monodon*. Aquaculture 464, 545–553. doi:10.1016/j.aquaculture.2016.08.003

Nguyen, T.T.T., Hayes, B.J., Ingram, B.A., 2014. Genetic parameters and response to selection in blue mussel (*Mytilus galloprovincialis*) using a SNP-based pedigree. Aquaculture 420–421, 295–301. doi:10.1016/j.aquaculture.2013.11.021

Nielsen, H.M., Sonesson, A.K., Yazdi, H., Meuwissen, T.H.E., 2009. Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes. Aquaculture 289, 259–264. doi:10.1016/j.aquaculture.2009.01.027

Nilsson, J., 1992. Genetic Variation in Resistance of Arctic Char to Fungal Infection. J. Aquat. Anim. Health 4, 126–128. doi:10.1577/1548-8667(1992)004<0126:GVIROA>2.3.CO;2

Norris, a T., Bradley, D.G., Cunningham, E.P., 2000. Parentage and relatedness determination in farmed Atlantic salmon (*Salmo salar*) using microsatellite markers. Aquaculture 182, 73–83. doi:10.1016/S0044-8486(99)00247-1

Novel, P., Porta, J.M., Porta, J., Béjar, J., Alvarez, M.C., 2010. PCR multiplex tool with 10 microsatellites for the European seabass (*Dicentrarchus labrax*) — Applications in genetic differentiation of populations and parental assignment. Aquaculture 308, S34–S38. doi:10.1016/j.aquaculture.2010.06.032

O'Flynn, F.M., Bailey, J.K., Friars, G.W., 1999. Responses to two generations of index selection in Atlantic salmon (*Salmo salar*). Aquaculture 173, 143–147. doi:10.1016/S0044-8486(98)00482-7

O'Reilly, P.T., Herbinger, C., Wright, J.M., 1998. Analysis of Parentage Determination in Atlantic Salmon (*Salmo Salar*). Anim. Genet. 29, 363–370.

O'Reilly, P.T., Hamilton, L.C., McConnell, S.K., Wright, J.M., 1996. Rapid analysis of genetic variation in Atlantic salmon (*Salmo salar*) by PCR multiplexing of dinucleotide and tetra-nucleotide microsatellites. Can. J. Fish. Aquat. Sci. 53, 2292–2298. doi:10.1139/cjfas-53-10-2292

Ødegård, J., Moen, T., Santi, N., Korsvoll, S.A., Kjøglum, S., Meuwisse, T.H.E., 2014. Genomic prediction in an admixed population of Atlantic salmon (*Salmo salar*). Front. Genet. 5, 1–8. doi:10.3389/fgene.2014.00402

Palti, Y., Gao, G., Liu, S., Kent, M.P., Lien, S., Miller, M.R., Rexroad, C.E., Moen, T., 2015. The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. Mol. Ecol. Resour. 15, 662–672. doi:10.1111/1755-0998.12337

Pante, M.J.R., Gjerde, B., McMillan, I., 2001. Inbreeding levels in selected populations of rainbow trout, *Oncorhynchus mykiss*. Aquaculture 192, 213–224. doi:10.1016/S0044-8486(00)00466-X

Pardo, B.G., Machordom, A., Foresti, F., Porto-Foresti, F., Azevedo, M.F.C., Bañon, R., Sánchez, L., Martínez, P., 2005. Phylogenetic analysis of flatfish (Order Pleuronectiformes) based on mitochondrial 16s rDNA sequences. Sci. Mar. 69, 531–543. doi:10.3989/scimar.2005.69n4531

Paterson, S., Piertney, S.B., Knox, D., Gilbey, J., Verspoor, E., 2004. Characterization and PCR multiplexing of novel highly variable tetranucleotide Atlantic salmon (*Salmo salar* L.) microsatellites. Mol. Ecol. Notes 4, 160–162. doi:10.1111/j.1471-8286.2004.00598.x

Perry, G.M., Danzmann, R.G., Ferguson, M.M., Gibson, J.P., 2001. Quantitative trait loci for upper thermal tolerance in outbred strains of rainbow trout (*Oncorhynchus mykiss*). Heredity (Edinb). 86, 333–341. doi:doi:10.1046/j.1365-2540.2001.00838.x

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7. doi:10.1371/journal.pone.0037135

Pompanon, F., Bonin, A., Bellemain, E., Taberlet, P., 2005. Genotyping errors: causes, consequences and solutions. Nat. Rev. Genet. 6, 847–859. doi:10.1038/nrg1707

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. doi:10.1086/519795

Quinton, C.D., McMillan, I., Glebe, B.D., 2005. Development of an Atlantic salmon (*Salmo salar*) genetic improvement program: Genetic parameters of harvest body weight and carcass quality traits estimated with animal models. Aquaculture 247, 211–217. doi:10.1016/j.aquaculture.2005.02.030

Ren, P., Peng, W., You, W., Huang, Z., Guo, Q., Chen, N., He, P., Ke, J., Gwo, J.C., Ke, C., 2016. Genetic mapping and quantitative trait loci analysis of growth-related traits in the small abalone *Haliotis diversicolor* using restriction-site-associated DNA sequencing. Aquaculture 454, 163–170. doi:10.1016/j.aquaculture.2015.12.026

Robinson, N.A., Gopikrishna, G., Baranski, M., Katneni, V., Shekhar, M.S., Shanmugakarthik, J., Jothivel, S., Gopal, C., Ravichandran, P., Gitterle, T., Ponniah, A.G., 2014. QTL for white spot syndrome virus resistance and the sex-determining locus in the Indian black tiger shrimp (*Penaeus monodon*). BMC Genomics 15, 731. doi:10.1186/1471-2164-15-731

Robinson, N.A., Schipp, G., Bosmans, J., Jerry, D.R., 2010a. Modelling selective breeding in protandrous, batch-reared Asian sea bass (*Lates calcarifer*, Bloch) using walkback selection. Aquac. Res. 41, 643–655. doi:10.1111/j.1365-2109.2010.02584.x

Robinson, N., Hayes, B., 2008. Modelling the use of gene expression profiles with selective breeding for improved disease resistance in Atlantic salmon (*Salmo salar*). Aquaculture 285, 38–46. doi:10.1016/j.aquaculture.2008.08.016

Robinson, N., Li, X., Hayes, B., 2010b. Testing options for the commercialization of abalone selective breeding using bioeconomic simulation modelling. Aquac. Res. 41. doi:10.1111/j.1365-2109.2010.02528.x

Robinson, P.N., 2012. Deep phenotyping for precision medicine. Hum. Mutat. 33, 777–780. doi:10.1002/humu.22080

Roff, D.A., 2002. Inbreeding depression: tests of the overdominance and partial dominance hypotheses. Evolution (N. Y). 56, 768–775. doi:10.1111/j.0014-3820.2002.tb01387.x

Rohde, D.L.T., Olson, S., Chang, J.T., 2004. Modelling the recent common ancestry of all living humans. Nature 431, 562–566. doi:10.1038/nature02842

Ross, C.T., Weise, J.A., Bonnar, S., Nolin, D., Satkoski Trask, J., Smith, D.G., Ferguson, B., Ha, J., Kubisch, H.M., Vinson, A., Kanthaswamy, S., 2014. An empirical comparison of short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs) for relatedness estimation in Chinese rhesus macaques (*Macaca mulatta*). Am. J. Primatol. 76, 313–324. doi:10.1002/ajp.22235

Rye, M., Refstie, T., 1995. Phenotypic and genetic parameters of body size traits in Atlantic salmon *Salmo Salar* L. Aquac. Res. 26, 875–885. doi:10.1111/j.1365-2109.1995.tb00882.x

Rye, M., Gjerde, B., 1996. Phenotypic and genetic parameters of body composition traits and flesh colour in Atlantic salmon, *Salmo salar* L. Aquac. Res. 27, 121–133. doi:10.1111/j.1365-2109.1996.tb00976.x

Rye, M., Lillevik, K.M., Gjerde, B., 1990. Survival in early life of Atlantic salmon and rainbow trout: estimates of heritabilities and genetic correlations. Aquaculture 89, 209–216. doi:10.1016/0044-8486(90)90126-8

Sánchez, C., Smith, T.P., Wiedmann, R.T., Vallejo, R.L., Salem, M., Yao, J., Rexroad, C.E., 2009. Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. BMC Genomics 10, 559. doi:10.1186/1471-2164-10-559

Sellars, M.J., Dierens, L., McWilliam, S., Little, B., Murphy, B., Coman, G.J., Barendse, W., Henshall, J., 2014. Comparison of microsatellite and SNP DNA markers for pedigree assignment in Black Tiger shrimp, *Penaeus monodon*. Aquac. Res. 45, 417–426. doi:10.1111/j.1365-2109.2012.03243.x

Shi, Y., Wang, S., Gu, Z., Lv, J., Zhan, X., Yu, C., Bao, Z., Wang, A., 2014. High-density single nucleotide polymorphisms linkage and quantitative trait locus mapping of the pearl oyster, *Pinctada fucata* martensii Dunker. Aquaculture 434, 376–384. doi:10.1016/j.aquaculture.2014.08.044

Skaarud, A., Woolliams, J.A., Gjøen, H.M., 2011. Strategies for controlling inbreeding in fish breeding programs; an applied approach using optimum contribution (OC) procedures. Aquaculture 311, 110–114. doi:10.1016/j.aquaculture.2010.11.023

Sonesson, A.K., Meuwissen, T.H.E., 2009. Testing strategies for genomic selection in aquaculture breeding programs. Genet. Sel. Evol. 41, 37. doi:10.1186/1297-9686-41-37

Sørensen, M.K., Sørensen, A.C., Baumung, R., Borchersen, S., Berg, P., 2008. Optimal genetic contribution selection in Danish Holstein depends on pedigree quality. Livest. Sci. 118, 212–222. doi:10.1016/j.livsci.2008.01.027

Takeuchi, T., Kawashima, T., Koyanagi, R., Gyoja, F., Tanaka, M., Ikuta, T., Shoguchi, E., Fujiwara, M., Shinzato, C., Hisata, K., Fujie, M., Usami, T., Nagai, K., Maeyama, K., Okamoto, K., Aoki, H., Ishikawa, T., Masaoka, T., Fujiwara, A., Endo, K., Endo, H., Nagasawa, H., Kinoshita, S., Asakawa, S., Watabe, S., Satoh, N., 2012. Draft genome of the pearl oyster Pinctada fucata: A platform for understanding bivalve biology. DNA Res. 19, 117–130. doi:10.1093/dnares/dss005

Tange, O., 2011. Gnu parallel-the command-line power tool. USENIX Mag. 36, 42–47.

Taylor, S., Wakem, M., Dijkman, G., Alsarraj, M., Nguyen, M., 2010. A practical approach to RT-qPCR-Publishing data that conform to the MIQE guidelines. Methods 50, S1–S5. doi:10.1016/j.ymeth.2010.01.005

Team, R.C., 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.

Teichert-Coddington, D.R., Smitherman, R.O., 1988. Lack of Response by Tilapia nilotica to Mass Selection for Rapid Early Growth. Trans. Am. Fish. Soc. 117, 297–300. doi:10.1577/1548-8659(1988)117<0297:LORBNT>2.3.CO;2

Ten Hoopen, P., Pesant, S., Kottmann, R., Kopf, A., Bicak, M., Claus, S., Deneudt, K., Borremans, C., Thijsse, P., Dekeyzer, S., Schaap, D.M., Bowler, C., Glöckner, F.O., Cochrane, G., 2015. Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards. Stand. Genomic Sci. 10, 20. doi:10.1186/s40793-015-0001-5

Thodesen, J., Rye, M., Wang, Y.X., Li, S.J., Bentsen, H.B., Yazdi, M.H., Gjedrem, T., 2013. Genetic improvement of tilapias in China: Genetic parameters and selection responses in growth, survival and external color traits of red tilapia (*Oreochromis* spp.) after four generations of multi-trait selection. Aquaculture 416–417, 354–366. doi:10.1016/j.aquaculture.2013.09.047

Thorpe, J.E., Miles, M.S., Keay, D.S., 1984. Developmental rate, fecundity and egg size in Atlantic salmon, *Salmo salar* L. Aquaculture 43, 289–305. doi:10.1016/0044-8486(84)90030-9

Toonen, R.J., Puritz, J.B., Forsman, Z.H., Whitney, J.L., Fernandez-Silva, I., Andrews, K.R., Bird, C.E., 2013. ezRAD: a simplified method for genomic genotyping in non-model organisms. PeerJ 1, e203. doi:10.7717/peerj.203

Trygve Gjedrem, Baranski, M., 1990. Selective breeding in aquaculture. Food Rev. Int. 6, 359–372. doi:10.1080/87559129009540877

Tsai, H.Y., Hamilton, A., Tinch, A.E., Guy, D.R., Gharbi, K., Stear, M.J., Matika, O., Bishop, S.C., Houston, R.D., 2015. Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. BMC Genomics 1–9. doi:10.1186/s12864-015-2117-9

Tsai, H.-Y., Hamilton, A., Tinch, A.E., Guy, D.R., Bron, J.E., Taggart, J.B., Gharbi, K., Stear, M., Matika, O., Pong-Wong, R., Bishop, S.C., Houston, R.D., 2016. Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations. Genet. Sel. Evol. 48, 47. doi:10.1186/s12711-016-0226-9

UN – United Nations, 2015. World Population Prospects: The 2015 Revision. United Nations Econ. Soc. Aff. XXXIII, 1–66. doi:10.1007/s13398-014-0173-7.2

Vallejo, R.L., Palti, Y., Liu, S., Marancik, D.P., Wiens, G.D., 2014. Validation of linked QTL for bacterial cold water disease resistance and spleen size on rainbow trout chromosome Omy19. Aquaculture 432, 139–143. doi:10.1016/j.aquaculture.2014.05.003

Vandeputte, M., Haffray, P., 2014. Parentage assignment with genomic markers: a major advance for understanding and exploiting genetic variation of quantitative traits in farmed aquatic animals. Front. Genet. 5, 1–8. doi:10.3389/fgene.2014.00432

Vardar, H., Yıldırım, Ş., 2012. Effects of long-term extended photoperiod on somatic growth and husbandry parameters on cultured gilthead seabream (*Sparus aurata*, L.) in the net cages. Turkish J. Fish. Aquat. Sci. 12.

Visscher, P.M., Woolliams, J. a, Smith, D., Williams, J.L., 2002. Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. J. Dairy Sci. 85, 2368–75. doi:10.3168/jds.S0022-0302(02)74317-8

Wang, C.M., Lo, L.C., Zhu, Z.Y., Lin, G., Feng, F., Li, J., Yang, W.T., Tan, J., Chou, R., Lim, H.S., Orban, L., Yue, G.H., 2008. Estimating reproductive success of brooders and heritability of growth traits in Asian sea bass (*Lates calcarifer*) using microsatellites. Aquac. Res. 39, 1612–1619. doi:10.1111/j.1365-2109.2008.02034.x

Wang, J., Santiago, E., Caballero, A., 2016. Prediction and estimation of effective population size. Heredity (Edinb). 1–14. doi:10.1038/hdy.2016.43

Wang, L., Feng, Z., Wang, X., Wang, X., Zhang, X., 2009. DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26, 136–138. doi:10.1093/bioinformatics/btp612

Wang, S., Meyer, E., McKay, J.K., Matz, M. V, 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat. Methods 9, 808–810. doi:10.1038/nmeth.2023

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63. doi:10.1038/nrg2484

Weinman, L.R., Solomon, J.W., Rubenstein, D.R., 2015. A comparison of single nucleotide polymorphism and microsatellite markers for analysis of parentage and kinship in a cooperatively breeding bird. Mol. Ecol. Resour. 15, 502–511. doi:10.1111/1755-0998.12330

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D., 2012. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS One 7, e29715. doi:10.1371/journal.pone.0029715

Wiggans, G.R., Sonstegard, T.S., Vanraden, P.M., Matukumalli, L.K., Schnabel, R.D., Taylor, J.F., Chesnais, J.P., Schenkel, F.S., Tassell, C.P. Van, Alliance, S., 2008. ICAR 2008 Genomic Evaluations in the United States and Canada : A Collaboration Field test, in: Animals. pp. 1–6.

Wiggans, G.R., Sonstegard, T.S., VanRaden, P.M., Matukumalli, L.K., Schnabel, R.D., Taylor, J.F., Chesnais, J.P., Schenkel, F.S., Van Tassell, C.P., 2009. Genomic evaluations in the United States and Canada: A collaboration, in: ICAR Technical Series. International Committee for Animal Recording (ICAR), Rome, pp. 347–353.

Wiggans, G.R., VanRaden, P.M., Cooper, T.A., 2011. The genomic evaluation system in the United States: Past, present, future. J. Dairy Sci. 94, 3202–3211. doi:10.3168/jds.2010-3866

Winkelman, A.M., Peterson, R.G., 1994. Genetic parameters (heritabilities, dominance ratios and genetic correlations) for body weight and length of chinook salmon after 9 and 22 months of saltwater rearing. Aquaculture 125, 31–36. doi:10.1016/0044-8486(94)90279-8

Yáñez, J.M., Naswa, S., López, M.E., Bassini, L., Correa, K., Gilbey, J., Bernatchez, L., Norris, A., Neira, R., Lhorente, J.P., Schnable, P.S., Newman, S., Mileham, A., Deeb, N., Di Genova, A., Maass, A., 2016. Genome-wide single nucleotide polymorphism (SNP) discovery in Atlantic salmon (*Salmo salar*): validation in wild and farmed American and European populations. Mol. Ecol. Resour. n/a-n/a. doi:10.1111/1755-0998.12503

Yano, A., Guyomard, R., Nicol, B., Jouanno, E., Quillet, E., Klopp, C., Cabau, C., Bouchez, O., Fostier, A., Guiguen, Y., 2012. An Immune-Related Gene Evolved into the Master Sex-Determining Gene in Rainbow Trout, *Oncorhynchus mykiss*. Curr. Biol. 22, 1423–1428. doi:10.1016/j.cub.2012.05.045

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J.R., Amaral-Zettler, L., Gilbert, J.A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., Birren, B.W., Blaser, M.J., Bonazzi, V., Booth, T., Bork, P., Bushman, F.D., Buttigieg, P.L., Chain, P.S.G., Charlson, E., Costello, E.K., Huot-Creasy, H., Dawyndt, P., Desantis, T., Fierer, N., Fuhrman, J.A., Gallery, R.E., Gevers, D., Gibbs, R.A., Gil, I.S., Gonzalez, A., Gordon, J.I., Guralnick, R., Hankeln, W., Highlander, S., Hugenholtz, P., Jansson, J., Kau, A.L., Kelley,

S.T., Kennedy, J., Knights, D., Koren, O., Kuczynski, J., Kyrpides, N., Larsen, R., Lauber, C.L., Legg, T., Ley, R.E., Lozupone, C.A., Ludwig, W., Lyons, D., Maguire, E., Methé, B.A., Meyer, F., Muegge, B., Nakielny, S., Nelson, K.E., Nemergut, D., Neufeld, J.D., Newbold, L.K., Oliver, A.E., Pace, N.R., Palanisamy, G., Peplies, J., Petrosino, J., Proctor, L., Pruesse, E., Quast, C., Raes, J., Ratnasingham, S., Ravel, J., Relman, D.A., Assunta-Sansone, S., Schloss, P.D., Schriml, L., Sinha, R., Smith, M.I., Sodergren, E., Spor, A., Stombaugh, J., Tiedje, J.M., Ward, D.V., Weinstock, G.M., Wendel, D., White, O., Whiteley, A., Wilke, A., Wortman, J.R., Yatsunenko, T., Glöckner, F.O., 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat. Biotechnol. 29, 415–420. doi:10.1038/nbt.1823

Yue, G.H., 2014. Recent advances of genome mapping and marker-assisted selection in aquaculture. Fish Fish. 15, 376–396. doi:10.1111/faf.12020

Zak, T., Deshev, R., Benet-Perlberg, A., Naor, A., Magen, I., Shapira, Y., Ponzoni, R.W., Hulata, G., 2014. Genetic improvement of Israeli blue (Jordan) tilapia, *Oreochromis aureus* (Steindachner), through selective breeding for harvest weight. Aquac. Res. 45, 546–557. doi:10.1111/are.12072

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., Xiong, Z., Que, H., Xie, Y., Holland, P.W.H., Paps, J., Zhu, Y., Wu, F., Chen, Y., Wang, J., Peng, C., Meng, J., Yang, L., Liu, J., Wen, B., Zhang, N., Huang, Z., Zhu, Q., Feng, Y., Mount, A., Hedgecock, D., Xu, Z., Liu, Y., Domazet-Lošo, T., Du, Y., Sun, X., Zhang, S., Liu, B., Cheng, P., Jiang, X., Li, J., Fan, D., Wang, W., Fu, W., Wang, T., Wang, B., Zhang, J., Peng, Z., Li, Y., Li, N., Wang, J., Chen, M., He, Y., Tan, F., Song, X., Zheng, Q., Huang, R., Yang, H., Du, X., Chen, L., Yang, M., Gaffney, P.M., Wang, S., Luo, L., She, Z., Ming, Y., Huang, W., Zhang, S., Huang, B., Zhang, Y., Qu, T., Ni, P., Miao, G., Wang, J., Wang, Q., Steinberg, C.E.W., Wang, H., Li, N., Qian, L., Zhang, G., Li, Y., Yang, H., Liu, X., Wang, J., Yin, Y., Wang, J., 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490, 49–54. doi:10.1038/nature11413

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., Weir, B.S., 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326–3328. doi:10.1093/bioinformatics/bts606

Ziemann, M., Eren, Y., El-Osta, A., 2016. Gene name errors are widespread in the scientific literature. Genome Biol. 17, 177. doi:10.1186/s13059-016-1044-7

Zohar, Y., Abraham, M., Gordin, H., 1978. The gonadal cycle of the captivity-reared hermaphroditic teleost *Sparus aurata* (L.) during the first two years of life. Ann. Biol. Anim. Biochim. Biophys. 18, 877–882. doi:10.1051/rnd:19780519