THE UNIVERSITY *of* EDINBURGH

Cost-effective Genomic Selection in Aquaculture Breeding Programmes:

Optimizing Genotype Imputation and Incorporation of Functional Annotation

Christina Kriaridou

THE UNIVERSITY *of* EDINBURGH

College of Medicine and Veterinary Medicine

The Royal (Dick) School of Veterinary Studies

Dissertation submitted for the degree of Doctor of Philosophy

Genetics and Genomics

The Roslin Institute

2024

Declaration of authentication

Edinburgh, May 2024

I hereby declare that this thesis, entitled “Cost-effective genomic selection in aquaculture breeding programmes: optimizing genotype imputation and incorporation of functional annotation” has been composed by myself. The work and illustrations contained herein are my own, except where explicitly acknowledged otherwise in the text. Any included publications is my own work, except where indicated throughout the thesis and summarised and clearly identified on the declarations page of the thesis. This work has not been submitted, either in whole or in part, for any other degree or professional qualification of any other institution.

The contributions of other individuals to this work have been explicitly indicated below.

Chapter 1: This chapter was written by me with feedback from Dr Diego Robledo and Dr Clémence Fraslin.

Chapter 2: This chapter was published, and author contributions are listed at the end of the document.

Chapter 3: John A. H. Benzie from Worldfish provided us with the Nile tilapia samples and DNA extractions were performed by Dr Carolina Peñaloza (The Roslin Institute). Dr Agustin Barría (The Roslin Institute) contributed to the generation and the down-sampling of the whole-genome sequencing (WGS) data.

Chapter 4: The traits used for the analyses in this chapter were inferred by Dr Jamie Prentice (The Roslin Institute) from Prof Andrea Wilson’s group using the SIRE model (Pooley *et al.*, 2020). Information on the position of regulatory elements and their overlap with single nucleotide polymorphisms and structural variants were the result of AQUA-FAANG project and provided to us by Paulino Martínez and Andrés Blanco from University of Santiago de Compostela. Additionally, Zexin Jiao from Prof Dan Macqueen’s group discovered the structural variants (SVs) of the WGS turbot parents, which were incorporated in the analysis

together with the SNPs.

Chapter 5: This chapter was written by me with feedback from Dr Diego Robledo and Dr Clémence Fraslin.

All the computational work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).

Christina Kriaridou

Dedication

It is often said that our family and the people we meet form one of the most important aspects of our lives. This thesis is dedicated to the precious people who have been part of my life's journey, and to those whose impact has been transformative.

Acknowledgements

First of all, I want to thank my supervisor Dr Diego Robledo for his invaluable support and encouragement throughout these years. Thank you for believing in me and supporting my pursuit of a PhD after completing my Master's degree. This decision has been a crucial turning point that has shaped my life over the past four years. Thank you for your positive attitude, for teaching me so much, and for being there when I needed it. Words cannot express how special you are to me and how grateful I am to have you as my mentor and friend.

I was also fortunate enough to receive mentorship from Professor Ross Houston, and I want to express my sincere gratitude for his support and guidance as my supervisor at the beginning of my PhD. I am equally appreciative for Professor Gregor Gorjanc's valuable contribution and thankful that he stepped in as my second supervisor. Additionally, I would like to wholeheartedly thank Professor Andrea Wilson and Professor Paulino Martínez for their collaboration. I gained valuable knowledge from working both alongside them and alongside the fantastic people in their labs.

I would like to thank Dr Smaragda Tsairidou for all the precious time and help she provided when I first arrived at The Roslin Institute for my Erasmus placement in 2019, and subsequently as a co-supervisor during the first year of my PhD studies. I have greatly benefited from and been inspired by your expertise, enthusiasm and support.

I am more than grateful to Dr Clémence Fraslin for being my co-supervisor. I appreciate all of your guidance, the knowledge you've shared, and your companionship. I've enjoyed our discussions about science and life in general. Thank you for being an amazing scientist, colleague, friend, and for all the experiences we have shared. I look forward to the moments we will share in the future.

I am extremely grateful to all the wonderful people in the aquaculture team for all the activities, conferences, lab meetings, trips, and happy moments we have shared together. A special thanks to my friends Dr Carolina Peñaloza, Dr Carolina Moraleda, Dr Yehwa Jin, Dr Sarah Salisbury, Dr Paula Rodriguez Villamayor, Dr

Agustin Barria, Assistant Professor Manu Kumar Gundappa, Dr Rose Ruiz Daniels, Dr Tim Regan, Dr Tim Bean, Dr Robert Mukiibi, Hannah Farley, Alex Florea, Dr Jennifer Nascimento-Schulze, Dr Robert Potts, Robert Stewart, Dr Remi Gratacap, Dr Athina Papadopoulou, Dr Panagiotis Kokkinias, Marina Mantsopoulou, Dr Maeve Ballantyne, Dr Nunticha Pankaew, Dana Albatesh, Dr Ambre Chapuis, Dr Lavanya Vythalingam, Dr Diego Perojil Morata, Oliver Eve, Dr Ophélie Gervais and Jiaqi Wang for being part of this unforgettable experience. You all made me feel welcome from the beginning, as if I were part of a bigger family. I feel incredibly fortunate to have you in my life. I hope we will continue to share more experiences in the future and meet again in this small world we live in.

Thank you to my friends and desk neighbours at Roslin, Dr Valentina Riggio, Dr Enrique Sanchez Molano, Dr Ricardo Pong-Wong, Dr Masoud Ghaderi Zefreh, Dr Juliane Friedrich, Dr Cammy Beyts, Dr Ismail Ozkaraca, Dr Mohamed Aboelela, Dr Mulya Agung, Dr June Bhak, Vasilios Raptis, Dr Meenu Bhati, Jing Qi Chong and Dr Anna Eleonora Karagianni, for your company during coffee and lunch breaks, for all the fun activities we planned together, the laughs and friendship.

A big thank you to my parents and family for their unconditional love and support. I am forever grateful for everything you have provided me. I couldn't ask for better support from my husband, who moved with me from Greece to be by my side from the beginning of this journey during the pandemic, and played a key role in making this dream happen. Thank you for your positive thoughts, your encouragement that kept me sane and your love that gave me strength to continue in difficult times.

A special thank you to the people at Xelect Ltd for their warm welcome during my placement in St. Andrews. I am especially thankful to Professor Ian Johnston and Dr Mark Looseley for their collaboration and valuable contributions. I want to thank Dr Lidia de los Ríos Pérez for inviting me to her hospitable home with her beautiful cats, but also Chris Wallard, Dr Tom Ashton, Dr Marie Smedley, Dr José Mota-Velasco, Dr Carlos Diaz Gil, Dr Rachael Wilbourn, Dr Stelios Kyriakidis, Dr Kyriakos Varypatakis, Dr Kyriakos Varypatakis, Dr Paolo Ruggeri, Dr Max Coulter, Jeroen ven der Kaay and Anne Shaw. It was my pleasure to meet you all, and I

am very grateful for all the constructive discussions, the friendly talks, and the time we spent together.

I would like to express my gratitude to the University of Edinburgh and The Roslin Institute for awarding me the Principal's Career Development PhD Scholarship, which made my dream possible. Additionally, I extend my sincere appreciation to Xelect Ltd for partially funding this postgraduate studentship project.

Table of Contents

Lay summary

Selection of individuals using information of their genome holds promise for improving traits of economic importance in aquaculture. However, obtaining information about the genome is expensive, and therefore it has been limited to large companies and a few high value aquatic species. To tackle this challenge, a method called genotype imputation offers a possible solution by predicting the variation in the genomes of a target population based on sparse genomic information, which is cheaper to produce. In order to predict the missing information of a genome, a smaller population that is related to the target population and has more information available is used as reference for imputation. This project aims to explore genotype imputation strategies to make genomic selection more accessible for aquaculture farms.

In the first chapter of this project, I investigated genotype imputation using different densities of genome markers in four aquaculture species: Atlantic salmon, turbot, common carp, and Pacific oyster. For this, panels with different number of genome markers were created and three imputation software were used to test how accurately they can predict the missing information. Results showed that imputed panels performed similarly to panels containing a large number of variants for predicting of phenotypes, except in Pacific oysters. This suggests that optimizing variant selection methods for the creation of panels with a small number of variants together with imputation could lower genotyping costs for accurate prediction of individual phenotypes in breeding programmes.

Next, I examined imputation accuracy when using low coverage whole-genome sequencing in Nile tilapia. Different sequencing depths were tested, and imputation accuracy was compared between reference panels with different depths. Results showed higher accuracy with deeper reference sequencing. While low-coverage sequencing is cheaper than whole-genome sequencing and can provide a considerable amount of information for subsequent analyses after imputation, it remains more expensive than single nucleotide polymorphism (SNP) arrays and is therefore unlikely to be a viable strategy for the industry.

Finally, I analysed a dataset from an experiment involving a turbot population exposed to a parasite. Offspring with limited genome information were imputed to

their parents' whole-genome sequences, and the influence of integrating functional information into genomic prediction models was explored. Functional information can help us better understand how DNA regions influence traits and in theory could improve the ability of the models to choose individuals with the best characteristics. However, the results of this study showed that adding functional annotation information did not improve prediction accuracy compared to the models without annotation. Further research is needed to refine annotation categories for different traits and optimize predictive models.

Abstract

Genomic selection has the potential to significantly enhance genetic progress in aquaculture breeding programmes. However, its widespread adoption has been limited to large companies and a few high value species, mainly due to the high costs associated with genotyping. To address this issue, genotype imputation emerges as a promising solution, offering the possibility of reducing genotyping expenses by predicting ungenotyped variants in low-density and low-coverage datasets. Additionally, it has been hypothesized that the prioritisation of functional variants could allow accurate selection across distant relatives, which could reduce the need for genotyping every generation. This study aims to explore genotype imputation strategies and variant prioritization with the goal of democratizing genomic selection in aquaculture.

In the first chapter I investigated genotype imputation of low-density panels across four aquaculture species: Atlantic salmon, turbot, common carp, and Pacific oyster. A total of eight low-density panels were constructed *in silico* for each species, ranging from 300 to 6000 SNPs, and imputation to high-density was tested using three available genotype imputation software (AlphaImpute v.2, FImpute v.3 and findhap v.4). Subsequently, the genomic prediction accuracy of the various densities was evaluated for each species using the imputed genotypes. Results revealed that FImpute v.3 is the best performing software for parents-to-offspring imputation in aquaculture populations. In terms of prediction accuracy, the low-density and imputed panels generally performed comparably to the high-density panels in fish species. However, for Pacific oyster imputation and genomic prediction accuracy results were significantly lower. Nonetheless, the optimisation of SNP selection for the design of low-density panels may be sufficient to achieve near maximum prediction accuracy in most fish species/populations, suggesting a potential opportunity for reducing genotyping costs.

In the second chapter, I examined the accuracy of imputation from low-coverage re-sequencing to whole-genome sequencing in a Nile tilapia population. For the target offspring, we tested six down-sampled datasets representing varying sequencing depths from 0.1X to 5X. These datasets were imputed using

GLIMPSE v.1 to two reference panels with whole-genome sequence data: one at 5X sequencing depth and another at 26X sequencing depth. Finally, the cost of the different low-coverage whole-genome sequenced datasets was compared to that of SNP arrays in a hypothetical scenario involving 140 parents and 2100 offspring. Results revealed that imputation accuracy and the number of retained SNPs were higher with a 26X reference sequencing depth compared to 5X. Additionally, imputation accuracy exceeded 90% for all down-sampled target datasets, with higher accuracy at homozygous compared to heterozygous sites for coverages below 5X. While imputation of low-coverage whole-genome sequencing is cheaper than whole-genome sequencing and holds potential benefits for discovering causative variants as well as other genomic analyses between populations, it still remains more expensive than SNP arrays, and therefore is probably not a viable strategy for aquaculture breeding programmes.

In the final experimental chapter, a dataset involving a turbot population exposed to the parasite *Plilasterides dicentrarchi* was analysed. The genotypes of the offspring were imputed to whole-genome genotypes using the whole-genome re-sequenced parents as reference. Various approaches for integrating functional information data into genomic prediction were explored to study the potential advantages of integrating such information. The methodology involved categorizing markers into functional annotations by examining their overlap with regions of the genome potentially influencing protein function or promoter and enhancer regions. Two Bayesian models were tested with and without annotation for comparison alongside GBLUP. It was observed that the integration of functional annotation data did not enhance genomic prediction accuracy, with BayesR and GBLUP demonstrating superior performance or comparable results in certain scenarios. Future investigations should explore different approaches to defining annotation categories for traits with different architectures to optimize the predictive models.

List of Abbreviations

Abbreviation	Full name
AMBP	Aquaculture molecular breeding platform
ANN	Artificial neural networks
AQUA-FAANG	Advancing European Aquaculture by Genome Functional Annotation of Animal Genomes
ATAC-seq	Assay for transposase-accessible chromatin with sequencing
BLUP	Best Linear unbiased prediction
CDS	Coding sequence
ChiP-seq	Chromatin immunoprecipitation assays with sequencing
CV	Coefficient of variation
DGRP	Drosophila genetic reference panel
DL	Deep learning
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded ribonucleic acid
DT	Decision trees
EBV	Estimated breeding values
FAANG	Functional annotation of animal genomes
GBLUP	Genomic best linear unbiased prediction
GBM	Gradient boost machine
GBS	Genotyping-by-sequencing
GEBV	Genomic estimated breeding value
GIFT	Genetically improved farmed tilapia
GL	Genotype likelihood
GMBLUP	Genomic best linear unbiased prediction based on metabolite data
GO	Gene ontology
GRBLUP	Genomic best linear unbiased prediction with a transcriptome-based Gaussian kernel
GRM	Genomic relationship matrix
GS	Genomic selection

GTBLUP	Genomic best linear unbiased prediction with a transcriptome-based linear kernel
GTiBLUP	Best linear unbiased prediction based on the genome, the transcriptome and/or their interaction
GTMBLUP	Best linear unbiased prediction based on the genome and metabolite data
GWAS	Genome wide association study
hcWGS	High-coverage whole-genome sequencing
HMM	Hidden Markov model
INFO/DP	Combined sequencing depth across samples
IPN	Infectious pancreatic necrosis
lcWGS	Low-coverage whole-genome sequencing
LD	Low-density
LOCO	Leave-one-chromosome-out
MAF	Minor allele frequency
MAS	Marker assisted selection
MBLUP	Best linear unbiased prediction based on metabolite data
MCMC	Markov chain Monte Carlo
MLMA	Mixed linear model association
MME	Mixed model equations
MQ	Mapping quality
PHG	Practical haplotype graph
PIT tags	Passive integrated transporter tags
PPP	Pathogens, parasites and pests
QTL	Quantitative trait loci
QUAL	Quality score
RAD-seq	Restriction-site associated DNA sequencing
2bRAD-seq	Restriction-site associated DNA sequencing using type IIB restriction enzymes (BsaXI or AlfI) that cleave upstream and downstream of a recognition site

RF	Random Forest
RKHS	Reproducing kernel Hilbert spaces
RNA	Ribonucleic acid
sd	Standard deviation
SIP	Selection index procedure
SIRE	Susceptibility, infectivity and recoverability estimation
SNP	Single nucleotide polymorphism
SPF	Specific pathogen free
ssGBLUP	Single-step genomic best unbiased prediction
SV	Structural variant
SVM	Support vector machine
TBLUP	Best linear unbiased prediction based on gene transcript data
TBV	True breeding value
UTR	Untranslated region
VCF	Variant call format
VNN	Viral nervous necrosis
VIE tags	Visible implant Elastomer
WGS	Whole-genome sequencing

List of Figures

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

Supplementary figures

Chapter 2:

List of Tables

Chapter 2:

Chapter 3:

Chapter 4:

Supplementary tables

Chapter 2:

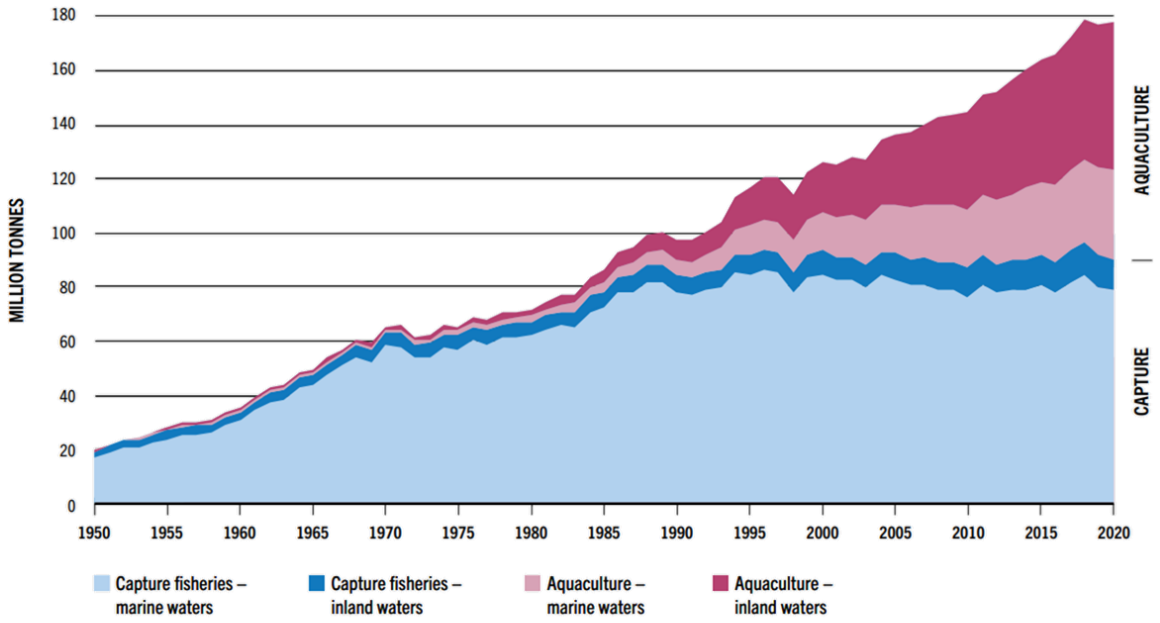
Chapter 4:

Chapter 1
General Introduction

1.1 Aquaculture

1.1.1 Global Growth

Aquaculture has experienced significant growth in recent years and has become a major source of seafood production globally. As one of the fastest growing food production sectors in the world, aquaculture production experienced a rise of over 600% in annual output in the period between 1990 and 2020 (around 400% if we exclude aquatic mammals, crocodiles, alligators, caimans and algae[1]), with an average annual growth rate of 6.7% since 1976 (FAO, 2022). The total global production of aquatic animals was estimated at 177.8 million tonnes in 2020, and aquaculture represented 49% of the total (87.5 million tonnes in live weight), with a production value of US $264.8 billion (FAO, 2022) (**Figure 1.1**).



NOTES: Excluding aquatic mammals, crocodiles, alligators, caimans and algae. Data expressed in live weight equivalent.
SOURCE: FAO.

Figure 1.1 World capture fisheries and aquaculture production (1950-2020), modified figure from FAO's 2022 report.

[1] Include multicellular macroalgae (e.g., seaweeds), unicellular microalgae (e.g., *Chlorella* spp.) and Cyanobacteria, not true algae but informally known as "blue-green algae" (e.g., *Spirulina* spp.).

Aquaculture production has finally reached practically the same levels as capture fisheries production. However, capture fisheries are still exerting serious pressures on wild stocks, and in fact, production has been almost stagnant since the end of the 1990s (FAO, 2022) (**Figure 1.1**). Overfishing combined with other human related interventions and the impact of climate change have adversely affected many aquatic species and ecosystems (Longo *et al.*, 2019). Moreover, in 2019, the Food and Agriculture Organization of the United Nations (FAO) reports that 35.4% of stocks were fished at biologically unsustainable levels (FAO, 2022). The world is now facing a challenge to feed the growing population but at the same time to maintain steady levels of natural resources without exhausting them and causing ecological damage. To protect wild fish stocks, aquaculture has to increase its production to supply the growing demand of aquatic foods. Projections from FAO, based on specific assumptions, estimate that the share of farmed aquatic species production will grow from 49% to 53% in 2030, surpassing capture fisheries production (FAO, 2022).

Aquaculture growth is not only fundamental for the preservation of aquatic habitats across all continents, but also for the provision of a vital source of nutritious and accessible animal protein for human diets. Projections show that the increasing world population will continue to drive an increased demand for protein in the coming decades, and proportionally this demand is expected to be higher for aquatic foods (FAO, 2022). Global apparent[2] consumption of aquatic foods increased at an average annual rate of 3.0 percent from 1961 to 2019, a rate almost twice that of annual world population growth (1.6 percent) for the same period (FAO, 2022). The rapid rise in aquaculture production resulted in increased availability of seafood and a decline in prices, particularly for the species that are mainly farmed rather than wild-caught. Through this continued aquaculture growth, people from both low and high-income countries have benefited from consistent access to aquatic foods throughout the year. Aquaculture production is expected to grow 22% by 2030, an increase of 106 million tonnes compared to 2020 (FAO,

[2] Proxy measure to indicate the supply of food available in a country for the indicated reference period. It refers to the amount available for human consumption and not to the effective food consumption, i.e., the actual quantity of food eaten, which can be measured through household or individual food consumption surveys. Apparent food consumption data are derived from FAO Food Balance Sheets and have been available on an annual basis at country level since 1961.

2022). The sustainable growth of global aquaculture production in the following years is required to meet the rising demand for aquatic food while concurrently creating employment opportunities and ensuring income stability.

Blue foods have the potential to provide essential nourishment to vulnerable populations and address issues of malnutrition. Aquatic species are not only an excellent source of protein, but they are also a healthy, lean source of micronutrients and essential fatty acids proven to offer many health benefits (Golden *et al.*, 2021). Fish stands out as a nutritious option, surpassing other animal protein sources (Maulu *et al.*, 2021). Its richness in long-chained n−3 polyunsaturated fatty acids (n-3 PUFAs) suggests that a diet abundant in fish may lower the likelihood of coronary heart diseases and certain types of cancer. Additionally, the presence of vitamins and minerals in seafood, in relatively high concentrations, is especially interesting for their role in preventing lifestyle disorders. The incorporation of fish into one's diet could potentially aid in alleviating instances of depression and anxiety (Gjedrem, Robinson and Rye, 2012a). Promoting healthy diets inclusively is critical and demands programmes and initiatives that enhance consumer awareness, as well as increase the availability of healthy, safe, and nutritious aquatic foods. This is especially crucial in areas with low food and nutrition security.

1.1.2 Diversity of Species

Aquaculture encompasses a wide variety of species, including fish, crustaceans, molluscs, other aquatic invertebrates, frogs, reptiles and even aquatic plants raised in different types of aquaculture farming systems (**Figure 1.2**). During 2021, 513 aquaculture species listed in ASFIS[3] were cultivated across 201 countries (FAO, 2023). While this richly diverse pool of aquatic species exists, several selected ‘core’ species or groups of species dominate aquaculture production. Approximately 70 species are responsible for supporting 80% of the total global aquaculture production (Houston *et al.*, 2020). Specifically, the major

[3] ASFIS = Aquatic Sciences and Fisheries Information System. ASFIS species items could refer to either individual species, hybrids, or groups of related species, such as genera (when identification to species is impossible).

species and species groups that contribute significantly to worldwide aquaculture production are: carps (*Ctenopharyngodon idellus*, *Hypophthalmichthys molitrix*, *Cyprinus carpio*, *Hypophthalmichthys nobilis*, *Mylopharyngodon piceus*), accounting for 38.3% of finfish production in inland aquaculture; Atlantic salmon (*Salmo salar*), representing 32.6% of finfish production in marine and coastal aquaculture; whiteleg shrimp (*Penaeus vannamei*), contributing 51.7% to total crustacean production; cupped oysters (*Crassostrea* spp.), constituting 30.7% of total mollusc production; and Japanese kelp (*Laminaria japonica*), ranking as the top species among algae with a 35.5% share of total production (FAO, 2022).

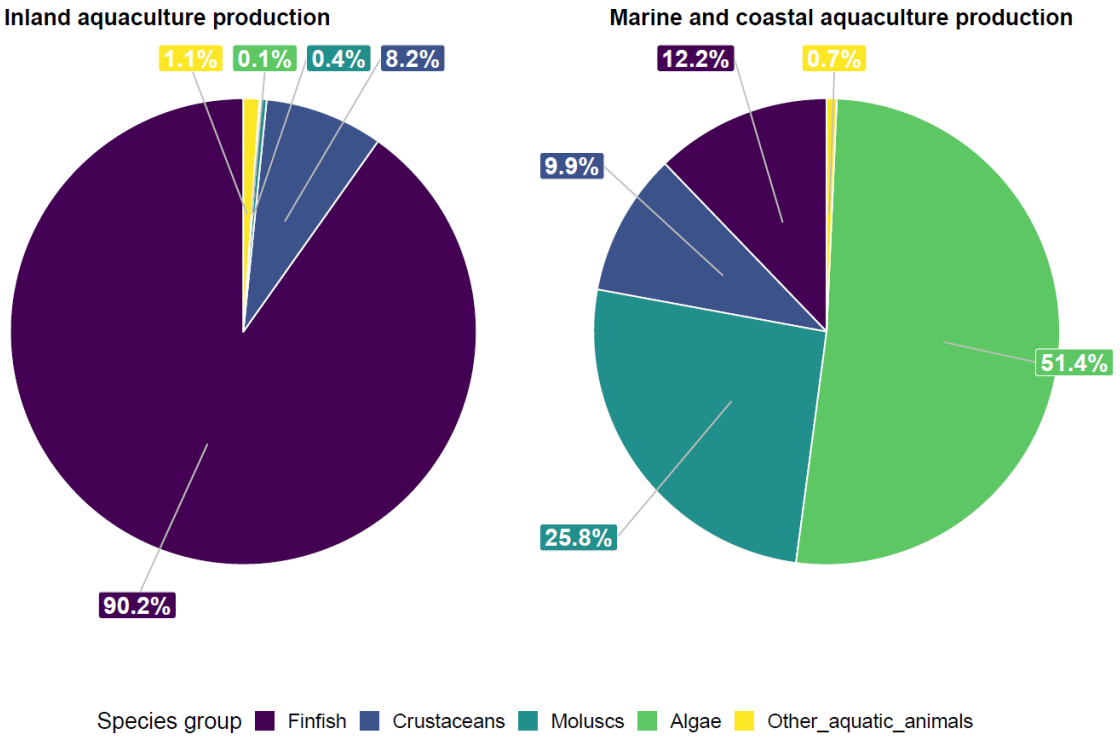


Figure 1.2 Inland aquaculture and marine and coastal aquaculture production by species group. This figure was created in R with data from Table 8 of FAO's report (FAO, 2022).

Although the addition of new species to further diversify aquaculture can effectively mitigate both biological and financial risks, the private sector often lacks the necessary incentives to diversify their farming practices (Cai, Yan and Leung, 2022). This hesitation arises from the considerable costs and associated risks of developing or adopting new species. However, the diversification of aquaculture not only offers a potential solution for risk management but also holds the promise

of enhancing production efficiency. This can be achieved through practices such as polyculture or adjusting species selection according to seasonality (Thomas *et al.*, 2021). By doing so, diversified aquaculture has the capacity to reduce both biological risks, such as diseases, and financial risks, such as price variations (Wilson and Archer, 2010). In light of increasing concerns related to climate change, disease outbreaks, market fluctuations, and other uncertainties, the strategy of species diversification has gained importance as a key approach for achieving sustainable aquaculture development (FAO, 2017).

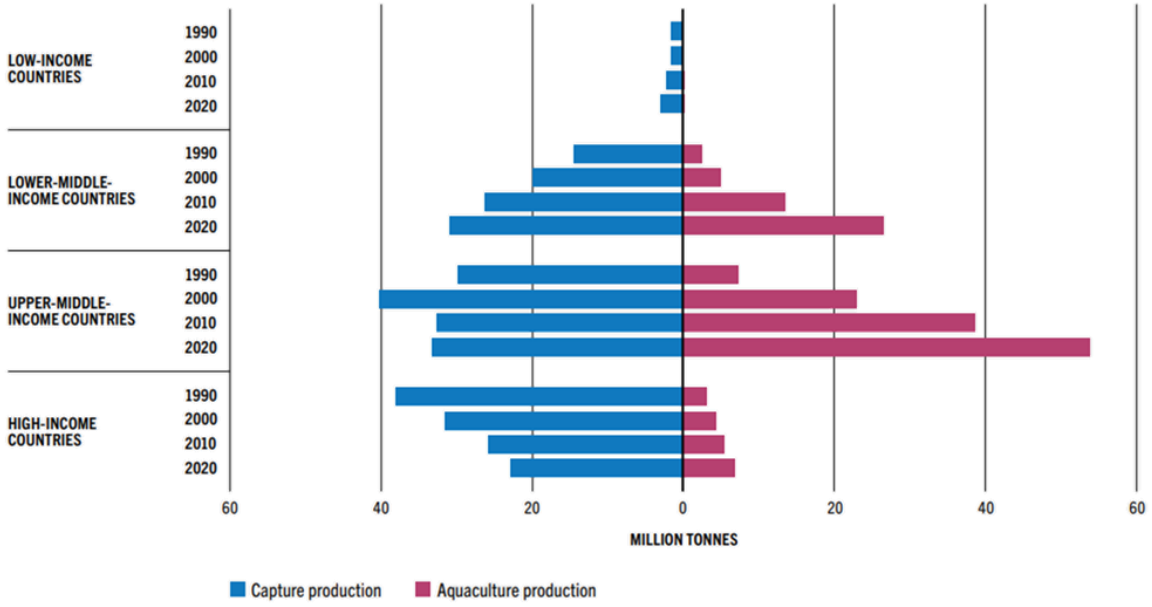
1.1.3 Aquaculture production per country

There is enormous variation across regions and countries concerning the production of aquaculture species. Among the 201[4] countries contributing to global aquaculture production, just ten countries account for 90 percent of the total production (FAO, 2023). Asia remains the dominant force in global aquaculture, accounting for 91.6 percent of the overall production of aquatic animals and seaweeds in 2020 (70 percent when considering only aquatic animals) (FAO, 2022). China stands out as the largest contributor, producing 49.90 million tonnes of fish from aquaculture, representing 57.03 percent of the global total (FAO, 2022). Following Asia, the major global continents in terms of production volume by decreasing size are the Americas, Europe, Africa, and Oceania.

According to FAO in 2020 the share of aquaculture in the total production exceeds capture fisheries for all major species groups except marine finfish (FAO, 2022) (Figure 3). If we look at the figures of aquaculture production of the different countries in relation to the World Bank's income level classification, we can see that aquaculture production in middle-income countries surpasses capture fisheries production, whereas in low and high-income countries capture fisheries exceed aquaculture (Mair *et al.*, 2023) (**Figure 1.3**). Aquaculture production in middle-income countries contributed to 61.7 percent of the total, a significant increase since 1990, when the same percentage was 19.8 (FAO, 2022). For lower-income countries the percentage of aquaculture production between the period of 1990 and 2020 increased from 14.7 to 46.2 (FAO, 2022). Aquaculture has been

[4] Including non-sovereign territories.

playing an increasing role in supporting the food security and livelihoods of low- and middle-income countries.



NOTE: Data expressed in live weight equivalent.
SOURCE: FAO.

Figure 1.3 Fisheries and aquaculture growth comparison by country group by income level (excluding algae), 1990-2020.

1.1.4 Challenges

In the past two decades, aquaculture has shown positive trends in production and per capita seafood consumption. Compared to land animal farming, aquaculture is one of the most resource-efficient and environmentally friendly methods for producing protein for human consumption (Boyd *et al.*, 2020; MacLeod *et al.*, 2020). However, persistent challenges such as the impact of pathogens, parasites, and pests (PPP) (Naylor *et al.*, 2021), pollution, harmful algal blooms (Díaz *et al.*, 2019), habitat destruction (Ahmed, Thompson and Glaser, 2019), and climate change (Holsman *et al.*, 2018) are becoming more pronounced as the aquaculture industry rapidly expands and becomes increasingly interconnected with its environment and global food systems.

In high-value species, advances in PPP identification and treatment over the past 20 years, but also the adoption of better management and breeding practises (Boudry *et al.*, 2021; Houston, Kriaridou and Robledo, 2022), have minimized the risks across all types of production systems (Holsman *et al.*, 2018).

However, there is a limited availability of such practices for low-value species and regions with low income (Naylor *et al.*, 2021; Houston, Kriaridou and Robledo, 2022). The aquaculture industry's responses to PPP pressures include the implementation of biosecurity measures, advanced water management, vaccine development, feed supplementation, and the use of chemical substances, although misuse of antimicrobials can lead to antimicrobial resistance issues (Naylor *et al.*, 2021). A more sustainable strategy involves selective breeding for disease resistance. This leads to fish that are more resilient to disease and reduces the losses farmers face from pathogens.

1.2 Selective breeding

In the field of genetics and agriculture, the practice of selective breeding stands as a fundamental mechanism for harnessing and enhancing desirable traits in plant and animal populations. Darwin was likely the first to employ the term “artificial selection” in his work “On the Origin of Species” (Darwin, 1859). The term artificial selection or selective breeding describes the process by which animals or plants are crossed by humans in a controlled manner to produce offspring with desirable traits, ultimately shaping the characteristics of subsequent generations. Selective breeding effectively exploits the genetic variation present in a population, transmitting permanent and cumulative genetic gains.

Variation observed in phenotypes within a population arises fundamentally from the underlying diversity in genotypes among individuals. Different combinations of alleles influence particular traits in different degrees, with some alleles conferring advantageous characteristics while others may be neutral or deleterious. However, the observed phenotypic variation is not just due to genetics, since the environment also has a large role in determining individual phenotypes, and there are also interactions between genetic and environmental factors. Therefore, selective breeding needs to take into account that only a proportion of the observed phenotypes is determined directly by genetic factors – this proportion of the phenotypic variance explained by genetic factors is known as heritability. The heritability determines how quickly traits can be improved via selective breeding.

The genomic basis of improvement of heritable traits from one generation to the next is the change of allelic frequencies in the population. However, in very few cases the favourable alleles are directly selected, since they are difficult to pinpoint for most traits. Instead, animals are ranked based on their ability to produce high performing offspring, a value known as breeding value, which represents the additive genetic merit of an individual for a trait of interest. Every individual receives half of its genes from each parent, so the additive genetic variance is the sum of the average additive effects of all the alleles the individual receives from both parents (Falconer and Mackay, 1996). The true breeding value (TBV) of an individual is impossible to determine since it is masked by environmental effects and non-additive genetic effects (such as dominance and epistasis), and therefore is indirectly estimated from the phenotypic values for the trait of interest, producing what we know as estimated breeding values (EBVs).

Phenotypic observations are defined by the following linear model, usually employed in problems of breeding value prediction in animal breeding:

$$y_{ij} = \mu_i + g_{ai} + e_{ij}$$

where y_{ij} is the record j of the i^{th} animal, μ_i refers to environmental fixed effects, g_{ai} is the additive genetic value of the genotype of animal i, and e_{ij} is the sum of the random environmental effects, dominance and epistatic genetic values. It is assumed that y follows a multivariate normal distribution, suggesting that traits are determined by an infinite number of additive genes with infinitesimal effects at unlinked loci, commonly referred to as the infinitesimal model (Mrode, 2014).

Phenotypic differences between individuals form the basis of estimation of breeding values and they are measured as variance. Thus, phenotypic variance is composed of genetic and non-genetic variance. The part of genetic variance we measure is the additive genetic variance (variance of breeding values), which is the heritable part that can be easily estimated and the only component that can be selected for. The non-genetic effects, which consist of the non-additive genetic effects and environmental differences, form the residual variance. Thus, the phenotypic variance can be divided in four components:

$$var(y) = var(a) + var(d) + var(i) + var(e)$$

$var(a)$: additive genetic variance

$var(d)$: genetic variance due to dominance

$var(i)$: genetic variance due to epistatic effects

$var(e)$: environmental variance

The animal's own phenotype is the direct way to estimate its breeding value and it can be calculated as:

$$\hat{a}_i = b(y_i - \mu)$$

where b is the regression of the true breeding value on phenotype, and μ is the mean performance of individuals in the same group. The slope of the regression (b) is equal to the heritability which is derived from the quantitative genetic theory:

$$b = \frac{cov(a,y)}{var(y)} = \frac{cov(a,a+e)}{var(y)} = \frac{\sigma_a^2}{\sigma_y^2} = h^2$$

Heritability is a constant value that can only vary between traits or between the same trait for different populations reared in distinctly different environments. The larger the heritability of a trait, the more confidently we can say that the observed phenotypic differences are due to additive genetic variance.

The precise estimation of the breeding values is a crucial element in any breeding programme, as the success of genetic improvement relies on accurately identifying and selecting individuals with the highest breeding value. Prediction accuracy is calculated as the correlation between the phenotypic value, and the true breeding value $(r_{a,y}^2)$. The expected response to selection per individual is calculated as follows:

$$R = ir_{a,y}^2\sigma_y = ih^2\sigma_y$$

and since: $h^2\sigma_y = \left(\frac{\sigma_a^2}{\sigma_y^2}\right)\sigma_y = \sigma_a\left(\frac{\sigma_a}{\sigma_y}\right) = h\sigma_a$

$$R = ih\sigma_a$$

where i is the selection intensity, which is the standardized mean of the selected group.

Genetic evaluation heavily relies on the genetic covariance among individuals to build the genetic covariance matrix. Higher accuracy and unbiased results can be achieved if the genetic relationship among individuals is recorded with molecular markers compared to pedigree. This is due to the fact that the similarity between relatives varies based on the level of genetic relatedness. The concept of resemblance between relatives plays a central role in estimating the amount of additive genetic variance and the overall phenotypic variance.

1.3 Selective breeding in aquaculture

The aquaculture industry is much more diverse in terms of species than terrestrial farming. Despite this high diversity, most of the species were only domesticated very recently (<100 years ago, Houston *et al.*, 2020), and most breeding populations have high levels of genetic diversity. This genetic diversity is the substrate of selective breeding (Driscoll, Macdonald and O'Brien, 2009), and its correct management is fundamental to ensure the long-term preservation of the genetic health of aquaculture breeding populations. The typical high fecundity of aquatic species provides high flexibility, facilitating the design of breeding programmes that preserve genetic diversity while maximising selection for the different traits of interest.

Taking advantage of this flexibility to establish selective breeding programmes for the plethora of available aquaculture species is fundamental to reach FAO's production target by 2030. Despite the rapid and consistent increase of aquaculture production during the last 20 years, only a few species have breeding programmes. While high-value species have established breeding programmes with a high level of sophistication (Houston, Kriaridou and Robledo, 2022), the vast majority of fish and shellfish producers still use stocks and seed sourced directly from the wild or stocks that were very recently domesticated (Teletchea, 2021). Establishing well-managed selective breeding programmes of lower-value species is crucial for food security.

Many studies have reported the response to selection in aquaculture breeding programmes. A review study conducted by Gjedrem and Rye (Gjedrem

and Rye, 2018) provides a summary of the reported genetic gains achieved for various traits in various aquaculture species. For example, estimates of genetic gain for body weight per generation range from 2.3% for whiteleg shrimp (Sui *et al.*, 2016) to 42% for sea bass (Vandeputte *et al.*, 2009), with an overall average gain per generation of 12.7% for the reviewed species. Another example is the response to selection for bacterial cold water disease resistance in rainbow trout with 19% genetic gains per generation (Leeds *et al.*, 2010). These results clearly show the power of selection to improve economically important traits. High genetic gains reported in many fish and shellfish breeding programmes are due to the relatively high heritabilities and high selection intensities made possible by the high fecundity of aquaculture species (Gjedrem, Robinson and Rye, 2012b).

1.3.1 Selection methods

Since the beginning of the first large scale breeding programme for Atlantic salmon in Norway in 1975 (Gjøen and Bentsen, 1997), new selective breeding methodologies have been developed to maximise genetic improvement, which are variably applied across different aquaculture settings. These methods can be classified depending on which individuals provide information used for selection decisions in individual (mass selection), family selection or combined selection.

Individual selection

In individual or mass selection, the breeding candidates are selected according to their own phenotypic performance and family relationships are not monitored. Individual selection is relatively easy to perform; however, it can only be applied for traits that can be recorded on the breeding candidates themselves, while still alive. In this method, all candidates are measured and only the best ranked individuals are selected to produce the next generation.

Individual selection presents two large drawbacks. First, the number of animals in the breeding nucleus has to be high (> 200 per generation) to ensure a large effective population size, decrease the chance of crossings between relatives, and avoid inbreeding (Bentsen and Olesen, 2002). High representation of individuals from a few specific families can lead to very high inbreeding in a few

generations, resulting in poor seed quality, survival and growth. While introducing individuals from a different stock can mitigate the adverse effects of inbreeding, it may simultaneously hinder the genetic improvement of the cultured stocks depending on the genetic merit of the introduced stock for the selected trait(s).

Secondly, it is impossible to disentangle the impact of genetics and the environment on the measured phenotypes, which means that individual selection is only efficient for highly heritable traits. For certain traits, the environmental component of the phenotypic variance can be larger than the genetic component. In such a scenario, the chosen broodstock might exhibit superior phenotypes due to favourable environmental conditions rather than being genetically superior, resulting in ineffective selection. To maximise the number of individuals that are correctly ranked it is important to keep all of them under identical environmental conditions to minimise its impact on the phenotypic differences between animals.

Family-based selection

In family-based selection schemes, the family origin of the breeding candidates needs to be traced. Pedigree management is performed using physical tags (e.g., Passive Integrated Transporter (PIT) tags or Visible Implant Elastomer (VIE) tags) or molecular markers (Ren et al., 2022). There are two family-based selection approaches: between-family selection and within-family selection.

In between-family selection, mean phenotypic values for each family are calculated and ranked, determining whether entire families are kept or removed (Lush, 1947). Generally, there is no predetermined minimum threshold value in this type of selection. Instead, farmers typically opt to keep a certain number of top families based on their performance. The greater the family size, the more closely the average phenotypic value aligns with the average genotypic value in family selection. In order to maintain a low rate of inbreeding while achieving a satisfactory level of genetic improvement, it is necessary to test a substantial number of family groups when implementing family selection (Dupont-Nivet *et al.*, 2006). Additionally, to minimise potential differences between families due to environmental effects, all families should be reared together and individually tagged as early as possible.

An important application of between-family selection is for target traits that

cannot be measured on live individuals (e.g., carcass quality traits, meat quality, disease resistance or age at sexual maturity). For traits of this nature, as most aquaculture species are highly fecund, it is possible to sacrifice some members of each family to obtain measurements of phenotypes that require the death or potential death (e.g., disease challenge experiments) of the fish. This information can be used to estimate the breeding value for the entire family, where the breeding candidates are assessed by considering the performance records of their full- and half-siblings. This is possible because these relatives share extensive segments of DNA. It is important to note that records from more distant relatives, beyond half-siblings, have a limited impact on the accuracy of the estimated breeding values.

In within-family selection, each family is treated as a distinct subpopulation, and selection is performed separately within each family. Fish within each family are ranked based on the deviation of each individual from the family average, and the farmer retains the best-performing individuals. In cases where there is sexual dimorphism, fish must be selected separately for each sex within each family.

The within-family selection provides an advantage in the case when there is a big component of environmental variance common to individuals of the same family. The lower limit of inbreeding in the programme is influenced by the number of families involved, and this can be effectively managed by implementing a rotational mating system (Farias, César and Silva, 2017).

Combined selection

Combined selection is employed to indicate the utilization of multiple selection methods within a breeding strategy. It incorporates information derived from individuals but also information on the performance of their relatives in the best way. By combining information from full and half-siblings, the additive genetic variance will be utilised in an optimal way and increase the accuracy of estimates of individual breeding values (Gjedrem, 2005). This method maximises the rate of genetic gain and is therefore considered as the optimal selection method when applicable.

1.3.2 Best linear unbiased prediction

The estimation of breeding values is central to all plant and animal breeding programmes. Over the years, with advances in the theory of quantitative genetics, breeders have proposed different genetic evaluation procedures. Selection Index Procedure (SIP) was the most commonly used method for prediction of breeding values until the early 70's (Gjedrem, 2005). Later on, a more powerful procedure called best linear unbiased prediction (BLUP) model was suggested. Henderson was the one who discovered BLUP and mixed model equations (MME) by solving a problem he was assigned in a statistics class using Lush's most probable producing ability methods (Henderson, 1973; Schaeffer, 1991). The origin of mixed model equations by Henderson appears in two abstracts he published in 1949 and 1950 (Henderson, 1949, 1950). However, Goldberger was the first to use the term "best linear unbiased predictor" in 1962 (Goldberger, 1962).

BLUP is a method that combines pedigree and phenotype records using all relevant information to estimate the breeding value of individuals (EBV). Genetic evaluations with BLUP use a relationship matrix in the model derived from the pedigree (Henderson, 1976), and they also provide the ability to correct breeding values for fixed effects. BLUP linear models take the following form (Robinson, 1991):

$$y = Xb + Za + e$$

where y is the vector of observed phenotypes, b is the vector of fixed effects, a is the vector of random animal effects to be predicted, e is the vector of non-observable random residual effects, X is the design matrix, which relates records to fixed effects, Z is a design matrix, which relates records to random animal effects and

$$\begin{bmatrix} \mathrm{a} \\ \mathrm{e} \end{bmatrix} \sim \mathrm{N} \begin{bmatrix} \mathrm{A}\sigma_{\mathrm{a}}^2 & 0 \\ 0 & \mathrm{I}\sigma_{\mathrm{e}}^2 \end{bmatrix}$$

where σ_a^2 and σ_e^2 are the additive genetic variance and the error variance, respectively, A is the pedigree-based numerator relationship matrix and I is identity matrix.

The prediction equation is first established in a training population, also

called reference, in which individuals are phenotyped. The vector a can be extended then to animals with no recorded phenotypes, comprising a set of individuals for whom we aim to predict the phenotypes; this set constitutes the validation set.

With the development of computational technologies and the ability to analyse large data sets, BLUP has been widely used to improve various economic traits in many aquaculture species like Nile tilapia (Barría *et al.*, 2020), Pacific oyster (Zhai *et al.*, 2021), Chinese mitten crab (Li *et al.*, 2021), black tiger shrimp (Noble *et al.*, 2020), rohu carp (Gjerde *et al.*, 2019) or Russian sturgeon (Song *et al.*, 2022).

While well-managed traditional selection programmes and BLUP have played a critical role in increasing the production of over 45 important aquaculture species (Gjedrem and Rye, 2018; Shen and Yue, 2019), they are not efficient for the improvement of traits that cannot be measured in the selection candidates. Traditional BLUP methods assign identical breeding values to the non-genotyped candidates within the same family and tend to lead to an increased rate of inbreeding per generation (Daetwyler *et al.*, 2007; Sonesson and Ødegård, 2016). Therefore, breeders started exploring new, more efficient, breeding technologies to accelerate the genetic improvement of aquaculture species.

1.4 Application of genomic tools to selective breeding

Since the advent of massive parallel sequencing technologies around 2005 (Heather and Chain, 2016), the advances in the field of genomics have been spectacular. The continued reduction of the cost of sequencing has enabled the generation of large amounts of genomic data, which is now readily available for most important aquaculture species. In the context of selective breeding, these developments have enabled the incorporation of genomic information into breeding programmes, mainly in the form of molecular markers.

In aquaculture breeding programmes, genomic information is utilized through two major breeding methods: marker-assisted selection (MAS) and genomic selection (GS).

1.4.1 Marker-assisted selection

Marker-assisted selection is the process of indirectly selecting individuals by using DNA markers associated with traits of commercial interest. This method relies on the previous detection of quantitative trait loci (QTL), regions of the genome associated with differences in the trait of interest. A QTL analysis basically searches for non-random associations between phenotypes and DNA markers across the genome. The markers themselves usually do not have any biological effect; they are linked to the causative mutation underlying the QTL through linkage disequilibrium (Guimaraes *et al.*, 2007). This linkage disequilibrium between genetic markers and causative mutations is the underlying principle of marker-assisted selection, as the linked genetic markers to the QTL can be used as proxy to select animals with the favourable causative alleles.

MAS is useful only when the identified QTL have large effects on the trait. An exceptional example of MAS is that for infectious pancreatic necrosis (IPN) resistance in Atlantic salmon. A major QTL for resistance to IPN virus was identified on chromosome 26, explaining 80-100% of the genetic variation of the trait in both seawater and freshwater trials (Houston *et al.*, 2008, 2009; Moen *et al.*, 2009; Gheyas *et al.*, 2010). MAS for this major QTL was rapidly incorporated by all major breeding programmes, which resulted in a significant decrease of IPN outbreaks and associated mortalities (Norris, 2017).

Nevertheless, applying this method to polygenic traits, which involve many genes with small individual effects, remains challenging.

1.4.2 Genomic selection and GBLUP

Genomic selection refers to a group of methods that estimate the breeding value of individuals based on genome-wide genetic markers. Studies proposing genomic selection were initially theoretical, but technological advances, such as high-throughput sequencing and SNP arrays, have made genotyping hundreds of animals for thousands of SNPs less costly compared to the past, when these technologies were first discovered.

While the genomic selection methodology was first proposed by

Meuwissen, Hayes and Goddard (Meuwissen, Hayes and Goddard, 2001), the incorporation of molecular markers to genomic evaluations was first suggested by Nejati-Javaremi *et al.* (1997). Nejati-Javaremi *et al.* (1997) showed that relationship matrices calculated using genetic markers (i.e., based on the genomic similarity between individuals) record relatedness more accurately than the pedigree, as they account for Mendelian sampling. If we do not include genomic information, it is assumed that all full-siblings share 50% of their genome, which for traits that cannot be measured directly on the candidates themselves results in identical breeding values for a whole family. However, in reality, because of Mendelian sampling, the actual percentage of the genome shared between full-siblings can significantly vary from the average of 50%. The marker-based relationship matrix is called genomic relationship matrix (GRM) and relatedness estimates of this matrix can deviate significantly when compared to the pedigree relationship matrix (A) (Hill and Weir, 2011). Thus, genomic selection can discriminate the differences between full-sibs and utilise both the between and within-family components of genetic variation which leads to increased selection accuracy (Boudry *et al.*, 2021; Fugeray-Scarbel *et al.*, 2021; Joshi *et al.*, 2021; Vallejo *et al.*, 2021; Song and Hu, 2022a).

Genomic best linear unbiased prediction (GBLUP) is the most popular genomic selection method. It was introduced by VanRaden (VanRaden, 2008) and Habier (Habier, Fernando and Dekkers, 2007), and combines genomic information into BLUP by using a genomic relationship matrix instead of a pedigree-derived relationship matrix. The general model for GBLUP is:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zg} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the vector of observed phenotypes, $\boldsymbol{b}$ is the vector of fixed effects, $\boldsymbol{g}$ is the vector of additive genetic effects with variance $var(\boldsymbol{g}) = G\sigma_a^2$ where G is the genomic relationship matrix, $\boldsymbol{e}$ is the vector residual effects with variance σ_e, $\boldsymbol{X}$ is the design matrix which relates records to fixed effects**,** $\boldsymbol{Z}$ is a design matrix which relates records to additive effects.

In GBLUP the marker effects are assumed to have a normal distribution and the same variance is assigned to all loci treating them all as equally important (Ren *et al.*, 2021). This assumption is suitable for traits that conform to the infinitesimal

model. Once the prediction equation is established, breeding candidates can then be selected on the basis of their genomic estimated breeding value (GEBV).

While traditional BLUP methodology requires pedigree information to be recorded to define the covariance (resemblance) between relatives, in genomic selection schemes, pedigree records are not needed as the relationships between individuals are calculated based on genetic markers covering the whole-genome. This means that families do not have to be kept separately until tagging for genetic evaluation, which is useful for species that reproduce in groups (mass spawners) where rearing each family separately is impossible (Sonesson, Meuwissen and Goddard, 2010).

Genomic selection has been playing an ever-increasing role in aquaculture breeding programmes, resulting in an increase in the accuracy of breeding value prediction and subsequent genetic gain, improving the performance of farmed populations (Houston *et al.*, 2020).

Other genomic selection methods

Additional methods have been proposed for genomic selection. Howard *et al.* (2022) provides an overview of the most commonly used models for predicting single traits under single environments. In their paper Zhang *et al.* (2011) classifies these models in two broad groups of direct and indirect methods.

Direct methods calculate the GEBVs using mixed model equations, as described above in GBLUP. A method termed as “single step”, which combines the use of both genomic and pedigree matrices, was proposed in four papers around the same time (Legarra, Aguilar and Misztal, 2009; Misztal, Legarra and Aguilar, 2009; Aguilar *et al.*, 2010; Christensen and Lund, 2010). Legarra and Ducrocq referred to this method as single-step genomic BLUP (ssGBLUP) (Legarra and Ducrocq, 2012). The model allows the inclusion of phenotypes and pedigree of genotyped and non-genotyped animals. ssGBLUP is similar to GBLUP, but the G matrix is replaced with the H matrix, which takes the following inverse form:

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

H is a unified relationship matrix, where A_{22}^{-1} corresponds to the inverse of the numerator relationship matrix for genotyped individuals and G^{-1} is the inverse of the genomic relationship matrix.

Indirect methods for calculating the GEBV of candidates involve estimating marker effects in a reference population of genotyped and phenotyped individuals. The GEBVs of genotyped candidates are then determined by summing the pre-estimated effects of all markers for each individual. The following general model is used for the estimation of marker effects (Zhang, Zhang and Ding, 2011):

$$y = Xb + \sum_{i=1}^{m} Z_i g_i + e$$

where y is a vector of the phenotypic values of all individuals in the reference population, b is a vector of fixed effects, X is the design matrix for b, g_i is the effect of the ith marker, m is the total number of markers, Z_i is an index vector representing the genotypes of all individuals at marker i in the reference population, and e is a vector of residual errors with variance-covariance matrix $I\sigma_e^2$ (σ_e^2 is residual variance).

Different methods have been proposed for the estimation of marker effects based on the model above. These methods include Bayesian methods, which were originally developed by Meuwissen *et al.* (2001). Bayesian methods allow for the effects of all genotyped markers to be fitted at the same time, assume different prior distributions of marker effects, allow for the use of variable selection procedures to identify important markers and the incorporation of additional information like omics data or functional annotation (Wolc and Dekkers, 2022). As opposed to BLUP methods, different weights are assigned to different markers and a limited number of markers are assumed to have effects on the trait.

Two of the first Bayesian methods used for genomic selection are Bayes A, which assumes all markers have a nonzero effect, with different variances that follow an inverse chi-square distribution, and Bayes B, in which a proportion of the markers have a zero effect and the rest have different effects as in Bayes A (Meuwissen, Hayes and Goddard, 2001). The computational approach is to implement Markov chain Monte Carlo (MCMC) methods for sampling using the

prior distributions for the marker effects. These methods quickly evolved to Bayes Cπ and Bayes Dπ methods (Habier *et al.*, 2011), or Bayes-C0 (Kizilkaya, Fernando and Garrick, 2010), which is actually a special case of (G)BLUP where a single normal prior is used for the distribution of marker effects. There is also Bayesian Lasso (Park and Casella, 2008), Bayes R (Erbe *et al.*, 2012), Bayes RS (Froberg Brøndum *et al.*, 2012), Bayes U (Pong-Wong and Woolliams, 2014) and many variations named with other letters known as the Bayesian alphabet (Gianola *et al.*, 2009).

Numerous investigations have been carried out to compare the predictive ability of Bayesian and GBLUP methods, summarized in Song *et al.* (2023). Generally, GBLUP methods achieve the same or higher accuracy for polygenic traits controlled by many QTLs with small effects. On the contrary, Bayesian methods perform better when major QTLs partially or entirely govern the trait. While both GBLUP and Bayesian models are commonly used in studies involving aquaculture species, the performance of the models varies depending on the information available in each study, and neither of them is the best in all situations. As there is currently a lack of evidence in favour of Bayesian models, and these methods are computationally demanding, its application is in most cases not practical.

Apart from the methods mentioned above, a third group of methods consists of machine-learning-based algorithms (Nayeri, Sargolzaei and Tulpan, 2019). These algorithms include artificial neural networks (ANN), support vector machine (SVM), random forest (RF), gradient boosting machine (GBM), reproducing Kernel Hilbert spaces (RKHS), deep learning (DL), and decision trees (DT). Some of these methods were tested for prediction of disease resistance and growth traits in aquaculture species and resulted in an improvement of genomic prediction accuracy values (Bargelloni *et al.*, 2021; Palaiokostas, 2021; Zhu *et al.*, 2021). While machine-learning approaches can provide similar results, or potentially even surpass other approaches, and are also more scalable, allowing the storage and computation of all the information accumulated over years of breeding programmes, more evidence and research are needed before transitioning to these novel, mostly untested methods.

1.4.3 Genome wide association studies and weighted models

Genome wide association studies (GWAS) can be used to identify genomic regions that explain a great proportion of the genetic variation, and are fundamental for the identification of QTLs for marker-assisted selection (Wolc and Dekkers, 2022; Yáñez *et al.*, 2023a). A classic GWAS approach requires a population that is phenotyped and genotyped for thousands of markers across the genome. A regression-type of analysis is applied for each marker, one at a time, and the method estimates the individual effect of each marker on the phenotype of interest. GWAS exploits linkage disequilibrium between markers located closely within the same chromosome with the goal of finding genetic markers statistically associated with a particular trait. Unlike single-SNP GWAS approaches, Bayesian GWAS methods simultaneously account for the effects of all SNPs by taking into account the specified prior distributions of marker effects. Bayesian methods can enhance GWAS by integrating prior information, such as functional annotation and gene expression data, to gain meaningful insights from the biology and the genetic architecture of the trait (Wolc and Dekkers, 2022).

While GBLUP models are in principle more suitable for polygenic traits since they assume equal variance for all markers, alternative methods of genomic relationship matrix construction can reflect both the genomic relationship between individuals and the genetic architecture of the trait. For instance, a weighted G matrix can incorporate information from GWAS or Bayesian analysis (Zhang *et al.*, 2015, 2016; Fragomeni *et al.*, 2017; Ren *et al.*, 2021). These methods give different weights to markers that are in linkage disequilibrium with the causative marker or large-effect QTL. For instance, weighted single-step GBLUP puts weight to the markers depending on the variance they explain, which can make it suitable for oligogenic traits as well (Wang *et al.*, 2014). Studies using different strategies for the construction of the weighted G matrix showed that it can significantly improve the accuracy of GBLUP model prediction (Zhang *et al.*, 2010; Yoshida and Yáñez, 2021; Fraslin, Koskinen, *et al.*, 2022; Song and Hu, 2022a).

Regardless of the method employed, genomic selection has consistently demonstrated superior performance compared to pedigree-based selection methods. Its integration into the aquaculture industry is crucial for realizing

sustainable growth objectives.

1.4.4 Incorporation of functional annotation into breeding programmes

Since the early 2000's, accompanied by the fast developments of genotyping and sequencing technologies, various genomic prediction models have been proposed, as already discussed in the previous sections. These methods have improved the accuracy of predicting genomic breeding values to varying degrees, depending on the specific trait of interest, through the utilization of genomic markers distributed across the genome. An interesting approach that has the potential to further improve genomic prediction models, and subsequently prediction accuracy, is the utilisation of both genomic marker information and other biological knowledge (e.g., in the form of omics data or functional annotation).

To extract functional information from DNA sequencing data, it is essential to annotate the genome with structural elements like promoters, introns, exons, protein coding sequences (CDS), transcriptional start sites, etc. This structural annotation is followed by functional annotation where the function of the different regions is discovered. These annotations can be derived from computational predictions, frequently based in comparative biology, or experimental data. In recent years, an increasing amount of biological information is becoming publicly available in the form of functional or structural annotation, gene expression, and trait mapping information from large-scale projects like the Functional Annotation of Animal Genomes (FAANG) or AQUA-FAANG to unlock the genotype to phenotype connection in farmed animals (https://www.animalgenome.org/community/FAANG/index).

Several studies have tried to introduce modified versions of the existing genomic prediction models to incorporate functional information. Examples include modifications of the most routinely used GBLUP models, such as GTBLUP (genomic best linear unbiased prediction with a transcriptome-based linear kernel), GRBLUP (genomic best linear unbiased prediction with a transcriptome-based Gaussian kernel) and GTiBLUP, in which the models are fitted by considering the genome, the transcriptome and/or their interaction (Li *et al.*, 2019; Morgante *et al.*,

2020). Another group of models discussed previously are the Bayesian models, which are based on different assumptions about how individual markers contribute to the overall genomic variance by attributing different prior distributions to the vector of the marker effects. For example, in BayesR (Erbe *et al.*, 2012) a single prior annotation category can be defined with marker effects following a mixture of normal distributions including zero, small, medium or large variance. BayesRC is an extension of BayesR in which prior biological information can be incorporated into more than one disjoint annotation categories and the markers in each annotation category are modelled according to the mixture prior, defined as in BayesR (MacLeod *et al.*, 2016). Mollandin *et al*. (2022a) proposed two different models that allow the assignment of markers in more than one annotation categories to exploit the information that multi-annotated markers may provide. The first one is the BayesRCπ model, which allows markers to overlap when assigned to categories and marker effects follow a mixture of mixtures prior distribution allowing markers to be *a posteriori* assigned to the annotation category which maximizes its likelihood according to the estimates of the other parameters in the model (Mollandin *et al.*, 2022a). The second model is called BayesRC+, and assumes that multiple annotation categories cumulatively impact the estimated marker effects. Thus, during each iteration of the Gibbs sampler, the effect of a marker is the sum of its conditional effects estimated for each annotation category it is associated with (Mollandin *et al.*, 2022a).

Results from studies are not consistent and this additional information not always results in a positive impact in the predictive ability of the models. Ehsani *et al.* (2012) discovered that Bayesian models incorporating transcriptomic information were better at predicting phenotypes and explained more of the phenotypic variance than the models that were using the SNPs alone. Li *et al*. (2019) concluded in their study that the predictive ability of the models when they included expression data in the models was generally similar to GBLUP, but the amount of phenotypic variance explained was higher. In another study with inbred lines of maize, information from gene expression level and metabolite level was used for the prediction of complex traits (Guo *et al.*, 2016). Predictive abilities from GBLUP and TBLUP (BLUP based on gene transcript data) were higher than those from MBLUP (BLUP based on metabolite data) for each subpopulation that they

tested, but when averaging across all traits the predictive ability of TBLUP was lower than GBLUP. In the same study, when expression or metabolite data was combined with SNPs the GTBLUP and GMBLUP models achieved greater or similar results compared to GBLUP and TBLUP. Overall, within each subpopulation GTMBLUP achieved on average the highest predictions across all traits but in many situations the improvement was minimal. Two other studies in Drosophila melanogaster used metabolites for the prediction of complex traits achieving greater accuracy compared to when using genetic variants alone (Zhou *et al.*, 2020; Rohde *et al.*, 2021).

Different Bayesian models incorporate the available biological information in different ways and their performance from studies with different species has shown to be highly dependent on the genetic architecture of the trait, the way markers are assigned into annotation categories, and the relevance or the nature of the incorporated information. Bayesian methods have been shown to perform better than GBLUP for traits controlled by genes of moderate to large effects (VanRaden *et al.*, 2009; Legarra *et al.*, 2011; Ma *et al.*, 2019). Practical difficulties exist in defining the optimal way of defining annotation categories and how to include this biological knowledge, relevant to the trait of interest, without introducing noise or redundant information that does not contribute to the model. For example, omics data should be derived from the challenged population used for the prediction of phenotypes and has to be measured from the tissue that is relevant to the trait of interest at the required developmental stage (Morgante *et al.*, 2020). Such big omics datasets, from experiments that are designed specifically for the studied population, are rare to find. Other prior biological information such as structural or functional annotation can be collected on the same individuals used for genomic prediction, or accessed from publicly available databases from studies on different populations. Biological information readily available from public sources can then be utilized to categorize variants into annotation groups for the prediction models.

There are no studies incorporating biological information into prediction models for aquaculture species thus far. Future investigations aimed at aquatic species should explore the advantages of incorporating functional or structural annotation data specifically collected for a population into genomic prediction

models. Additionally, the manner in which this data is integrated and defined within these models needs to be examined to ensure its relevance for the targeted trait. Our understanding of functional mechanisms and the link between genotype and phenotype is limited and should be prioritized in future research.

1.4.5 Genomic selection in aquaculture breeding programmes

Genomic selection was incorporated into aquaculture relatively recently. The first genomic selection study was reported only 10 years ago on an admixed population of Atlantic salmon (Ødegård *et al.*, 2014), after the development of the first high-density SNP array (Houston *et al.*, 2014). With the drop in sequencing prices and the realization that genomic prediction models outperform pedigree-based models (Houston *et al.*, 2020), the application of genomic selection gradually expanded to other species.

According to Song *et al.* (2023) genomic selection has been applied to about 20 important aquaculture species. These species include Atlantic salmon, rainbow trout, large yellow croaker, gilthead seabream, Nile tilapia, Pacific white shrimp, yellowtail kingfish, common carp, Pacific oyster, Japanese flounder, European seabass, Yesso scallop, Zhikong scallop, channel catfish, coho salmon, yellow drum, banana shrimp, turbot, rock bream and Pacific abalone. However, breeding programmes of the majority of aquaculture species do not take advantage of the available genomic technologies, including some of the top species by production volume globally.

Houston *et al*. (2022) reviewed the status of the breeding programmes for the top 10 species by production volume. Only three of these species, namely Atlantic salmon, whiteleg shrimp and Nile tilapia, have advanced genome-assisted breeding programmes for a substantial proportion of their production.

Atlantic salmon

The Atlantic salmon (*Salmo salar*) industry stands out globally in aquaculture production, particularly in its advanced use of genetic technologies. Atlantic salmon commercial breeding programmes operate at level 4, which is the most advanced level of sophistication (Houston, Kriaridou and Robledo, 2022),

and play a crucial role in supplying a significant portion of global salmon production. At level 4, genomic tools like MAS and genomic selection are routinely applied and breeding programmes focus on sibling testing targeting many traits via selection index (Houston, Kriaridou and Robledo, 2022). Originating from Norway in the 1960s, the first commercial-scale salmon farming was followed by family-based trials of salmon collected from 40 Norwegian rivers in the early 1970s (Gjeren and Bentsen, 1997). Nowadays, large international companies dominate the market and supply seeds to various salmon-producing countries. These companies improve several traits simultaneously, mainly focusing on growth and resistance to pathogens and parasites. The success and advanced status of the Atlantic salmon industry have facilitated technology transfer to other aquaculture species, with some of these companies now involved in breeding programmes for tilapia, whiteleg shrimp, and rainbow trout (Houston, Kriaridou and Robledo, 2022).

Carp and catfish

Carp and catfish play a significant role in global aquaculture, comprising half of the top 10 most produced species worldwide (Houston, Kriaridou and Robledo, 2022). Primarily cultivated in Asia in extensive or semi-intensive freshwater systems, often in polyculture systems, these species, particularly carp, have a long history of domestication and breeding (Chen *et al.*, 2021). Despite their economic importance, these species are mostly at Level 1 (wild-collected seed with minimal or no genetic management of stock) or Level 2 (seed is supplied from hatcheries with advanced genetic management and maybe some basic directional selection) of breeding programme sophistication remaining behind in the application of genetic technologies compared to other species (Houston, Kriaridou and Robledo, 2022). The common carp (*Cyprinus carpio*) stands out as the most advanced in terms of genomic resources and available genetic tools, with China leading in production and breeding advancements (Hu *et al.*, 2018; Yang *et al.*, 2024). Chen *et al*. (2021) in his study summarises the identified loci/genes for traits such as growth, body shape, resistance, muscle quality and sex. Breeding is still at the traditional stage for most of these important economic traits due to the lack of major QTL for MAS selection and inadequate investments for incorporating genomic selection in the breeding process (Chen *et al.*, 2021). In recent years, genomic

tools and studies on genetic parameters have become available for other top carp and catfish species. However, there is limited evidence supporting the existence of comprehensive, commercially driven breeding programmes. There is great potential for genetic improvement in these species to enhance global aquaculture, but the low economic value of this group of species, coupled with the large investment required and the slow realization of breeding programme benefits pose uncertainties about who will drive such advancements. Public sector programmes are likely contributors to at least kick-start the application of genetic technologies to tackle challenges related to economic and food insecurity in these species (Houston, Kriaridou and Robledo, 2022).

Whiteleg shrimp

Over the past two decades, whiteleg shrimp (*Penaeus vannamei*) aquaculture has experienced rapid growth, surpassing the giant tiger prawn in production (Gorjan, 2022), mainly due to its enhanced resistance to infectious diseases. Whiteleg shrimp underwent domestication and selective breeding starting in the 1980s in the United States of America and Latin America (Alday-Sanz *et al.*, 2020). The existence of well-managed selective breeding programmes, ranging from technology Level 3 (family-based selective breeding programmes that supply seed and pedigree is routinely recorded for the estimation of breeding values for a small number of traits) to 4, contribute significantly to global production. The expansion of these programmes is driven by challenges posed by diseases like white spot syndrome, Taura syndrome, and early mortality syndrome (Castillo-Juárez *et al.*, 2015), which cause significant financial loses to farmers. The pressing need to avoid infectious diseases has led to the use of specific pathogen-free (SPF) seed from hatcheries. Major companies are now supplying germplasm with disease resistance and improved performance for other traits (Houston, Kriaridou and Robledo, 2022).

Nile tilapia

Nile tilapia (*Oreochromis niloticus*) holds significant global importance in aquaculture, providing essential protein and nutrients in low- and middle-income countries such as South-East Asia and Africa. Farming of this species is successful due to its capacity for adaptation, durability, fast growth, and ability to thrive in

varied environmental circumstances (Yáñez, Joshi and Yoshida, 2020). Semi-intensive or intensive production systems for Nile tilapia vary widely, ranging from small farms to large-scale commercial operations. Attempts of mass selection to improve growth rate for this species resulted in a 2-3% gain per generation (Bentsen *et al.*, 2017). The first selective breeding programme for Nile tilapia was designed in 1988 by the International Centre for Living Aquatic Resources Management together with AKVAFORSK in Norway and they developed the 'Genetically Improved Farmed Tilapia' (GIFT) strain (Yáñez, Joshi and Yoshida, 2020). The GIFT strain originated from diverse populations reared in the Philippines and wild populations imported from Africa. It is considered a major success story in aquaculture as it resulted in an accumulated growth rate of 86% in five generations (Bentsen *et al.*, 2017). Today the GIFT strain dominates global tilapia aquaculture production, including GIFT-derived strains used by major commercial breeding programmes on level 4 of sophistication (Houston, Kriaridou and Robledo, 2022).

Bivalves

Until now, genomic resources have been established for various commercially significant bivalve species, such as the Pacific oyster (*Crassostrea gigas*), pearl oyster (*Pinctada fucata*), blue mussel (*Mytilus galloprovincialis*), Eastern oyster (*Crassostrea virginica*), Yesso scallop (*Patinopecten yessoensis*), Zhikong scallop (*Chlamys farreri*), Manila clam (*Ruditapes philippinarum*), and Snout otter clam (*Lutraria rhynchaena*) (Tan, Zhang and Zheng, 2020). Selective breeding programmes commonly used for bivalves are based on mass selection, family and combined selection (Tan, Zhang and Zheng, 2020; Nascimento-Schulze *et al.*, 2021). Despite various studies showing that genomic selection can achieve higher prediction accuracies than traditional methods (Tan, Zhang and Zheng, 2020), modern selective breeding programmes and large-scale commercial investment into genetic technologies have been absent and limited to public sector, academic efforts and collaborative organisations involving farmers (Houston, Kriaridou and Robledo, 2022). The progress of genomic selection in shellfish aquaculture has been partly hindered because effective selective breeding programmes require pedigree information and precise phenotype

records. This task of keeping records of relatedness is more challenging in shellfish species due to the small size of juveniles at hatching, typically a few micrograms, making physical tagging impractical (Yáñez, Newman and Houston, 2015). Consequently, keeping records of families would mean that large numbers of juveniles should be raised separately, which is highly costly, requires extensive infrastructure, and introduces potential confounding factors like "tank effect" and "family effect". These factors can significantly reduce the accuracy of estimated breeding values.

The Manila clam is a major cultured shellfish species that stands out as the sole bivalve in the global top 10 by production volume (Houston, Kriaridou and Robledo, 2022). There are several studies on the genetic basis of commercially important traits but the available genetic tools necessary for the initiation of selective breeding programmes are not as advanced as in other cultured fish and shellfish (Smits *et al.*, 2020). The abundance of wild seed and the low individual value of each animal contribute to the underutilization of genetic technologies. Smits *et al.* (2020) developed a SNP panel for Manila clam consisting of 245 SNPs but achieved only partial parental assignment (41%), hence further improvement in genotyping and assignment rate is required for more accurate estimates of genetic parameters. Pedigree reconstruction is also important for inbreeding management and sib testing in breeding designs to test traits that are lethal or require the sacrifice of the animal.

Although large-scale commercial investment in bivalve genetic technologies is currently limited, emerging specialized programmes suggest a growing uptake of genetic technologies in bivalve aquaculture, promising enhanced sustainable seed supply and improved production traits in the future (Houston, Kriaridou and Robledo, 2022).

1.4.6 Barriers for the widespread implementation of genomic selection

The establishment of modern breeding programmes and the incorporation of genomic tools (i.e. genomic selection) holds great potential for increasing aquaculture production by accelerating the improvement of desirable traits

(Houston *et al.*, 2020). The implementation of genomic selection depends on several factors, such as the ability of companies to allocate resources towards employee training, development and modernisation of breeding programmes, but it can also be limited by the reproductive biology of each species and the existing genomic resources (Houston *et al.*, 2020; Boudry *et al.*, 2021). Suggestions for improvement of genomic selection are necessary to not only increase production but also reduce the cost and make aquaculture products available to everyone.

One of the main factors limiting the widespread adoption of genomic selection in aquaculture is the cost of genotyping. Genotyping cost depends on various factors, such as the genotyping panel marker density, genotyping technology and crucially on the number of samples genotyped. Large companies genotype thousands of animals per year, which substantially reduces genotyping cost per sample. On the other hand, the costs of genotyping can be prohibitively expensive for small and medium aquaculture operations, making it more challenging for them to adopt genomic selection practises (Boudry *et al.*, 2021). For these industries to benefit from genomic selection, low-cost genotyping strategies that do not significantly compromise the prediction accuracy of breeding values are required.

As a cost-effective alternative to high-density panels, a number of studies have looked into the use of low-density SNP panels for genomic selection. Generally, these studies have reported that generally SNP densities can be significantly reduced without a significant loss of prediction accuracy (Tsai *et al.*, 2016; Palaiokostas *et al.*, 2018, 2019; Robledo, Matika, *et al.*, 2018; Yoshida *et al.*, 2019; Zenger *et al.*, 2019; Gutierrez *et al.*, 2020; Kriaridou *et al.*, 2020; Tsairidou *et al.*, 2020; Al-Tobasei *et al.*, 2021). For example, studies on Atlantic salmon (Tsai *et al.*, 2016; Robledo, Matika, *et al.*, 2018; Kriaridou *et al.*, 2020; Tsairidou *et al.*, 2020) showed that densities between 1,000 and 5,000 SNPs can be sufficient to reach prediction accuracies similar to those obtained with a high-density panel. Similar results were reported for common carp (Palaiokostas *et al.*, 2018, 2019). For fillet yield in Nile tilapia (Yoshida *et al.*, 2019) and resistance to *Flavobacterium columnare* in rainbow trout (Fraslin *et al.*, 2023), low-density panels of about 3,000 SNPs resulted in comparable accuracies to the full genotyped dataset. In Pacific oyster and sea bream populations around 2,000 to

2,500 SNPs were sufficient to achieve near maximum genomic prediction accuracy results (Gutierrez *et al.*, 2020; Kriaridou *et al.*, 2020). While the use of low-density panels has shown to be effective and very beneficial in terms of costs in aquaculture breeding programmes there are other strategies, like genotype imputation, that can further improve prediction accuracy of low-density panels for low-cost genomic selection.

1.5 Genotype imputation

Genotype imputation is a promising method that is commonly used to predict genotypes for untyped loci in individuals genotyped with a low-density SNP panel, using a reference population genotyped for a high-density panel (Sargolzaei, Chesnais and Schenkel, 2010). By combining the use of low-density panels with genotype imputation for the selection candidates in a breeding programme, the genotyping cost can be reduced significantly. Genotype imputation uses the markers in the low-density panel that are common to both the low-density and the reference population as anchors to impute the missing genotypes, relying on linkage and linkage disequilibrium structure within the population.

Genotype imputation can also be applied to datasets after quality control to fill in missing genotypes, or can be used to impute missing genotypes after combining two datasets genotyped with different but partially overlapping SNP panels (e.g. for meta-analysis) (Browning, 2008; Marchini and Howie, 2010a). Other applications of genotype imputation include the use of low-coverage whole-genome sequence data. Low-coverage whole-genome sequence has been proposed as an alternative approach to genotyping arrays. With this method, each site is sequenced at a lower coverage, which has the downfall of increased genotype calling uncertainty and introduces errors, whereas at the same time more individuals can be sequenced with the same budget.

Different imputation methods can be used to impute missing genotypes (Browning, 2008). The general idea of genotype imputation is that related individuals share long haplotype blocks (set of markers in linkage disequilibrium

that segregate together). The closer the individuals are related to one another the longer are the haplotypes they share. However, recombination events occurring from one generation to another break these haplotype blocks, hence two animals that are not closely related share shorter haplotype blocks. Imputation algorithms can be generally classified in two main categories: i) population-based and ii) pedigree-based (Browning, 2008; Bouwman *et al.*, 2014; Sargolzaei, Chesnais and Schenkel, 2014; Wang *et al.*, 2016; Antolín *et al.*, 2017; Lashmar, Muchadeyi and Visser, 2019).

Population-based methods utilize linkage disequilibrium information between markers by modelling haplotype frequencies in populations of not closely related individuals. They mainly use Hidden Markov Model (HMM) approaches to model genotype variation and rely on population-wide linkage disequilibrium between markers (short shared haplotypes) (Sargolzaei, Chesnais and Schenkel, 2014). Population-based imputation methods can be slower, computationally intensive, and sometimes less accurate than pedigree-based methods (Antolín *et al.*, 2017). However, previous studies have shown that accuracy can be increased by increasing the number of reference individuals and SNPs (Sargolzaei, Chesnais and Schenkel, 2014; Tsai *et al.*, 2017); as the low panel density increases, the likelihood of finding short segments of shared haplotypes also increases.

Pedigree-based methods incorporate information from both linkage-disequilibrium and pedigree relationships for imputation. These methods take advantage of the long-haplotypes shared by closely related individuals such as parent-offspring or full-sibs, but also Mendelian inheritance rules to infer missing genotypes (Antolín *et al.*, 2017). Pedigree information is of great importance especially as the low-density panel becomes sparser, because it helps capture the long-range haplotype blocks shared between related animals. Pedigree-based imputation software perform better in terms of accuracy when imputing rare variants than population-based software, because variants with low minor allele frequency are not in high linkage disequilibrium with common variants (Hickey *et al.*, 2012). Further, a rare variant present in a specific population might be absent or very rare in a more distantly related population, hence making this variant more difficult to impute (Hickey *et al.*, 2012).

Most imputation software now can use a combination of population and pedigree-based methods (e.g., AlphaImpute, FImpute, findhap etc.), using population-based algorithms to phase the genotypes of the animals at the top of the pedigree (or in general those without genotyped relatives), and pedigree-based algorithms to impute the genotypes of their descendants.

1.5.1 Factors affecting genotype imputation

Several factors affect genotype imputation accuracy, namely the imputation software used, the number of SNPs in the low-density panel, their distribution along the chromosomes, the specific set of SNPs selected, SNP minor allele frequency, the number of individuals in the reference (high-density) set and the population structure.

The choice of software can have a great impact on the results; different algorithms make use of the available information in different ways, so that the optimal imputation software may be different depending on the population of interest and the available genotypic data. Three popular software for low density panel imputation are AlphaImpute v.2 (Whalen and Hickey, 2020), FImpute v.3 (Sargolzaei, Chesnais and Schenkel, 2014) and findhap v.4 (VanRaden *et al.*, 2013) commonly used in livestock and aquaculture populations. All three software have the option to infer missing genotypes by combining pedigree and population imputation.

AlphaImpute v.2 employs a three-step imputation approach combining pedigree- and population-based methods. Initially, pedigree imputation utilizes an approximate multi-locus iterative peeling method (a method that models the haplotypes of an individual based on the haplotypes of its relatives (Whalen *et al.*, 2018)), followed by phasing high-density individuals using a population-based algorithm. The phased haplotypes form a reference library for imputing low-density individuals, with subsequent rounds of multi-locus iterative peeling refining the imputation process.

FImpute v.3 also begins with pedigree-based imputation, utilizing relationship information for phasing and imputation. Missing SNPs are filled by

matching progeny haplotypes to parental ones, with subsequent imputation using an overlapping sliding window approach. the window size decreases by a constant factor in each chromosomal scan to capture information from distant relatives and account for shorter haplotype similarity (Sargolzaei, Chesnais and Schenkel, 2014). This method assumes varying degrees of relatedness among individuals, with haplotype length influencing accuracy.

Findhap v.4 similarly combines population and pedigree haplotyping. Chromosomes are segmented, and missing genotypes are imputed based on the two most likely haplotypes. Posterior allele probabilities are updated repeatedly from the prior probabilities within those haplotypes as new animal sequences are processed. Once population haplotyping is completed the programme examines the pedigree to resolve parent-progeny haplotype conflicts, detect new haplotypes that were created by crossovers and impute non-genotyped founders from their genotyped progeny (Vanraden *et al*., 2011).

Low-coverage sequencing data imputation software include GeneImp (Spiliopoulou *et al.*, 2017), GLIMPSE (Rubinacci *et al.*, 2021), QUILT (Davies *et al.*, 2021), LOIMPUTE (Wasik *et al.*, 2021) and STITCH (Davies *et al.*, 2016). In a study by Rubinacci *et al.* (2021) using human data, GLIMPSE was compared to Beagle, GeneImp, LOIMPUTE and STITCH. The results showed that for common variants GLIMPSE performed slightly better than BEAGLE and outperformed GeneImp, LOIMPUTE and STITCH. GLIMPSE also outperformed all tested methods providing higher accuracy for rare variants and shorter running times. While these software have been tested in human and livestock species, there are only two recent studies that have evaluated STICH and GLIMPSE in yellow croaker and Atlantic salmon, respectively (Zhang *et al.*, 2021; Gundappa *et al.*, 2023).

GLIMPSE uses as input a matrix of genotype likelihoods calculated from the low-coverage reference panel at the different positions of the genome. The genotype likelihoods are refined by running iterations of genotype imputation and haplotype phasing using Gibbs sampling. During every iteration, GLIMPSE estimates a pair of haplotypes for each sample according to its genotype likelihoods, the reference panel of haplotypes, and the previously estimated haplotypes of the other individuals. In the end GLIMPSE outputs the haplotype

calls and genotype posterior probabilities for every position.

While software choice is an important factor, as mentioned above there are other parameters that affect the accuracy of imputation. For instance, minor allele frequency (MAF) has a large impact on imputation accuracy for all imputation methods; as MAF increases, the accuracy of imputation of the minor allele increases because it is more frequently found in the haplotype library of the reference individuals (Wang *et al.*, 2016).

The degree of relationship between individuals in the tested population and the reference set as well as the size of the reference population also affects imputation accuracy. There are several studies in livestock (Druet, Schrooten and de Roos, 2010; Zhang and Druet, 2010; Carvalheiro *et al.*, 2014; Cleveland and Hickey, 2014) and fish species (Yoshida *et al.*, 2018; Fraslin, Yáñez, *et al.*, 2022) recognizing the importance of the relationships and the number of individuals in the reference set, especially when lower density panels are used for the analysis. The existence of more ancestors who are closely related to the individuals of interest (e.g., their parents) increases the chance of finding their haplotypes, stretching over longer distances, in the reference database. The higher the relatedness between the target and the reference population is and the bigger the reference group is, the more accurate the imputation accuracy will be.

The impact of the selection of SNPs in the low-density panel also deserves attention. Other studies have previously reported specific regions of the genome with very low imputation accuracy (Erbe *et al.*, 2012; Carvalheiro *et al.*, 2014; Yoshida *et al.*, 2018), which probably consist of markers in low linkage disequilibrium with their neighbours. Another possibility is incorrect mapping or issues in the assembly of the reference genome. In an attempt to select the best variants for imputation, different methods have been proposed for the design of low-density SNP panels. Tsairidou *et al.* (2020) selected the SNPs randomly across the genome or within the chromosomes. In another study SNPs were selected to be evenly spaced according to position and chromosome size or based on linkage disequilibrium patterns (Yoshida, Yáñez and De Ciencias, 2021). Other studies attempted the selection of highly polymorphic SNPs explaining most of the phenotypic variance of a trait (Aliloo *et al.*, 2018; Wu *et al.*, 2020), or the design of

multi-trait specific SNP panels (He *et al.*, 2018). The results of these studies demonstrate that depending on the trait of interest the variant selection method for the construction of the low-density panel can significantly influence the prediction accuracy.

1.5.2 Previous aquaculture imputation studies

There have been several studies comparing the performance of imputation software and the different parameters affecting genotype imputation in human, plant and livestock populations. However, the number of studies for aquaculture species is limited, only testing FImpute and AlphaImpute software in Atlantic salmon, rainbow trout and Nile tilapia populations (Kijas *et al.*, 2017; Tsai *et al.*, 2017; Yoshida *et al.*, 2018, 2019; Kjetså, Ødegård and Meuwissen, 2020; Tsairidou *et al.*, 2020; Yoshida, Yáñez and De Ciencias, 2021).

Atlantic salmon has been the main focus of imputation studies in aquaculture. The only published study for aquaculture species using AlphaImpute 1.3.2 software is by Tsai *et al.* (2017). In this study, imputing from nearly 8K to 78K SNPs in Atlantic salmon resulted in an imputation accuracy of 0.9 (calculated as the correlation between the allele dosage of the true genotype and the most likely imputed genotype), and genomic prediction accuracies very close to the ones achieved with real genotypes for both body weight and sea lice count. A study in a Tasmanian Atlantic salmon population by Kijas *et al.* (2017) revealed imputation accuracies >95% when imputing from 3K to 78K SNPs with FImpute 2.2 (in this study a big reference panel of 574 individuals was used and multiple generations were present). Similar results were observed by Yoshida *et al.* (2018), suggesting that imputation from 3K to a high-density panel of 50K SNPs could lower the cost for the application of genomic selection in Atlantic salmon without a significant reduction in genomic prediction accuracy. In another Atlantic salmon study by Tsairidou *et al.* (2020), genotyping offspring at the very low-density of 200 SNPs and imputing them with FImpute 2.2 to their parents' medium-density panel (5K SNPs) achieved almost the same genomic prediction accuracy as the true medium-density panel; this scenario greatly reduced the estimated genotyping cost by 62% compared to genotyping the whole population for 5000 SNPs.

More recently, a preprint study attempted imputation of low-coverage whole-genome sequencing data of SNPs and structural variants (SVs) in Atlantic salmon. In this study imputation using 1X whole-genome sequencing was performed with GLIMPSE which resulted in an accuracy of 95% (measured as the percentage of the ratio between the number of correct genotype calls divided by the sum of correct and incorrect genotype calls) (Gundappa *et al.*, 2023).

Only one study has evaluated imputation accuracy in rainbow trout. In this study, Yoshida *et al*. (2021) found that increased genomic prediction accuracies were achieved, close to that of the high density panel, when using the 50K and 1K panels imputed to whole-genome sequence genotypes. Genomic prediction accuracy was dependent on the SNP selection method and the genetic architecture of the trait. The SNP selection method based on Genome Wide Association (GWA) summary statistics (most important 50K SNPs capturing 78% of genetic variance) had an advantage over the selection of SNPs pruned with PLINK software (50K evenly spaced SNPs, distributed according to chromosome size). Likewise, a single study has been performed in Nile tilapia (Yoshida *et al.*, 2019), where imputation achieved accuracies above 90% (0.9 accuracy for the 0.5K low-density panel and 0.98 for the 3K, imputed to 32K), with the reference panel consisting of parents but also 20% of the offspring. Genomic prediction accuracies were always higher for the imputed panels compared to using the low-density genotypes alone, and quite close to or the same as if using the full high-density panel.

The promising results of these studies suggest that the combination of low-density SNP panels with genotype imputation can achieve genomic prediction accuracies similar to those of high-density panels. This combination can decrease genotyping costs in aquaculture species, enabling the wider implementation of genomics in aquaculture breeding programmes. However, in many cases the results of these studies, conducted in a limited number of aquaculture species, are not directly comparable because they use different metrics for the assessment of results and test different densities, some of them using relatively dense low-density panels. For genotype imputation to be routinely implemented in aquaculture genomic selection programmes worldwide, further testing and optimisation is needed to understand the ideal parameters to minimise cost and maximise genetic

gain.

1.6 Aims and Objectives

There is a growing interest in transitioning from wild fisheries towards aquaculture, but there are still challenges for the smaller aquaculture settings, which hold back the development of this sector. The success of this so-called 'blue transition' depends on different factors, such as the political and economic situation, legislation, or social and environmental conditions, but also on the availability and application of new technologies. Technological advances in the field of genetics have not yet been widely adopted by the aquaculture industry for most farmed species, and their application could contribute to the sustainable development of this growing industry. In particular, the development of new breeding programmes, leading to high-performing, locally adapted stocks are fundamental for future-proofing aquaculture. The majority of small and medium enterprises, dominating aquaculture in low- and middle-income countries, do not have well-managed breeding programmes for directional selection and improvement of the desirable traits. The establishment of breeding programmes for small farms is expensive and where basic breeding programmes exist, they lag behind in the implementation of the available genomic tools required by modern breeding programmes due to the high cost compared to their relatively small production. Enabling the use of genomics can significantly improve production cost-efficiency and is key to unlock the potential of aquaculture stocks and ensure food security.

The focus of this PhD project is to design and evaluate cost-effective genotyping strategies and breeding programme designs for different species, which will enable aquaculture to fully benefit from genomic selection. These strategies will be developed by exploiting low-density genotyping, imputation, and the incorporation of putative functional variants to establish best practices for their use in aquaculture breeding programmes. The specific objectives of the project are:

1. To establish best practices for the use of low-density panels and genotype imputation in different aquaculture breeding schemes.

2. To assess the potential of low-coverage whole-genome sequencing (lcWGS), combined with imputation as an alternative to whole-genome sequencing (WGS) and genotyping.
3. To test the impact of incorporation of putative functional variants on the genomic prediction accuracy.

Chapter 2
Assessment of Cost-Effective Genomic Selection through Imputation of Low-Density SNP Panels

This chapter is based on the published paper:

Kriaridou Christina, Smaragda, Tsairidou, Clémence Fraslin, Gregor Gorjanc, Mark E. Looseley, Ian A. Johnston, Ross D. Houston, and Diego Robledo. 2023. “Evaluation of Low-Density SNP Panels and Imputation for Cost-Effective Genomic Selection in Four Aquaculture Species.” Frontiers in Genetics 14 (May): 1194266. https://doi.org/10.3389/fgene.2023.1194266.

The original document is included within this chapter's body of work, along with an introduction, conclusion, and supplementary data.

2.1 Introduction to Chapter 2

Selective breeding is fundamental to meeting the globally growing demand for seafood by increasing aquaculture production and improving traits related to animal health. Unlike terrestrial livestock, selective breeding practices are not widespread in aquaculture. Farmed aquatic species are still in early stages of domestication, indicating that there is substantial genetic variation that we can select for in traits of economic importance such as resistance to diseases, growth rate, meat quality and maturation age. The establishment of modern breeding programmes, coupled with the integration of genomic tools like genomic selection, holds great potential for accelerating the improvement of desirable traits in aquaculture production. However, the implementation of genomic selection poses challenges, particularly for small and medium-sized enterprises, as it requires a substantial budget allocation for genotyping. In this chapter, we propose a potential cost-effective alternative strategy for the widespread adoption of genomic selection. This strategy involves the utilization of low-density SNP panels in combination with genotype imputation. Our study's first objective was to establish best practices for using low-density panels and genotype imputation in aquaculture breeding programmes by comparing various imputation methods, marker densities, imputation software, and marker selection strategies across several aquaculture species.

2.2 Original published manuscript as it appears in https://doi.org/10.3389/fgene.2023.1194266

Evaluation of low-density SNP panels and imputation for cost-effective genomic selection in four aquaculture species

Christina Kriaridou[1], Smaragda Tsairidou[2], Clémence Fraslin[1], Gregor Gorjanc[1], Mark Looseley[3], Ian A. Johnston[3], Ross D. Houston[1, 4] and Diego Robledo[1,*]

1. The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, UK

2. Global Academy of Agriculture and Food Systems, University of Edinburgh, UK

3. Xelect Ltd, Horizon House, St Andrews, Scotland, KY16 9LB, UK

4. Benchmark Genetics, 1 Pioneer Building, Edinburgh Technopole, Penicuik, UK

*Corresponding author: diego.robledo@roslin.ed.ac.uk

Key words: selective breeding, imputation, genomic prediction, aquaculture, fish, bivalve

Abstract

Genomic selection can accelerate genetic progress in aquaculture breeding programmes, particularly for traits measured on siblings of selection candidates. However, it is not widely implemented in most aquaculture species, and remains expensive due to high genotyping costs. Genotype imputation is a promising strategy that can reduce genotyping costs and facilitate the broader uptake of genomic selection in aquaculture breeding programmes. Genotype imputation can predict ungenotyped SNPs in populations genotyped at a low-density (LD), using a reference population genotyped at a high-density (HD). In this study, we used datasets of four aquaculture species (Atlantic salmon, turbot, common carp and Pacific oyster), phenotyped for different traits, to investigate the efficacy of genotype imputation for cost-effective genomic selection. The four

datasets had been genotyped at HD, and eight LD panels (300-6000 SNPs) were generated *in silico*. SNPs were selected to be: i) evenly distributed according to physical position or ii) selected to minimise the linkage disequilibrium between adjacent SNPs. Imputation was performed with three different software packages (AlphaImpute2, FImpute v.3 and findhap v.4). The results revealed that FImpute v.3 was faster and achieved higher imputation accuracies. Imputation accuracy increased with increasing panel density for both SNP selection methods, reaching correlations greater than 0.95 in the three fish species and 0.80 in Pacific oyster. In terms of genomic prediction accuracy, the LD and the imputed panels performed similarly, reaching values very close to the HD panels, except in the pacific oyster dataset, where the LD panel performed better than the imputed panel. In the fish species, when LD panels were used for genomic prediction without imputation, selection of markers based on either physical or genetic distance (instead of randomly) resulted in a high prediction accuracy, whereas imputation achieved near maximal prediction accuracy independently of the LD panel, showing higher reliability. Our results suggests that, in fish species, well-selected LD panels may achieve near maximal genomic selection prediction accuracy, and that the addition of imputation will result in maximal accuracy independently of the LD panel. These strategies represent effective and affordable methods to incorporate genomic selection into most aquaculture settings.

1 Introduction

Aquaculture has been the fastest-growing food production sector in recent decades, with a 609% rise in the total annual output from 1990 to 2020 (FAO, 2022). This growth has revolutionised the supply of seafood products across the planet, providing nutritious seafood to a growing human population and significantly contributing to meeting food security objectives in many regions. However, the development of aquaculture in different countries has been uneven, and seafood production still needs to be increased to ensure food security and reduce the effect of fishing on wild populations, offsetting the environmental impacts of overexploitation (Cottrell *et al.*, 2021).

In 2016 over 95% of the global aquaculture output originated from low and middle-income countries (Stentiford *et al.*, 2020). The rapid expansion of aquaculture in these countries is primarily due to the adoption of aquaculture by small and medium-sized enterprises, but there are still challenges that hold back the development of smaller aquaculture settings (Kumar, Engle and Tucker, 2018; FAO, 2020). A significant restriction is the lack of well-managed breeding programmes for directional selection and improvement of desirable traits. In addition, the establishment of breeding programmes for small farms is expensive. Therefore, where basic breeding programmes exist, they lag behind in the implementation of the available genomic tools utilised by modern breeding programmes due to the high cost compared to their relatively small production. The use of genomics can improve selection intensity and breeding value prediction accuracy, particularly for traits not possible to measure directly on selection candidates. In turn, this can then lead to a more efficient production, benefiting the entire supply chain, which is essential to unlock the potential of aquaculture stocks and ensure food security (Houston *et al.*, 2020; FAO, 2022).

Genomic selection uses genetic markers to more accurately predict the breeding values of individuals compared to pedigree-based approaches, leading to higher rates of genetic gain and better management of inbreeding (Houston *et al.*, 2020; Boudry *et al.*, 2021; Regan *et al.*, 2021). Despite its potential, genomic selection has only been implemented in the most advanced aquaculture sectors, and only for a small number of aquatic species, such as Atlantic salmon, rainbow

trout, American catfish, whiteleg shrimp or Nile tilapia (Lillehammer *et al.*, 2020; Yáñez, Joshi and Yoshida, 2020; Boudry *et al.*, 2021; Houston, Kriaridou and Robledo, 2022). One of the barriers to the widespread adoption of genomic selection is the high cost of genotyping. Genotyping can be prohibitively expensive for small and medium aquaculture operations, making it more challenging for them to adopt genomic selection practises (Boudry *et al.*, 2021). For these industries to benefit from genomic selection, low-cost genotyping strategies that do not significantly compromise the prediction accuracy of breeding values are required.

Several studies have looked into the use of low-density (LD) SNP panels as a cost-effective alternative, with only a few thousands or even hundreds of SNPs used for genomic selection, in contrast to high-density (HD) panels, usually containing tens of thousands of SNPs. Generally, studies on aquaculture species have reported that SNP densities can be reduced from tens of thousands to thousands without a significant loss of prediction accuracy (Tsai *et al.*, 2016; Palaiokostas *et al.*, 2018, 2019; Robledo, Matika, *et al.*, 2018; Yoshida *et al.*, 2019; Gutierrez *et al.*, 2020; Kriaridou *et al.*, 2020; Tsairidou *et al.*, 2020; Al-Tobasei *et al.*, 2021). Additionally, complementary strategies such as genotype imputation can be used to further reduce the cost and improve the accuracy of low-cost genomic selection.

Genotype imputation is a method that can be used to predict missing genotypes in an individual based on the genotypes of other individuals of the same species. A common imputation strategy is to use a group of individuals genotyped with a HD panel (reference population) to infer the missing genotypes of other individuals (target population) genotyped with a LD panel, which is composed of a subset of markers from the HD panel (Marchini and Howie, 2010b; Sargolzaei, Chesnais and Schenkel, 2010). The reference and target populations need to be related to some degree as imputation relies on linkage and linkage disequilibrium within those populations. The general idea of genotype imputation is that related individuals share long haplotype blocks (set of markers in linkage disequilibrium segregating together). These haplotype blocks are broken by recombination events occurring from one generation to the next; hence two animals will share longer haplotypes the more related they are.

Imputation algorithms can use a combination of population and pedigree-based methods (Browning, 2008; Bouwman *et al.*, 2014; Sargolzaei, Chesnais and Schenkel, 2014; Wang *et al.*, 2016; Antolín *et al.*, 2017; Lashmar, Muchadeyi and Visser, 2019; Phocas, 2022). FImpute (Sargolzaei, Chesnais and Schenkel, 2014) and AlphaImpute (Whalen and Hickey, 2020) are popular algorithms developed for animals and plants, combining population and pedigree-based imputation methods. Population-based methods utilise linkage disequilibrium information between markers in various ways. Generally, they use Hidden Markov Model (HMM) approaches to model genotype and underlying haplotype variation relying on population-wide linkage disequilibrium between markers (short shared haplotypes) (Sargolzaei, Chesnais and Schenkel, 2014; Whalen *et al.*, 2018). Pedigree-based methods incorporate information from linkage and pedigree relationships for imputation. These methods take advantage of the long-haplotypes shared by closely related individuals, such as parent-offspring or full-sibs, as well as using Mendelian inheritance rules to infer missing genotypes (Antolín *et al.*, 2017). Pedigree information increases in importance as the LD panel becomes sparser, because it enables capturing the long-range haplotype blocks shared between relatives. Studies where imputation is applied to a population of related individuals (family studies) are more powerful and effective in identifying low-frequency variants (Sargolzaei, Chesnais and Schenkel, 2014; Liu *et al.*, 2019). The choice of software can also impact the results; different algorithms make use of the available information differently, so the optimal imputation software may differ depending on the population of interest.

In addition to the imputation method, there are several other factors affecting genotype imputation accuracy, namely SNP minor allele frequency (MAF), the selection of SNPs for the LD panel (number of SNPs and their chromosomal distribution), the number of individuals in the reference population and the population structure. MAF significantly impacts imputation accuracy for all imputation methods; as MAF increases, the accuracy of imputation of the minor allele increases (Wang *et al.*, 2016). Imputation of rare alleles is important because variants with low frequency may have large effects, linked to the “missing heritability” in some complex traits (Manolio *et al.*, 2009; Sargolzaei, Chesnais and Schenkel, 2014; Gonzalez-Recio *et al.*, 2015). The size of the reference

population also affects imputation; the greater the number of individuals in the reference panel, and the more closely related they are to the target individuals, the more accurate is genotype imputation (Garcia *et al.*, 2022). Finally, one aspect that requires further investigation is the impact of SNP selection strategy for the LD panel. Various methods have been proposed for the design of LD SNP panels, such as: i) randomly selected SNPs across the genome or within the chromosome (Tsairidou *et al.*, 2020), ii) evenly spaced according to position and chromosome size (Yoshida, Yáñez and De Ciencias, 2021), iii) based on linkage disequilibrium patterns (Yoshida, Yáñez and De Ciencias, 2021), iv) selection of highly polymorphic SNPs explaining most of the phenotypic variance of a trait (Aliloo *et al.*, 2018; Wu *et al.*, 2020), v) or even the design of multi-trait-specific SNP panels (He *et al.*, 2018) and family-specific SNP panels (Whalen, Gorjanc and Hickey, 2019). These studies have shown that for some traits, the SNP selection method for the LD panel plays an important role.

Several studies have compared the performance of imputation software and the different parameters affecting genotype imputation in human, plant and livestock populations. However, aquaculture broodstock populations are typically comprised of relatively few (but large) full and half sib families, with limited population structure and, as such, might be expected to show a different response to imputation strategies. Despite this, the number of studies testing imputation performance in aquaculture species is limited and they mainly use either FImpute or AlphaImpute software in Atlantic salmon (Kijas *et al.*, 2017; Tsai *et al.*, 2017; Yoshida *et al.*, 2018; Kjetså, Ødegård and Meuwissen, 2020; Tsairidou *et al.*, 2020), rainbow trout (Vallejo *et al.*, 2021; Yoshida, Yáñez and De Ciencias, 2021) and Nile tilapia (Yoshida *et al.*, 2019; Garcia *et al.*, 2022). Only one recent study has tested Beagle imputation software in Atlantic salmon, common carp, sea bream and rainbow trout (Song and Hu, 2022b). The promising results of these studies suggest that the combination of LD SNP panels with genotype imputation can achieve similar genomic prediction accuracies to HD panels. This combination can decrease the genotyping cost in aquaculture species, enabling the broader implementation of genomics in breeding programmes. However, in many cases the results of these studies are not directly comparable because they use different metrics to assess results and test different parameters. Therefore, further testing

and optimisation of imputation algorithms and SNP selection methods is needed, across a range of aquaculture species and traits with the use of common assessment methods for genotype imputation to be routinely implemented in aquaculture selection programmes worldwide.

The objectives of this study were to (i) evaluate the performance of three imputation software packages, FImpute v.3, AlphaImpute2 and findhap v.4 in breeding populations from four diverse aquaculture species; (ii) investigate the impact on imputation accuracy of the number of markers in the LD panel and their selection method; and (iii) evaluate the genomic prediction accuracy of imputed vs LD genotypes for different traits in the four species. Our results contribute towards the definition of best practices for the broader application of genotype imputation and cost-effective genomic selection in aquaculture.

2 Materials and Methods

2.1 Datasets

This study used previously published datasets from four species. Specifically:

- A farmed Atlantic salmon (*Salmo salar*) population of 624 individuals (90 parents and 534 offspring), belonging to 61 full-sib families as described in (Tsai *et al.*, 2015). This population was challenged with *Lepeophtheirus salmonis* and sea lice counts on the fish were recorded for all the offspring. This trait had a positively skewed distribution and was logarithmically transformed. All individuals were genotyped with a 132K SNP array, and 78,035 SNPs distributed across 29 pairs of chromosomes were retained after quality control for further analysis.
- A turbot (*Scophthalmus maximus*) population of 1,445 fish (47 parents and 1,398 offspring), distributed across 36 full-sib families as described in Anacleto *et al.* (2019). The gonads of the fish were checked for the presence or absence of a parasite causing Scuticociliatosis (*Philasterides dicentrarchi*). Individuals were genotyped using RAD-seq and after quality control 11,069 SNPs were successfully mapped to the 22 pairs of

chromosomes.

- A common carp (*Cyprinus carpio*) population of 1,319 individuals (60 parents and 1259 offspring), comprising 195 full-sib families. This population was challenged with koi herpesvirus as described in Palaiokostas *et al.* (2018) and phenotypic records of body weight were obtained. Individuals were genotyped using RAD-Seq sequencing method and 15,615 SNPs were retained for downstream analysis (Palaiokostas *et al.*, 2019). The positions of these markers were updated according to the latest reference genome (GenBank assembly accession number GCA_018340385.1) by using standard nucleotide BLAST (Altschul *et al.*, 1990) and 8,506 SNPs were successfully assigned to 50 pairs of chromosomes from which 8,103 SNPs were retained after quality control.
- A Pacific oyster (*Crassostrea gigas*) population of 762 individuals (44 parents and 718 offspring), belonging to 30 full-sib families. Individuals in this study were challenged with ostreid herpesvirus (OsHV-1), measured for time to death, and genotyped using a SNP array with ~27K informative Pacific oyster SNPs (Gutierrez *et al.*, 2020). After updating the SNP positions according to the latest genome assembly (Peñaloza *et al.*, 2020) and quality control, 16,447 SNPs remained, distributed across the 10 chromosome pairs.

2.2 Quality control

All datasets were filtered using PLINK v.1.9 (Purcell *et al.*, 2007). Individuals with just one of their two parents genotyped or > 20% missing genotypes were excluded from the analysis. SNPs with > 10% missing genotypes; significant deviation from Hardy–Weinberg Equilibrium (P-value < 10^{-6}); MAF < 0.05; or Mendelian error rates > 10% were also excluded from subsequent imputation analyses. A summary of the data for the different species before and after quality control can be found in **Table 1**. After imputation, all the datasets were filtered again for MAF (< 0.05).

Table 1 Summary of the datasets.

Species	SNPs before and after QC		Individuals before and after QC		Full-sib families	Phenotypes	Study with available dataset
Salmo salar	78,362	78,035	624	606	57	Sea lice (*Lepeophtheirus salmonis)* count	Tsai *et al*. 2015
Scophthalmus maximus	17,690	11,069	1,445	1,396	38	Presence of parasites (*Philasterides dicentrarchi*) in the gonads	Anacleto *et al*. 2019
Cyprinus carpio	8,506	8,103	1,319	1,172	195	Body weight	Palaiokostas *et al*. 2019
Crassostrea gigas	22,994	16,447	762	701	30	Resistance to oyster herpesvirus (OsHV-1)	Gutierrez *et al*. 2020

2.3 SNP selection methods for the low-density panels

The LD SNP panels were generated *in silico* by selecting 300, 500, 700, 1,000, 2,000, 3,000, 5,000 and 6,000 SNPs using the two methods described below. The LD panels were created by masking (i.e., setting to missing) all the SNPs not selected by each method.

Physical-distance-based method: The selection of SNPs for the LD panels was implemented with a custom R script (available in https://github.com/Roslin-Aquaculture/Select-SNPs-to-generate-low-density-panels), considering the total number of SNPs and the length of each chromosome. For each density, a single panel was created with the number of markers selected being proportional to chromosome length and evenly distributed across the chromosomes according to position (physical distance). For this SNP selection method, the first and the last SNP on each chromosome were always selected and included in the LD panel. When no SNPs were available in the required position to achieve an even distribution, the closest available SNP was selected to obtain a LD panel with the desired number of markers. If a chromosome did not have enough SNPs (e.g., for densities ≥ 5,000 SNPs), all of the SNPs on that chromosome were selected and the final panel density was allowed to be slightly lower than expected (i.e., no additional SNPs were selected on the other chromosomes).

Genetic-distance-based method: For the SNP selection method based on linkage disequilibrium, PLINK 1.9 (Purcell *et al.*, 2007) was used to generate pruned SNP subsets based on variable window size, step size and squared correlation (r^2) threshold values, to achieve the desired number of SNPs for each density. SNP pruning was performed using the "--indep-pairwise" command. In brief, at each step, squared correlation was calculated between each pair of SNPs within a genomic window, specified using SNP count ("variant ct"). All SNPs with squared correlation greater than the given r^2 threshold were removed from the window until there were no such pairs. At the end of each step, the window was shifted forward by a "step size (variant ct)", and the procedure was repeated. A single LD panel was created for each target density.

Randomly selected SNPs: Additionally, four LD panels were generated

by randomly choosing 300, 500, 700 and 1,000 SNPs throughout the genome to test prediction accuracy before and after imputation with FImpute v.3.

2.4 Genotype imputation

Imputation of the offspring's LD genotypes was performed using their parents as reference population (genotyped for the HD panels) with three software packages: AlphaImpute2 (Whalen and Hickey, 2020), FImpute v.3 (Sargolzaei, Chesnais and Schenkel, 2014) and findhap v.4 (VanRaden *et al.*, 2013); a two-generation pedigree was available for all datasets, therefore pedigree and population-based imputation were performed.

AlphaImpute2 (Whalen and Hickey, 2020) imputation was performed separately for each chromosome using the default parameters, which are listed below, and SNPs in the genotype input file were ordered according to position on the chromosome. In the first step of pedigree imputation, five rounds of multi-locus iterative peeling were performed. The genotype calling threshold for the first round of peeling before phasing was 0.9. In the second step, where the algorithm builds the reference haplotype library, five rounds of phasing were conducted. Finally, for the 3rd step of pedigree imputation another five rounds of multi-locus iterative peeling were performed, using the phased genotypes in the second step, and genotypes were set to the best-guess.

FImpute v.3 (Sargolzaei, Chesnais and Schenkel, 2014) uses a single genotype file with all the chromosomes present, and also requires information of the genomic location of the SNPs, provided in a map file, to model recombination. The 'parentage_test' parameter was used to check for parentage errors with an error rate threshold of 0.05 to find progeny-parent mismatches. When a progeny-parent Mendelian inconsistency was detected, in most cases, genotypes of progeny and parents were set to missing and re-imputed. For this analysis, the conflicting parents were set to missing and original genotypes were not adjusted. In the results presented here, random filling of genotypes based on allele frequency was used to allow for a better comparison with AlphaImpute2.

For Findhap v.4 (VanRaden *et al.*, 2013), the maximum and minimum length of haplotype segments were defined as 600 and 65, respectively, with an

overlapping length of 10 and an error rate of 0.004. The number of different haplotypes within any segment was set to 1,000 for the lower densities, and it was increased to 2,000 for the 5,000 and 6,000 SNPs densities to consider all the possible haplotypes.

For all three methods, imputation accuracy was measured as the average Pearson correlation between the original and the imputed genotypes for each test individual. To test the effect of MAF on imputation accuracy, we calculated minor allele frequencies with PLINK v.1.9 and divided the SNPs into five MAF bins: (0-0.1], (0.1-0.2], (0.2-0.3], (0.3-0.4] and (0.4-0.5].

2.5 Estimation of genetic parameters

For each trait in the different datasets, heritabilities were estimated using ASReml 4.2 (Gilmour, Gogel and Welham, 2021) using a linear mixed model as follows:

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

where $\boldsymbol{y}$ is a vector of observed phenotypes, $\boldsymbol{\mu}$ is the overall mean of phenotype records, $\boldsymbol{b}$ is the vector of fixed effects, $\boldsymbol{a}$ is a vector of additive genetic effects distributed as $\boldsymbol{a} \sim \boldsymbol{N}(0, \boldsymbol{G}\sigma_a^2)$, where σ_a^2 is the additive genomic variance and $\boldsymbol{G}$ is the genomic relationship matrix, while $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the corresponding incidence matrices for fixed and additive effects, respectively, and $\boldsymbol{e}$ is a vector of residuals.

Gonad parasite trait in the turbot dataset was binary, thus we used the generalized linear mixed model with the logit link function that links the probability of observing an event to the underlying linear model:

$$P(\boldsymbol{y}_i = 1) = \frac{\exp\left(\boldsymbol{\mu} + \boldsymbol{Xb}_i + \boldsymbol{Za}_i + \boldsymbol{e}_i\right)}{1 + \exp\left(\boldsymbol{\mu} + \boldsymbol{Xb}_i + \boldsymbol{Za}_i + \boldsymbol{e}_i\right)}$$

The fixed effects included in the different models for each species were i) body weight in Atlantic salmon, ii) factorial-cross group (four levels) in carp, iii) box (36 levels) in turbot, and iv) tank (two levels) in oyster.

The genomic relationship matrix between pairs of individuals j and k (gjk)

was calculated using the GCTA software (Yang *et al.*, 2011) as follows:

$$g_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}$$

where N is the total number of SNPs, x_{ij} and x_{ik} are the number of copies of the reference allele for the i^{th} SNP for the j^{th} and k^{th} fish, respectively, and p_i is the frequency of the reference allele estimated from the markers.

2.6 Cross-validation for genomic-based prediction accuracy

The accuracy of genomic prediction was estimated by 20 replicates of fivefold cross-validation analysis (80% of individuals in the training set and 20% in the validation set; 'CVrep' GitHub statistical R package (Tsairidou 2019), available at https://github.com/SmaragdaT/CVrep). The phenotypes in the validation set were masked, and genomic best linear unbiased prediction (GBLUP) was applied to predict the breeding values of the validation set individuals in ASReml 4.2 (Gilmour, Gogel and Welham, 2021), using the linear mixed model described above. Prediction accuracy was calculated as the correlation between the predicted breeding values of the validation set and the actual phenotypes divided by the square root of heritability, estimated from the full dataset for each trait [$\approx \frac{r(y,\hat{y})}{h}$].

3 Results

3.1 Trait summary and genetic parameters

A different phenotype was used in each dataset (**Table 2**): i) In Atlantic salmon, the log-transformed sea lice count was used as phenotype. Log-transformed sea lice counts had a mean of 3.11 ± 0.56 and a genomic heritability estimate of 0.19 ± 0.07. ii) In turbot, the binary trait of absence or presence of gonad parasites was used. Gonad parasites were present in 881 individuals, while 441 individuals were free of parasites. The estimated genomic heritability for this trait was 0.27 ± 0.08. iii) In Pacific oyster, we used the phenotype of days to death

after infection with OsHV-1-µvar, with survivors being assigned a value of eight days (end of the challenge). The mean and standard deviation of surviving days was 6.91 ± 1.82, and the estimated genomic heritability was 0.64 ± 0.05. iv) In common carp, the mean value for body weight was 16.36 ± 4.65, and the heritability estimate was 0.22 ± 0.04.

Table 2 Genomic heritability and prediction accuracy using HD panels.

Species	**Phenotypes**	**Genomic heritability estimates**	**HD panel genomic prediction accuracy (mean ± sd)**
Atlantic salmon	Log transformed sea lice (*Lepeophtheirus salmonis*) count	0.19 ± 0.07	0.54 ± 0.05
Turbot	Presence/absence of gonad parasites (*Philasterides dicentrarchi)*	0.27 ± 0.08	0.34 ± 0.02
Common carp	Body weight	0.22 ± 0.04	0.69 ± 0.02
Pacific oyster	Resistance to oyster herpesvirus (OsHV-1) measured as time to death	0.64 ± 0.05	0.62 ± 0.03

3.2 Accuracy of Imputation

Imputation accuracy increased with increasing panel density for all software (**Figure 1**). Overall, the results revealed that FImpute v.3 was more accurate for most of the densities in all the species, and findhap v.4 was mostly second in the ranking. Although AlphaImpute2 was generally ranked last between the three software, it outperformed findhap v.4 in terms of accuracy for the five lowest densities (300-2,000 SNPs) in the Atlantic salmon dataset. It also outperformed FImpute v.3 at the lowest density of 300 SNPs (**Figure 1A**). Imputation accuracy for the lowest density of 300 SNPs, when imputing with FImpute v.3, ranged between 0.61 (Pacific oyster) and 0.76 (Atlantic salmon and

turbot). For the 6,000 SNPs density, the fish species reached very high imputation accuracies (0.95-0.98), but the accuracy value was noticeably lower for Pacific oyster (0.80) (**Figure 1**).

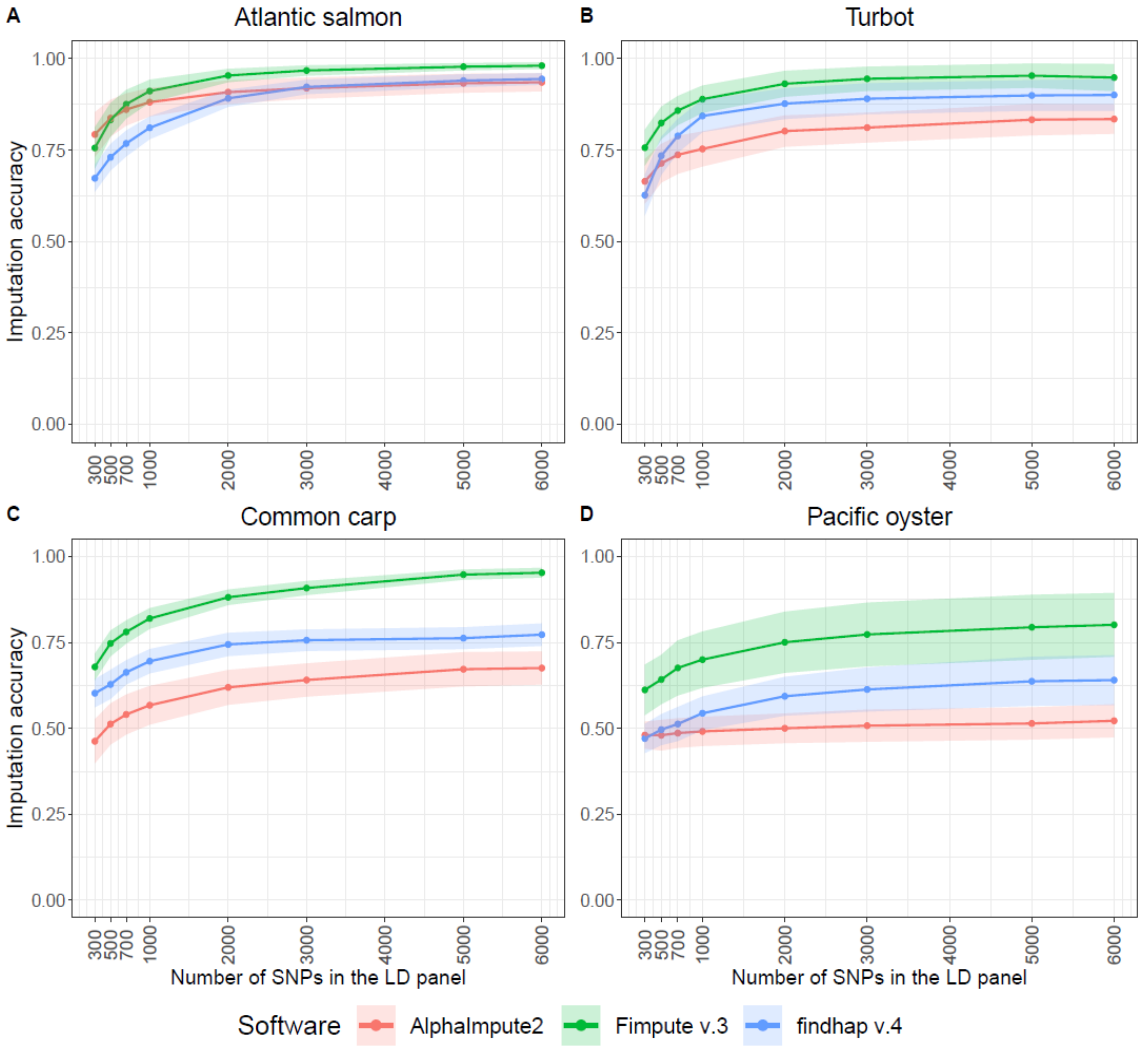


Figure 1 Genotype imputation accuracy in four aquaculture species. Average genotype imputation accuracy (correlation between true and imputed genotypes) for the three imputation software in each of the four species. The ribbons represent the standard deviation of the average imputation accuracy across all individuals. The SNP selection method based on physical distance was used to impute the LD panels in these graphs. The Atlantic salmon LD panels (A) were imputed to 78,035 SNPs, the turbot (B) to 11,069 SNPs, the common carp (C) to 8,103 SNPs and the Pacific oyster (D) to 16,447 SNPs.

Regarding computing time, FImpute v.3 was faster than the other two software tested. Running time results of the three software when imputing the 300 SNPs panel density for each species are shown in **Table 3**. The average computational time across the four species for the LD panel of 300 SNPs when imputing with FImpute v.3 was 1 min and 13 sec, with findhap v.4 showing a similar average running time of 1 min 55 sec, and AlphaImpute2 considerably longer running times of 24 min 56 sec in average.

Table 3 Computational time for each software to impute from the 300 SNPs density panel.

Species	FImpute v.3	findhap v.4	AlphaImpute2
Salmo salar	1 min 16 sec	2 min 49 sec	30 min 36 sec
Scophthalmus maximus	47 sec	1 min 45 sec	16 min 46 sec
Cyprinus carpio	27 sec	1 min 17 sec	44 min 01 sec
Crassostrea gigas	1 min	1 min 7 sec	7 min 41 sec

In Figure 2, the genetic distance method based on linkage disequilibrium slightly increased the accuracy of imputation for most of the very low densities in Atlantic salmon, turbot and Pacific oyster (300-2,000 SNPs), while in the common carp dataset it improved the imputation accuracy of the higher densities (2,000-6,000 SNPs) (**Figure 2**). However, the differences observed in imputation accuracy between the two LD panel SNP selection methods were mostly non-significant. Since both the imputation and prediction accuracy results of the imputed panels were similar when the SNPs were selected with the genetic or the physical-distance-based method, the results we present below are with the physical-distance-based method and imputed with FImpute v.3 software package.

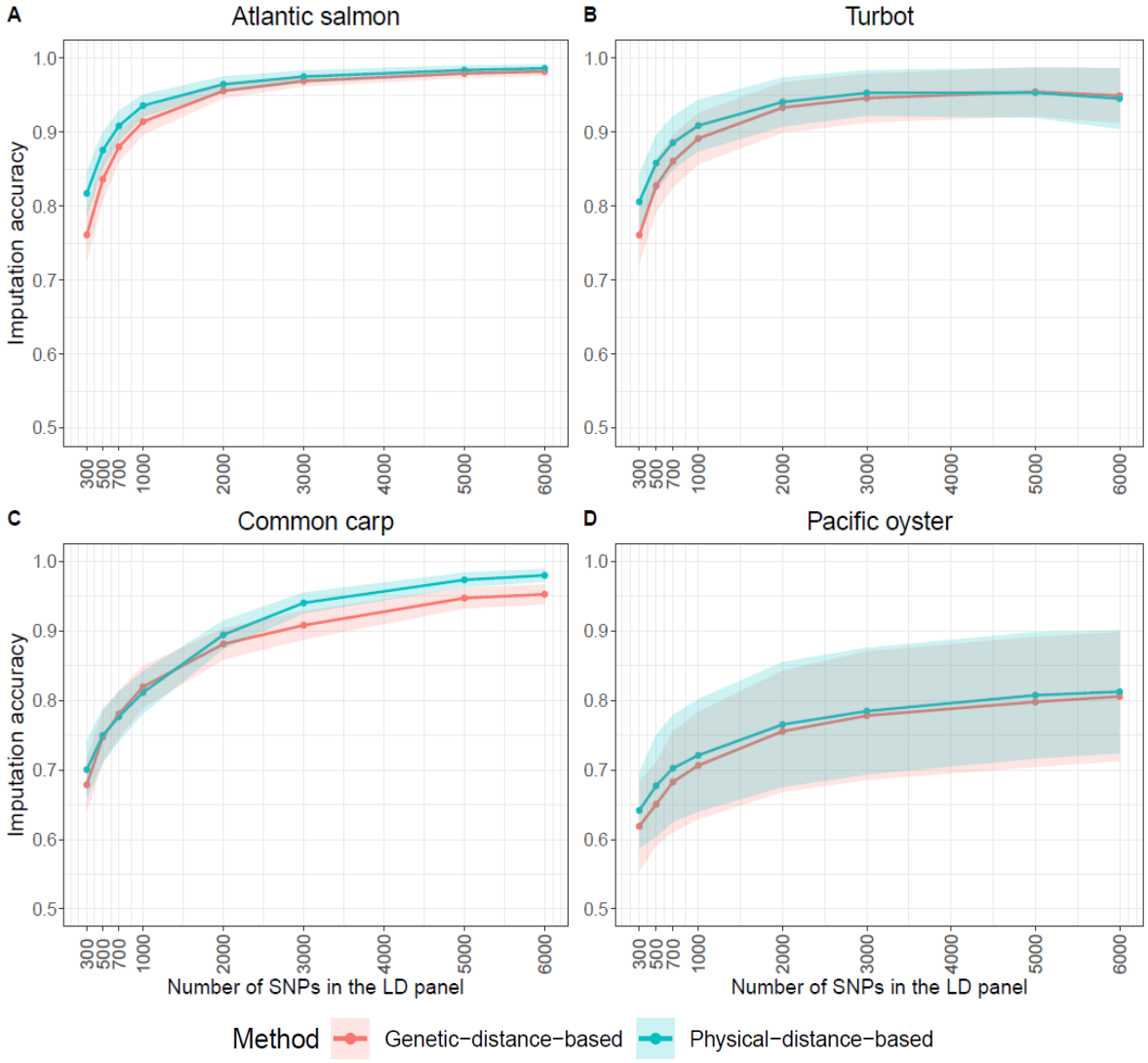


Figure 2 Influence of LD SNP panel design on imputation accuracy. Average genotype imputation accuracy (correlation between true and imputed genotypes) using FImpute v.3 in each of the four species for the two SNP selection methods: physical and genetic distance-based. The ribbons represent the standard deviation of the average imputation accuracy across all individuals. The y-axis in these graphs ranges from 0.5 to 1 to facilitate the comparison of the two methods.

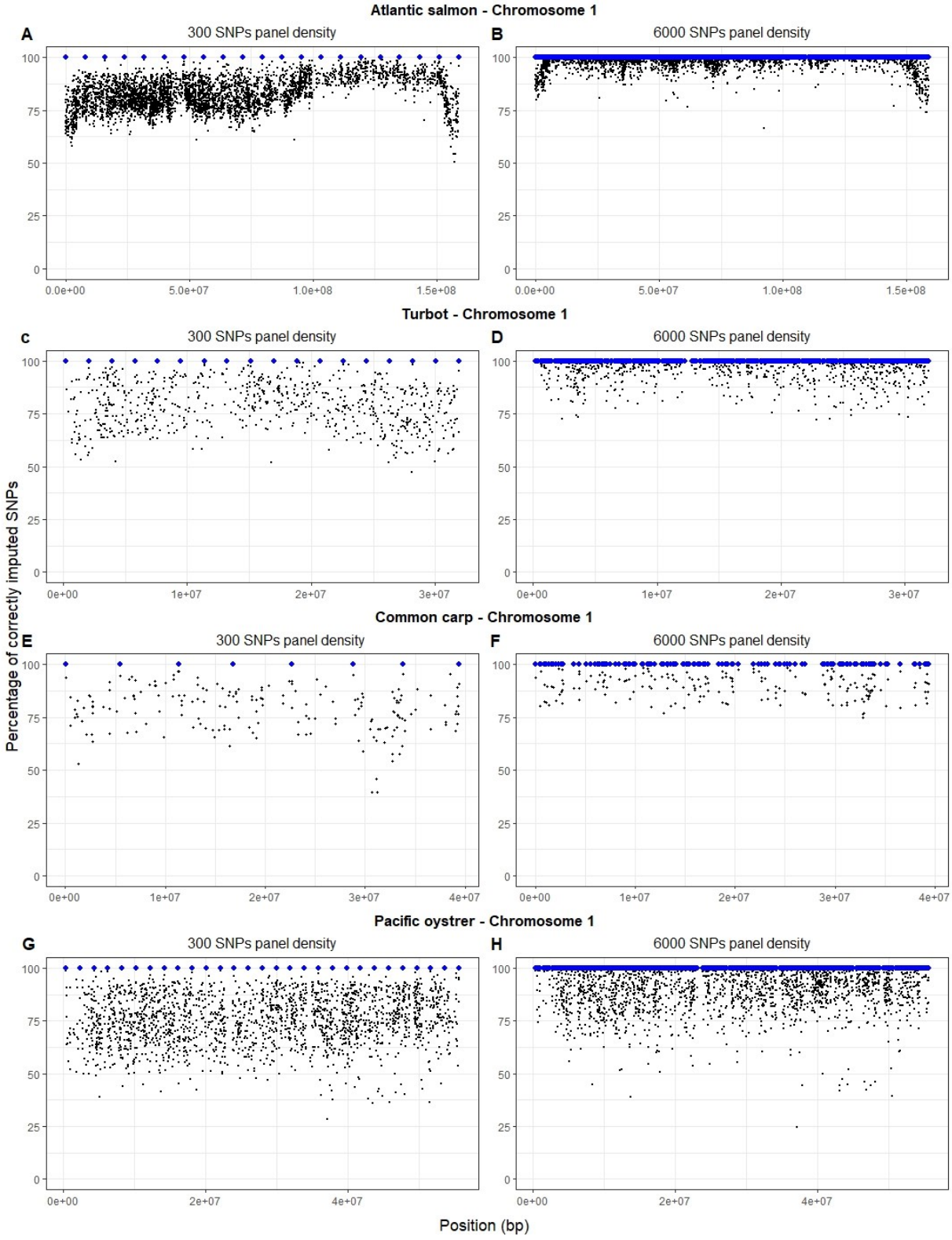


Figure 3 Percentage of correctly imputed genotypes with FImpute v.3 for each SNP of chromosome 1 in each of the four species, using the LD panels of 300 (A, C, E, G) and 6,000 (B, D, F, H) SNPs (selected with the physical-distance-based method). The blue dots indicate the physical position of the SNPs in the LD panel, whereas the black dots indicate the imputed SNPs.

There is a visible pattern of slightly decreased imputation accuracy at the ends of the chromosomes of the four species (**Figure 3**), but this was not consistent for all chromosomes (**Supplementary Material**). This phenomenon is clearer in Atlantic salmon (**Figure 3A and B**), possibly due to the higher number of SNPs in the HD panel. Increasing the SNP density of the LD panel from 300 SNPs to 6,000 SNPs substantially improved imputation accuracy throughout the chromosome and especially at chromosomal ends (**Figure 3**). In the oyster dataset, there were poorly imputed SNPs throughout the chromosome, and for some of these SNPs accuracy did not improve when the panel density was increased (**Figure 3G, H**).

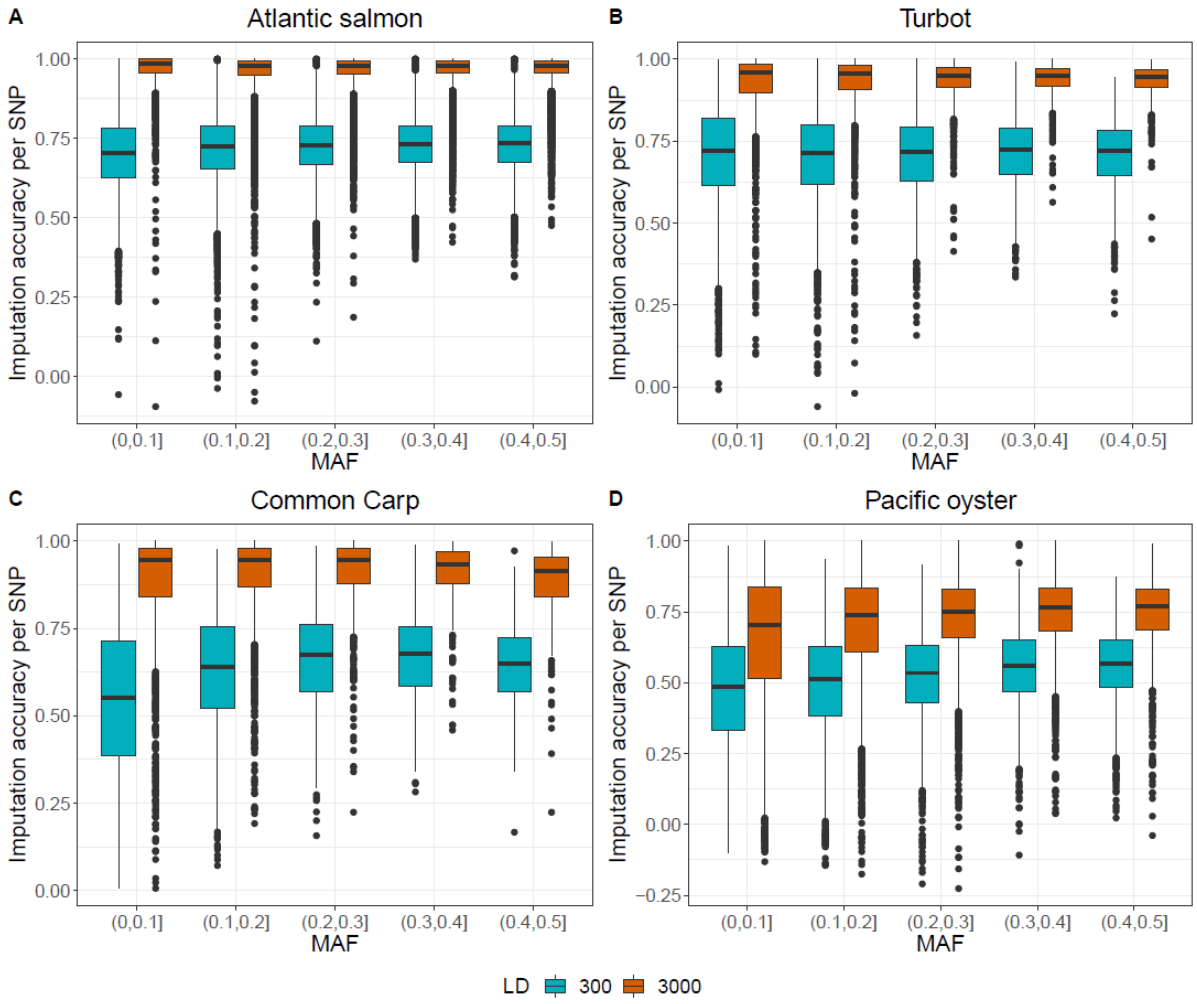


Figure 4 Correlation between the original and the imputed genotypes for each SNP plotted against MAF, for the two LD panels of 300 and 3,000 SNPs. Genotypes of the Atlantic salmon (A), turbot (B), common carp (C) and Pacific oyster (D) dataset were imputed with FImpute v.3.

Figure 4 shows the effect of MAF on imputation accuracy using FImpute v.3. The density of the LD panel did not seem to have a MAF-dependant impact on the imputation accuracy. However, there is a wider distribution of imputation accuracy values in the (0-0.1] MAF bin compared to the other bins, suggesting that there were more SNPs with very low MAF that were poorly imputed.

3.3 Genomic prediction using imputed SNP panels

The HD panel was used to estimate the genomic heritability and obtain genomic prediction accuracies for each species (**Table 2**), which were compared to those obtained using the LD panels (**Figure 5)**. Prediction accuracies were estimated for the LD panels with and without imputation. For Atlantic salmon, turbot and common carp, genomic prediction using the LD and the imputed panels gave comparable accuracies, which were very close to the accuracies obtained with the HD panel (**Figures 5A-C**). However, in the Pacific oyster, all the LD panels (300 to 6,000 SNPs) outperformed the imputed panels (**Figure 5D**), reaching maximal prediction accuracy when the LD panel consisted of 2,000 SNPs.

Since these results were unexpected according to previous reports, which showed that the accuracy of genomic prediction post imputation was higher than using the LD panels, we wanted to further investigate whether the SNP selection methods were responsible for the high prediction accuracy of the LD panels without imputation. Therefore, we randomly sampled SNPs throughout the genome to generate LD panels and perform imputation to compare their prediction accuracy with the other SNP selection methods. **Figure 6** shows the prediction accuracy of four LD panels (300, 500, 700 and 1000 SNPs) with and without imputation. For all four species, the prediction accuracy of the randomly designed LD SNP panels was considerably lower than the accuracy achieved with the HD SNP panel. Imputation of these LD SNP panels improved the predictive ability for Atlantic salmon, turbot and common carp with accuracy values very close to the maximal. However, the imputation of the Pacific oyster's random LD SNP panel did not improve prediction accuracy. Both the randomly designed LD panel and the imputed one achieved similar results that were lower than the accuracy of the

HD panel (**Figure 6D**).

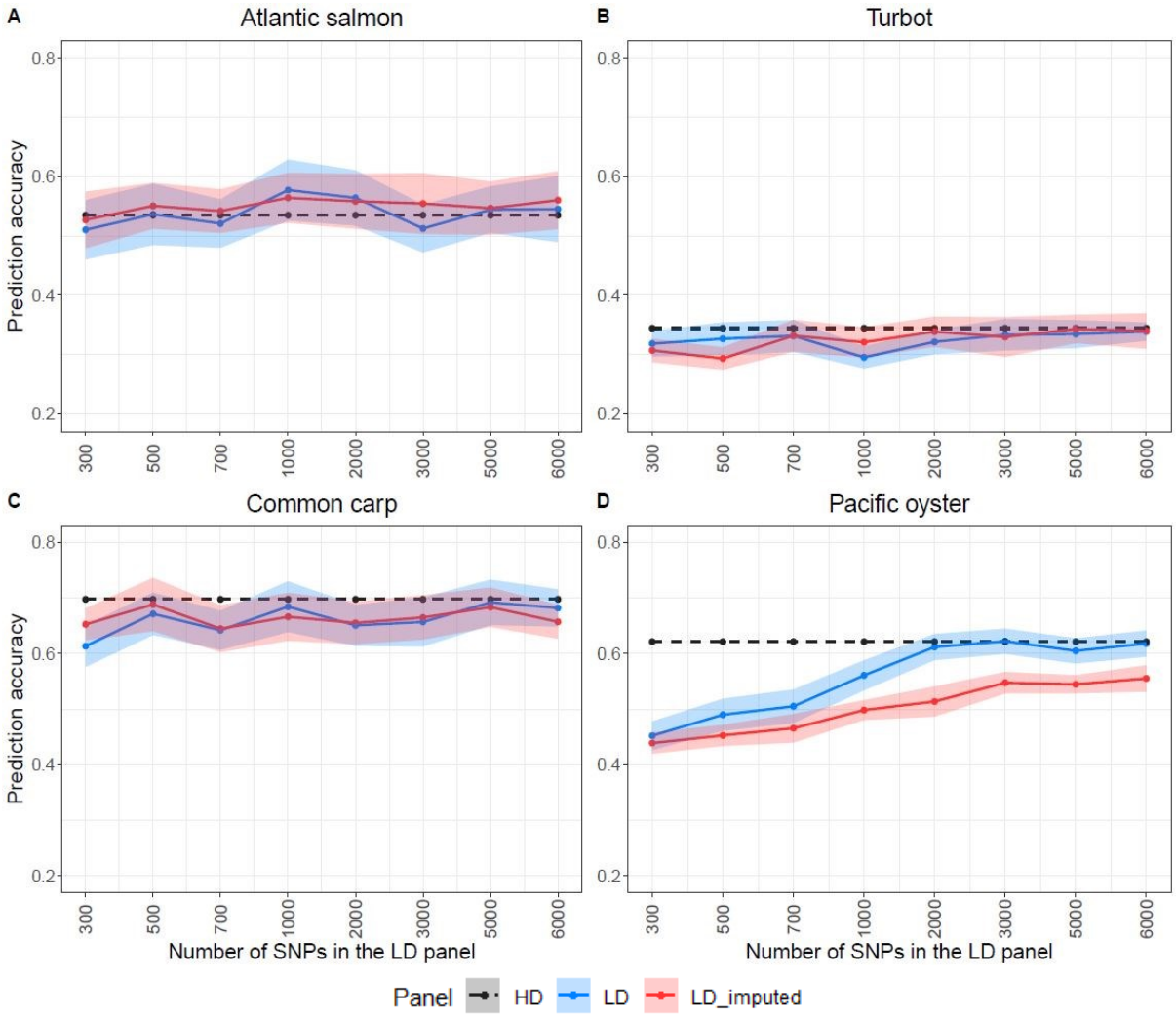


Figure 5 Prediction accuracies estimated for the high-density (HD), the low-density (LD) and the imputed LD panels (LD-imputed) for the four species. The LD panels were designed with the physical-distance-based method. The ribbons represent the standard deviations over 20 replicates of fivefold cross-validation analyses. The y-axis in these graphs ranges from 0.2 to 0.8 to facilitate the comparison between the LD and LD-imputed prediction accuracies. The Atlantic salmon LD panels (A) were imputed to 78,035 SNPs, the turbot (B) to 11,069 SNPs, the common carp (C) to 8,103 SNPs and the Pacific oyster (D) to 16,447 SNPs with FImpute v.3 software.

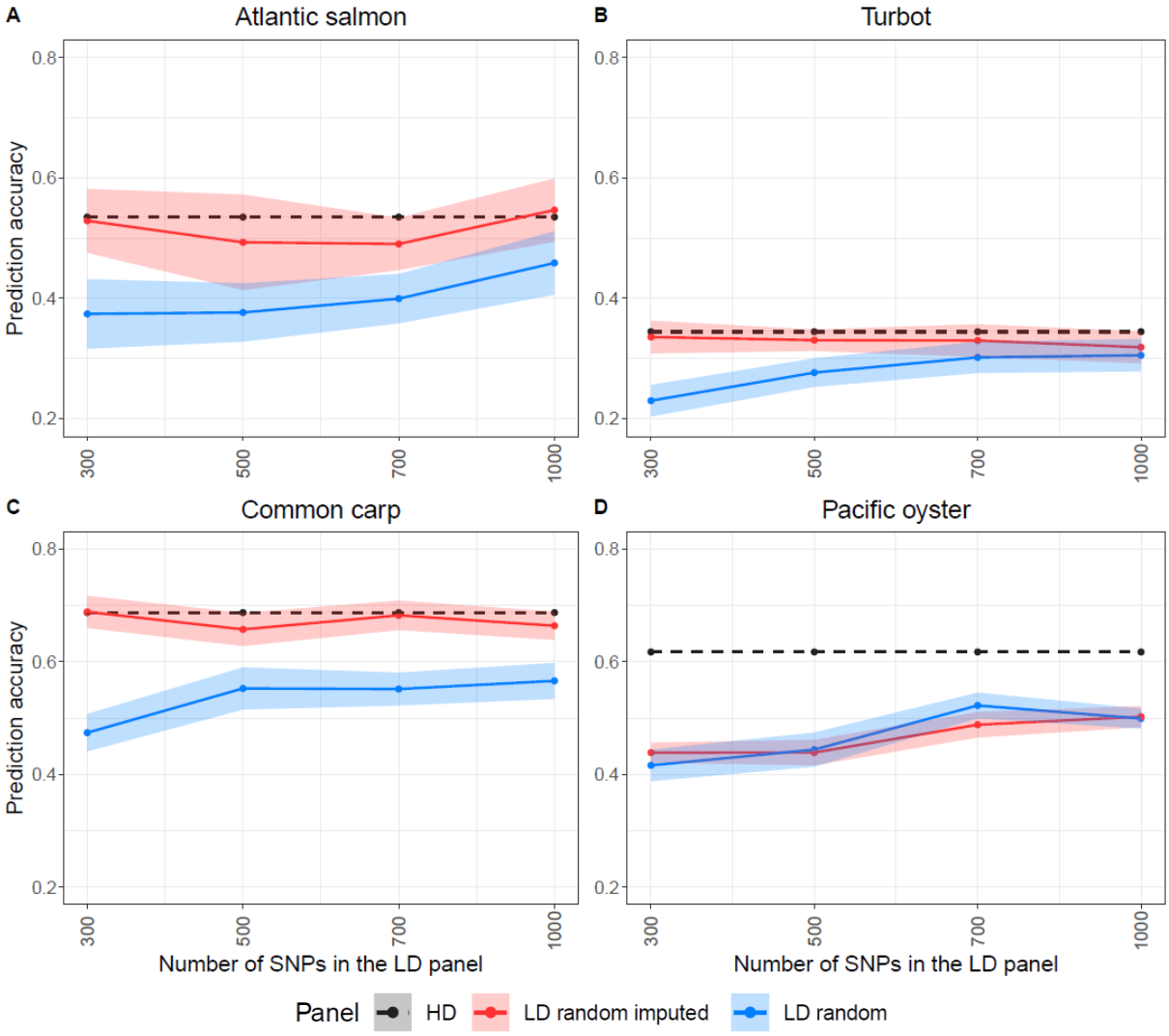


Figure 6 Prediction accuracies estimated for the high-density (HD), the low-density (LD random) and the imputed LD panels (LD random imputed), when SNPs were randomly selected for the four species. The y-axis in these graphs ranges from 0.2 to 0.8 to facilitate the comparison. The Atlantic salmon LD panels (A) were imputed to 78,035 SNPs, the turbot (B) to 11,069 SNPs, the common carp (C) to 8,103 SNPs and the Pacific oyster (D) to 16,447 SNPs with FImpute v.3 software.

.

4 Discussion

Genotype imputation is a powerful tool that has the potential to reduce the genotyping cost of genomic selection in aquaculture breeding programmes without a dramatic loss of prediction accuracy. In this study, we investigated some of the main factors affecting the accuracy of imputation and genomic prediction to contribute towards the establishment of best practices for the wider application of this method in the aquaculture sector.

4.1 Choice of imputation software

Three genotype imputation software were tested for their performance and compared between four populations of different aquatic species. All three software packages used a combination of population and pedigree-based imputation methods, and both parents' genotypes were present for all individuals in the datasets. The existence of pedigree information and close relatives in the dataset becomes more important as the number of markers in the LD panels decreases, as it becomes difficult to find the truly shared haplotypes between the reference and the target individuals.

FImpute showed the best performance across the four species in our study, with highest imputation accuracies for most LD panels and a shorter running time. FImpute shows extremely fast computational times when compared to other imputation software (e.g., Beagle, findhap, AlphaImpute, PHASEBOOK, Eagle-Minimac4 approach) for populations where pedigree information was available (Johnston, Kistemaker and Sullivan, 2011; Chud *et al.*, 2015; Ventura *et al.*, 2016; Wang *et al.*, 2016; Pausch *et al.*, 2017; Ye *et al.*, 2018; Fernandes Júnior *et al.*, 2021). Compared to AlphaImpute2, which uses a probabilistic algorithm (Whalen and Hickey, 2020), FImpute and findhap are faster in speed because they directly search for haplotypes in descending size and frequency order (Vanraden *et al.*, 2011). FImpute is also known to infer rare alleles with higher accuracy (Ma *et al.*, 2013; Wang *et al.*, 2016; Fernandes Júnior *et al.*, 2021) because the process starts by effectively matching long haplotypes between closely related individuals (Sargolzaei, Chesnais and Schenkel, 2014). This is pertinent because in a population with closely related individuals, the long haplotypes shared between

them usually carry rare alleles (Kamatani *et al.*, 2004) which can be frequent in families with a common ancestor who had the variant (Liu *et al.*, 2019).

4.2 Composition of the low-density panels

The number of SNPs in the LD panel and the linkage disequilibrium between adjacent SNPs was found to substantially affect imputation accuracy; by increasing the number of SNPs in the LD panels, we observed an increase in imputation accuracy (**Figure 1**). As previously discussed by Sargolzaei *et al.* (2014) this is because it becomes more likely to find shorter haplotype segments shared between related individuals due to the improved crossover resolution (Sargolzaei, Chesnais and Schenkel, 2014). However, there was a number of SNPs in the LD panel above which imputation accuracy improved only slightly (**Figure 1**). For Atlantic salmon, turbot and Pacific oyster the number of SNPs to reach this plateau was between 2,000 and 3,000.

In Pacific oyster, imputation accuracy was lower for all the LD panels compared to the fish species. Previous studies have found that some Pacific oyster populations exhibit rapid decay of linkage disequilibrium (Gutierrez *et al.*, 2017; Zhong *et al.*, 2017). This means that recombination between markers at each generation is high and therefore higher SNP densities might be required to achieve the same imputation accuracy results achieved in the other species. Additionally, the oyster genome, and in general bivalves' genomes, is highly polymorphic. Studies have shown that the Pacific oyster genome exhibits high levels of heterozygosity and is abundant in repetitive sequences, with some active transposable elements shaping this genomic variation (Zhang *et al.*, 2012; Hedgecock *et al.*, 2015; Gutierrez *et al.*, 2017). These highly polymorphic regions hinder the construction of the genome assembly (Gutierrez *et al.*, 2020) and can lead to a pronounced decrease in imputation accuracy (Fernandes Júnior *et al.*, 2021), possibly due to errors in marker order. Other characteristics of their genome that may be impairing mapping and consequently imputation accuracy are the putative high rate of de novo mutations during meiosis or larval development, which contribute to unusual segregation patterns and deviations from Mendelian inheritance patterns (Hedgecock *et al.*, 2015; Soledad Peñaloza

Navarro, 2017). Imputation of bivalve genomes requires further research in different populations and species to discover which parameters can contribute towards the improvement of imputation accuracy and their resulting prediction accuracy.

Regarding chromosomal position, we observed a lower number of correctly imputed SNPs at the beginning and at the end of Atlantic salmon chromosome 1. However, this decreased imputation accuracy at chromosomal ends was not evident in all the species. The lower number of SNPs available in the HD panel for some species may have had an effect in our ability to discern drops in imputation accuracy in certain regions of the genome, recombination and linkage disequilibrium can also explain the differences in imputation accuracy. Poorly imputed SNPs can be found in chromosomal regions with high recombination rates (Hozé *et al.*, 2013), such as the beginning and the end of chromosomes in some species (Druet, Schrooten and de Roos, 2010; Ventura *et al.*, 2016), or in regions difficult to assemble, but it can also be related to patterns of linkage disequilibrium throughout the genome. For example, recombination hot spots make the precise reconstruction of haplotypes difficult; consequently, imputation accuracy is low in these regions (Yoshida *et al.*, 2018). Centromeres also tend to show low imputation accuracies because they are difficult to assemble, potentially leading to incorrect order of markers. If we exclude centromeres and telomeres, regions with high imputation errors can be related to the patterns of linkage disequilibrium throughout the genome. SNPs with incorrect positions on the genetic map or SNPs wrongly assigned to chromosomes are challenging to impute, because they are not in linkage disequilibrium with the neighbouring markers on the map (Druet, Schrooten and de Roos, 2010; Yoshida *et al.*, 2018). Overall, as the density of the LD panels increased, imputation accuracy at the extremes and throughout the chromosomes increased due to the increased resolution of recombination patterns (Yoshida *et al.*, 2018; Fernandes Júnior *et al.*, 2021).

4.3 Genomic prediction accuracy

Low-cost genomic selection is successful when the genotype data of LD

panels accurately capture the genetic variation among the training and prediction individuals, resulting in no or minor loss of prediction accuracy when compared to HD genotypes. In this study, we achieved highly accurate genomic breeding value estimates for SNP densities as low as 300 SNPs for the Atlantic salmon, turbot and common carp populations. Small numbers of markers were sufficient probably because the shared haplotypes and linkage blocks between the reference and target individuals are long (full and half-sibs of the test population present in the reference population), and therefore their effects can be captured even with a small number of markers. Further, the number of families in a standard aquaculture breeding programme is small (100-200 families). The small effective population size and the degree of relatedness between individuals can explain the good performance of extremely low-density SNP panels.

Other studies have shown that a small number of markers and imputation are sufficient for accurate genomic prediction. For example, Gorjanc et al. (2017) suggested that 200 SNPs (20 SNPs per chromosome for a 10 chromosome simulated genome of 20,000 SNPs in total) imputed to HD can result in prediction accuracies comparable to HD panels in plant populations with a structure similar to that of aquaculture populations. Delomas et al. (2023), in a simulation study in oysters, achieved nearly maximal accuracy of genomic estimated breeding values by using 250-500 LD panels imputed to 40,000 SNPs. However, we did not observe similar high prediction accuracy results in our study with the Pacific oyster population we tested. In a study in Atlantic salmon, imputed genotype data from a ~250 LD SNP panel achieved comparable genomic prediction accuracy results to the true genotype data in Tsai et al. (2017). Yoshida et al. (2018) studied a two-generation Atlantic salmon population and suggested a genotyping strategy which combines genotyping all the parents and 10% of offspring with a HD panel, while the rest of the progeny are genotyped with a 500 SNPs panel and imputed to HD to achieve identical genomic prediction accuracies as with the 50,000 SNP panel. In another Atlantic salmon study, genotyping offspring at the very LD of 200 SNPs and imputing them with FImpute 2.2 to their parents' medium-density panel (5,000 SNPs) achieved almost the same genomic prediction accuracy as the true medium-density panel (Tsairidou *et al.*, 2020). There is a general consensus that imputation leads to close to maximal prediction accuracy.

Our findings demonstrate that for three out of the four species tested, the accuracy of genomic prediction is heavily dependent on the choice of SNPs when using the LD panels without imputation. The selection of evenly distributed SNPs in the LD panels resulted in markedly higher prediction accuracies when compared to that obtained with randomly selected SNPs. Whilst evenly distributed SNPs did not benefit from imputation, since the accuracy was already similar to that obtained with HD panels, imputation significantly increased the accuracy of randomly selected LD panels, bringing it in line with HD genotypes. In conclusion, the choice of SNPs in the LD panel is crucial when they are used without imputation for genomic selection; however, if imputation is used the choice of SNPs in the panel is irrelevant. Considering that the LD panel would have to be designed specifically for the target population, and that its performance might decrease as the genetic makeup of the population changes with each generation of selection, imputation is an exceptional tool to ensure that near-maximum prediction accuracies are obtained in every scenario.

Imputation accuracy did not affect prediction accuracy in the three fish species tested, with imputation accuracies of 0.76 to 0.98 depending on the number of SNPs in the LD panel resulting in similar prediction accuracies. However, this is not true in oysters, where the prediction accuracy of the imputed LD panel was significantly lower than that achieved with the LD panel alone, even when the number of SNPs in the LD panel was increased to 6,000 (Figure 5D). In this dataset, imputation accuracy was lower compared to the other species (Fig. 1), which can probably explain why the LD panels outperformed the imputed panels. Because of the rapid decay of linkage disequilibrium in the Pacific oyster, breeding candidates require regular testing on close relatives to preserve high accuracy levels between generations in a breeding programme (Gutierrez *et al.*, 2020). Nonetheless, more studies in bivalve species are necessary to determine if this is a general phenomenon or rather specific to the dataset studied here.

4.4 Cost reduction by using LD panels and genotype imputation

A significant cost reduction can be achieved by sequencing the target population with a very low-density panel (300-500 SNPs), which should still

provide maximal prediction accuracy when combined with imputation to HD, using a reference population containing relatives of the target population. While using the LD panels alone could result in a further reduction of the cost of genomic selection, we consider that the potential risk is not worth it since the number of animals that have to be genotyped at HD for imputation is low (i.e., the number of animals in aquaculture broodstock populations is usually around 100). In any case, if we estimate the cost of HD genotyping at $15 and the cost of LD genotyping at $12, for a relatively small population of 5,000 animals, the use of LD panels would result in a reduction of the cost in the application of genomic selection of 20% ($75,000 vs $60,000). Considering that the cost of HD genotyping is usually higher for most aquaculture species and that most species require the use of genetic tools to reconstruct the pedigree, LD panels and imputation can play an important role in the incorporation of genomic selection into aquaculture breeding programmes worldwide.

5 Conclusions

In this study, we explored the use of LD panels and imputation to reduce the cost of genomic selection in aquaculture breeding programmes, exploring different imputation software and SNP selection methods. Imputation accuracies were very high for the three fish species tested, while the performance of imputation was markedly lower in our oyster dataset. FImpute v.3 was the fastest and most accurate imputation method in almost all scenarios tested. When the LD panels were used without imputation, LD panels with the SNPs evenly distributed across the chromosomes achieved prediction accuracies very similar to the HD panel in the three fish species, even with just 300 SNPs, while randomly selected LD panels resulted in markedly lower prediction accuracies. However, imputation significantly increased the prediction accuracy of the randomly selected LD panels, reaching values similar to those of the HD panel in the fish species. Our results indicate that genotyping cost for the implementation of genomic selection can be reduced by the use of LD panels or a combination of LD panels in combination with imputation. While the use of appropriately selected LD SNP panels would be more cost-effective, we suggest the use of imputation to eliminate

the risk from potential changes in performance of the LD panels. This manuscript will help facilitate the widespread adoption of genomic selection in commercial aquaculture, leading to increased production and stability.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Ethics statement

Ethical review and approval was not required for the animal study because this study uses datasets generated in previous studies, which had appropriate approvals.

Author contributions

DR, RH, IJ, and ML were responsible for the concept and design of this work. CK, ST, and CF performed the analyses. CK plotted the figures, CK, DR, and GG drafted the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by BBSRC Institute Strategic Funding Grants to the Roslin Institute (BBS/E/D/20002172, BBS/E/D/ 30002275, and BBS/E/D/10002070).

Acknowledgments

This work was supported by BBSRC Institute Strategic Funding Grants to the Roslin Institute (BBS/E/D/20002172, BBS/E/D/30002275, and BBS/E/D/10002070). This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/). The authors acknowledge the contribution of Alastair Hamilton and Hendrix Genetics for generation of the Atlantic salmon data; Christos Palaiokostas, Martin Kocour and Martin Prchal for generation of the carp data; Paulino Martinez, Miguel Hermida, Carmen Bouza, Andres Blanco, Francesco Maroso and Adrian Millan

for generation of the turbot data; and Alejandro P. Gutiérrez, Jane Symonds, Nick King, Konstanze Steiner and Tim P. Bean for generation of the oyster data, which was funded by Cawthron's MBIE-funded Cultured Shellfish Programme, CAWX1315.

Conflict of interest

Authors ML and IJ were employed by Xelect Ltd., and author RH was employed by Benchmark Genetics.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

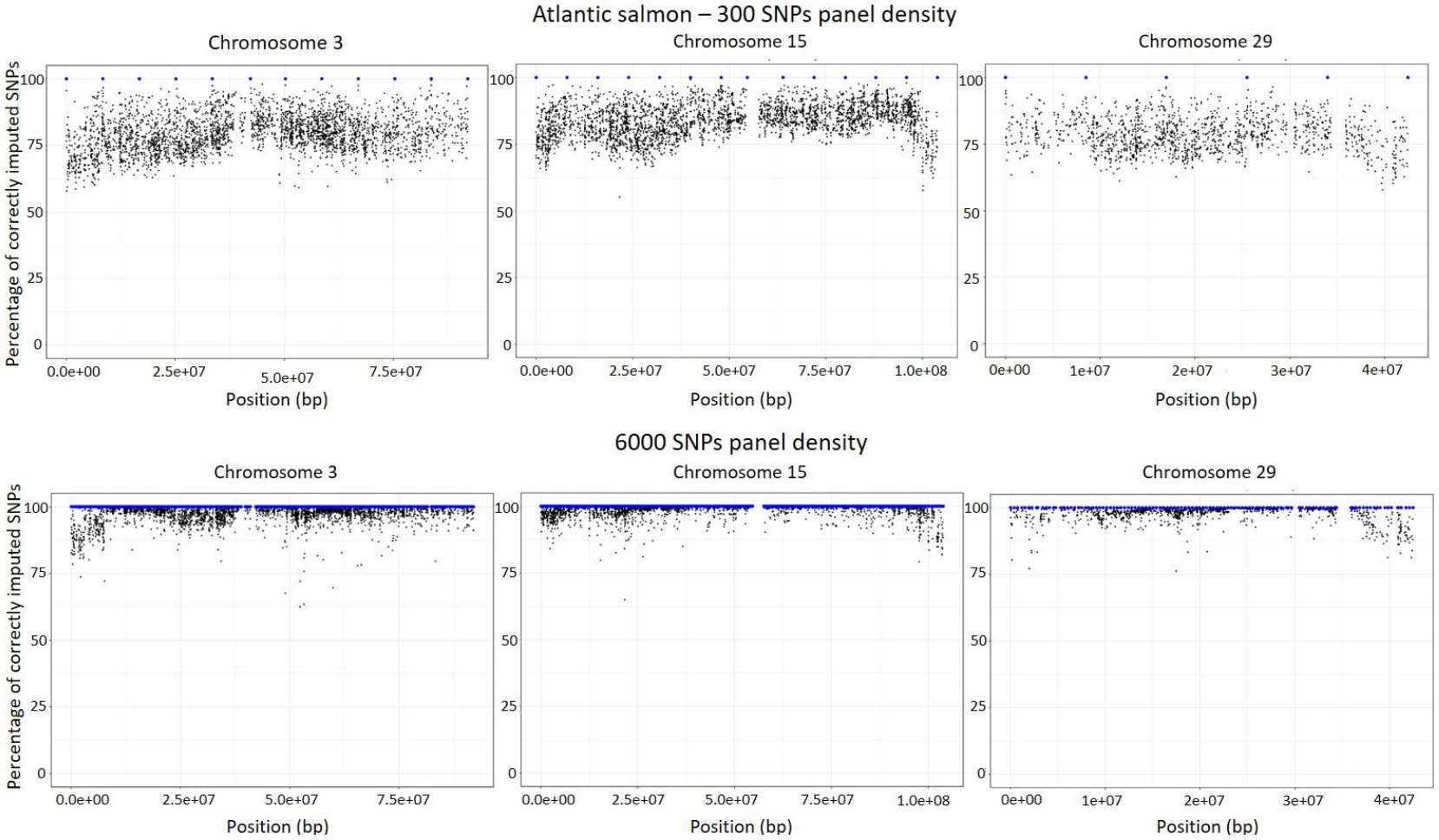
Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

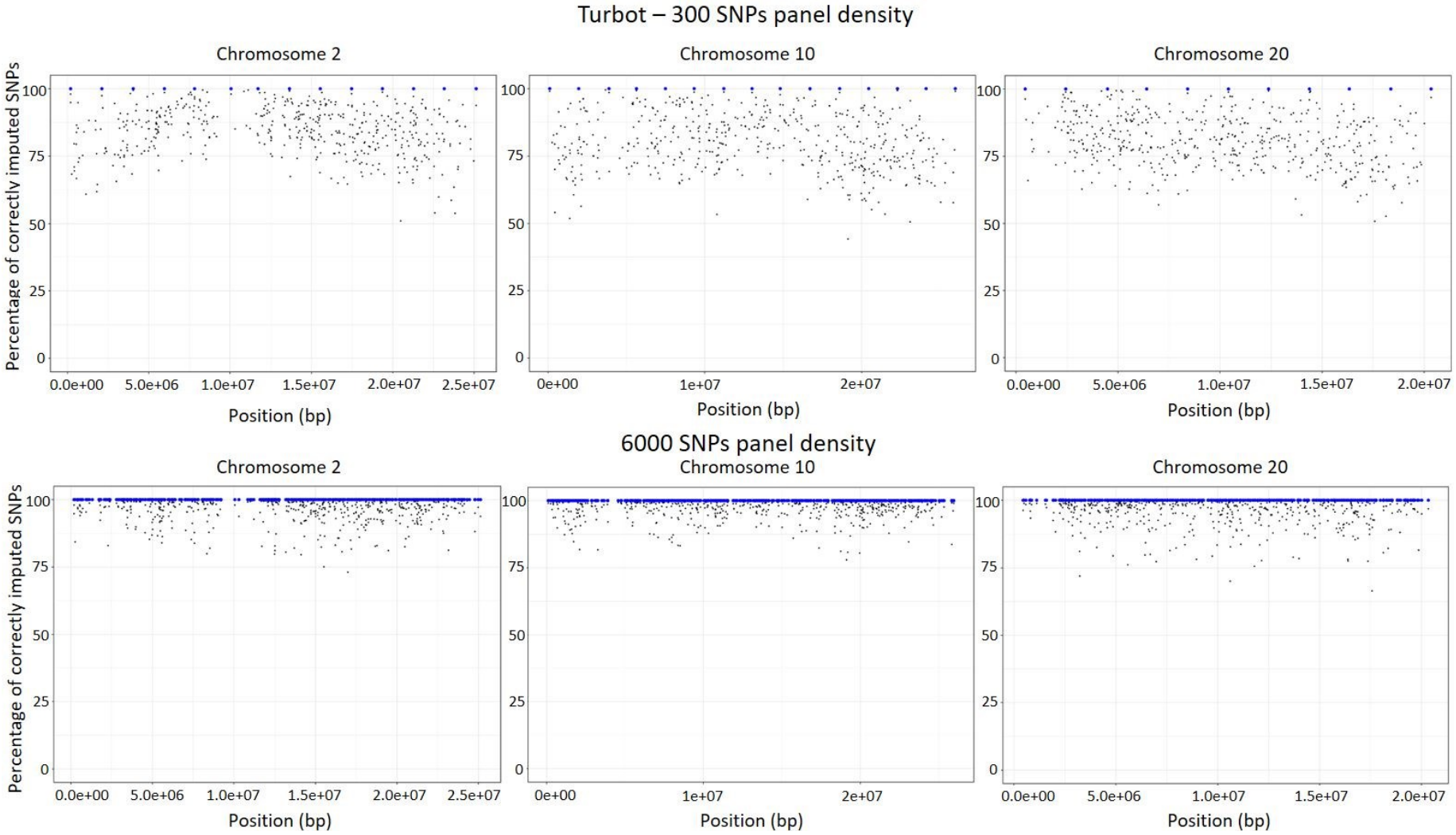
Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1194266/full#supplementary-material

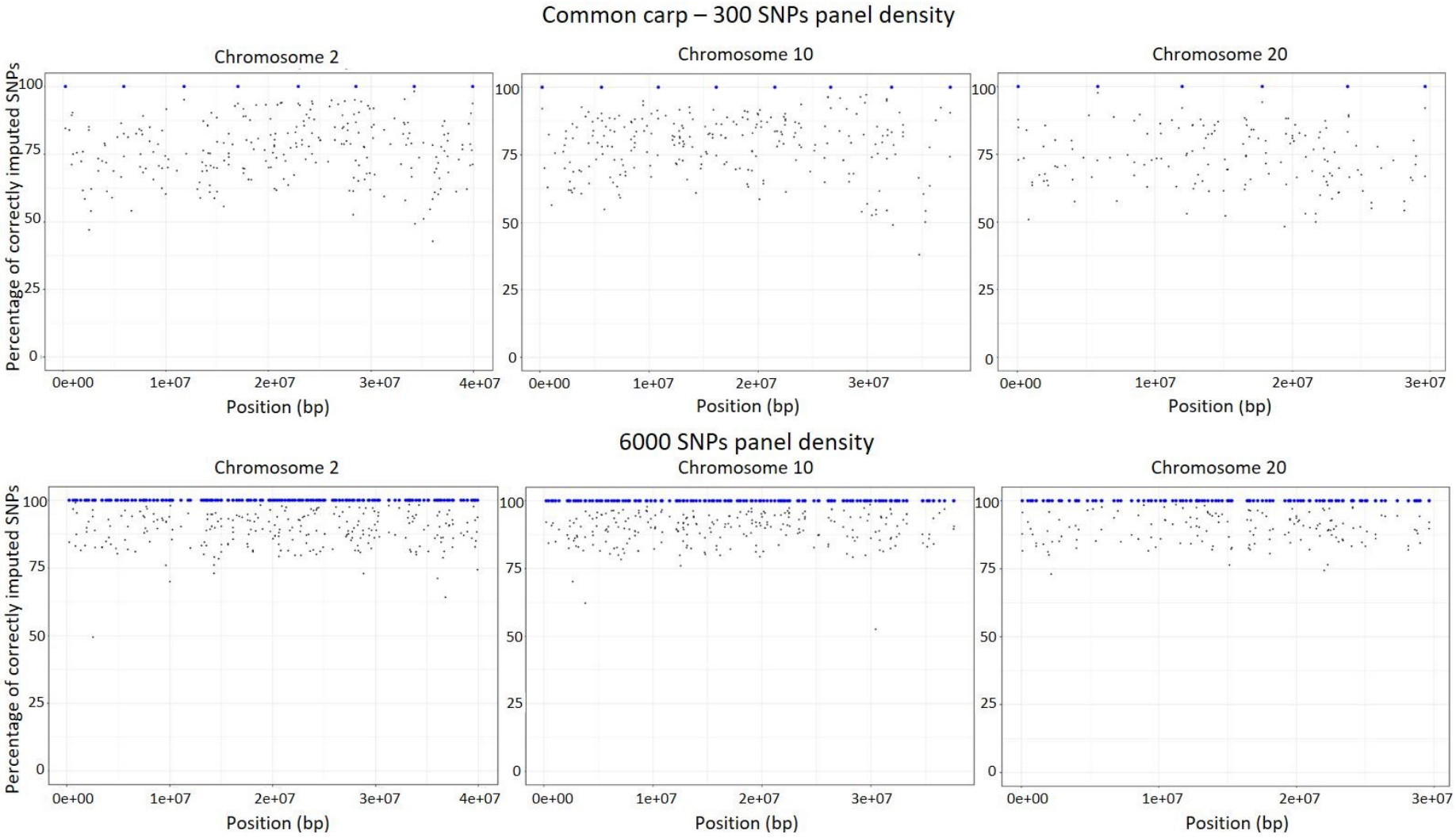
Supplementary material



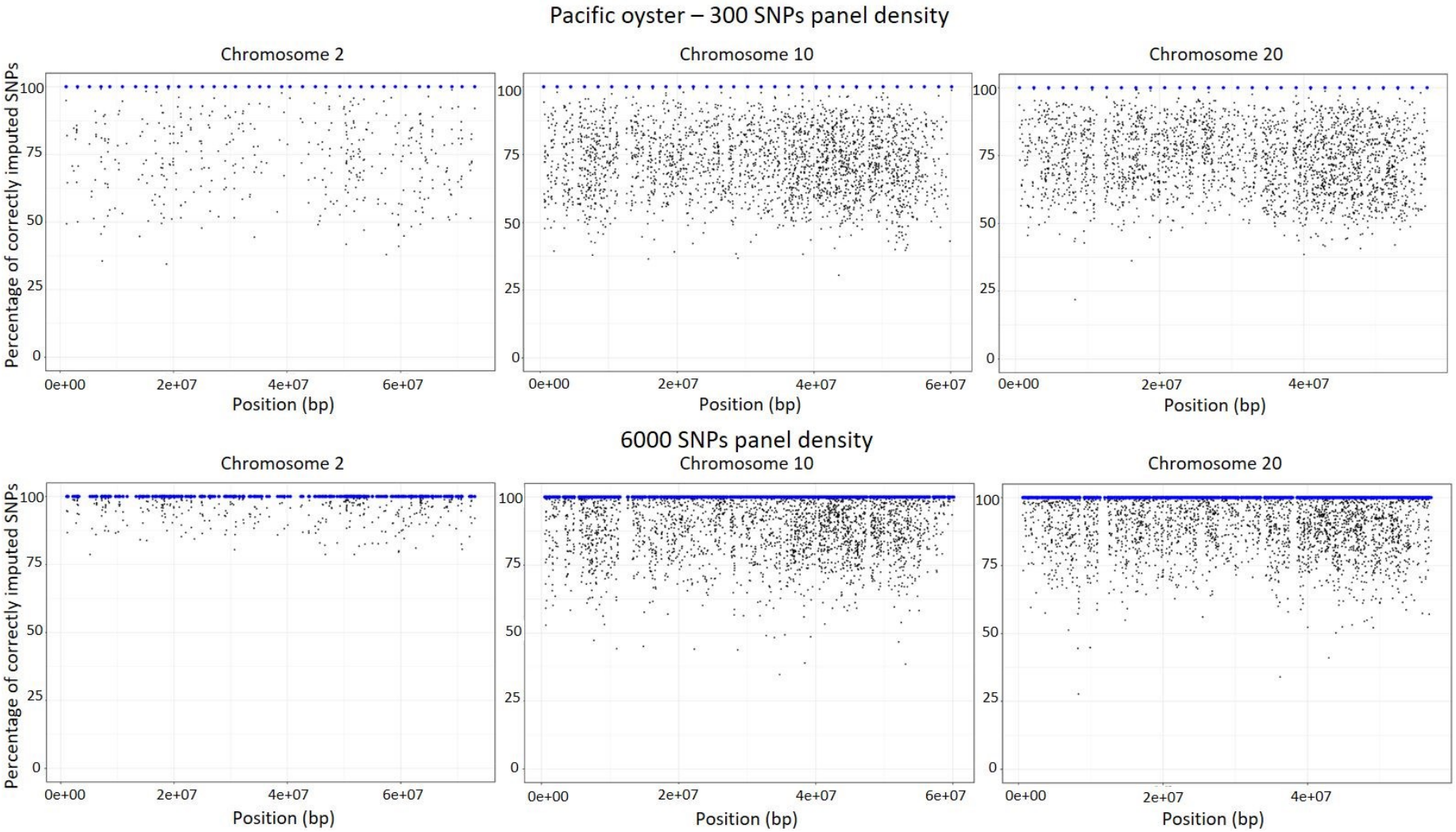
Supplementary figure 1. Percentage of correctly imputed genotypes with FImpute v.3 for each SNP of chromosome 3, 15 and 29 in Atlantic salmon dataset, using the LD panels of 300 and 6,000 SNPs (selected with the genetic-distance-based method). The blue dots indicate the physical position of the SNPs in the LD panel, whereas the black dots indicate the imputed SNPs.



Supplementary figure 2. Percentage of correctly imputed genotypes with FImpute v.3 for each SNP of chromosome 2, 10 and 20 in turbot dataset, using the LD panels of 300 and 6,000 SNPs (selected with the genetic-distance-based method). The blue dots indicate the physical position of the SNPs in the LD panel, whereas the black dots indicate the imputed SNPs.



Supplementary figure 3. Percentage of correctly imputed genotypes with FImpute v.3 for each SNP of chromosome 2, 10 and 20 in the common carp dataset, using the LD panels of 300 and 6,000 SNPs (selected with the genetic-distance-based method). The blue dots indicate the physical position of the SNPs in the LD panel, whereas the black dots indicate the imputed SNPs.



Supplementary figure 4. Percentage of correctly imputed genotypes with FImpute v.3 for each SNP of chromosome 2, 10 and 20 in the Pacific oyster dataset, using the LD panels of 300 and 6,000 SNPs (selected with the genetic-distance-based method). The blue dots indicate the physical position of the SNPs in the LD panel, whereas the black dots indicate the imputed SNPs.

Supplementary table 1. The datasets presented in this study were previously published and their availability status can be found in the articles mentioned below.

Species	Study with available dataset	DOI
Salmo salar	Tsai *et al*. 2015	https://doi.org/10.1186/s12864-015-2117-9
Scophthalmus maximus	Anacleto *et al*. 2019	https://doi.org/10.1038/s41598-019-40567-w
Cyprinus carpio	Palaiokostas *et al*. 2019	https://doi.org/10.3389/fgene.2019.00543
Crassostrea gigas	Gutierrez *et al*. 2020	https://doi.org/10.1111/age.12909

2.3 Concluding remarks

This chapter investigated factors influencing imputation accuracy across four aquaculture species to determine if utilizing low-density panels followed by imputation is an effective strategy for cost-effective genomic selection. We highlighted several parameters pivotal in this process. Firstly, the choice of imputation software, which can impact accuracy and computational efficiency. Secondly, the number of SNPs and the method of SNP selection for designing the low-density panels. Lastly, the unique genomic characteristics of each species. Once these parameters are optimized, the imputed panels must undergo validation to assess genomic prediction accuracy. Our results suggest that near-maximum genomic prediction accuracy can be achieved when all aforementioned parameters are considered and optimized for each species. The application of customized low-density panels in aquaculture breeding programmes is promising, given their cost-effectiveness—potentially offering around 20% savings compared to high-density panels —and their current utilization in parentage assignment. In the next chapters we will utilise genotype imputation as a tool to obtain whole genome sequences from low-coverage sequences and also to use the imputed WGS individuals for functional annotation enrichment of genomic selection models.

2.4 Clarification notes

Below are some clarification notes regarding this published chapter:

Regarding the naming of the genetic-distance-based method: This term is used to refer to genetic linkage. It is used as an extension of measurement of the linkage distance between two markers or genes on the chromosome. When the distance of two genetic markers is small (adjacent markers) there is a lower frequency of recombination and thus the correlation between them is higher. With this method we ended up keeping markers with low correlation (independent markers), meaning recombination between them is higher than with the physical-distance method. This can be interpreted as a maximisation of the centimorgans

separating two adjacent markers, and a centimorgan is a unit of measurement of the linkage distance between two markers.

In Figure 2.2 of this chapter, the colours assigned to the two SNP selection methods in the legend and the figure are inverted. Specifically, red should denote the physical distance-based method, while blue should denote the genetic distance-based method.

Lastly, in Section 3.2 of the results in this chapter, it is stated that “there is a visible pattern of slightly decreased imputation accuracy at the ends of the chromosomes of the four species”. To clarify, this is slightly visible for turbot and carp (Supplementary figure 2 and 3) but not for oyster and not for all chromosomes.

Chapter 3
Assessment of Low-Coverage Whole Genome Sequencing Imputation Performance as a Cost-Effective Alternative to Whole Genome Sequencing in Nile Tilapia (*Oreochromis niloticus*)

3.1 Introduction

The increase in aquaculture production witnessed across various species has been driven by the proliferation of public and private breeding programmes globally, a trend similarly observed in Nile tilapia (*Oreochromis niloticus*). With an approximate production of 4.4 million tonnes in 2020 (FAO, 2022), Nile tilapia ranks as the fourth most farmed fin fish species worldwide (Houston, Kriaridou and Robledo, 2022). Notably, the Genetically Improved Farmed Tilapia (GIFT), which is the pioneering strain developed for this species, serves as the cornerstone of numerous breeding programmes. Its significance has facilitated the recent development of linkage maps (Joshi *et al.*, 2018; Etherington *et al.*, 2022), reference genomes (Conte *et al.*, 2017; Etherington *et al.*, 2022), and SNP arrays (Joshi *et al.*, 2018; Peñaloza *et al.*, 2020; Yáñez *et al.*, 2020), achieved through collaborative efforts between the private and public sectors.

These genomic tools are continuously developing and improving, mainly due to the reduction in sequencing costs and the rapid advancements in sequencing technologies. SNP arrays are more affordable than whole-genome sequencing (WGS), but in most cases they are still unlikely to contain all causal mutations. The greater information content discovered through WGS, coupled with the absence of ascertainment bias[1] of the genotypes present at particular SNP arrays, gives an advantage to the use of WGS in genomic analyses. In particular, one of the major incentives for utilizing WGS data lie in enhancing genomic prediction accuracy across populations and generations (Hayes *et al.*, 2013; VanRaden *et al.*, 2017). This is achieved by directly incorporating causative

[1] In contrast to whole-genome re-sequencing data, arrays lack representation of a significant portion of globally rare variants and tend to exhibit bias towards variants found in the populations used during the array's development. This phenomenon influences population genetic estimations and is commonly referred to as SNP ascertainment bias. (Geibel *et al.*, 2021).

mutations affecting the trait of interest (Yoshida and Yáñez, 2022), thereby avoiding reliance on linkage disequilibrium between SNPs and mutations, which tends to decay rapidly with recombination over time. Additionally, WGS can be used for population structure genomic studies (Brække, 2023), and high resolution genome-wide association studies (GWAS). For instance, the identification of rare and/or novel variants has been linked to important improvements in association power and fine-mapping analyses (Höglund *et al.*, 2019; Uffelmann *et al.*, 2021; Chen, Coombes and Larson, 2022).

However, despite the steep reduction in sequencing prices, high-coverage sequencing remains expensive for genomic studies and even more for selective breeding programmes, which typically demand a large number of sequenced individuals. To address this challenge, a more cost-effective approach, low-coverage whole-genome sequencing (lcWGS), has emerged as an attractive option when combined with genotype imputation. This technology involves sequencing of the entire genome at a reduced depth, providing in most cases adequate coverage for haplotype reconstruction while significantly reducing the associated costs compared to standard WGS.

To date, several low-coverage imputation methods have been developed and several studies have assessed lcWGS imputation in humans (Pasaniuc *et al.*, 2012; Luo *et al.*, 2017; Spiliopoulou *et al.*, 2017; Hui *et al.*, 2020; Davies *et al.*, 2021; Rubinacci *et al.*, 2021, 2023; Chat *et al.*, 2022), cattle (Lamb, Nguyen, Briody, *et al.*, 2023; Lloret-Villas, Pausch and Leonard, 2023), as well as in other species such as donkeys (Zhao *et al.*, 2021), dogs (Buckley *et al.*, 2022; Wragg *et al.*, 2024), chicken (Li *et al.*, 2022), pigs (Nosková *et al.*, 2021), large yellow croaker (Zhang *et al.*, 2021), and recently one study in salmon (Gundappa *et al.*, 2023). However, there is a lack of studies evaluating the feasibility and effectiveness of adopting lcWGS instead of SNP arrays for other key species within the aquaculture industry.

In the present chapter we investigated the second objective of this PhD. For this purpose, a cohort of 166 Nile tilapia individuals from two consecutive generations (comprising 126 individuals from generation 15 and 40 individuals from generation 16) belonging to the GIFT breeding programme were subjected to high-coverage whole-genome sequencing (hcWGS). Subsequently, down-

sampling was conducted *in silico* to generate lcWGS datasets for individuals from generation 16. These lcWGS datasets were then imputed and assessed as an alternative to an open-access 65K SNP array previously developed based on the GIFT breeding nucleus population by our team (Peñaloza *et al.*, 2020). The objectives of this study were to i) evaluate the imputation accuracy of six *in silico* down-sampled datasets (0.1X, 0.2X, 0.5X, 1X, 2X, and 5X sequencing depth), ii) assess the impact of reducing the sequencing depth of the reference panel to 5X on imputation accuracy, and ii) evaluate the cost-benefit of employing a lcWGS approach for the target population combined with imputation as an alternative to WGS. These findings will offer valuable insights into the implications of utilizing lcWGS in genomic analyses such as genome-wide association studies (GWAS) and genomic selection (GS).

3.2 Materials and Methods

3.2.1 Nile tilapia population

The animals used in the current study belong to the Genetically Improved Farmed Tilapia (GIFT) strain from Malaysia, managed and owned by WorldFish, and selected for growth along 17 generations. In total 166 fish were sampled from that programme for the current study: 40 fish from generation 16, representing data from 10 full-sib families (4 fish per family), along with 126 fish from generation 15, which included both parents (dam and sire) for 6 of the 10 families of generation 16, and one parent of the remaining 4 families. Therefore, generation 16 had 24 individuals who had both parents and 16 individuals who had only one parent sequenced. The remaining individuals from generation 15 are related to some extend to generation 16, with a maximum relationship of 0.25 (uncles/aunts). Fin clips from all fish were sampled, placed in ethanol 95% and stored at -20°C until DNA extraction.

3.2.2 Nucleic acid extraction and DNA sequencing

Total DNA was extracted through a modified salt-based extraction protocol,

described by Aljanabi and Martinez, (1997). Modifications to this protocol as described in Taslima et al., (2016), were followed. Quality of the extracted DNA was assessed through an agarose gel electrophoresis and also by the 260/280 and 260/230 ratios on a NanoDrop 100 UV spectrophotometer. The concentration of the genomic DNA was measured using a Qubit dsDNA BR assay kit (Invitrogen, Life technologies).

All samples were sequenced in an Illumina Novaseq 6000 platform with paired-end sequencing at 150 base pairs (PE150). Raw sequencing data underwent quality control, trimming adapters and low-quality reads using Trim Galore wrapper tool (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with default parameters. The output reads were aligned to the GIFT reference genome (Etherington *et al.*, 2022) using BWA-MEM v.0.7.17 algorithm (Li and Durbin, 2009). Finally, the aligned reads were transformed to binary format, sorted and indexed with Samtools v.1.9 (Li and Durbin, 2009).

3.2.3 Average coverage and down-sampling of WGS data

The average coverage for each sample was estimated by assessing the alignment files using Picard (http://broadinstitute.github.io/picard/). For the target population (generation 16), an average coverage of 12.78X (ranging from 7.72X to 16.08X) was estimated as the high coverage used for imputation accuracy evaluation. Subsequently, this data was *in silico* down-sampled to 0.1X, 0.2X, 0.5X, 1X, 2X, and 5X (step 4a in **Figure 3.1**) with Samtools v.1.9. For the reference panel (generation 15), an average coverage of 26.19X (ranging from 21.99X to 33.25X) was estimated as the hcWGS and also down-sampled to 5X to test the impact of the coverage of the reference on imputation accuracy (1st step in **Figure 3.1)**. A statistical summary of the sequencing depth for each of the different datasets is presented in **Table 3.1**.

3.2.4 SNP calling and imputation analyses

The SNP discovery and imputation analyses were conducted on chromosomes 3, 8, and 17, which represent the largest, shortest and medium sized chromosomes across the *Oreochromis niloticus* reference genome, with

sizes 116.7, 31.4 and 41.9 Mb, respectively. These three chromosomes account for approximately 19% of the total genome.

BCFtools v.1.9 (Li, 2011) was used for SNP calling. Following SNP discovery, the SNPs obtained from the two panels generated for the 15th generation (26X and 5X sequencing depth datasets) were filtered to retain high-quality variants. The filtering thresholds were: mapping quality (MQ) >30, quality score (QUAL) >300, combined depth across samples (INFO/DP) <7000 and >1000 for the 26X and >200 for the 5X dataset. The next step of filtering with BCFtools removed multiallelic SNPs, indels and monomorphic SNPs. Finally, SNPs with more than 10% missing genotypes and less than 0.01 minor allele frequency (MAF) were removed. The variants in these panels were then used as reference for imputation.

For the imputation analyses we used GLIMPSE v.1 (Genotype Likelihoods Imputation and PhaSing mEthod) (Rubinacci *et al.*, 2021). The workflow used here was similar to the GLIMPSE tutorial available online (https://odelaneau.github.io/GLIMPSE/glimpse1/tutorial_b38.html). After filtering the reference dataset, we extracted the variable sites separately for each chromosome and for the two sequencing depths (26X and 5X) with BCFtools. The set of SNPs extracted from the two datasets were used as reference to perform imputation of the lcWGS target dataset (2nd step in **Figure 3.1**). A genetic map was created for each chromosome, assuming a ratio of 1Mb per centimorgan (cM), using a custom script. Following this, the genome was segmented into chunks using the "GLIMPSE_chunk" command (3rd step in **Figure 3.1**). This command considers both the level of missing data and the length of the defined regions, as longer regions can prolong running time while smaller regions may compromise accuracy. For this step a window size of 10 Mb and a buffer size of 200Kb was used to produce the chunks for the imputation and phasing step. Next, genotype likelihoods (GLs) were computed for all low-coverage target individuals and for all the SNP sites present in the reference panel with BCFtools “mpileup” and “call” command (step 5 in **Figure 3.1**). This is the input file format GLIMPSE requires for the next step of imputing the lcWGS target population to the reference panels. The next step of imputation was implemented with the “GLIMPSE_phase” command using i) the low-coverage GL files produced previously, ii) the reference

panel (26X or 5X), iii) the genetic map and iv) chromosome chunks (6th step in **Figure 3.1**). In this step the algorithm is iteratively refining the GLs of the lcWGS target individuals and outputs among other information the best guess genotypes and their genotype probabilities. After imputation the final step is to merge together the different chunks of chromosomes to produce one file per chromosome (7th step in **Figure 3.1**) and check imputation accuracy.

3.2.5 Imputation accuracy estimation

For the target population (16th generation), the 13X dataset was considered as the high-coverage dataset containing the true genotypes. Consequently, all lcWGS samples post-imputation were evaluated against this dataset. Furthermore, the influence of the sequencing depth in the reference population on imputation accuracy was examined. Therefore, the imputed lcWGS dataset was separately evaluated in the first scenario when imputed to the 26X and in the second scenario when imputed to the 5X reference population.

The imputed genotypes of the target population were extracted and filtered using a custom R script to calculate imputation accuracies. The imputed Variant Call Format (VCF) files (after the ligation of chunks, 7th step in **Figure 3.1)** for each sample were processed to extract posterior probability values for each genotype. The imputed genotypes were filtered at a 0.75 genotype posterior probability cut-off value (8th step in **Figure 3.1**) and were subsequently compared against the 13X true genotype calls to generate a matrix containing the number of genotypes that were imputed correctly and incorrectly (9th step in **Figure 3.1**). The output for each lcWGS dataset was used to calculate the percentage of correctly imputed genotypes. This percentage was determined by dividing the number of correctly imputed genotype calls by the sum of correct and incorrect genotype calls. The percentage of correctly imputed genotypes for the three chromosomes and the mean across chromosomes for the different depths were visualised using the ggplot2 package (Wickham, 2016) in R.

3.2.6 Cost effective analyses

To assess the economic advantages of employing lcWGS in candidate animals for subsequent genomic analyses (such as GWAS and/or genomic predictions) over SNP arrays or WGS, we conducted a cost-benefit analysis. Four strategies were considered; i) genotyping all animals (reference and target) with the 65K open access SNP array, ii) lcWGS the offspring and the parents at 5X sequencing depth, iii) lcWGS the offspring at 1X and the parents at 5X, and iv) lcWGS the offspring at 0.5X and the parents at 5X. A population size of 140 parents and 2,100 offspring was assumed, representing 70 full-sib families with 30 fish per family.

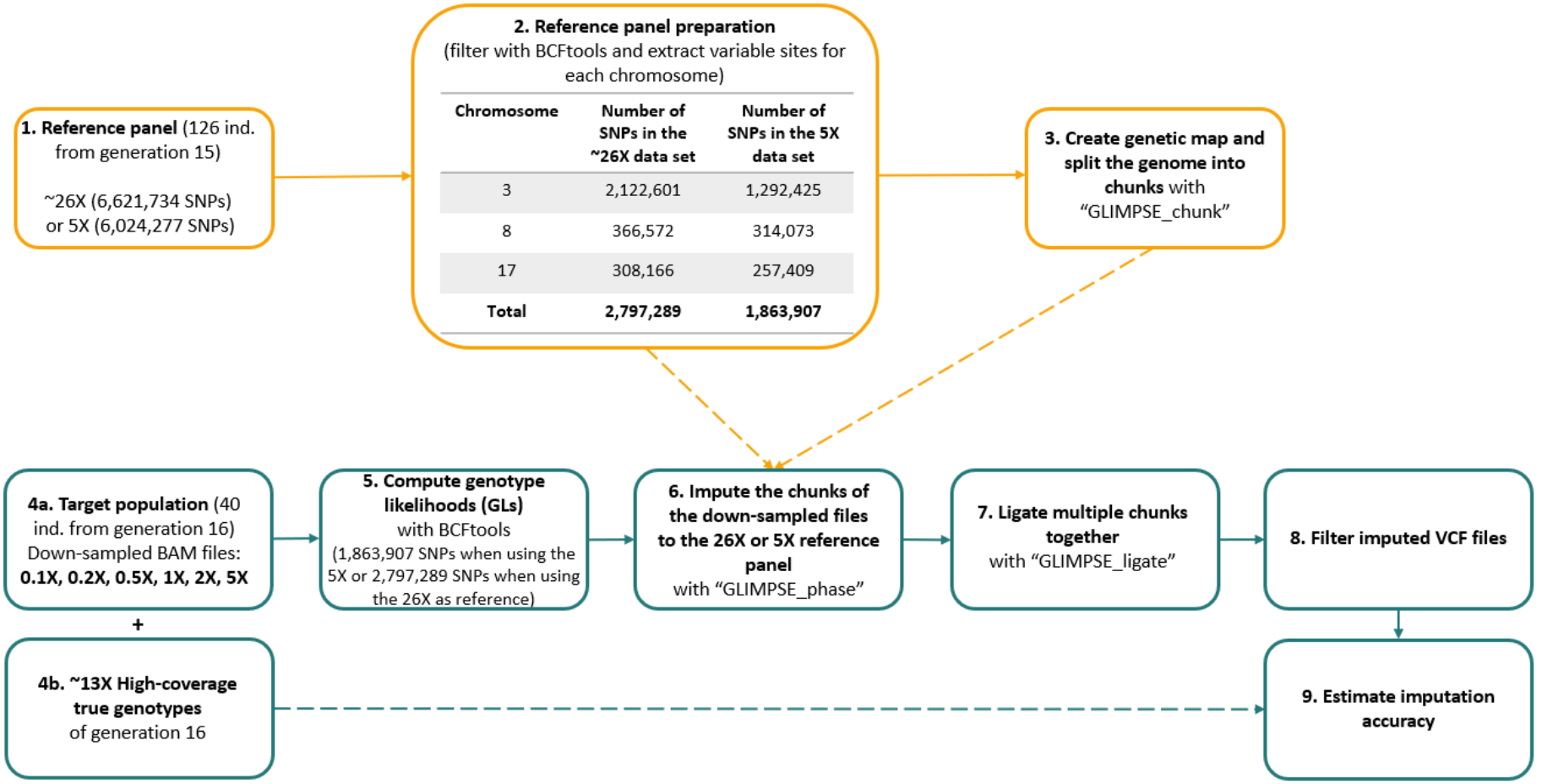


Figure 3.1 Genotype imputation workflow with GLIMPSE v.1. The down-sampled target population (16th generation) was imputed to the reference population (15th generation) to assess imputation accuracy. Imputation accuracy was measured against the true genotypes of the target population called using ~13X whole-genome sequencing depth.

3.3 Results

3.3.1 Data summary

A summary of the coverage of the different whole-genome sequencing datasets used in this study is presented in **Table 3.1** (including subsampled datasets). The initial number of SNPs in the reference panel (generation 15) before filtering was 6,621,739 in the 26X and 6,024,277 in the 5X scenario. After filtering, 2,797,289 SNPs remained in the 26X and 1,863,907 SNPs in the 5X panels.

Table 3.1 Summary statistics for each high and low coverage WGS across offspring and parents for a Nile tilapia (Oreochromis niloticus) breeding population. CV refers to the coefficient of variation (%) (calculated as sd/Mean × 100) and is used to provide an idea about the range or dispersion of the data.

Depth	Reference panel (generation 15)				
	Min	**Mean**	**sd**	**Max**	**CV**
26X	21.99	26.19	2.84	33.25	10.84
5X	4.19	5	0.54	12.7	10.80
Depth	**Target population (generation 16)**				
	Min	**Mean**	**sd**	**Max**	**CV**
13X	7.72	12.78	1.8	17.13	14.08
5X	3.02	5.00	0.70	6.71	14.00
2X	1.21	2.00	0.28	2.68	14.00
1X	0.61	1.00	0.14	1.34	14.00
0.5X	0.3	0.50	0.07	0.67	14.00
0.2X	0.12	0.20	0.03	0.27	15.00
0.1X	0.06	0.01	0.10	0.13	14.05

3.3.2 Number of SNPs retained post-imputation

After imputation, genotypes with a genotype posterior probability below 0.75 were considered not reliable and therefore removed. The remaining genotypes were compared against the “true” genotype calls (target population 13X coverage). This filtering threshold can be adjusted: lowering it will retain more genotypes but may decrease imputation accuracy, while increasing it will result in fewer genotypes remaining for subsequent analyses but more accurate results.

The total number of imputed SNPs retained after filtering is shown in **Figure 3.2** for each scenario. When imputing to the 26X reference panel, the number of SNPs obtained ranged from 2,177,476 SNPs for the 0.1X sequencing depth to 2,636,092 SNPs for the 5X sequencing depth of the target population, whereas when imputing to the 5X reference panel the number of SNPs retained for the same densities was lower, ranging from 1,367,969 to 1,784,481 for the 0.1X and 5X respectively.

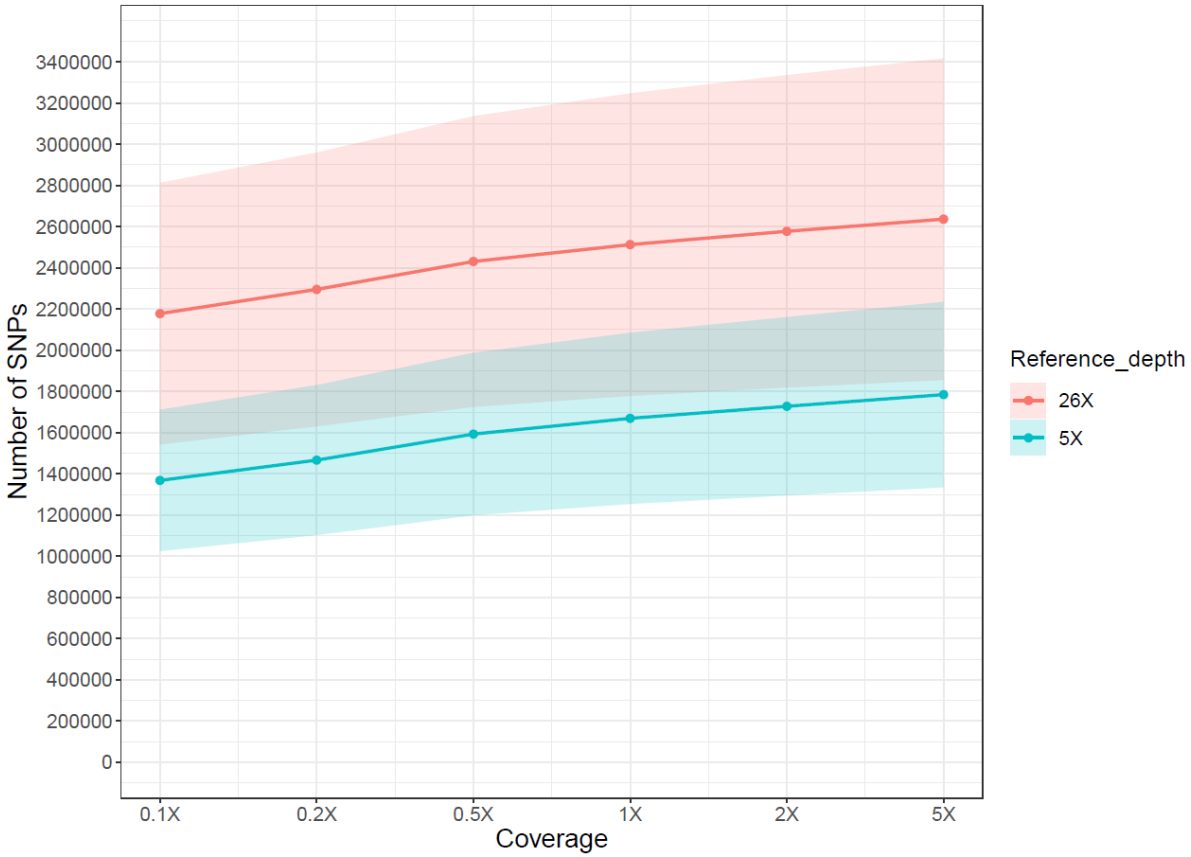


Figure 3.2 Total number of SNPs after imputation and filtering (0.75 genotype posterior probability). The red line depicts the SNPs retained after filtering the imputed down-sampled target data to the 26X reference panel and the blue line is for imputation to 5X. The ribbon represents the standard deviation of the number of SNPs between samples for the sum of the SNPs in the three chromosomes.

3.3.3 Imputation accuracy results

All scenarios tested had similar imputation accuracies ranging between 90.23% and 94.09% (**Figure 3.3**). Overall, imputation accuracy increased as the

mean sequencing depth of the target dataset increased. The mean imputation accuracy for the target population lcWGS datasets when imputing to the 26X reference dataset (red line in **Figure 3.3**) was higher compared to imputation to the 5X reference dataset (blue line in **Figure 3.3**) for all low-coverages, except when using the highest depth for the target population (5X), where imputation accuracies to the 5X was slightly higher. Nonetheless, the differences were marginal, with only 1-1.5% difference in imputation accuracy when imputing using the 26X or the 5X reference panels (**Table 3.2**). The target population coverage showed a similar pattern, with increasing imputation accuracies with increasing coverage, however the difference in imputation accuracy between 0.1X and 5X scenarios was less than 4%. When comparing imputation accuracy between the three chromosomes, we observed that chromosome 3 consistently had a considerably lower imputation accuracy than the other two chromosomes, and chromosome 8 showed a marginally lower imputation accuracy than chromosome 17 (**Figure 3.4**).

Table 3.2 Average imputation accuracy of the lcWGS datasets when imputing to 5X and 26X reference coverage.

Reference sequencing depth	Target sequencing depth	Mean imputation accuracy	Standard deviation
5X	0.1X	90.67	± 3.02
5X	0.2X	90.38	± 3.05
5X	0.5X	90.23	± 3.16
5X	1X	90.81	± 3.11
5X	2X	92.07	± 2.92
5X	5X	94.09	± 2.52
26X	0.1X	91.76	± 2.62
26X	0.2X	91.76	± 2.83
26X	0.5X	91.75	± 3.07
26X	1X	91.99	± 3.13
26X	2X	92.54	± 3.03
26X	5X	93.54	± 2.75

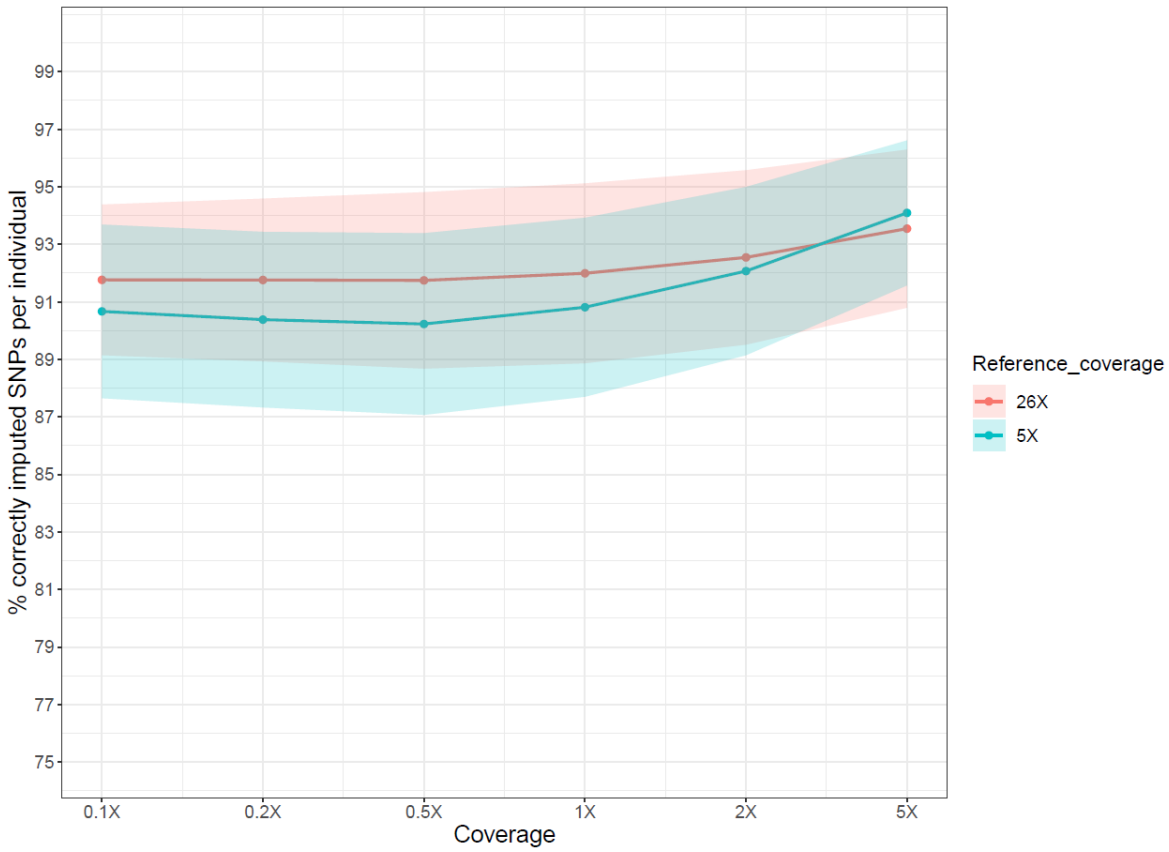


Figure 3.3. Comparison of the mean imputation accuracy for each sequencing depth when imputing the lcWGS datasets to the 26X and 5X reference panel.

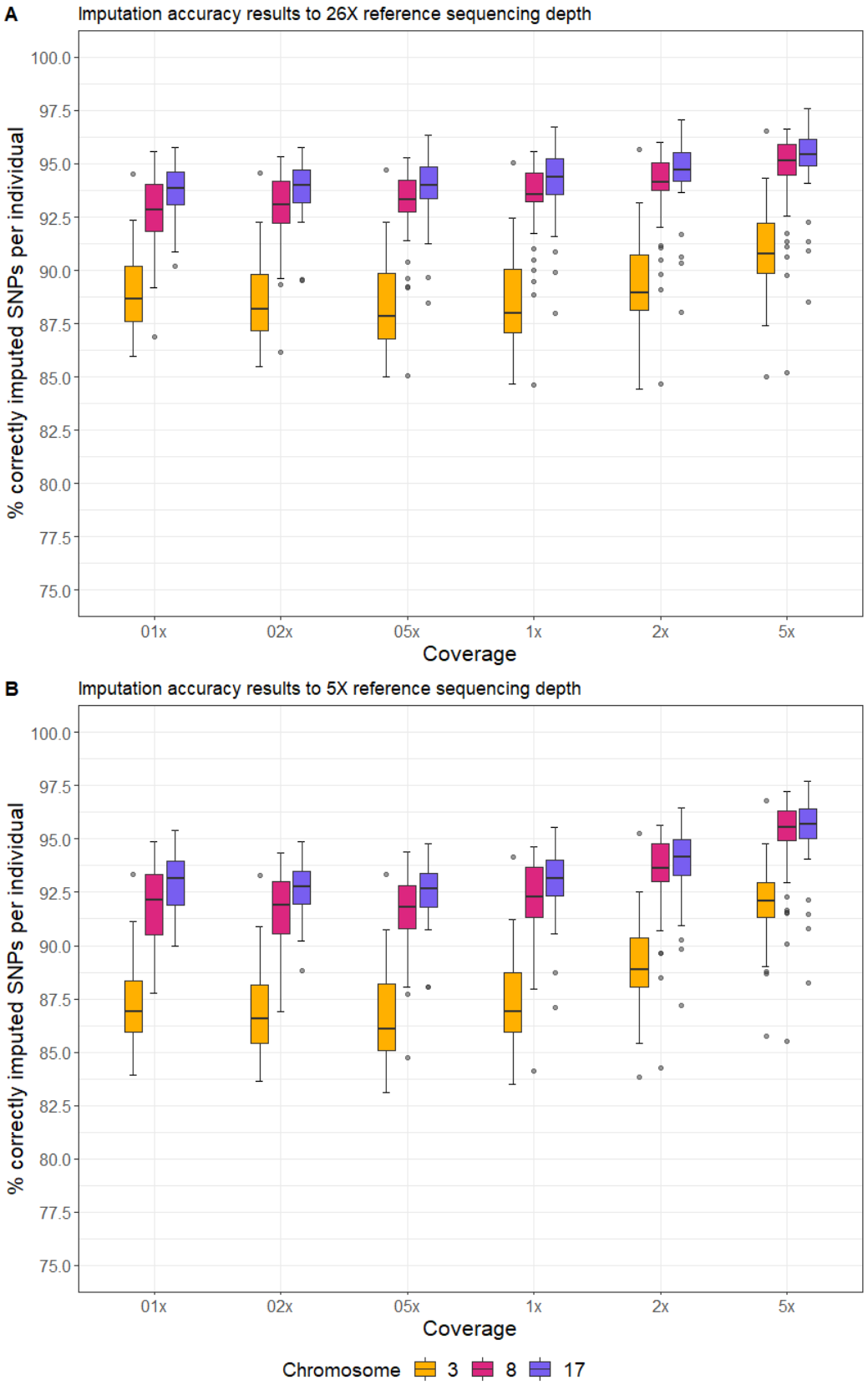


Figure 3.4. Boxplots with imputation accuracy results when imputing the lcWGS target population to the 26X (A) and 5X (B) hcWGS reference panel. Imputation accuracy is measured as percentage of correctly imputed SNPs per individual (y axis) for each of the three chromosomes across the different depths (x axis).

3.3.4 Imputation accuracy of heterozygous and homozygous sites

Imputation accuracy at homozygous sites was much higher than at heterozygous sites for coverages 0.1X, 0.2X, 0.5X, 1X and 2X, but not at 5X (**Figure 3.5** and **Figure 3.6**). Accuracy at heterozygous sites was particularly low for chromosome 3 compared to the other two chromosomes. The difference in imputation accuracy between homozygous and heterozygous sites progressively reduced with increasing sequencing depth in the target population. For the 2X coverage imputation accuracy was similar between heterozygous and homozygous sites for chromosomes 8 and 17 but still lower for chromosome 3. For the 5X coverage heterozygous sites had slightly higher accuracy than homozygous sites.

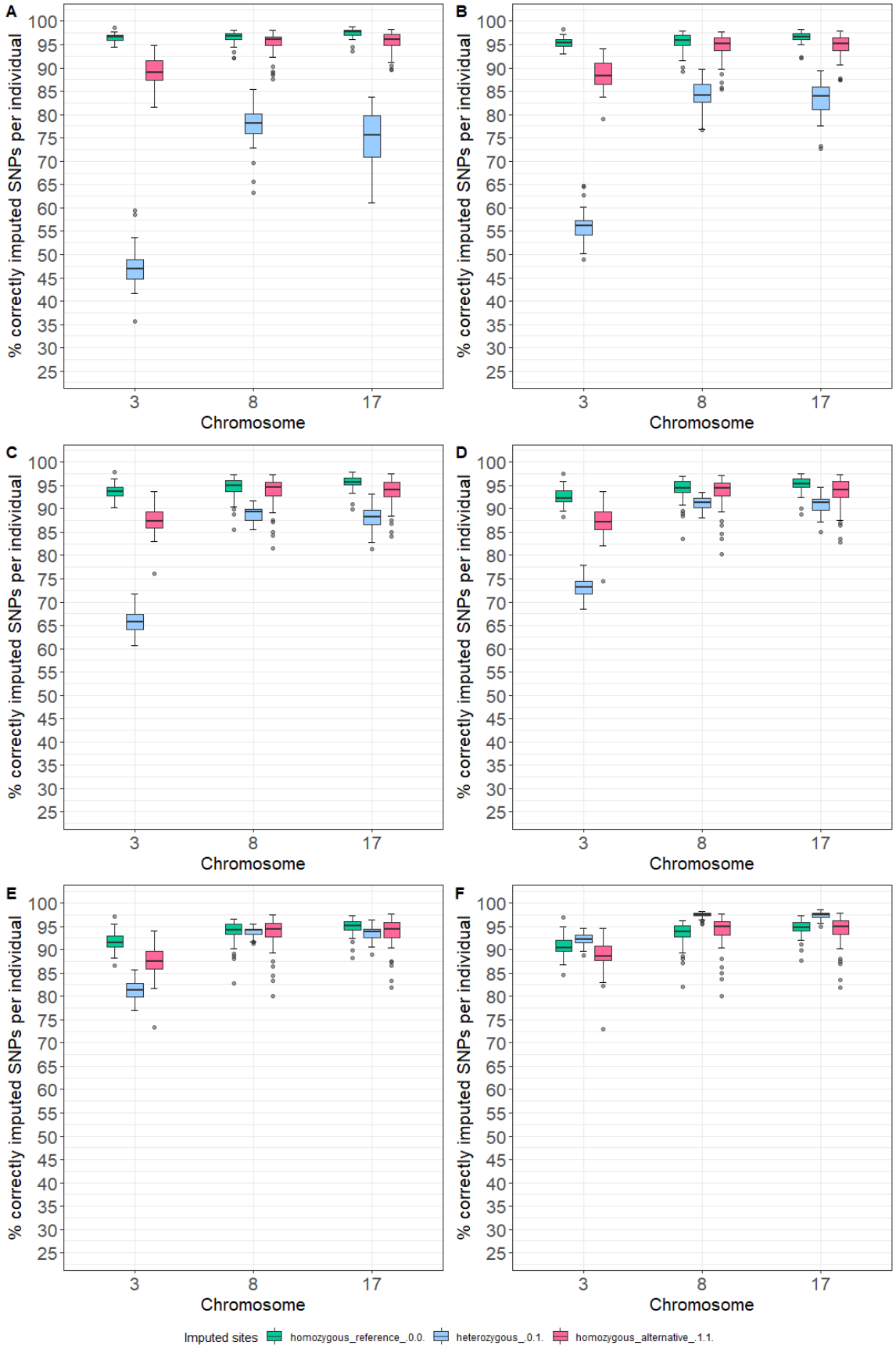


Figure 3.5. Imputation accuracy at homozygous and heterozygous sites of chromosome 3, 8 and 17 when imputing to the 26X reference panel. Figures A, B, C, D, E and F compare accuracy of 0.1X, 0.2X, 0.5X, 1X, 2X and 5X sequencing depths respectively.

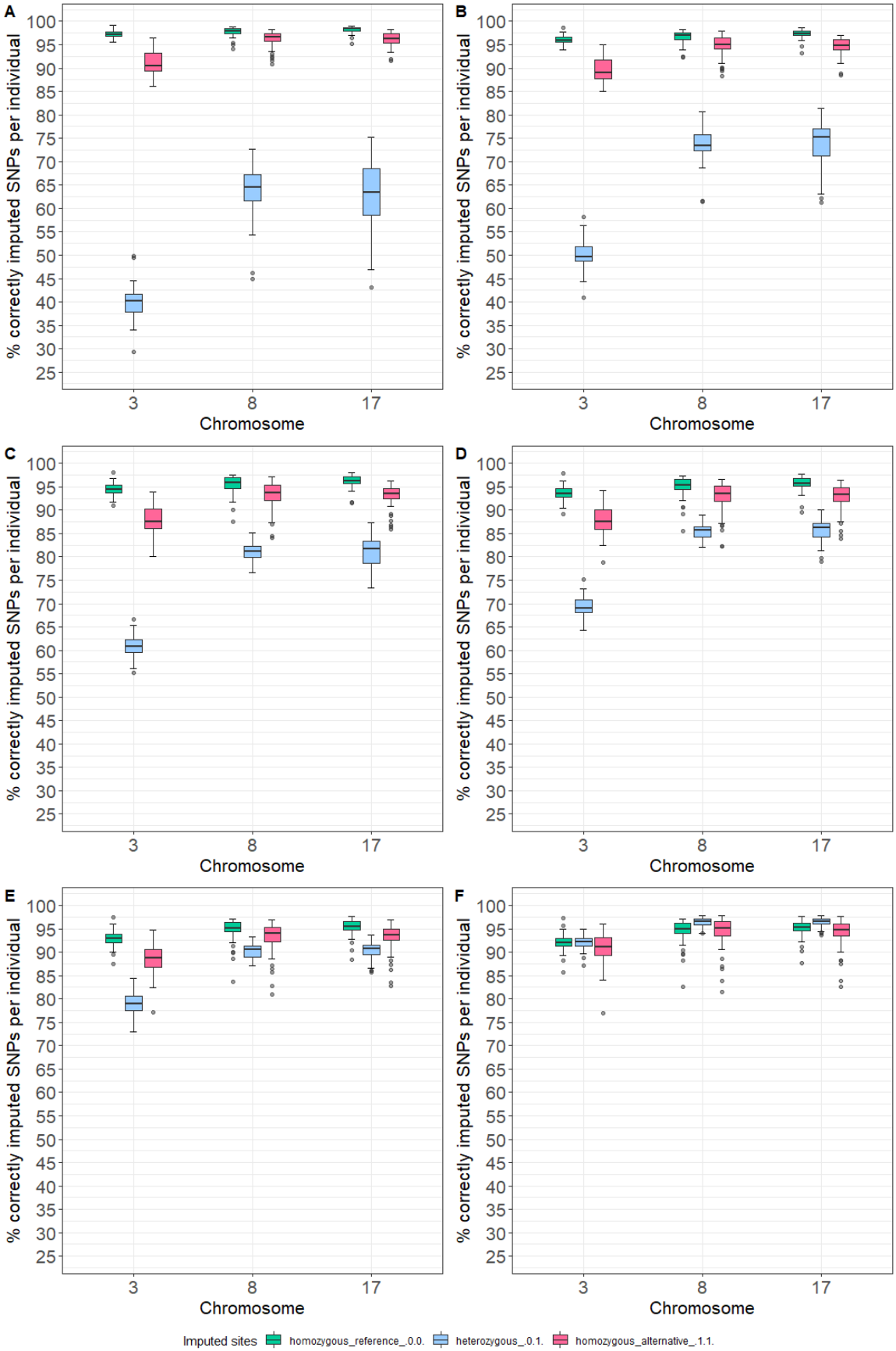


Figure 3.6. Imputation accuracy at homozygous and heterozygous sites of chromosome 3, 8 and 17 when imputing to the 5X reference panel. Figures A, B, C, D, E and F compare accuracy of 0.1X, 0.2X, 0.5X, 1X, 2X and 5X sequencing depths respectively.

3.3.5 Cost-benefit analyses

In **Table 3.3** we list the prices for the SNP array and lcWGS based on recent informal quotes from typical service providers[2]. While we acknowledge that these costs can vary considerably depending on the provider and volume of samples, we consider they are realistic and valid for a theoretical comparison. An approximate cost of £56,000 was estimated for the first strategy, where all parents and offspring are genotyped with a 65K array (£25 per sample). The second strategy, in which both parents and offspring are sequenced at 5X coverage, is the most expensive, totalling to £129,920 (£58 per sample). The third strategy, which involves sequencing parents at 5X coverage and offspring at 1X coverage, is 6% more expensive than the first strategy. The cost for the final strategy, with parents sequenced at 5X coverage and offspring at 0.5X coverage, is estimated at £52,220, making it approximately 7% cheaper than the first strategy in which all individuals are genotyped with a 65K SNP array.

Table 3.3 Summary table of the costs of genotyping for the different scenarios.

Strategy	Approximate cost per individual	Total approximate cost for 140 parents and 2100 offspring
1. All ind. genotyped with a 65K SNP array	£25	£56,000
2. All ind. sequenced at 5X sequencing depth	£58	£129,920
3. Sequence parents with 5X and offspring with 1X	£58 for 5X £24.4 for 1X	£59,360
4. Sequence parents with 5X and offspring with 0.5X	£58 for 5X £21 for 0.5X	£52,220

3.4 Discussion and future prospects

Despite the increasing popularity of other genotyping technologies over the past 10 years, such as genotyping by sequencing (GBS) (Robledo, Palaiokostas,

[2] These prices include sample and library preparation costs.

et al., 2018), SNP arrays remain the current standard practice for genomic analyses in aquaculture species. In our study, we have investigated a new alternative approach, involving a combination of lcWGS and imputation. This strategy offers a cost-effective method for fine-mapping GWAS, population genetic studies, and potential improvement of genomic prediction accuracy across populations and generations in aquaculture species. The data generated from imputation of lcWGS contains variants unbiased to any specific population and thus new population-specific variants can be discovered and used to update the haplotype database of the reference panels (Martin *et al.*, 2021). Compared to the routinely used SNP arrays within the aquaculture sector, this approach has the potential to significantly increase the number of discovered SNPs, which may contribute to elucidating population specific traits. We assessed two scenarios of imputing different lcWGS to two densities of the reference panel. Our findings provide valuable insights into the potential of incorporating lcWGS into future aquaculture genetic studies.

3.4.1 Imputation accuracy

In the dataset used in this study we did not observe a large difference in imputation accuracy when using 26X or 5X coverage in the reference panels, and the differences between different coverages in the target population were relatively small. However, the 26X reference panel resulted in ~800,000 more SNPs per individual after imputation for all down-sampled target datasets. This difference is mainly due to the larger number of SNPs initially present in the 26X reference panel compared to the 5X panel. Additionally, it is influenced by post-imputation filtering, where genotypes below the 0.75 genotype posterior probability cut-off value were removed. A more stringent filtering criterion could result in more accurately imputed genotypes; however, it would also retain a smaller number of SNPs for subsequent analyses.

We assessed imputation accuracy in three chromosomes (largest, smallest and medium length) and observed that chromosome 3 exhibited the lowest accuracy compared to chromosomes 8 and 17, thereby lowering the mean accuracy calculated for all three chromosomes. Chromosome 3 is the largest

chromosome in *O. niloticus* (~3 times longer than the other chromosomes) and has been proposed to be the result of a fusion event (Chew *et al.*, 2002; Conte *et al.*, 2021), with a large part of it (>40Mbp) associated with sex determination (Conte *et al.*, 2019). The content of this chromosome is highly repetitive (Conte *et al.*, 2019, 2021). In fact, Conte et al. (2019) mention that 54.7% of this mega chromosome was annotated as repetitive compared to 37% of the assembly across the genome. The repetitive content of this chromosome is possibly making it difficult to assemble and could be one of the reasons that imputation accuracy is low.

By examining imputation accuracy of heterozygous and homozygous sites separately, it becomes evident that imputation accuracy is significantly reduced in heterozygous sites when coverage falls below 5X. This phenomenon has also been noted in other studies (Hui *et al.*, 2020; Triay *et al.*, 2024). Triay *et al*. (2024) elaborate on this issue in their study, explaining that there is a confounding effect when only one or very few sequencing reads are obtained for a heterozygous locus, as it is very likely that only one of the two possible alleles will be revealed. With two reads, there is a 50% chance that only one of the alleles is revealed, and the probability of missing the second allele is still 25% with three reads (i.e., 3X coverage). The level of heterozygosity varies between genomes, and therefore the overall imputation accuracy when using lcWGS can vary greatly depending on the population or species.

3.4.2 Cost-effective strategy for genomic analyses

Based on our results, lcWGS followed by imputation can be considered a cost-effective alternative to WGS for genomic studies involving a large number of animals, especially those interested in the genomic basis of complex traits. Despite that, lcWGS is still more expensive than the currently used SNP arrays. If we compare the prices, the cost of 5X whole-genome sequencing was more than two times the cost of genotyping with a 65K SNP array, and if we factor the extra requirements in the form of computing time and expertise, it is unlikely to be applied in aquaculture breeding programmes. Nonetheless lcWGS renders 50-100x more SNPs than SNP arrays. Although the saved cost seems negligible for

the fourth strategy of sequencing the parents with 5X coverage and the offspring with 0.5X coverage, we must consider the significant increase in discovered markers that might be useful for research studies. Considering the low cost associated with very sparse lcWGS (≤ 1X for offspring), it's possible to increase the sample size for genomic studies compared to using SNP arrays, thereby achieving a higher statistical power.

3.4.3 Other low-coverage imputation studies

Studies in other species report promising imputation accuracy results. In a study of yellow croaker by Zhang *et al.* (2021) they used STICH for imputation and showed imputation accuracy results (measured as Pearson correlation coefficient) of 0.82 for 0.5X, 0.92 for 1X, and 0.97 for the 4X coverage, when they applied a cut-off filtering at a 0.7 genotype posterior probability. However, for the depth of 0.1X, imputation accuracy dropped dramatically to 0.04. In the cassava root crop, Long *et al.* (2022) found that imputation accuracies achieved with 0.5X sequencing coverage using Oxford Nanopore long-read sequencing data and the practical haplotype graph (PHG) tool for imputation, were comparable to those achieved with 5X sequencing coverage (r^2>0.8). In addition, PHG was able to distinguish between the available homozygous and heterozygous genotypes at the low-coverage of 0.5X and adding additional sequence data did not increase the accuracy. In another study involving cattle conducted by Teng *et al.* (2022), various imputation software programmes were compared, testing diverse combinations of reference population sizes, sample sizes, and sequencing depths. With a reference size of 1,059 individuals, sample size exhibited minimal impact on GLIMPSE. To achieve an accuracy exceeding r^2=0.90, a sequencing depth greater than 0.2X was necessary for GLIMPSE. Two other notable software used for imputation in this study, GeneImp and QUILT, demonstrated relative robustness across varying sample sizes and sequencing depths. Across all sample sizes (100–800) and sequencing depths (0.1X to 1X), GeneImp yielded imputation accuracies ranging from 0.94 to 0.96, while QUILT provided imputation accuracies of ranging from 0.97 to 0.98. Finally, in another study with cattle (Lamb, Nguyen, Copley, *et al.*, 2023) imputation accuracy (measured as correctly imputed

genotypes divided by the total) at 0.5X sequencing depth using QUILT was 0.96, while using GLIMPSE, accuracy was 0.8.

In our study we did not test genomic prediction accuracy due to the lack of available phenotypic data. However, assessing this accuracy will be necessary to determine the feasibility of implementing this method in genomic selection. Other studies have shown that when WGS coverage is lower or equal to 1X, the use of genotype imputation can give similar genomic prediction accuracy to higher-coverage resequencing in pigs (Yang *et al.*, 2021), dairy cattle (Teng *et al.*, 2022) and yellow croaker (Zhang *et al.*, 2021). Zhang et al. (2021) compared the effect of different sequencing depths on the accuracy of genomic prediction and noted that the genomic prediction accuracy obtained using 0.5X was similar to that obtained using 8X. In another study by Lamb, Nguyen, Copley, et al. (2023), a comparison was made between the accuracy of imputed genomic estimated breeding values (GEBVs) using QUILT and GLIMPSE and GEBVs derived from the bovine SNP array. Results indicated no difference between GEBVs obtained from the 35K SNP array and Oxford Nanopore sequencing coverages as low as 0.1X for QUILT and 0.5X for GLIMPSE. Using QUILT, the correlations between genomic breeding values obtained from Oxford Nanopore sequencing and those from low-density SNP arrays surpassed 0.91, reaching numbers as high as 0.97 even with sequencing coverages as low as 0.1X, with a reference panel comprising 48 million SNPs. These studies illustrate that with an appropriate imputation strategy genomic prediction can be accurately performed.

3.4.4 Future prospects

The findings in this study suggest that lcWGS with sequencing depths as low as 0.5X or 1X could be sufficient to produce cost-comparable information to SNP arrays. However, caution should be exercised because the low imputation accuracy of heterozygous sites may affect subsequent analyses and methods that are sensitive to rare variant genotypes. Compared to SNP arrays, ultra-low-coverage sequencing data present the challenge of the uncertainty of the genotypes we obtain.

More studies should be conducted in the future using the whole-genome of

Oreochromis niloticus, but other aquaculture species as well to obtain a more complete picture of the possibilities of using lcWGS. Additionally, in order to use low-coverage panels for genomic selection, prediction accuracy of the imputed panels should be tested and compared to WGS and high-density SNP arrays.

Chapter 4
Incorporating Functional Annotation into Genomic Prediction: Impact on a Turbot (*Scophthalmus maximus*) Population Challenged with the Parasite *Philasterides dicentrarchi*

4.1 Introduction

Infectious diseases pose a significant threat to farmed animals, impacting their health, welfare, and production, consequently jeopardizing food security. Effective solutions are needed to reduce the spread of infectious diseases. Despite efforts with conventional control methods such as chemical treatments, antibiotics, vaccination or implementation of biosecurity measures, the persistence of infectious diseases remains a challenge. In recent years, there has been a growing interest in enhancing host responses to infectious pathogens to control disease. Genomic selection can prevent or reduce disease spread by selecting for increased survival in a population, offering an alternative or supplementary approach to conventional control methods (Pooley *et al.*, 2020; Hulst, de Jong and Bijma, 2021). While selecting for crude survival is sufficient to reduce the impact of infectious diseases, using refined phenotypes that take into account the epidemiology of the disease and incorporating prior functional information into prediction models can be more efficient, but is rarely studied in aquaculture.

To effectively select animals with the best performance for a complex trait, we need to characterize the genetic makeup of the trait and enhance the accuracy of genomic prediction, which are the two primary objectives of quantitative genetics. Genomic prediction accuracy can be improved by the use of causal variants (Meuwissen *et al.*, 2022). The identification of causal variants can help avoid the reduction in prediction accuracy due to recombination, which causes reduction in correlations between SNPs and causal variants. Knowledge of the causal variants can thus lead to more accurate estimation of marker effects and the ability to capture more of the genetic variance (Meuwissen *et al.*, 2022).

A typical approach for pinpointing causal variants associated with complex traits begins with conducting a genome-wide association study (GWAS). However, achieving precise mapping demands a denser marker coverage than what's typically used in genomic prediction. Ideally, sequencing genotypes would fulfil this need but not all individuals can be sequenced due to high cost, thus imputation of genotypes can help achieve this goal. With all sequence variants incorporated into the dataset, the analysis holds potential to uncover the causal variants. However, even with sequencing the dataset rarely includes all of the variants because they are either difficult to genotype, filtered out during quality control or poorly imputed. Success to identify causal mutations for complex traits in aquatic species is very limited and has generally been restricted to variants that have large effects or the identification of segments of the genome where the variant exists rather than the causal variant itself (Yáñez *et al.*, 2023b).

It is widely known that most disease traits are regulated by many genes with small effects (Fraslin *et al.*, 2020), and most of the causal variants are located in poorly annotated non-coding regions (Dunham *et al.*, 2012; Albert and Kruglyak, 2015; Pan *et al.*, 2021; Prowse-Wilkins *et al.*, 2021). Finding the causative gene/mutation underlying a quantitative trait locus (QTL) can be challenging when there is a number of SNPs in linkage disequilibrium present in a region. To address the issue of linkage disequilibrium, the annotation of the genome can be utilized. Improved functional annotation information from projects such as AQUA-FAANG (Advancing European Aquaculture by Genome Functional Annotation), where active regulatory elements have been described, can help identify the candidate causative gene/mutation. AQUA-FAANG results, namely functional annotation maps, are freely available and easily accessible through Ensemble (https://www.ensembl.org/index.html). These can be overlapped with genome-wide genetic markers to identify genetic variants within protein coding genes and non-coding regulatory elements that could directly impact phenotypes. This overlap would in theory enrich a set of genetic variants in causal variants that can contribute to discriminating causative mutations of all those others in linkage disequilibrium detected in GWAS, discovering the underlying genomic basis of variation in traits of interest. Utilizing sophisticated methods to leverage biological data becomes essential for effectively connecting genotypic information with

phenotypes (Allayee *et al.*, 2023); however, it is crucial to employ appropriate statistical models to link such information (Tonner, Pressman and Ross, 2022; Yao *et al.*, 2024).

In standard genomic selection (GBLUP), the genetic similarity between individuals determines how similar their performance will be, which is true under the infinitesimal model, in which every region of the genome influences the trait with equal weight (Barton, Etheridge and Véber, 2017). In the case of oligogenic traits or traits that are influenced by a small number of QTLs explaining a big proportion of the variation, similarities between the genomes of two individuals may not accurately capture similarities in performance, as all the variants in the genome will be given the same weight (Daetwyler *et al.*, 2010). Several Bayesian methods have been developed to optimise the performance of prediction models, where relevant markers can be prioritised and categorised according to their functional annotation (Mollandin *et al.*, 2022a).

In aquaculture, disease resistance traits are essential components of selective breeding programmes. These traits necessitate frequent disease challenges every generation, resulting in significant economic and welfare costs. The identification and incorporation of putative causative variants for complex traits into the model, can mitigate concerns about the breakdown of linkage disequilibrium between the identified (from GWAS) variants and the causative variants occurring across generations (Habier, Fernando and Dekkers, 2007; Ma *et al.*, 2019). As a result, the genetic distance between individuals, which is crucial for accurate across generation prediction, becomes of minimal importance because the breakdown of linkage disequilibrium between the marker and the responsible gene or regulatory region is no longer a concern (Habier, Fernando and Dekkers, 2007; MacLeod, Hayes and Goddard, 2014). The combination of functional annotation information with emerging methods developed to incorporate such information can lead to more accurate selection improving the portability of genetic data across several generations (Johnston *et al.*, 2024). This could reduce the need for disease challenges in commercial farms.

In addition to the incorporation of functional annotation information, the refinement of the phenotypes we use to measure disease resistance could also lead to increased genetic progress. A previously published study by Anacleto *et*

al. (2019), was specifically designed to disentangle the different epidemiological components of a turbot (*Scophthalmus maximus*) population response to infection with *Philasterides dicentrarchi*, a ciliate parasite that causes high mortality in farmed flatfish. Parasites are naturally present in every healthy ecosystem, typically causing minimal impact on the overall fitness of healthy fish, but they can become a serious problem under the stressful conditions when fish are reared in crowded captive environments, greatly increasing the rates of transmission between individuals (Lieke *et al.*, 2020). The disease caused by this ciliate is called scuticociliatosis and clinical signs include haemorrhagic skin ulcers, darkened skin, swimming behaviour alterations, exophthalmos, and/or abdominal distension as a result of accumulation of ascitic fluid in the body cavity (Iglesias *et al.*, 2001; Piazzon, Leiro and Lamas, 2013). This disease is responsible for very important economic losses not only for the turbot industry, but also for other aquaculture species.

In their study, Anacleto *et al.* (2019) designed a disease transmission experiment using an infection model to provide direct evidence for genetic variation in host infectivity, resistance and endurance. In our study, we used the restriction-site associated DNA sequencing (RAD-seq) genotypes of the turbot challenged population mentioned above and imputed them to whole-genome genotypes using their whole-genome sequenced parents. Subsequently, the imputed genotypes were used to test the impact of incorporation of putative functional variants on the genomic prediction accuracy. The objectives of this chapter were to i) estimate the genetic variance of the disease traits using the imputed genotypes with GBLUP and Bayesian models ii) calculate genomic prediction accuracy of the different models and iii) assess changes in genomic prediction accuracy when functional annotation is incorporated in the Bayesian models.

4.2 Materials and methods

4.2.1 Experimental design

The transmission experimental design to detect genetic differences in host resistance, infectivity and tolerance for one of the trials is summarized in **Figure 4.1** (from Anacleto *et al*., 2019). In brief, there were two consecutive trials with 72 tanks in total (36 tanks in each trial). Each tank consisted of five artificially inoculated fish (shedders) from one family, and 20 susceptible fish (recipients) from four different families. The recipient families were distributed in tanks, resulting in the creation of nine combinations. Each combination included one of each of the four shedder families (S1 to S4). Fish were inspected twice a day over the duration of the experiment for visual signs of infection and for mortality.

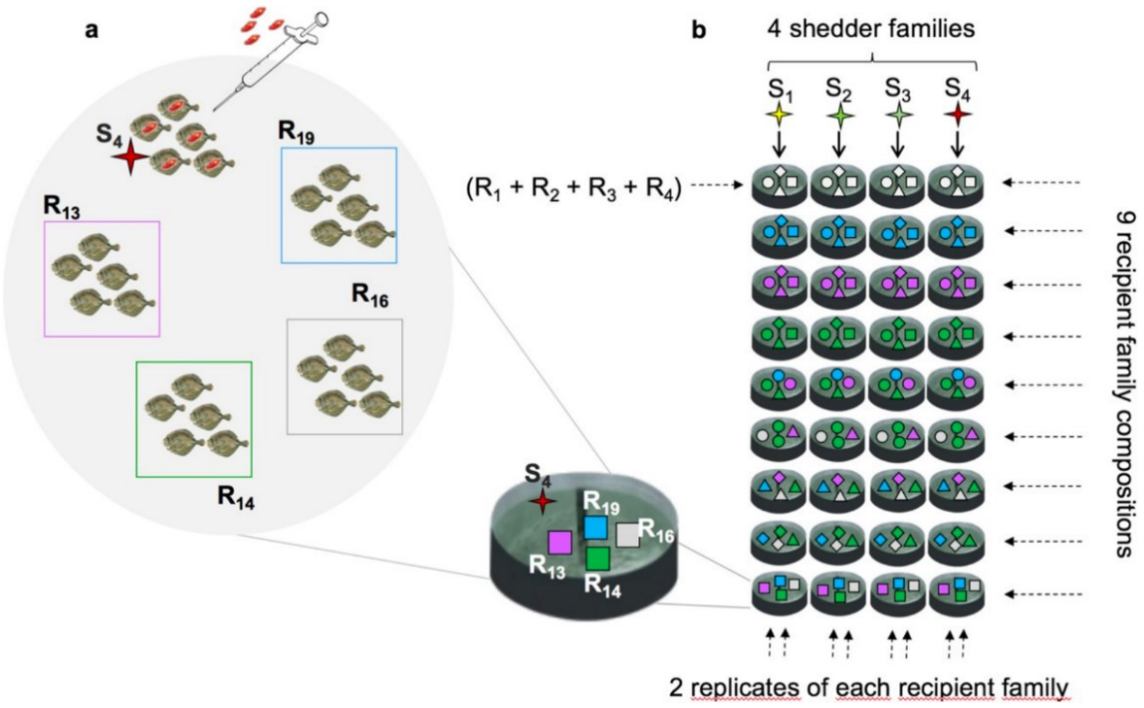


Figure 4.1 The figure illustrates the composition of the tanks used in the experimental design for studying the transmission of Philasterides dicentrarchi in turbot (from Anacleto et al. 2019). "S" refers to shedder and "R" to recipient fish.

4.2.2 Epidemiological traits

Three key epidemiological host traits were used for the analysis: susceptibility (propensity to acquire infection), infectivity (propensity to pass on infection to others), and recoverability or mortality post infection (time from

symptoms until death), which affects the duration of an infected animal to transmit the infection (Pooley *et al.*, 2020). A fourth trait named R0, which is a composite trait of the three above (R0 ~ susceptibility * infectivity * recoverability) was also evaluated. These four traits were inferred using the SIRE model (Pooley *et al.*, 2020). In addition to these inferred traits, we also used the “days to death” as a trait, after transformation to scale and centre for the two trials.

4.2.3 Genotypes and imputation to whole-genome sequence

The turbot dataset consisted of 1,445 animals from 39 full-sib families, challenged with *Philasterides dicentrarchi* using a semi-factorial design. This design involved donor and recipient families as depicted in **Figure 4.1** (Anacleto *et al*., 2019) that were genotyped using 2bRAD-seq resulting in the acquisition of 17,690 SNPs (Maroso *et al*., 2018).

For the current study, the 54 parents of the challenged fish were whole-genome sequenced using a Novaseq 6000 sequencing platform (150 bases of paired-end reads). Adapter sequences were trimmed and quality control of the raw paired-end read files was conducted using Trim Galore v.0.6.3 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). This step runs a paired-end validation on both files once the trimming is complete and removes entire read pairs if at least one of the two sequences became shorter than 20bp in length. The filtered reads were aligned to the new high-quality chromosome level genome assembly of turbot (RefSeq GCA_013347765.1) (Martínez *et al.*, 2021) with BWA-MEM v.0.7.17 using the default parameters. PCR duplicates and alignments with a minimum mapping quality of 10 were discarded from the subsequent analysis using Picard (http://broadinstitute.github.io/picard/).

Single nucleotide polymorphisms (SNPs) were called with BCFtools v.1.9, and genotypes were filtered in a two-step process using the same software. First, only SNPs with mapping quality (MQ) >40, quality score (QUAL) >300, and combined depth across samples (INFO/DP) <1200 and >350 were retained. Then, multiallelic SNPs and indels, monomorphic SNPs and SNPs in close proximity of indels were removed as alignment of sequences around indels can often be problematic. Finally, in addition to the SNP calling, collaborator Zexin Jiao

genotyped structural variants (SVs) on the same dataset using a custom pipeline (Bertolotti *et al.*, 2020).

The resulting dataset of 54 WGS parents was further filtered with PLINK v.1.9 (Purcell *et al.*, 2007), excluding individuals with >20% missing genotypes and variants with >10% missing genotypes, MAF<0.05, and significant deviation from Hardy–Weinberg Equilibrium (P-value < 10^{-6}). After merging the genotypes of the parents and the offspring, these were additionally filtered in PLINK for Mendelian error rates to remove SNPs that exceed 10%.

The final parental genotypes were used as the reference population to impute the 2bRAD-seq genotypes of their offspring with FImpute v.3 (Sargolzaei, Chesnais and Schenkel, 2014). For the imputation analysis, random filling based on allele frequency was turned off, and genotypes with Mendelian inconsistencies between the progeny and the parents were set to missing and re-imputed. Post-imputation, quality control was applied to the imputed genotypes using PLINK, excluding individuals with >20% missing genotypes, variants with >10% missing genotypes, MAF<0.05, and significant deviation from Hardy–Weinberg Equilibrium (P-value <10^{-8}) from subsequent analyses.

4.2.4 Estimation of genetic parameters and genomic relationship matrix

After quality control, the filtered imputed dataset was used to estimate the genetic parameters of the traits with ASREML v.4.2 (Gilmour, Gogel and Welham, 2021), using the following linear mixed model (Eq. 4.1):

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Za} + \boldsymbol{e} \quad \text{(Eq. 4.1)}$$

where $\boldsymbol{y}$ is a vector of observed phenotypes, $\boldsymbol{\mu}$ is the overall mean of phenotype records, $\boldsymbol{a}$ is a vector of additive genetic effects distributed as $\boldsymbol{a} \sim \boldsymbol{N}(0, \boldsymbol{G}\sigma_a^2)$, where σ_a^2 is the additive (genetic) variance and $\boldsymbol{G}$ is the genomic relationship matrix. $\boldsymbol{Z}$ is the incidence matrix for the additive effects, and $\boldsymbol{e}$ is a vector of residuals. The models did not include tank as fixed effect, as it was already taken into account when the epidemiological traits were inferred.

The genomic relationship matrix between pairs of individuals j and k (gjk)

was calculated using GCTA v.1.24.7 (Yang *et al.*, 2011) as follows :

$$g_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)} \quad \text{(Eq. 4.2)}$$

where N is the total number of SNPs, x_{ij} and x_{ik} are the number of copies of the reference allele for the i^{th} SNP for the j^{th} and k^{th} fish, respectively, and p_i is the frequency of the reference allele estimated from the markers. Two genomic relationship matrices were constructed, one using the filtered 2bRAd-seq genotypes and another one after imputation to whole-genome genotypes.

4.2.5 Genome-wide association study

Genome-wide association study (GWAS) was conducted for each trait, before and after imputation. The GWAS was performed using a mixed linear model association (MLMA) (Eq. 4.3) with the leave-one-chromosome-out (LOCO) option implemented in GCTA v.1.24.7:

$$\boldsymbol{y_i = \mu + a_j g_{ij} + u_i + e_i} \quad \text{(Eq.4.3)}$$

in which y_i is the observed phenotype of the i^{th} individual, μ the overall mean in the population, a_j is the additive genetic effect of the reference allele for the j^{th} SNP with its genotype for individual i (g_{ij}) coded as 0, 1 or 2, and e_i is the residual effect following a normal and independent distribution $e \sim N(0, I\sigma_e^2)$, with σ_e^2 the residual variance. Finally, u_i the random vector of polygenic effects followed a normal distribution $u \sim N(0, G\sigma_g^2)$ with σ_g^2 the estimated genetic variance and G a partial GRM constructed with 21 chromosomes after removing the chromosome containing the j^{th} SNP since the analysis was performed using the leave-one-chromosome-out (MLMA-LOCO) approach.

4.2.6 Estimation of genetic parameters with BayesRCO

BayesRCO (Mollandin *et al.*, 2022a) software was used for genomic prediction, which utilises a Bayesian Gaussian mixture model and the different parameters are estimated using a Markov Chain Monte Carlo (MCMC) algorithm.

BayesRCO implements different Bayesian hierarchical models, where the effects of the variants are assumed to follow a Gaussian mixture distribution with different number of components. The models can integrate prior biological information (e.g., functional annotations, candidate gene lists, known causal variants) with the variants being grouped into potentially overlapping annotation categories.

All Bayesian models in BayesRCO exploit the general statistical model below for genomic prediction by best estimating a vector of SNP effects *β*:

$$\boldsymbol{y} = \boldsymbol{\mu} 1_n + \boldsymbol{X\beta} + \boldsymbol{e},$$

$$e \sim N(0, I_{n\sigma_e^2})$$

The various models are differentiated by defining distinct prior distributions on the SNP effect vector *β*.

BayesR

BayesR includes a single prior annotation category to which all SNPs are assigned and assumes that SNP effects follow a four-component normal mixture:

$$f(\beta_i) = \sum_{k=1}^{4} \pi_k f_k(\cdot \,|\theta_k)$$

$$f_k = \begin{cases} \delta(0)\ if\ k = 1 \\ \varphi(\cdot\,|0, \theta_k)\ otherwise \end{cases}$$

where $\boldsymbol{\theta} = (\theta_2, \theta_3, \theta_4) = (0.0001\sigma_g^2, 0.001\sigma_g^2, 0.01\sigma_g^2)$, $\sum_{k=1}^{4} \pi_{k,c} = 1, \delta(0)$ represents a point mass at 0, and φ is the centred Gaussian probability density function. The BayesR model implies that markers are assigned to one of four different effect size classes: null, small, medium or large, corresponding respectively to 0%, 0.01%, 0.1% and 1% of the total additive genetic variance σ_g^2.

BayesRCπ

The BayesRCπ model enables the prior assignment of markers to multiple categories that may overlap, and these various annotation categories cumulatively influence the estimation of the marker effects. Each variant (variant i) has a corresponding set of annotations $Ci \subseteq \{c_1, c_1, \ldots, c_m\}$. Depending on the case, variant i can have a single annotation (i.e., $|Ci|$= 1) or be multi-annotated (i.e.,

$|Ci|\geq 1$). Specifically, this model defines a mixture of mixtures prior distribution for SNP effects:

$$f(\beta_i|C_i) = \sum_{c\in C_i} p_{i,c} \sum_{k=1}^{4} \pi_{k,c} f_k(\cdot|\theta_k)$$

$$\text{such that } f(k) = \begin{cases} \delta(0), \text{ if } k = 1 \\ \varphi(\cdot|0, \theta k) \text{ otherwise} \end{cases}$$

where $\boldsymbol{\theta} = (\theta_2, \theta_3, \theta_4) = (0.0001\sigma_g^2, 0.001\sigma_g^2, 0.01\sigma_g^2)$, $\sum_{k=1}^{4} \pi_{k,c} = 1$ for all $c \in \{c_1, c_2, \ldots, c_m\}$, $\delta(0)$ represents a point mass at 0, and φ is the centred Gaussian probability density function. The markers are assigned to one of four different effect size classes: null, small, medium or large, corresponding respectively to the default values of 0%, 0.01%, 0.1% and 1% of the total additive genetic variance σ_a^2. The mixing parameter $\boldsymbol{p}_i \in]0,1]^{|C_i|}$ is introduced for SNP i in its set of annotations $\boldsymbol{C}_i$, such that $\sum_{c\in \boldsymbol{C}_i} p_{i,c} = 1$ for all i. The mixing proportions π_c are assumed to follow a Dirichlet prior, giving the posterior $f(\pi_c|\cdot) \sim \text{Dirichlet}(\boldsymbol{\alpha} + \gamma_c)$, with $\boldsymbol{\alpha} = (1,1,1,1)$. The mixing proportions p_i are assumed to follow a Dirichlet prior, with size depending on the cardinality of the annotation set of each SNP i.

BayesRC+

The BayesRC+ model assigns an additive impact of multiple annotation categories on estimated SNP effects. This model offers an alternative way of interpreting a multi-annotated marker by assuming that more weight should be attributed to markers with a high number of annotations. Variants with multiple annotations are more likely to be considered as non-null in the model, leading to a larger estimated effect. To address this, BayesRC+ introduces the following cumulative mixture prior distribution for the effect of SNP i:

$$\boldsymbol{f}(\boldsymbol{\beta}|\boldsymbol{C_i}) = \sum_{\boldsymbol{c\in C_i}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{4}} \boldsymbol{\pi_{k,c}} \boldsymbol{f_k}(\cdot|\boldsymbol{\theta_k})$$

$$\text{such that } f(k) = \begin{cases} \delta(0), \text{ if } k = 1 \\ \varphi(\cdot|0, \theta k) \text{ otherwise} \end{cases}$$

where θ, $\delta(0)$ and φ are defined as before and $\sum_{\boldsymbol{k=1}}^{\boldsymbol{4}} \boldsymbol{\pi_{k,c}} = \boldsymbol{1}$ for all $c\in\{c_1, c_2, \ldots, c_m\}$. Prior and posterior distributions for the mixing proportions π_c are

as described above in BayesRCπ model.

Gibbs sampler

Bayesian inference is performed in all cases by obtaining draws from the posterior distribution using a Gibbs sampler. Model parameters are subsequently estimated using the posterior mean across iterations, after excluding a burn-in phase and thinning draws. The Gibbs sampler parameters were set to 60,000 iterations, including 30,000 as a burn-in and a thinning rate of 10.

4.2.7 Cross-validation for genomic-based prediction accuracy

The accuracy of genomic prediction was estimated by five replicates of fivefold cross-validation analysis (80% of individuals in the training set and 20% in the validation set) (Tsairidou, 2019). For each replicate, different individuals were randomly allocated and partitioned into different cross-validation groups. The phenotypes in the validation set were masked, and genomic best linear unbiased prediction (GBLUP) was applied to predict the breeding values of the validation set individuals in ASReml 4.2, using the linear mixed model described in Eq. 4.1. The same five replicates of five groups split into training and validation sets were used for prediction of breeding values with the BayesRCO software (Mollandin *et al.*, 2022a).

Additionally, to test the ability to perform accurate prediction across unrelated individuals, the population was split in five groups where each group consisted of full and half-sib families, with fish between the different groups not sharing any of their parents. The table with the design of the families in each of the five groups for the cross-validation analyses, together with the number of offspring in each group can be found in the supplementary material (**Supplementary table 4.1**).

Prediction accuracy (r) was calculated for each trait and averaged across folds and replicates as the Pearson correlation between the predicted breeding values (ŷ) in the validation set and the actual phenotypes (y), divided by the square root of the trait's heritability (h) [$r \approx \frac{r(y,\hat{y})}{h}$] (Legarra *et al.*, 2008).

4.2.8 Annotation categories

For the different functional categories used (see below), variants were characterized as belonging to that category (1) or not (0), and variants not fitting to any of the functional categories were grouped into the "other" category. Each variant, depending on the scenario, was associated with a distinct set of annotations and could be categorized under either a single annotation or multiple annotations.

Non-synonymous and other high impact protein coding mutations

One of the scenarios involved variants affecting protein coding such as non-synonymous mutations, premature stop codons or new transcription start sites. SnpEff v.5.2 (Cingolani *et al.*, 2012) was used to annotate and predict the effects of genetic variants on genes and proteins (such as amino acid changes) using Sequence Ontology terms. SnpEff outputs, among other information, the putative impact/deleteriousness of the variant in four levels (high, moderate, low, modifier), the common gene name if the variant is in or close to a gene, and also the type of feature (e.g., transcript, missense variants, start lost, intergenic region, synonymous variant, etc.) in Sequence Ontology terms. Variants annotated with SnpEff were set as belonging to this category potentially affecting protein function, if they were not classified as intergenic, intron, non-coding, 5 prime UTR (untranslated region) variant, 3 prime UTR variant and synonymous variant. The Sequence Ontology terms describing the variants that were used in this category are summarized in a table at the supplementary material of this chapter (**Supplementary table 4.2** and **Supplementary table 4.3**).

Markers in regulatory elements

Regulatory elements such as promoters and enhancers that were annotated within the AQUA-FAANG project using data from ATAC-seq (assay for transposase-accessible chromatin with sequencing) and ChiP-seq (chromatin immunoprecipitation assays with sequencing) (Ernst and Kellis, 2017; Aramburu *et al*. 2024, in preparation) were used as another functional category in the present study. SNPs and structural variants overlapping with the regulatory elements were assigned to this functional category.

BayesRCO annotation scenarios

We tested 3 scenarios of SNP and SV categorization:

1. The first scenario tested the impact of prioritising genetic markers potentially affecting protein function. The model had three categories: i) SNPs with relevant SnpEff annotation, ii) SVs with relevant SnpEff annotation, and iii) all remaining SNPs and SVs.
2. The second scenario tested the impact of regulatory variants. The model had three categories: i) SNPs overlapping with regulatory elements (promoters and enhancers), ii) SVs overlapping with regulatory elements (enhancers and promoters), and iii) all remaining SNPs and SVs.
3. Third scenario was a combination of the two previous scenarios. The model had five categories: i) SNPs with relevant SnpEff annotation, ii) SVs with relevant SnpEff annotation, iii) SNPs overlapping with regulatory elements (promoters and enhancers), iv) SVs overlapping with regulatory elements (promoters and enhancers), and v) all remaining SNPs and SVs.

4.3 Results

4.3.1 Data summary and genetic parameters

The average sequencing coverage across the 54 whole-genome sequenced samples (parents of the challenged population) was ~14 ± 8.6. After SNPs and SV calling and filtering, 2,807,473 markers remained (2,800,888 SNPs and 6,585 SVs). The genotypes of the parents (WGS) were then merged with the genotypes of the offspring (2b-RADseq) and after removing Mendelian inconsistencies (1,207 SNPs), 8,694 common SNPs were retained from the offspring dataset, which were used as anchorage for imputation of the offspring to whole-genome genotypes. After imputation, the final dataset consisted of 1,370 offspring genotyped for 1,274,749 variants (1,271,532 SNPs and 3,217 SVs).

Heritability estimates for the five traits with ASReml before and after imputation can be found in **Table 4.1**. Although the estimates before imputation were lower compared to the ones after imputation the difference was only statistically significant for R0. **Table 4.1** also presents heritability estimates obtained using the BayesRCO software for the different Bayesian models. We

observed higher heritability estimates compared to GBLUP for most of the Bayesian models, with BayesRCπ estimates being slightly higher than those from BayesR. Heritability estimates for BayesRC+ were 1 for all traits, even after increasing the Gibbs sampler iterations to 80,000. According to the developers (personal communication with Dr Andrea Rau and Dr Fanny Mollandin), who observed similar results with this model, these estimates are often biased, particularly when dealing with a large number of genetic variants, such as WGS data. In such cases, the genetic variance tends to be overestimated relative to the residual variance leading to heritability estimates close to 1. These estimates are not reliable and thus prediction accuracy results from the BayesRC+ model were not included in the results of this chapter.

Table 4.1 Heritability estimates of the different traits before and after imputation with GBLUP and the different Bayesian models.

Trait	GBLUP		BayesR (no annotation)	BayesRCπ		
	Heritability estimate before imputation	Heritability estimate after imputation		Scenario 1[1] (SnpEff)	Scenario 2[2] (enhancers + promoters)	Scenario 3[3]
R0	0.63 ± 0.03	0.70 ± 0.03	0.67 ± 0.004	0.69 ± 0.001	0.71 ± 0.001	0.81 ± 0.002
Infectivity	0.18 ± 0.04	0.20 ± 0.04	0.22 ± 0.001	0.26 ± 0.001	0.26 ± 0.001	0.30 ± 0.001
Susceptibility	0.29 ± 0.04	0.32 ± 0.04	0.32 ± 0.001	0.35 ± 0.001	0.36 ± 0.001	0.38 ± 0.001
Recoverability	0.12 ± 0.03	0.14 ± 0.04	0.13 ± 0.001	0.21 ± 0.001	0.20 ± 0.001	0.25 ± 0.001
Transformed days to death	0.14 ± 0.03	0.15 ± 0.04	0.13 ± 0.001	0.23 ± 0.001	0.21 ± 0.001	0.27 ± 0.001

[1] Scenario 1 included SNPs and SVs with relevant SnpEff annotation
[2] Scenario 2 included SNPs and SVs overlapping with regulatory elements (promoters and enhancers).
[3] Scenario 3 combined the scenario 1 and 2 by including categories for SNPs and SVs with relevant SnpEff annotation and overlapping with regulatory elements.

4.3.2 Annotation categories

A total of 24,072 SNPs and 3,217 SVs were assigned to functional categories with a potential impact on protein function according to SnpEff. The predicted effects with their Sequence Ontology term and the four levels of their putative impact can be found in the supplementary material (**Supplementary table 4.2** and **Supplementary table 4.3**). A summary of the number of SNPs and SVs in each putative impact category after annotation with SnpEff can be found in **Table 4.2**. The total number of SVs in **Table 4.2** is 3,734 because some SVs were predicted by SnpEff to have more than one effect category (in Sequence Ontology terms, as shown in **Supplementary table 4.3**). Consequently, their putative impact may fall into more than one of the four impact levels: high, low, moderate, or modifier. A total of 13,387 SNPs and 119 SVs were overlapping with the position of promoters and enhancers in the turbot genome.

Table 4.2 Number of SNPs and SVs categorised according to their putative impact on protein function after annotation with SnpEff.

Putative impact	Number of SNPs	Number of SVs
HIGH	566	158
LOW	1,680	43
MODERATE	21,799	6
MODIFIER	27	3,527
Total	**24,072**	**3,734**

4.3.3 Genome-wide association analysis

GWAS using the non-imputed and the imputed data identified a polygenic genetic architecture for three of the traits (susceptibility, infectivity and transformed days to death), with no variant surpassing the genome-wide significance threshold (**Figure 4.5)**. However, some variants for these traits showed suggestive association (blue line in **Figure 4.5**), but not for the same chromosomes before and after imputation.

Post imputation there were three suggestive SNPs for the trait of susceptibility in chromosomes 8 and 16. Three suggestive SNPs were also detected on chromosome 10 for infectivity. For recoverability there were three suggestive SNPs on chromosomes 12, 16, and 17 and one on chromosome 12 surpassing the genome wide significance threshold. For the composite trait of R0, there were thirty-nine SNPs showing suggestive association on chromosomes 2, 3, 4, 5, 10, 12, 13, 14, 16, 17, 18, 19, 21 and 22. For the same trait five SNPs reached genome-wide significance in chromosome 5, 10, and 13. No SNP surpassed any of the two thresholds for the trait of the transformed days to death post imputation.

Tables in the supplementary materials at the end of this chapter (**Supplementary table 4.4** to **Supplementary table 4.8**) contain information about the number, position and p value for the variants surpassing the chromosome wide and genome wide significance thresholds before and after imputation.

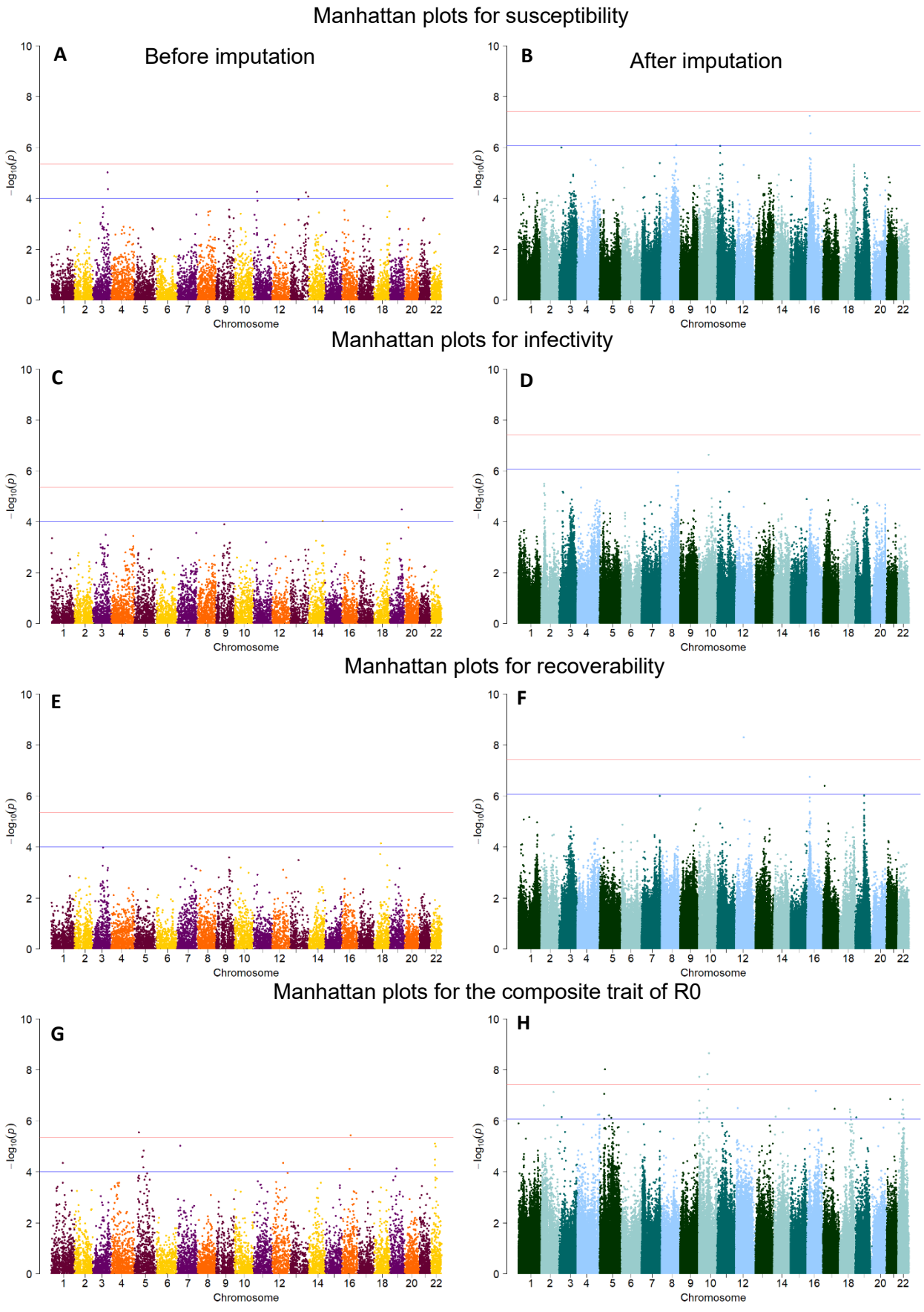
4.3.4 Incorporation of functional annotation into genomic prediction

The imputed genotypes of the offspring were used to estimate genomic prediction accuracies for the disease traits using GBLUP (fitted in ASReml) and Bayesian models (fitted in BayesRCO). BayesR, which does not incorporate annotation information, was used as a baseline for comparison with the BayesRCπ model, which included various annotation scenarios. Overall, the results showed that both BayesR and GBLUP outperformed most BayesRCπ models in terms of genomic prediction accuracy (**Table 4.3)**. When comparing GBLUP to BayesR, GBLUP slightly outperformed BayesR for most traits, except for recoverability, for which accuracies were similar, and transformed days to death, for which BayesR performed better.

More specifically, for the composite trait R0, GBLUP, BayesR and BayesRCπ model with annotation categories from SnpEff (scenario 1), performed quite similarly with prediction accuracies very close to 1. For the same trait, scenarios 2 and 3 achieved 2% and 10% lower prediction accuracy, respectively, compared to BayesR, and 4% and 12% lower accuracy compared to GBLUP. For infectivity, BayesR achieved 12% higher accuracy compared to scenario 1, 11%

higher than scenario 2 and 16% higher than scenario 3. For transformed days to death, BayesR was approximately 6.5% more accurate than GBLUP and 23% (scenario 2) to 32% (scenario 3) more accurate than the BayesRCπ models with annotations. For recoverability, BayesR showed similar accuracy to GBLUP whereas the three BayesRCπ scenarios had approximately 22% (scenario 2) to 31% (scenario 3) lower accuracy than BayesR.

A notable difference between the two approaches (GBLUP and Bayesian models) is the considerable variation in their running times. On average, ASReml GBLUP required just 3 minutes and 30 seconds to complete one group of the 5-fold cross-validation. In contrast, the BayesRCπ model took approximately 2 days and 12 hours to run a single group, meaning running 5 cross-validation replicates with 5 groups each would require a total of 62.5 days of computational time.


Manhattan plots for susceptibility
A
Before imputation
B
After imputation
Manhattan plots for infectivity
C
D
Manhattan plots for recoverability
E
F
Manhattan plots for the composite trait of R0
G
H
−log10(p)
Chromosome
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22
0 2 4 6 8 10

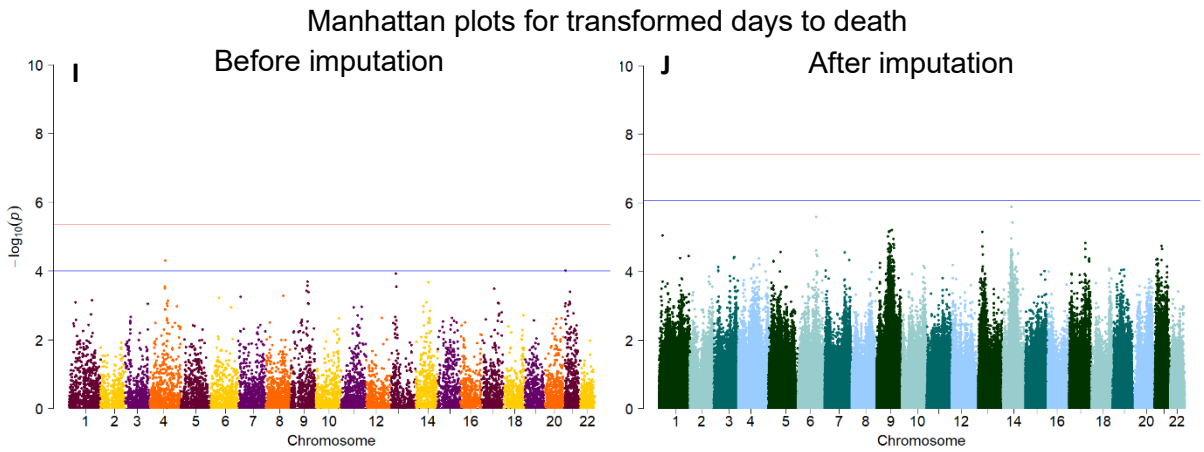


Figure 4.6 Manhattan plots of the GWAS with GCTA software, using the MLMA LOCO (leave-one-chromosome-out) approach for the traits of susceptibility, infectivity, recovery, R0 and transformed days to death. The values on the y-axis represent the −log10 of the P value and the x-axis the positions on the chromosomes. The red line is the 5% genome-wide significance threshold and the blue line is the 5% chromosome-wide significance threshold (Bonferroni correction).

Table 4.3 Genomic prediction accuracy results using imputed genotypes. The accuracy of the breeding values (including the standard deviation between replicates) was measured for results obtained with GBLUP and the Bayesian models for the different scenarios incorporating functional annotation information.

Trait	GBLUP (ASReml)	BayesR (no annotation)	BayesRCπ		
			Scenario 1[4] (SnpEff)	Scenario 2[5] (enhancers + promoters)	Scenario 3[6]
R0	1.00 ± 0.03	0.98 ± 0.10	0.99 ± 0.05	0.96 ± 0.09	0.88 ± 0.08
Infectivity	0.76 ± 0.16	0.73 ± 0.14	0.64 ± 0.14	0.65 ± 0.15	0.61 ± 0.11
Susceptibility	0.88 ± 0.09	0.85 ± 0.13	0.82 ± 0.11	0.70 ± 0.19	0.75 ± 0.13
Recoverability	0.64 ± 0.19	0.65 ± 0.17	0.49 ± 0.15	0.51 ± 0.14	0.45 ± 0.12
Transformed days to death	0.58 ± 0.16	0.62 ± 0.17	0.47 ± 0.13	0.48 ± 0.14	0.42 ± 0.12

[4] Scenario 1 included SNPs and SVs with relevant SnpEff annotation

[5] Scenario 2 included SNPs and SVs overlapping with regulatory elements (promoters and enhancers).

[6] Scenario 3 combined the scenario 1 and 2 by including categories for SNPs and SVs with relevant SnpEff annotation and overlapping with regulatory elements.

4.3.5 Accuracy of prediction across unrelated individuals

When the individuals were split in groups that contained separately full and half-sib families (fish in reference and validation groups did not share any common parents and so they were distantly related), no matter the trait the accuracy of prediction obtained with both GBLUP and BayesRCπ achieved were null or near zero (**Table 4.4**).

Table 4.4 Accuracy of prediction (± standard deviation) across unrelated individuals.

Trait	GBLUP (ASReml) Groups of distantly related individuals	BayesRCπ Groups of distantly related individuals Scenario 1 (SnpEff)
R0	-0.15 ± 0.21	-0.17 ± 0.19
Infectivity	0.01 ± 0.26	0.01 ± 0.22
Susceptibility	0.09 ± 0.42	0.09 ± 0.42
Recoverability	0.15 ± 0.22	0.13 ± 0.16
Transformed days to death	0.08 ± 0.30	0.08 ± 0.22

4.4 Discussion and future prospects

In this study, we utilized the imputed WGS genotypes of a turbot population to assess the predictive ability of GBLUP and Bayesian models by prioritising variants based on functional annotations for inferred epidemiological traits. We incorporated multiple overlapping annotation information into BayesRCπ models for this evaluation, and the BayesR model without annotations was used as a standard for comparison. Our observations revealed that GBLUP and BayesR generally outperformed Bayesian models with annotations in terms of prediction accuracy and GBLUP was much faster in terms of computational running time.

4.5 Accuracy of Bayesian and GBLUP models

The accuracy of genomic prediction is influenced by the model used due to its prior assumptions about SNP effects, as it reflects the underlying architecture of the trait (Daetwyler *et al.*, 2010). The assumption of Bayesian models that genetic variation is driven by a small subset of SNPs can offer an advantage over the GBLUP model when the trait's architecture is influenced by a few major QTL. This is showed in a study for survival to *Streptococcus sp*. in Tilapia (Joshi *et al.*, 2021) where prediction accuracy of the Bayesian models, including BayesR (0.67), was higher than GBLUP (0.56). The authors indicate that this may be because the trait is controlled by a limited number of major QTL, supported by their results and the results of other studies in different strains of the same pathogen (Weigel *et al.*, 2009). If the trait's architecture is polygenic, as in our case, GBLUP models may be just as accurate, or in some cases even outperform Bayesian models (MacLeod, Hayes and Goddard, 2014; Joshi *et al.*, 2021).

Several studies have been conducted to compare the predictive ability of Bayesian and GBLUP methods (Song *et al.*, 2023). For example, in a study on survival rates of Japanese flounder (Liu *et al.*, 2018), BayesCπ and GBLUP achieved similar predictive accuracies for the selection candidates with values of 0.61 and 0.60 respectively. In another study (Nguyen, Phuthaworn and Knibb, 2020) the authors report that GBLUP had similar predictive power to non-linear approaches such as BayesC and BayesCπ, particularly for growth and carcass traits, which are known to be controlled by many genes, each with small additive genetic effect. To the best of our knowledge, the current study is the first to compare Bayesian models with and without the incorporation of annotated information in an aquaculture species.

Apart from the genetic architecture of a trait, heritability can also affect prediction accuracy (Desta and Ortiz, 2014). By comparing the prediction accuracies between the different traits, we observed that some of them, such as the composite trait R0 and susceptibility, had considerably higher values. Heritability estimates for these two traits were moderate (~0.35 ± 0.03 for susceptibility, calculated as the mean across the five models from **Table 4.1**) and high (~ 0.72 ± 0.05 for R0, also calculated as the mean across the 5 models from

Table 4.1). Traits with high heritability tend to have higher prediction accuracies (Merrick and Carter, 2021). Consequently, higher prediction accuracies observed for these two traits may be due to the higher heritability estimates compared to the other traits tested here.

In our study, GBLUP performed better that BayesR in most cases for the polygenic traits that we tested. The performance of BayesRCπ model differed across scenarios with no clear pattern except that scenario 3 underperformed the other two scenarios for most traits. The BayesR model performed better for four of the five traits, indicating that the annotation categories used in this study, derived from SnpEff and active promoters and enhancers, did not result in any improvement in prediction accuracy for these traits in the tested population.

4.5.1 Tailored variant annotation strategies could improve the accuracy of Bayesian models

One strategy that could improve the predictive ability of the models tested here, and could be explored in future studies, is the incorporation of neighbouring variants by using extended windows both upstream and downstream of the variants assigned to annotation categories. This approach proves particularly beneficial when the precise location of causal mutations remains unknown. However, it is important to be cautious when adding too many markers to annotations, as this may risk diluting the information they contribute. This strategy was employed by Mollandin *et al.* (2022b) and resulted, in some instances, in modest gains in prediction accuracy.

The choice of annotation categories and the way variants are chosen to be included in them can affect prediction accuracy estimates, likely in a trait dependent manner. Two types of functional annotation information were used in this study. First, we predicted the effects of genetic variants at the protein level for all the genes in the genome using SnpEff. Perhaps, it would be desirable to restrict the annotation to SnpEff categories with a high impact (although in our case this would significantly lower the number of markers assigned to annotation categories), or to give higher weight to variants affecting proteins related to biological processes or molecular functions specific to *Philasterides dicentrarchi*

infection. Secondly, all variants overlapping with regulatory elements throughout the genome, were identified using immune challenges (*Vibrio* spp) carried both *in vivo* (by intra peritoneal injection) and *in vitro* (by immunostimulation of head-kidney primary leukocytes). Therefore, there is a possibility that these bacterial stimulations overlooked contributions from genes whose expression levels may have varied due to infection with the parasite *Philasterides dicentrarchi*. In summary, adding other levels of information or annotations more specific to the interactions between turbot and *Philasterides dicentrarchi* could potentially provide better results, as here we used genome-wide annotations agnostic to the trait. To the best of our knowledge, no study has yet compared the impact on prediction accuracy of annotations derived from tailored vs non-tailored experiments.

4.5.2 Impact of annotation category number and type on predictive performance

It would be interesting for future studies to refine and increase the number of annotation categories by adding extra information from other sources (e.g., GWAS, QTLs) to explore different strategies when constructing annotations. These annotation categories still have to contain a substantial number of markers (>1,000) explaining a considerable portion of the total variance (MacLeod *et al.*, 2016). For the current study, we did not have markers with large effects for the different traits tested. The architecture of these traits was polygenic thus, as observed in other studies, we cannot expect a substantial increase in the Bayesian compared to GBLUP prediction accuracy (MacLeod *et al.*, 2016).

In their study, Mollandin *et al.* (2022a) observed that for BayesRCπ and BayesRC+, markers with fewer annotations and smaller overlap tend to have underestimated posterior variances. They observed a similar trend for both models in which multi-annotated markers can navigate between annotations avoiding an underestimation of their effect in case they were incorrectly assigned. For BayesRCπ they tested this on large-effect QTL variance estimation for different annotation enrichment scenarios where it yielded larger estimated posterior variances for multi-annotated QTLs. For BayesRC+ they demonstrated

this phenomenon by testing various annotation scenarios ranging from 2 to 7. Their findings indicated that a sufficient number of annotations for large QTLs was approximately 4, while for medium QTLs, it was higher, requiring 7. Consequently, in order for a marker to be assigned a strong or a medium effect in BayesRC+ model, it has to be included in more than one highly enriched annotations.

We did not include prediction accuracy results from the BayesRC+ model in this study because the heritability estimates were 1 or close to 1, making them unreliable. Although our dataset had no trait controlled by any large or medium-effect QTLs, comparing the performance of BayesRC+ with BayesRCπ but also the other models would be valuable in future studies with different traits. It is possible that we had too few annotation categories and testing more annotation categories would be valuable. However, the construction and testing of different scenarios with relevant annotation categories is a challenging task, especially when runtime is long. In our case, the maximum number of annotation categories tested in scenario 3 was five. Running time can increase significantly with the increase of the number of categories requiring careful and strategic grouping.

4.5.3 Predictions across generations or populations

If we want to reduce the economic and welfare costs associated with frequent disease challenges to select for disease resistance, we need to develop accurate prediction methods that work well across distantly related populations. This is not straightforward, as genomic prediction relies on the linkage disequilibrium between causal loci and nearby variants, but this is lost across distant populations (Fraslin, Yáñez, *et al.*, 2022). Identifying the causal loci underlying traits of interest would enable direct selection in any population, but this is unrealistic especially for polygenic traits. However, prioritizing or giving higher weight to variants that are more likely to be causal could improve prediction accuracies when the validation population is genetically distant from the training population.

To test this concept, we organized animals into five distinct groups based on their parents, ensuring that none of the parents were shared across the groups. Consequently, individuals in the validation set exhibited greater genetic similarity

to one another and were more distant from the training set. When individuals in the validation set are genetically distant from the training set, the expected value of genomic relationship will be close to zero, resulting in lower prediction accuracy estimates (Karaman *et al.*, 2016; Fraslin, Yáñez, *et al.*, 2022). Indeed, our prediction accuracy results with both GBLUP and BayesRCπ were close to zero, with the caveat that we could only test one cross validation replicate using this design, and the resulting standard deviations across the five groups were very high.

Other studies obtained similar results when the tested individuals were not closely related to the training set. A study on Atlantic salmon by Fraslin *et al.* (2022) investigated the impact of reducing the genomic relationship between the training and validation set on genomic prediction accuracy. Using the year class of 2010 to predict sea lice count and body weight for the 2014 year class, resulted in near-zero and highly biased genomic prediction accuracy. However, there is a possibility that these results were obtained due to absence of genetic correlation between the different year classes for the same trait.

Studies in Drosophila showed similar results. Ober *et al.* (2015) used a small number of inbred lines from the Drosophila Genetic Reference Panel (DGRP) with low linkage disequilibrium and low average genomic relatedness, initially finding prediction accuracies of zero for chill coma recovery time using GBLUP. However, after accounting for genetic architecture (from GWAS results conducted with the training set) by using the top SNPs to build the genomic relationship matrix, they managed to significantly improve the predictive ability of the model (from 0.07 in females and 0 in males to 0.43 and 0.48 for females and males, respectively). Another study by Morgante *et al.* (2018) further investigated the low relatedness and linkage disequilibrium of the DGRP lines with simulated data to test a bigger number of individuals. When all variants were utilized in GBLUP, accuracy results were poor, sometimes even approaching zero depending on the simulated scenario. However, accounting for the genetic architecture of complex traits in the construction of the GRM matrix, significantly improved prediction accuracy. Finally, for the same species and inbred lines, Morgante *et al.* (2020) used WGS and RNA sequencing for three different traits. For two of the three traits, incorporating expression data by using a transcriptomic

relationship matrix, led to higher prediction accuracy compared to using genotypes alone, while the opposite outcome was observed for the third trait. Additionally, accuracy substantially improved for all traits when gene ontology (GO) was included as a third category (through a GO-specific GRM) in addition to genotypes and transcriptomic information. Although these two approaches resulted in increased improvements of predictions for this generation, the use of expression data to build the relationship matrix incorporates part of the environmental effects which might lead to decreased accuracy in the next generation if such information is not utilized again.

Currently, there are no studies in fish that compare the impact of incorporating functional or other biological information into Bayesian models. Assessing this impact across traits with varying genetic architectures can assist us in making informed decisions regarding whether it is beneficial to include such information in future breeding programmes instead of GBLUP models.

4.5.4 Concluding remarks

The knowledge obtained from omics studies offers great potential in unravelling underlying cellular mechanisms and trait aetiology. Bayesian models, with their ability to incorporate priors, provide a straightforward approach to integrating known functional information into genomic prediction models for complex traits. Integrating this valuable information into genomic prediction models has the potential to improve prediction accuracy but may also offer valuable insights into the genomic architecture of these traits. Nevertheless, future improvements to the experimental design are necessary to evaluate whether these methods can reliably increase prediction accuracies, and which kind of functional information is necessary to achieve this goal.

4.7 Supplementary material

Supplementary table 4.1 The five groups of full and half-sib families used to test prediction accuracy with ASReml. The offspring in these groups did not share any common parent, in contrast with the groups of the five replicate five-fold cross validation, in which offspring were randomly assigned and consiquently some of them shared a parent with individuals from another group.

Group 1			**Group 4**		
Father ID	Mother ID	Number of offspring	Father ID	Mother ID	Number of offspring
R9	R23	40	**R3**	**R26**	40
T11	R23	38	R7	R27	39
T11	T35	1	R7	T30	39
R4	T35	34	R14	T30	38
T13	T35	40	R14	R31	34
T6	T35	40	R8	R31	40
R4	T34	35	R8	T54	33
R15	T34	40			2
Total		**268**	**Total**		**265**
Group 2			**Group 5**		
FB-Y-92	R24	35	**T16**	**R36**	38
R12	R24	40	R19	T38	1
R12	R25	40	T17	T38	39
T1	R25	30	R19	R37	40
T2	R25	35	R19	R41	43
R12	R39	35	T17	R41	1
T21	R39	38	T17	T44	33
T1	T28	40	R18	T44	40
Total		**293**	R18	R42	39
Group 3			**Total**		**236**
T22	**T43**	**40**			
T22	**R29**	**38**			
T5	**R29**	**40**			
T10	R32	39			
T10	R33	35			
T5	R33	40			
T20	R40	38			
Total		**270**			

Supplementary table 4.2 Number of SNPs in each effect category predicted by SnpEff and the level indicating the putative impact of the effect.

SNP effects	Putative impact	Number of SNPs
5_prime_UTR_premature_start_codon_gain_variant	LOW	1,515
initiator_codon_variant	LOW	9
intragenic_variant	MODIFIER	27
missense_variant	MODERATE	21,013
missense_variant&splice_region_variant	MODERATE	786
splice_region_variant	LOW	124
splice_region_variant&stop_retained_variant	LOW	17
start_lost	HIGH	71
start_lost&splice_region_variant	HIGH	4
stop_gained	HIGH	384
stop_gained&splice_region_variant	HIGH	23
stop_lost	HIGH	8
stop_lost&splice_region_variant	HIGH	76
stop_retained_variant	LOW	15
Total	**-**	**24,072**

Supplementary table 4.3 Number of SVs in each effect category predicted by SnpEff and the level indicating the putative impact of the effect.

SV effects	Putative impact	Number of SVs
3_prime_UTR_variant	MODIFIER	50
5_prime_UTR_variant	MODIFIER	20
bidirectional_gene_fusion	HIGH	7
conservative_inframe_deletion	MODERATE	3
disruptive_inframe_deletion	MODERATE	2
duplication	LOW	3
duplication	MODERATE	1
exon_loss_variant	HIGH	21
exon_loss_variant&splice_acceptor_variant&splice_donor_variant&splice_region_variant&intron_variant	HIGH	14
exon_loss_variant&splice_acceptor_variant&splice_region_variant&intron_variant	HIGH	12
exon_loss_variant&splice_donor_variant&splice_region_variant&intron_variant	HIGH	22
exon_region	MODIFIER	2
frameshift_variant	HIGH	9
frameshift_variant&splice_acceptor_variant&splice_region_variant&intron_variant	HIGH	11
frameshift_variant&splice_donor_variant&splice_region_variant&intron_variant	HIGH	8
frameshift_variant&start_lost	HIGH	4
frameshift_variant&start_lost&splice_region_variant	HIGH	2
frameshift_variant&stop_gained&splice_region_variant	HIGH	3
frameshift_variant&stop_lost	HIGH	1
frameshift_variant&stop_lost&splice_region_variant	HIGH	4
gene_fusion	HIGH	4

intragenic_variant	MODIFIER	19
intron_variant	MODIFIER	3432
non_coding_transcript_exon_variant	MODIFIER	4
splice_acceptor_variant&conservative_inframe_deletion&splice_region_variant&intron_variant	HIGH	1
splice_acceptor_variant&disruptive_inframe_deletion&splice_region_variant&intron_variant	HIGH	2
splice_acceptor_variant&splice_donor_variant&splice_region_variant&intron_variant	HIGH	1
splice_acceptor_variant&splice_donor_variant&splice_region_variant&intron_variant&non_coding_transcript_exon_variant	HIGH	2
splice_acceptor_variant&splice_region_variant&3_prime_UTR_variant&intron_variant	HIGH	1
splice_acceptor_variant&splice_region_variant&5_prime_UTR_variant&intron_variant	HIGH	2
splice_acceptor_variant&splice_region_variant&intron_variant	HIGH	1
splice_acceptor_variant&splice_region_variant&intron_variant&non_coding_transcript_exon_variant	HIGH	2
splice_donor_variant&conservative_inframe_deletion&splice_region_variant&intron_variant	HIGH	3
splice_donor_variant&disruptive_inframe_deletion&splice_region_variant&intron_variant	HIGH	2
splice_donor_variant&duplication&splice_region_variant&intron_variant	HIGH	2
splice_donor_variant&splice_region_variant&5_prime_UTR_variant&intron_variant	HIGH	1
splice_donor_variant&splice_region_variant&intron_variant	HIGH	2
splice_donor_variant&splice_region_variant&intron_variant&non_coding_transcript_exon_variant	HIGH	5
splice_region_variant&intron_variant	LOW	19
splice_region_variant	LOW	15

splice_region_variant&non_coding_transcript_exon_variant	LOW	6
start_lost&conservative_inframe_deletion	HIGH	3
stop_gained&duplication	HIGH	1
stop_gained&duplication&splice_region_variant	HIGH	1
stop_lost&conservative_inframe_deletion	HIGH	1
stop_lost&conservative_inframe_deletion&splice_region_variant	HIGH	1
transcript_ablation	HIGH	2
Total	-	**3,734**

Supplementary table 4.4 Variants surpassing the chromosome wide significance thresholds before and after imputation for susceptibility.

Before imputation									
susceptibility									
Chr	SNPs surpassing chromosome wide significance threshold	bp	A1	A2	Freq	b	se	p	-log10(p)
3	sm5_s00031_485393_85811	20128391	A	C	0.239657	0.028415	0.006416	9.49E-06	5.02256
3	sm5_s00031_987084_86000	20626442	A	G	0.174779	0.02799	0.006842	4.30E-05	4.3663
11	sm5_s00021_1285176_68598	5401845	T	C	0.148849	0.040007	0.009921	5.52E-05	4.25791
13	sm5_s00173_262681_153268	21730120	G	C	0.226528	0.025966	0.006455	5.76E-05	4.23964
13	sm5_s00075_1308160_125439	24556969	T	G	0.234432	0.024062	0.006127	8.59E-05	4.06604
After imputation									
Chr	SNPs surpassing chromosome wide significance threshold	bp	A1	A2	Freq	b	se	p	-log10(p)
8	8:20367063	20367063	G	A	0.47214	0.0299849	0.00607176	7.88E-07	6.10368
16	16:3538536	3538536	T	A	0.283419	0.0394148	0.00726029	5.67E-08	7.24618
16	16:4182181	4182181	A	C	0.331967	-0.0306906	0.00597085	2.75E-07	6.56124

Supplementary table 4.5 Variants surpassing the chromosome wide significance thresholds before and after imputation for infectivity.

Before imputation									
infectivity									
Chr	SNPs surpassing chromosome wide significance threshold	bp	A1	A2	Freq	b	se	p	-log10(p)
14	sm5_s00057_537400_115290	19057853	A	C	0.20566	-0.0429367	0.0109929	9.39E-05	4.02738
19	sm5_s00068_1397055_121681	16955270	C	G	0.0806065	0.0726692	0.0174992	3.29E-05	4.48341
After imputation									
Chr	SNPs surpassing chromosome wide significance threshold	bp	A1	A2	Freq	b	se	p	-log10(p)
10	10:13797535	13797535	C	A	0.166538	0.0840318	0.0162475	2.32E-07	6.63519
10	10:13798097	13798097	A	G	0.166538	0.0840318	0.0162475	2.32E-07	6.63519
10	10:13799310	13799310	T	C	0.166538	0.0840318	0.0162475	2.32E-07	6.63519

Supplementary table 4.6 Variants surpassing the chromosome wide and genome wide significance thresholds before and after imputation for recoverability.

Before imputation									
recoverability									
Chr	SNPs surpassing chromosome wide significance threshold	bp	A1	A2	Freq	b	se	p	-log10(p)
18	sm5_s00011_4179188_42842	9651783	G	A	0.120463	0.061614	0.015501	7.05E-05	4.1521
After imputation									
Chr	SNPs surpassing chromosome wide significance threshold	bp	A1	A2	Freq	b	se	p	-log10(p)
12	12:11392898	11392898	C	T	0.056158	0.136042	0.0232668	5.00E-09	8.30066
16	16:3538536	3538536	T	A	0.283419	0.0731539	0.0140045	1.75E-07	6.75582
17	17:1640074	1640074	G	A	0.0588235	0.101377	0.0200017	4.01E-07	6.39669
Chr	SNPs surpassing genome wide significance threshold	bp	A1	A2	Freq	b	se	p	-log10(p)
12	12:11392898	11392898	C	T	0.056158	0.136042	0.0232668	5.00E-09	8.30066

Supplementary table 4.7 Variants surpassing the chromosome wide and genome wide significance thresholds before and after imputation for the composite trait of R0.

Before imputation									
Composite trait R0									
Chr	SNPs surpassing chromosome wide significance threshold	bp	A1	A2	Freq	b	se	p	-log10(p)
1	sm5_s00078_363281_127394	15891058	T	A	0.390922	-0.0159503	0.00391012	4.52E-05	4.34501
5	sm5_s00029_3213590_82580	6347181	T	C	0.186202	0.0215648	0.00460102	2.77E-06	5.55704
5	sm5_s00012_6454671_47098	11077840	T	C	0.101471	0.0272735	0.00648698	2.62E-05	4.58196
5	sm5_s00012_6062515_46950	11470504	C	T	0.0991947	0.0276439	0.00655637	2.48E-05	4.60504
5	sm5_s00012_5115246_46584	12418275	C	G	0.096714	0.0256968	0.00644451	6.68E-05	4.17519
5	sm5_s00012_4383062_46302	13148207	G	C	0.335434	-0.0195966	0.00452071	1.46E-05	4.83606
7	sm5_s00001_18165629_2999	4575553	C	A	0.263623	-0.0193991	0.00437909	9.43E-06	5.02569
12	sm5_s00005_1836021_22347	15541830	C	T	0.0902285	-0.0279626	0.00684805	4.44E-05	4.3526
16	sm5_s00019_165485_64379	10622794	A	T	0.310578	-0.0156941	0.00396866	7.67E-05	4.11523
16	sm5_s00019_1609329_64352	12060162	C	G	0.0643745	0.0389977	0.00842576	3.69E-06	5.43354
19	sm5_s00016_5697601_59039	9793715	A	G	0.189209	0.0172752	0.00436142	7.47E-05	4.12688
22	sm5_s00010_4272108_39575	5761037	C	T	0.220225	0.0217794	0.00539492	5.41E-05	4.26653
22	sm5_s00010_4460849_39645	5949572	T	C	0.0777126	0.0311205	0.00749869	3.32E-05	4.47846
22	sm5_s00010_4478159_39653	5966838	C	G	0.0752212	0.0339907	0.00759519	7.63E-06	5.11744
22	sm5_s00010_5462508_40021	6945445	A	G	0.0776557	0.0327323	0.00741247	1.01E-05	4.99734
After imputation									
Chr	SNPs surpassing chromosome wide significance threshold	bp	A1	A2	Freq	b	se	p	-log10(p)
2	2:3400912	3400912	T	A	0.490458	0.0302352	0.00586126	2.49E-07	6.60388
2	2:17358153	17358153	C	A	0.0888078	0.049626	0.00922755	7.53E-08	7.12315
3	3:3566759	3566759	T	G	0.115633	0.0386432	0.00779347	7.11E-07	6.14831
4	4:28821644	28821644	T	A	0.0623041	0.0518482	0.0103756	5.82E-07	6.2352

4	4:31100401	31100401	T	G	0.100478	-0.0321616	0.00642968	5.67E-07	6.24625
5	5:6153566	6153566	T	A	0.446332	0.0283929	0.0057623	8.33E-07	6.0791
5	5:6297020	6297020	A	G	0.253475	0.0252276	0.00471638	8.85E-08	7.05321
5	5:7134213	7134213	G	A	0.146426	0.0410071	0.00714817	9.65E-09	8.01537
5	5:12751269	12751269	T	G	0.12576	0.0334497	0.00670827	6.15E-07	6.21088
5	5:15894893	15894893	C	T	0.467037	0.0217443	0.00439462	7.50E-07	6.1249
10	10:661363	661363	T	A	0.0478102	0.0516688	0.0091882	1.87E-08	7.72759
10	10:663415	663415	A	G	0.0569343	0.0526889	0.0100551	1.61E-07	6.79442
10	10:663669	663669	T	C	0.0569343	0.0526889	0.0100551	1.61E-07	6.79442
10	10:1503157	1503157	T	C	0.0718978	0.0340216	0.00689866	8.15E-07	6.08858
10	10:1548987	1548987	C	G	0.0686131	0.0342879	0.00682921	5.15E-07	6.28848
10	7544	6707395	<DEL>	N	0.292139	-0.0325965	0.00647535	4.80E-07	6.31831
10	10:11490853	11490853	T	G	0.0890511	0.0623842	0.0126017	7.40E-07	6.13055
10	10:11515187	11515187	C	T	0.118293	0.0395679	0.0077409	3.20E-07	6.49547
10	10:12409510	12409510	A	C	0.0580292	0.122697	0.0216516	1.45E-08	7.83738
10	10:13438259	13438259	G	A	0.0813869	0.055501	0.0102368	5.90E-08	7.22889
10	10:14216163	14216163	A	T	0.070438	0.156749	0.0262038	2.21E-09	8.65657
12	12:2662150	2662150	T	C	0.158029	-0.0471495	0.00921926	3.15E-07	6.50169
13	13:14879334	14879334	A	C	0.0993426	0.0941999	0.0138094	9.01E-12	11.0451
14	14:2362865	2362865	A	G	0.0493373	0.0929734	0.0187328	6.94E-07	6.15886
14	14:20558701	20558701	C	T	0.0618637	0.0946212	0.0185153	3.21E-07	6.49292
16	16:11633401	11633401	G	T	0.0705128	0.089573	0.0166006	6.82E-08	7.16606
17	17:16206442	16206442	G	A	0.0547445	-0.0632831	0.012408	3.39E-07	6.46939
18	18:14398572	14398572	C	G	0.0497259	0.0911024	0.0180726	4.63E-07	6.33409
18	18:14562579	14562579	T	A	0.0591673	0.0452981	0.00890109	3.60E-07	6.44386
18	18:15518152	15518152	T	A	0.0559414	0.0474836	0.00957842	7.15E-07	6.14593
19	19:1764966	1764966	G	A	0.294205	0.0256977	0.00518831	7.31E-07	6.13619
21	24705	4938871	<DEL>	N	0.0613839	0.0809107	0.0153742	1.42E-07	6.84796
22	22:5199414	5199414	A	T	0.0791971	0.0442886	0.00884222	5.48E-07	6.26137
22	22:7182797	7182797	T	A	0.190876	0.0229822	0.00437685	1.51E-07	6.81988

22	22:7183067	7183067	T	C	0.263869	0.024745	0.00486368	3.62E-07	6.44079
22	22:7183407	7183407	G	A	0.263869	0.024745	0.00486368	3.62E-07	6.44079
22	22:7671810	7671810	G	T	0.25372	0.0255971	0.00511172	5.51E-07	6.2586
22	22:8100135	8100135	T	A	0.108394	0.0297954	0.0060253	7.61E-07	6.11846
22	22:8100306	8100306	T	C	0.108394	0.0297954	0.0060253	7.61E-07	6.11846
Chr	SNPs surpassing genome wide significance threshold	bp	A1	A2	Freq	b	se	p	-log10(p)
5	5:7134213	7134213	G	A	0.146426	0.0410071	0.00714817	9.65E-09	8.01537
10	10:661363	661363	T	A	0.0478102	0.0516688	0.0091882	1.87E-08	7.72759
10	10:12409510	12409510	A	C	0.0580292	0.122697	0.0216516	1.45E-08	7.83738
10	10:14216163	14216163	A	T	0.070438	0.156749	0.0262038	2.21E-09	8.65657
13	13:14879334	14879334	A	C	0.0993426	0.0941999	0.0138094	9.01E-12	11.0451

Supplementary table 4.8 Variants surpassing the chromosome wide significance thresholds before and after imputation for the transformed days to death.

Before imputation									
Transformed days to death									
Chr	SNPs surpassing chromosome wide significance threshold	bp	A1	A2	Freq	b	se	p	-log10(p)
4	sm5_s00014_1892815_51232	15659631	T	G	0.207965	-7.19724	1.77461	5.00E-05	4.30112
21	sm5_s00003_12661603_12092	1489898	T	G	0.179384	6.92802	1.77544	9.53E-05	4.0207

Chapter 5
General Discussion

With the global human population on the rise, the demand for enhanced aquaculture productivity is growing, since millions of people worldwide rely on farmed fish as a vital nutritional source. Genomic selection has proven to be more efficient at improving aquaculture production than traditional pedigree-based methods, providing more accurate estimates of genomic relationships between animals and consequently increased accuracy of breeding value prediction, leading to faster genetic gains. However, the majority of aquaculture producers in developing countries are deterred from employing genomic selection as a routine practice for breeding due to the high costs associated with genotyping or sequencing. To expedite the adoption of genomic selection in aquaculture breeding programmes, particularly for small and medium-sized companies and species with low or medium individual value, new cost-effective strategies are required to reduce these expenses while maintaining high genomic prediction accuracy, thereby increasing genetic gain.

The aim of my PhD thesis was to give valuable insights for the design and evaluation of cost-effective genotyping strategies and breeding programme designs for different species, which will enable aquaculture to fully benefit from genomic selection. To accomplish that aim, I explored and tested different methods to lower the cost associated with genotyping for genomic selection. This thesis exploited low-density genotyping and low-coverage WGS imputation to establish best practices for their use in aquaculture breeding programmes and other genomic studies. In addition to genotype imputation, this thesis also investigated the incorporation of functional genomic information to improve prediction models. This topic remains largely unexplored in aquatic species and is still developing in general.

The main findings and concluding remarks of the three chapters are summarised in the following sections.

5.1 Cost-effective genomic selection by imputation of LD panels

In the first results chapter, I explored genotype imputation of LD panels in four aquaculture species, namely Atlantic salmon, turbot, common carp and Pacific oyster. In total, eight LD panels were generated for each of the three SNP selection scenarios based on linkage disequilibrium (genetic distance-based method), evenly distributed according to physical position (physical distance-based method) and randomly selected. These panels were subsequently imputed using three available genotype imputation software. Genomic prediction accuracy of different traits was tested for each species with the imputed genotypes of the most favourable scenario.

I observed promising results that can be applied to the design of LD panels for commercial use. The key findings were:

- Imputation accuracies were very high for the three fish species (0.76 - 0.98 for the 300 and 6000 SNPs LD panel, respectively), while the performance of imputation was lower in our oyster dataset (0.61 - 0.80 for the 300 and 6000 SNPs LD panel, respectively).
- For the populations tested, no significant differences were observed in imputation accuracy between the physical and genetic distance-based LD panel SNP selection methods.
- In most scenarios tested, FImpute v.3 was the fastest and most accurate imputation method, when compared to findhap v.4 and AlphaImpute v.2.
- When LD panels were used without imputation, those with SNPs evenly distributed across chromosomes achieved prediction accuracies similar to the HD panel in three fish species, even with just 300 SNPs, while randomly selected LD panels yielded lower prediction accuracies.
- Imputation significantly increased the prediction accuracy of the randomly selected LD panels, reaching values similar to those of the HD panel in the fish species.

5.2 Imputation of lcWGS as a cost-effective alternative to WGS

In this chapter I assessed the accuracy of imputation to WGS using different sequencing depths in a Nile tilapia population. Six *in silico* down-sampled datasets (0.1X, 0.2X, 0.5X, 1X, 2X, and 5X sequencing depth) of the target offspring were imputed with GLIMPSE v.1 to the 5X and 26X reference panels containing, among others, the parents of the target population. Finally, the cost of the various lcWGS datasets was compared to that of SNP arrays in a hypothetical scenario involving 140 parents and 2100 offspring.

The key findings were:

- The total number of retained SNPs after imputation and filtering was higher with a reference sequencing depth of 26X compared to 5X.
- Imputation accuracy for all down-sampled target datasets exceeded 0.9, with slightly higher accuracy observed when genotypes were imputed to the 26X reference panel compared to 5X.
- Higher accuracies were observed at homozygous sites compared to heterozygous sites for target population coverages below 5X.
- While lcWGS is cheaper than WGS, it remains more expensive than SNP arrays.
- Imputation of lcWGS to higher coverage yields a greater number of SNPs compared to SNP arrays, facilitating the identification of causative variants for subsequent GWAS.

5.3 Assessing the influence of functional annotation on genomic prediction

For this chapter I used a dataset of a turbot population challenged with the parasite *Plilasterides dicentrarchi* from a previous project (Anacleto *et al.*, 2019). The genotypes of the offspring were imputed to the WGS parents and different scenarios of incorporating functional information into genomic prediction were tested. For the creation of functional annotation categories, I prioritised markers overlapping with regions of the genome potentially affecting protein function or promoter and enhancer regions. Two Bayesian models were tested with the

different scenarios and GBLUP was used as a reference for comparison.

The key findings were:

- Incorporating functional annotation information into the Bayesian models did not improve genomic prediction accuracy.
- GBLUP outperformed or slightly surpassed the Bayesian models in some cases.
- Alternative methods of defining annotation categories or applying the same method to different traits with varying architectures should be tested in the future.

5.4 Future prospects and limitations of imputation

Studies in the past have indicated that genomic prediction accuracy tends to improve with higher marker density. However, while this generally holds true for livestock populations with a limited number of closely related individuals, it tends not to be the case for aquatic species. Aquatic species produce numerous offspring, with individuals from different families sharing the same parent and thus exhibiting some degree of relatedness. For routine genomic predictions, only a subset of these variants is necessary, and a higher number of markers typically leads to a plateau in prediction accuracy (Hickey *et al.*, 2014; VanRaden *et al.*, 2017). In our previous study, we observed that accurate predictions could be achieved for closely related individuals across different aquaculture species using a relatively small number of markers, approximately 1000-2000 SNPs (Kriaridou *et al.*, 2020).

The use of imputation in genomic prediction has been explored across diverse farmed crops and animals and is now widely employed in the genetic evaluation of cattle to increase the available genotype data to higher densities (Kamprasert *et al.*, 2024). Compared to other livestock, the integration of imputation as a standard practice in aquaculture breeding programmes has only recently gained attention. This method could be applied in farms that already use low-density panels (300-500 SNPs) for parentage assignment (to control inbreeding and perform pedigree-based selection) with minimum extra cost.

However, these low-density panels should be adapted and re-designed to be suitable for both parentage assignment and imputation. Ideally, they should contain SNPs spread throughout the genome and overlap with the available high-density panel. This approach significantly reduces costs compared to genotyping all individuals at high density. Although higher-density genotyping is necessary for parents, the reduced number of individuals minimizes overall costs. Considering the long-term genetic gain a farm will see in its production, the improved animals are likely to compensate the additional genotyping cost.

The first objective of my thesis was to investigate if we could further lower the number of markers found in our previous study (Kriaridou *et al.*, 2020) and use imputation to infer the missing information and achieve equally accurate prediction accuracies with the high-density panels. The results demonstrate that cost-effective genomic selection can be achieved through this integrated approach for some species and traits even with non-imputed panels, by selecting the SNPs to be equally distributed throughout the genome and proportionally to chromosome length. However, imputation can improve genomic prediction accuracy of low-density panels in cases where SNPs are not optimally selected or some of the SNPs in the LD panel are filtered out because their genotyping quality is low. Imputation accuracy can play a significant role in obtaining accurate genomic breeding values and for some species, such as the Pacific oyster, low imputation accuracy can result in lower prediction accuracies and subsequently reduced genetic gain. Further research is needed to advance this strategy of imputing genomes in different species, especially in marine invertebrates, whose genomes are highly repetitive, exhibit high recombination and mutation rates (Abdelrahman *et al.*, 2017; Hollenbeck and Johnston, 2018; Song, 2020; Durland, De Wit and Langdon, 2022; Zhang *et al.*, 2023), if we want to use it routinely in breeding programmes.

Another approach, which uses low-coverage WGS and imputation, was also evaluated as a cost-effective alternative to WGS for population genetics studies, fine mapping of quantitative trait loci and other genetic studies. The advantages of lcWGS include the detection of novel population-specific variants, unbiased data generation across populations, and the potential discovery of causal markers (Martin *et al.*, 2021). This approach also seems promising since

we achieved high accuracies with low-coverage sequencing that was much cheaper compared to standard WGS. However, I did not evaluate prediction accuracy in this study and tested only three of the twenty-two chromosomes of the Nile tilapia population. Like in the case of imputation of low-density panels, imputation of low-coverage WGS can also yield low imputation accuracy results, particularly at heterozygous loci when using really low-coverages (<2X). Therefore, caution should be exercised when using imputed data in subsequent analyses, as it may compromise results, especially those dependent on allele frequencies.

Numerous studies aimed to achieve highly accurate imputed genotypes, recognizing that inaccuracies can significantly impact subsequent analyses. Given the potential downstream negative impact of low imputation accuracy, researchers have explored various factors influencing the performance of genotype imputation. In this thesis, I investigated the impact of imputation software, SNP selection method, number of SNPs and low-coverage whole-genome sequencing in different aquaculture species and traits. One of the factors that we have not tested but has been tested substantially across the different imputation studies of other species, is the impact of the reference panel on imputation accuracy.

The reference panel composition, along with the imputation software methodology, plays a crucial role in accurately inferring genotypes. When the reference panel comprises immediate ancestors, such as parents of the target population, haplotype construction becomes more relevant and precise for these closely related individuals. This is because longer haplotype blocks are shared, and even a small number of animals suffice for highly accurate imputation (Hayes *et al.*, 2012; Sargolzaei, Chesnais and Schenkel, 2014). If parents are unavailable for imputation, but data from individuals within the same or different populations are available, factors to explore, previously addressed in other studies, involve determining the optimal number of animals to include in the reference panel and their relatedness to the target individuals. For instance, Garcia *et al.* (2022) evaluated the reference size and origin as factors that may influence genotype imputation accuracy in Nile tilapia using WGS data. They observed that increasing the number of animals in the reference with animals from the sample population significantly enhanced imputation accuracy. Nevertheless, animals from diverse

populations also provided valuable information and can be utilized for WGS imputation. However, individuals from the same population still needed to be present to achieve high imputation accuracies. Increasing the number of animals used as reference, especially those from ancestral lines, can facilitate for a more accurate imputation process.

Bouwman and Veerkamp (2014) proposed the idea of the creation of a cattle “reference bank” for multibreed imputation of high-density SNP panels to WGS, when the reference population is small and there is a limited budget for genotyping or sequencing. This philosophy was implemented by the 1000 Bull Genomes project which is frequently used to impute large cohorts of genotyped cattle, enabling powerful genome-wide analyses (Daetwyler *et al.*, 2014). Examples of similar applications in other species include the 1000 Genomes project (Auton *et al.*, 2015) and the Haplotype Reference Consortium (McCarthy *et al.*, 2016), which serve as reference databases comprising various human ancestry populations. These resources were incorporated into a web-based service for computationally efficient imputation (Das *et al.*, 2016). Similarly in pigs, a haplotype reference panel named PHARP was developed for imputation (Wang *et al.*, 2022). This panel incorporates data from over 49 studies covering 71 pig breeds. Additionally, a recently launched public web server called SWine IMputation (SWIM), which includes more individuals from more pig breeds, aims to streamline imputation and genetic mapping processes (Ding *et al.*, 2023). In this study Ding *et al.* (2023) demonstrated the ability of this new public server to achieve more accurate imputation compared to previous databases. They also showcase two studies as an example of the possible uses of this database. In the first they successfully discovered the suggestive causative mutation for backfat thickness after imputation of a SNP chip, from which the causative SNP was absent, to whole-genome genotypes with the accuracy of the imputed genotypes for this SNP reaching 99.71%. In the second example, they used the imputed to whole-genome genotypes to perform a GWAS for body length in Yorkshire boars. They discovered a regulatory variant upstream the BMP2 gene, which explained 13.65% of the total variance compared to the most significant SNP from the SNP chip based GWAS which explained 8.22%.

Researchers have also created multispecies databases. Animal-ImputeDB

is another public database with reference panels of 13 species (with 19 to 658 individuals per species) providing the ability to utilise this service for online genotype imputation (Yang *et al.*, 2020). For aquaculture species there is a web server for genotype imputation and genetic analysis created from available datasets of 18 aquaculture species with 40 to 218 individuals per species, called Aquaculture Molecular Breeding Platform (AMBP) (Zeng *et al.*, 2022).

Databases such as those for humans or bovines encompass a vast number of individuals. However, most of the databases for other species lack diversity across different breeds and populations worldwide, with the number of individuals often being too small to serve as representative references for imputation, particularly for local or rare populations. My suggestion for future studies would be the collaboration of Institutions and companies towards the implementation of a big project, similar to the 1000 Genomes projects mentioned above, for aquatic species. Combining data from multiple populations to create sufficiently large reference populations, representative of various populations within each aquaculture species, may enhance relevant platforms for more accurate imputation. Such a big project will also contribute towards fine mapping and the potential for discovery of causative mutations for the different traits of interest. Nonetheless, the construction and utilization of such large multi-population reference panels should be carefully considered before their design, as they may introduce other issues. This is because useful multiallelic markers will be usually filtered out before imputation (Lloret-Villas, Pausch and Leonard, 2023).

5.5 Future prospects and limitations of the incorporation of functional annotation in prediction models

Functional mechanisms are the intermediate steps connecting genotypes to phenotypes. Our understanding of these mechanisms is still limited and this sometimes hinders our ability to accurately predict the different complex traits, such as disease resistance, that are important for aquaculture breeding. Prediction models that prioritise variants playing a major role in the genetic architecture of complex traits have potential to increase the accuracy of genomic selection (Xiang *et al.*, 2021). However, to investigate this hypothesis the

generation of functional annotation information is needed for the non-coding genomic sequences that control gene expression (Albert and Kruglyak, 2015), enabling the prediction of the impact of genetic variation and prioritisation of genetic markers.

Many aquaculture species lack functional annotation information, and only recently the AQUA-FAANG (https://www.aqua-faang.eu/) project aimed to address this issue for six important aquaculture species. Consequently, the hypothesis regarding the incorporation of functional annotation can now be explored for these species using the data provided by this project. However, to make such research feasible, the availability of WGS data for a large number of individuals is necessary to identify causative variants for each trait. While WGS remains costly, the combination of low-coverage WGS with previously described imputation methods provides a means to reduce this cost and recover missing information.

The main challenge with imputation, as discussed previously, is imputation accuracy, and we are particularly interested in imputation accuracy of causal variants, if we want to capture them correctly. Imputing to WGS presents particular challenges due to the high number of markers involved. To address this, an intermediate step of imputation to medium density is often employed in other studies, especially when the marker density of the target population is very low, to avoid poor imputation (Carvalheiro *et al.*, 2014; Garcia *et al.*, 2022). Stepwise imputation from low-density panels to medium or high, and finally to sequence level, can considerably improve imputation accuracy. This is achieved by reducing the difference in marker numbers between the reference and the target densities. The extra information about where linkage disequilibrium blocks break, provided by the markers of the medium density, contributes towards more accurate results. Nevertheless, even this imputation scenario is currently unrealistic or impractical for routine application in aquaculture breeding programmes. The reason is that it requires genotyping of individuals with a medium to high density SNP arrays and WGS of the reference individuals, thereby dramatically increasing costs.

Multiple potential benefits are associated with the advancement of aquaculture breeding through the utilization of functional annotation. These include the development of customised genotyping arrays informed by the biology

of the traits, improved portability of genetic data across generations to reduce animal testing in disease challenges, and identification of possible targets for genome editing (Johnston *et al.*, 2024).

Recent years have seen attempts to design custom arrays informed by the biology of the underlying mechanisms affecting the desired traits. Xiang *et al.* (2021) in their paper, fine-mapped genome-wide informative sequence variants with pleiotropic effects and functional significance. They then selected ~50K markers and used them to predict multiple traits in different cattle breeds and populations. Their panel showed increased accuracy in predicting genetic values for multiple traits compared to the standard array across multiple cattle breeds. Another study in humans by Amariuta *et al.* (2020), constructed a resource of IMPACT[1] regulatory annotations and prioritised functional variants, shared across populations. The prioritisation of variants using IMPACT, improved the accuracy of polygenic risk score predictions from Europeans to East Asians for all the 21 tested traits by a 49.9% mean increase. This is a very important outcome as it shows the potential of the incorporation of functional annotation information into the panels resulting in improved genomic selection across populations or generations.

Despite these advantages, we are still in the early stages of research. To systematically incorporate functional information into prediction models for various traits, requires extensive research to develop models that can effectively incorporate and interpret such information, while also generating more information for the genomes of various aquaculture species. For these strategies to be incorporated into future breeding programmes, they must outperform the standard GBLUP, which currently performs quite well at a significantly lower cost and computational time.

[1] A genome annotation strategy that identifies regulatory elements defined by cell-state-specific transcription factor binding profiles (Amariuta *et al.*, 2019).

5.6 Concluding remarks

In this thesis, I investigated three approaches to reduce the cost of implementing genomic information in breeding programmes and increase the accuracy of genomic prediction. My research will have a substantial impact on LD genotyping and the incorporation of imputation into breeding programmes to reduce genotyping costs. Genotyping prices have dropped dramatically since the beginning of my PhD, making it more feasible to genotype only a few individuals with high or medium density panels, serving as reference for imputation. A recent example of such an application is the development of a custom LD SNP panel by the AQUA-FAANG consortium for European sea bass (Johnston *et al.*, 2024). This panel, consisting of 350 to 400 SNPs, is mainly aimed for pedigree-based selection, but it can serve as a cheap LD panel for imputation. Similar projects can be designed and tested for their effectiveness in practice.

Despite the higher price of lcWGS compared to SNP arrays, which makes it expensive for routine use in genomic selection, it offers several advantages. After imputation to WGS genotypes, we can obtain a plethora of SNPs, some of which will be unique to the sequenced population. A combined method of low-coverage sequencing and imputation, investigated in this thesis, can help identify causal variants which can subsequently expand our annotation categories for their use into Bayesian models. These discoveries can serve as a starting point for future projects related to the design of custom SNP arrays. These arrays will utilize all this valuable information for more accurate prediction of breeding values.

The attempt to incorporate functional annotation into the Bayesian models in this thesis did not improve genomic prediction accuracy in the tested scenarios. Nonetheless, there were several limitations to the available information and datasets that were used. To enhance similar experiments in the future, it would be beneficial to impute the target individuals from a higher density. This would ensure highly accurate imputation results, especially considering the substantial difference in SNP numbers between the RAD-seq data and SNPs called from WGS. There is also an enormous possibility of combinations and scenarios concerning the biological information we incorporate into the prediction models. We only investigated a small number of possible scenarios, as these models

require a significant amount of computational time, which scales depending on the amount of information provided to them. However, according to the encouraging results of some other studies, the potential of functional annotation should be further investigated for other aquaculture species, using a combination of different sources of information (like QTLs, gene expression and other omics data) testing different traits and populations. This is now possible with the availability of open access information provided by the AQUA-FAANG project for six important aquaculture species (European seabass, gilthead seabream, rainbow trout, Atlantic salmon, common carp and turbot).

If we want to utilise genomic selection in commercial breeding programmes, we have to determine strategies that are viable for most companies. The successful integration of genomic selection should offer return of investment and that is what future research should address. This will make possible to achieve higher gains with lower prices by utilising precision breeding tailored to the species, informed by the biology and the underlying mechanisms affecting the phenotype of individuals.

Appendix

6.1 Publications and conferences

Publications

Kriaridou, Christina, Tsairidou, Smaragda, Fraslin, Clémence, Gorjanc, Gregor, Looseley, Mark E., Johnston, Ian A., Houston, Ross D., and Robledo, Diego. 2023. “Evaluation of Low-Density SNP Panels and Imputation for Cost-Effective Genomic Selection in Four Aquaculture Species.” Frontiers in Genetics 14 (May): 1194266. https://doi.org/10.3389/fgene.2023.1194266.

Houston, Ross D., **Kriaridou,** Christina, and Robledo, Diego. 2022. “Animal Board Invited Review: Widespread Adoption of Genetic Technologies Is Key to Sustainable Expansion of Global Aquaculture.” Animal 16 (10): 100642. https://doi.org/10.1016/J.ANIMAL.2022.100642.

Oral presentations in conferences

Kriaridou C., Tsairidou S., Fraslin C., Looseley M. E., Johnston I. A., Houston R. D., Robledo D. 2022. Optimising Genomic Selection for Aquaculture Breeding Programmes in Small-Scale Operations and Developing Countries. 6th Genomics in Aquaculture (GIA) Symposium. Granada, Spain, 4th – 6th May.

Kriaridou C., Tsairidou S., Fraslin C., Gorjanc G., Looseley M. E., Johnston I. A., Houston R. D., Robledo D. 2022. Evaluation of Low-Density Panels and Imputation for Cost-Effective Genomic Selection in Four Aquaculture Species. 14th International Symposium on Genetics in Aquaculture. Puerto Varas, Chile, 27th November – 2nd December.

Kriaridou C., Fraslin C., Jiao Z., Prentice J., Aramburu O., Pong-Wong R., Gorjanc G., Looseley M. E., Johnston I. A., Saura M., Villanueva B., Millan A., Macqueen D. J., Martínez P., Doeschl-Wilson A. B., Robledo D. 2023. Incorporation of whole-genome sequencing and functional information to refine quantitative trait loci and improve prediction accuracy for controlling scuticociliatosis (*Philasterides dicentrarchi*) in turbot (Scophthalmus maximus). Aquaculture Europe 2023 (AE2023) - Balanced Diversity in Aquaculture Development. Vienna, Austria, 18th – 21st September.

Kriaridou C., Fraslin C., Jiao Z., Prentice J., Aramburu O., Pong-Wong R., Gorjanc G., Looseley M. E., Johnston I. A., Saura M., Villanueva B., Millan A., Macqueen D. J., Martínez P., Doeschl-Wilson A. B., Robledo D. 2024. Impact of the incorporation of functional annotation into genomic prediction in a turbot population challenged with *Philasterides dicentrarchi*. 7th Genomics in Aquaculture (GIA) Symposium. Thessaloniki, Greece, 22nd – 24th May.

6.2 Courses, workshops and awards

Courses

Attended the following Quantitative Genetics and Genome Analysis (MSc) courses from University of Edinburgh in the first year of my PhD:

- 1st Semester: Statistics and Data Analysis, Population and Quantitative Genetics, Genetic Interpretation
- 2nd Semester: Animal Genetic Improvement, Linkage and Association in Genome Analysis, Quantitative Genetic Models

Workshops

- Completed the University of Edinburgh “Developing your Data Skills Programme” (Level 3), November 2021 – May 2022.
- Attended the University of Georgia (UGA) course “Programming and computer algorithms in animal breeding with a focus on single-step GBLUP and genomic selection in practice”, Athens, Georgia, 25th May – 3rd June 2022.
- Supported the Easter Bush Science Outreach Centre (EBSOC) team for the hands-on career-long professional learning workshop, for secondary school science teachers “PCR Masterclass: A Question of Taste”, 22nd June 2022.
- Participated and assisted in the practical course “Increasing researcher´s skills on the use of omics tools”, Portugal, 16th -17th March 2023.
- Participated in the 2-week Workshop on Genomics, being held in Cesky Krumlov, Czech Republic, 14th – 27th May 2023.
- Contributed to the AquaIMPACT training course on “Genomic Innovations for Aquaculture Breeding“ in Wageningen University, Netherlands, 26th – 27th June 2023, with a Lecture and a Practical about Genotype Imputation: https://projects.luke.fi/aquaimpact/aquaimpact-training-course-on-genomic-innovations-for-aquaculture-breeding/

The workshop material I prepared for the training course can be found on my GitHub page: https://github.com/ChristinaKriaridou/Genotype-Imputation-with-FImpute

During my PhD I was recruited as a moderator for the online Global Academy's MSc programme in Agricultural Science. My role involved providing support to the teaching staff during online teaching sessions. Furthermore, I was appointed as an invigilator by the University for student exams at the School of Veterinary Studies. In May of 2023, I also became a member of the Roslin internal seminar committee, where I have taken on the responsibility of organising and managing calendar invites for speakers.

Awards

The School of Veterinary Studies has awarded me with the Birrell-Gray travelling scholarship fund. The award is to the value of £500 to attend the International Symposium on Genetics in Aquaculture - ISGA XIV, 27th Nov-2nd Dec 2022, held in Puerto Varas, Chile. Additionally, my abstract entitled "Evaluation of low-density panels and imputation for cost-effective genomic selection in four aquaculture species" was selected for a fellowship in the ISGA Symposium. This fellowship covered the registration and accommodation expenses for six days.

References

Abdelrahman, H. *et al.* (2017) 'Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research', *BMC Genomics 2017 18:1*, 18(1), pp. 1–23. doi: 10.1186/S12864-017-3557-1.

Aguilar, I. *et al.* (2010) 'Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score', *Journal of dairy science*, 93(2), pp. 743–752. doi: 10.3168/JDS.2009-2730.

Ahmed, N., Thompson, S. and Glaser, M. (2019) 'Global Aquaculture Productivity, Environmental Sustainability, and Climate Change Adaptability', *Environmental Management*, 63, pp. 159–172. doi: 10.1007/s00267-018-1117-3.

Al-Tobasei, R. *et al.* (2021) 'Genomic predictions for fillet yield and firmness in rainbow trout using reduced-density SNP panels', *BMC Genomics*, 22(1), pp. 1–11. doi: 10.1186/s12864-021-07404-9.

Albert, F. W. and Kruglyak, L. (2015) 'The role of regulatory variation in complex traits and disease'. doi: 10.1038/nrg3891.

Alday-Sanz, V. *et al.* (2020) 'Facts, truths and myths about SPF shrimp in Aquaculture', *Reviews in Aquaculture*, 12(1), pp. 76–84. doi: 10.1111/RAQ.12305.

Aliloo, H. *et al.* (2018) *Optimal design of low density marker panels for genotype imputation*, *Proceedings of the World Congress on Genetics Applied to Livestock Production*. Available at: https://cgspace.cgiar.org/handle/10568/98244 (Accessed: 2 November 2020).

Aljanabi, S. M. and Martinez, I. (1997) 'Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques.', *Nucleic Acids Reseach*, 25(22).

Allayee, H. *et al.* (2023) 'Systems genetics approaches for understanding complex traits with relevance for human disease', *eLife*, 12. doi: 10.7554/ELIFE.91004.

Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of molecular biology*, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Amariuta, T. *et al.* (2019) 'IMPACT: Genomic Annotation of Cell-State-Specific Regulatory Elements Inferred from the Epigenome of Bound Transcription Factors', *American Journal of Human Genetics*, 104(5), p. 879. doi: 10.1016/J.AJHG.2019.03.012.

Amariuta, T. *et al.* (2020) 'Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements', *Nature genetics*, 52(12), p. 1346. doi: 10.1038/S41588-020-00740-8.

Anacleto, O. *et al.* (2019) 'Genetic differences in host infectivity affect disease spread and survival in epidemics', *Scientific Reports*, 9(1), pp. 1–12. doi: 10.1038/s41598-019-40567-w.

Antolín, R. *et al.* (2017) ‘A hybrid method for the imputation of genomic data in livestock populations’, *Genetics Selection Evolution*, 49(1), p. 30. doi: 10.1186/s12711-017-0300-y.

Auton, A. *et al.* (2015) ‘A global reference for human genetic variation’, *Nature 2015 526:7571*, 526(7571), pp. 68–74. doi: 10.1038/nature15393.

Bargelloni, L. *et al.* (2021) ‘Data imputation and machine learning improve association analysis and genomic prediction for resistance to fish photobacteriosis in the gilthead sea bream’, *Aquaculture Reports*, 20, p. 100661. doi: 10.1016/J.AQREP.2021.100661.

Barría, A. *et al.* (2020) ‘Genetic parameters for resistance to Tilapia Lake Virus (TiLV) in Nile tilapia (Oreochromis niloticus)’, *Aquaculture*, 522, p. 735126. doi: 10.1016/J.AQUACULTURE.2020.735126.

Barton, N. H., Etheridge, A. M. and Véber, A. (2017) ‘The infinitesimal model: Definition, derivation, and implications’, *Theoretical Population Biology*, 118, pp. 50–73. doi: 10.1016/J.TPB.2017.06.001.

Bentsen, H. B. *et al.* (2017) ‘Genetic improvement of farmed tilapias: Response to five generations of selection for increased body weight at harvest in Oreochromis niloticus and the further impact of the project’, *Aquaculture*. doi: 10.1016/j.aquaculture.2016.10.018.

Bertolotti, A. C. *et al.* (2020) ‘The structural variation landscape in 492 Atlantic salmon genomes’. doi: 10.1038/s41467-020-18972-x.

Boudry, P. *et al.* (2021) ‘Current status and potential of genomic selection to improve selective breeding in the main aquaculture species of International Council for the Exploration of the Sea (ICES) member countries’, *Aquaculture Reports*, 20, p. 100700. doi: 10.1016/j.aqrep.2021.100700.

Bouwman, A. C. *et al.* (2014) ‘Imputation of non-genotyped individuals based on genotyped relatives: Assessing the imputation accuracy of a real case scenario in dairy cattle’, *Genetics Selection Evolution*, 46(1), pp. 1–11. doi: 10.1186/1297-9686-46-6.

Bouwman, A. C. and Veerkamp, R. F. (2014) ‘Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy’, *BMC Genetics*, 15(1), pp. 1–9. doi: 10.1186/S12863-014-0105-8/TABLES/6.

Boyd, C. E. *et al.* (2020) ‘Achieving sustainable aquaculture: Historical and current perspectives and future needs and challenges’, *Journal of the World Aquaculture Society*, 51(3), pp. 578–633. doi: 10.1111/JWAS.12714.

Brække, N. (2023) *Whole genome sequencing reveals development of structured salmon lice (Lepeophtheirus salmonis, Krøyer, 1838) populations among aquaculture net pens through production*. UiT The Arctic University of Norway. Available at: https://munin.uit.no/handle/10037/29364 (Accessed: 22 March 2024).

Browning, S. R. (2008) ‘Missing data imputation and haplotype phase inference for genome-wide association studies’, *Human Genetics*. Hum Genet, pp. 439–450. doi: 10.1007/s00439-008-0568-7.

Buckley, R. M. *et al.* (2022) 'Best practices for analyzing imputed genotypes from low-pass sequencing in dogs', *Mammalian genome : official journal of the International Mammalian Genome Society*, 33(1), pp. 213–229. doi: 10.1007/S00335-021-09914-Z.

Cai, J. N., Yan, X. and Leung, P. S. (2022) *Benchmarking species diversification in global aquaculture. Fisheries and Aquaculture Technical Paper No. 605. Rome, FAO*, *Benchmarking species diversification in global aquaculture*.

Carvalheiro, R. *et al.* (2014) 'Accuracy of genotype imputation in Nelore cattle', *Genetics Selection Evolution*, 46(1), pp. 1–11. doi: 10.1186/S12711-014-0069-1/FIGURES/3.

Castillo-Juárez, H. *et al.* (2015) 'Genetic improvement of Pacific white shrimp (Penaeus (Litopenaeus)vannamei): Perspectives for genomic selection', *Frontiers in Genetics*, 5(FEB), p. 93. doi: 10.3389/FGENE.2015.00093/BIBTEX.

Chat, V. *et al.* (2022) 'Ultra Low-Coverage Whole-Genome Sequencing as an Alternative to Genotyping Arrays in Genome-Wide Association Studies', *Frontiers in genetics*, 12. doi: 10.3389/FGENE.2021.790445.

Chen, L. *et al.* (2021) 'Research advances and future perspectives of genomics and genetic improvement in allotetraploid common carp', *Reviews in Aquaculture*. doi: 10.1111/RAQ.12636.

Chen, W., Coombes, B. J. and Larson, N. B. (2022) 'Recent advances and challenges of rare variant association analysis in the biobank sequencing era', *Frontiers in Genetics*, 13, p. 1014947. doi: 10.3389/FGENE.2022.1014947/BIBTEX.

Chew, J. S. K. *et al.* (2002) 'Molecular and cytogenetic analysis of the telomeric (TTAGGG)n repetitive sequences in the Nile tilapia, Oreochromis niloticus (Teleostei: Cichlidae)', *Chromosoma*, 111(1), pp. 45–52. doi: 10.1007/S00412-002-0187-3.

Christensen, O. F. and Lund, M. S. (2010) 'Genomic prediction when some animals are not genotyped', *Genetics Selection Evolution*, 42(3), pp. 1–8. doi: 10.1186/1297-9686-42-2/TABLES/1.

Chud, T. C. S. *et al.* (2015) 'Strategies for genotype imputation in composite beef cattle', *BMC Genetics*, 16(1), pp. 1–10. doi: 10.1186/s12863-015-0251-7.

Cingolani, P. *et al.* (2012) 'A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3', *Fly*, 6(2), pp. 80–92. doi: 10.4161/FLY.19695.

Cleveland, M. A. and Hickey, J. M. (2014) 'Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation', *Journal of Animal Science*, 91(8), pp. 3583–3592. doi: 10.2527/JAS.2013-6270.

Conte, M. A. *et al.* (2017) 'A high quality assembly of the Nile Tilapia (Oreochromis niloticus) genome reveals the structure of two sex determination regions', *BMC Genomics*, 18(1), pp. 1–19. doi: 10.1186/S12864-017-3723-5/TABLES/9.

Conte, M. A. *et al.* (2019) ‘Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes’, 8, pp. 1–20. doi: 10.1093/gigascience/giz030.

Conte, M. A. *et al.* (2021) ‘Origin of a Giant Sex Chromosome’, *Molecular biology and evolution*, 38(4), pp. 1554–1569. doi: 10.1093/MOLBEV/MSAA319.

Cottrell, R. S. *et al.* (2021) ‘The search for blue transitions in aquaculture-dominant countries’, *Fish and Fisheries*, pp. 1006–1023. doi: 10.1111/faf.12566.

Daetwyler, H. D. *et al.* (2007) ‘Inbreeding in genome-wide selection’, *Journal of Animal Breeding and Genetics*, 124(6), pp. 369–376. doi: 10.1111/J.1439-0388.2007.00693.X.

Daetwyler, H. D. *et al.* (2010) ‘The Impact of Genetic Architecture on Genome-Wide Evaluation Methods’, *Genetics*, 185(3), pp. 1021–1031. doi: 10.1534/GENETICS.110.116855.

Daetwyler, H. D. *et al.* (2014) ‘Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle’, *Nature Genetics*, 46(8), pp. 858–865. doi: 10.1038/NG.3034.

Darwin, C. (2009) ‘CLASSICS On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life’.

Das, S. *et al.* (2016) ‘Next-generation genotype imputation service and methods’, *Nature genetics*, 48(10), pp. 1284–1287. doi: 10.1038/NG.3656.

Davies, R. W. *et al.* (2016) ‘Rapid genotype imputation from sequence without reference panels’, *Nature genetics*, 48(8), p. 965. doi: 10.1038/NG.3594.

Davies, R. W. *et al.* (2021) ‘Rapid genotype imputation from sequence with reference panels’, *Nature genetics*, 53(7), pp. 1104–1111. doi: 10.1038/S41588-021-00877-0.

Delomas, T. A. *et al.* (2023) ‘Evaluating cost-effective genotyping strategies for genomic selection in oysters’, *Aquaculture*, 562, p. 738844. doi: 10.1016/J.AQUACULTURE.2022.738844.

Desta, Z. A. and Ortiz, R. (2014) ‘Genomic selection: genome-wide prediction in plant improvement’, *Trends in Plant Science*, 19(9), pp. 592–601. doi: 10.1016/J.TPLANTS.2014.05.006.

Díaz, P. A. *et al.* (2019) ‘Impacts of harmful algal blooms on the aquaculture industry: Chile as a case study’, *Perspectives in Phycology*, 6(2), pp. 39–50. doi: 10.1127/pip/2019/0081.

Ding, R. *et al.* (2023) ‘The SWine IMputation (SWIM) haplotype reference panel enables nucleotide resolution genetic mapping in pigs’, *Communications Biology 2023 6:1*, 6(1), pp. 1–10. doi: 10.1038/s42003-023-04933-9.

Driscoll, C. A., Macdonald, D. W. and O’Brien, S. J. (2009) ‘From wild animals to domestic pets, an evolutionary view of domestication.’, *Proceedings of the National Academy of Sciences of the United States of America*, 106 Suppl 1(supplement_1), pp. 9971–9978. doi: 10.1073/PNAS.0901586106/ASSET/2AC274AF-27B5-4BE2-AD4F-

952CAAA7E063/ASSETS/GRAPHIC/ZPQ9990983330002.JPEG.

Druet, T., Schrooten, C. and de Roos, A. P. W. (2010) 'Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle', *Journal of Dairy Science*, 93(11), pp. 5443–5454. doi: 10.3168/jds.2010-3255.

Dunham, I. *et al.* (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature 2012 489:7414*, 489(7414), pp. 57–74. doi: 10.1038/nature11247.

Dupont-Nivet, M. *et al.* (2006) 'Effect of different mating designs on inbreeding, genetic variance and response to selection when applying individual selection in fish breeding programs', *Aquaculture*, 252(2–4), pp. 161–170. doi: 10.1016/J.AQUACULTURE.2005.07.005.

Durland, E., De Wit, P. and Langdon, C. (2022) 'Genetic changes in larval oysters are more abundant and dynamic than can be explained by rare events or error: a response to Hedgecock (2022)', *Proceedings of the Royal Society B: Biological Sciences*, 289(1976). doi: 10.1098/RSPB.2022.0197.

Ehsani, A. *et al.* (2012) 'Inferring genetic architecture of complex traits using Bayesian integrative analysis of genome and transcriptome data', *BMC Genomics*, 13(1), pp. 1–9. doi: 10.1186/1471-2164-13-456/TABLES/3.

Erbe, M. *et al.* (2012) 'Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels', *Journal of Dairy Science*, 95(7), pp. 4114–4129. doi: 10.3168/jds.2011-5019.

Ernst, J. and Kellis, M. (2017) 'Chromatin-state discovery and genome annotation with ChromHMM', *Nature Protocols 2017 12:12*, 12(12), pp. 2478–2492. doi: 10.1038/nprot.2017.124.

Etherington, G. J. *et al.* (2022) 'Chromosome-level genome sequence of the Genetically Improved Farmed Tilapia (GIFT, Oreochromis niloticus) highlights regions of introgression with O. mossambicus', *BMC Genomics*, 23(1), pp. 1–16. doi: 10.1186/S12864-022-09065-8/TABLES/2.

Falconer, D. S. and Mackay, T. F. C. (1996) 'Introduction to Quantitative Genetics (Fourth Edition)', *Trends in Genetics*, 12, p. 464.

FAO (2017) 'FAO FISHERIES AND AQUACULTURE PROCEEDINGS Planning for aquaculture diversification: the importance of climate change and other drivers'.

FAO (2020) *The State of World Fisheries and Aquaculture 2020. Sustainability in action.* Rome. doi: 10.4060/ca9229en.

FAO (2022) *The State of World Fisheries and Aquaculture 2022. Towards Blue Transformation.* Rome. doi: 10.4060/cc0461en.

FAO (2023) 'Top 10 species groups in global, regional and national aquaculture 2021'. Available at: https://www.fao.org/3/cc6319en/cc6319en.pdf (Accessed: 3 October 2023).

Farias, T., César, J. and Silva, L. (2017) 'Methods of Selection Using the

Quantitative Genetics in Aquaculture-A Short Review', *Insights in Aquaculture and Biotechnology*, 1(1), p. 464. Available at: https://www.imedpub.com/articles/methods-of-selection-using-the-quantitative-genetics-in-aquaculturea-short-review.php?aid=18927 (Accessed: 16 December 2023).

Fernandes Júnior, G. A. *et al.* (2021) 'Imputation accuracy to whole-genome sequence in Nellore cattle', *Genetics Selection Evolution*, 53(1), pp. 1–10. doi: 10.1186/s12711-021-00622-5.

Fragomeni, B. O. *et al.* (2017) 'Incorporation of causative quantitative trait nucleotides in single-step GBLUP', *Genetics, selection, evolution : GSE*, 49(1). doi: 10.1186/S12711-017-0335-0.

Fraslin, C. *et al.* (2020) 'Combining Multiple Approaches and Models to Dissect the Genetic Architecture of Resistance to Infections in Fish', *Frontiers in Genetics*, 11, p. 540967. doi: 10.3389/FGENE.2020.00677/BIBTEX.

Fraslin, C., Koskinen, H., *et al.* (2022) 'Genome-wide association and genomic prediction of resistance to Flavobacterium columnare in a farmed rainbow trout population', *Aquaculture*, 557, p. 738332. doi: 10.1016/J.AQUACULTURE.2022.738332.

Fraslin, C., Yáñez, J. M., *et al.* (2022) 'The impact of genetic relationship between training and validation populations on genomic prediction accuracy in Atlantic salmon', *Aquaculture Reports*, 23, p. 101033. doi: 10.1016/J.AQREP.2022.101033.

Fraslin, C. *et al.* (2023) 'Potential of low-density genotype imputation for cost-efficient genomic selection for resistance to Flavobacterium columnare in rainbow trout (Oncorhynchus mykiss)', *Genetics, selection, evolution : GSE*, 55(1). doi: 10.1186/S12711-023-00832-Z.

Froberg Brøndum, R. *et al.* (2012) 'Genome position specific priors for genomic prediction', *BMC Genomics*, 13, p. 1. doi: 10.1186/1471-2164-13-543.

Fugeray-Scarbel, A. *et al.* (2021) 'Why and How to Switch to Genomic Selection: Lessons From Plant and Animal Breeding Experience', *Frontiers in Genetics*, 12. doi: 10.3389/fgene.2021.629737ï.

Garcia, B. F. *et al.* (2022) 'Accuracy of genotype imputation to whole genome sequencing level using different populations of Nile tilapia', *Aquaculture*, 551. doi: 10.1016/J.AQUACULTURE.2022.737947.

Geibel, J. *et al.* (2021) 'How array design creates SNP ascertainment bias', *PLOS ONE*, 16(3), p. e0245178. doi: 10.1371/JOURNAL.PONE.0245178.

Gheyas, A. A. *et al.* (2010) 'Segregation of infectious pancreatic necrosis resistance QTL in the early life cycle of Atlantic Salmon (Salmo salar)', *Animal genetics*, 41(5), pp. 531–536. doi: 10.1111/J.1365-2052.2010.02032.X.

Gianola, D. *et al.* (2009) 'Additive genetic variability and the Bayesian alphabet', *Genetics*, 183(1), pp. 347–363. doi: 10.1534/GENETICS.109.103952.

Gilmour, A. R., Gogel, B. J. and Welham, S. J. (2021) 'User Guide Release 4.2 Functional Specification'. Available at: www.vsni.co.uk (Accessed: 20 June

2022).

Gjedrem, T. (2005) *Selection and Breeding Programs in Aquaculture*. Dordrecht: Springer.

Gjedrem, T., Robinson, N. and Rye, M. (2012a) 'The importance of selective breeding in aquaculture to meet future demands for animal protein: A review', *Aquaculture*, 350–353, pp. 117–129. doi: 10.1016/J.AQUACULTURE.2012.04.008.

Gjedrem, T., Robinson, N. and Rye, M. (2012b) 'The importance of selective breeding in aquaculture to meet future demands for animal protein: A review', *Aquaculture*, 350–353(null), pp. 117–129. doi: 10.1016/j.aquaculture.2012.04.008.

Gjedrem, T. and Rye, M. (2018) 'Selection response in fish and shellfish: a review', *Reviews in Aquaculture*, 10(1), pp. 168–179. doi: 10.1111/RAQ.12154.

Gjerde, B. *et al.* (2019) 'Genetic parameters for growth and survival in rohu carp (Labeo rohita)', *Aquaculture*, 503, pp. 381–388. doi: 10.1016/J.AQUACULTURE.2019.01.029.

Gjeren, H. M. and Bentsen, H. B. (1997) 'Past, present, and future of genetic aquaculture', *ICES Journal of Marine Science*, 54, pp. 1009–1014. Available at: https://academic.oup.com/icesjms/article/54/6/1009/673889 (Accessed: 3 November 2023).

Goldberger, A. S. (1962) 'Best Linear Unbiased Prediction in the Generalized Linear Regression Model', *Journal of the American Statistical Association*, 57(298), p. 369. doi: 10.2307/2281645.

Golden, C. D. *et al.* (2021) 'Aquatic foods to nourish nations', *Nature 2021 598:7880*, 598(7880), pp. 315–320. doi: 10.1038/s41586-021-03917-1.

Gonzalez-Recio, O. *et al.* (2015) 'Rare variants in transcript and potential regulatory regions explain a small percentage of the missing heritability of complex traits in Cattle', *PLoS ONE*, 10(12), p. e0143945. doi: 10.1371/journal.pone.0143945.

Gorjan, N. (2022) *Global Shrimp Aquaculture Production Survey and Forecast Presented by the Global Seafood Alliance and Rabobank*.

Gorjanc, G. *et al.* (2017) 'Prospects for Cost-Effective Genomic Selection via Accurate Within-Family Imputation', *Crop Science*, 57(1), pp. 216–228. doi: 10.2135/CROPSCI2016.06.0526.

Guimaraes, E. P. *et al.* (2007) *Marker-assisted Selection: Current Status and Future Perspectives in Crops, Livestock and Fish*. Italy: FAO. Available at: https://books.google.co.uk/books?hl=el&lr=&id=-r3WvHj7cg4C&oi=fnd&pg=PR6&dq=Guimaraes,+et+al.,+2007+qtl&ots=FOiuC57pfW&sig=e4UQnRUrk4itERyOqLlkfVV0Ds8#v=onepage&q=Guimaraes%2C et al.%2C 2007 qtl&f=false (Accessed: 4 November 2023).

Gundappa, M. K. *et al.* (2023) 'High performance imputation of structural and single nucleotide variants in Atlantic salmon using low-coverage whole genome sequencing', *bioRxiv preprint*. doi: 10.1101/2023.03.05.531147.

Guo, Z. *et al.* (2016) ‘Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize’, *Theoretical and Applied Genetics*, 129(12), pp. 2413–2427. doi: 10.1007/S00122-016-2780-5/FIGURES/6.

Gutierrez, A. P. *et al.* (2017) ‘Development of a Medium Density Combined-Species SNP Array for Pacific and European Oysters (Crassostrea gigas and Ostrea edulis)’, *G3 Genes|Genomes|Genetics*, 7(7), pp. 2209–2218. doi: 10.1534/G3.117.041780.

Gutierrez, A. P. *et al.* (2020) ‘Potential of genomic selection for improvement of resistance to ostreid herpesvirus in Pacific oyster (Crassostrea gigas)’, *Animal Genetics*, 51(2), pp. 249–257. doi: 10.1111/age.12909.

Habier, D. *et al.* (2011) ‘Extension of the bayesian alphabet for genomic selection’, *BMC bioinformatics*, 12. doi: 10.1186/1471-2105-12-186.

Habier, D., Fernando, R. L. and Dekkers, J. C. M. (2007) ‘The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values’, *Genetics*, 177(4), p. 2389. doi: 10.1534/GENETICS.107.081190.

Hayes, B. *et al.* (2013) ‘1000 Bull Genomes - Toward genomic Selection from whole genome sequence Data in Dairy and Beef Cattle’, in. Available at: https://research.wur.nl/en/publications/1000-bull-genomes-toward-genomic-selectionf-from-whole-genome-seq (Accessed: 22 March 2024).

Hayes, B. J. *et al.* (2012) ‘Accuracy of genotype imputation in sheep breeds’, *Animal Genetics*, 43(1), pp. 72–80. doi: 10.1111/J.1365-2052.2011.02208.X.

He, J. *et al.* (2018) ‘Comparing strategies for selection of low-density SNPs for imputation-mediated genomic prediction in U. S. Holsteins’, *Genetica*, 146(2), pp. 137–149. doi: 10.1007/s10709-017-0004-9.

Heather, J. M. and Chain, B. (2016) ‘The sequence of sequencers: The history of sequencing DNA’, *Genomics*, 107(1), pp. 1–8. doi: 10.1016/j.ygeno.2015.11.003.

Hedgecock, D. *et al.* (2015) ‘Second-Generation Linkage Maps for the Pacific Oyster Crassostrea gigas Reveal Errors in Assembly of Genome Scaffolds’, *G3 (Bethesda, Md.)*, 5(10), pp. 2007–2019. doi: 10.1534/G3.115.019570.

Henderson, C. R. (1949) ‘Estimation of changes in herd environment’, *Journal of Dairy (Abstract)*, 21, pp. 309–310. Available at: http://morotalab.org/literature/pdf/henderson1949.pdf.

Henderson, C. R. (1950) ‘Estimation of Genetic Parameters’, *Journal of Paediatrics and Child Health*, 21, pp. 309–310. Available at: http://morotalab.org/literature/pdf/henderson1950.pdf.

Henderson, C. R. (1973) ‘Sire evaluation and genetic trends’, *Lush Symposium paper*, (April), pp. 10–41.

Henderson, C. R. (1976) ‘A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values’, *Biometrics*, 32(1), p. 69. doi: 10.2307/2529339.

Hickey, J. M. *et al.* (2012) ‘Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs’, *Crop Science*, 52(2), pp.

654–663. doi: 10.2135/cropsci2011.07.0358.

Hickey, J. M. *et al.* (2014) 'Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation', *Crop Science*, 54(4), pp. 1476–1488. doi: 10.2135/CROPSCI2013.03.0195.

Höglund, J. *et al.* (2019) 'Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers', *Scientific Reports 2019 9:1*, 9(1), pp. 1–14. doi: 10.1038/s41598-019-53111-7.

Hollenbeck, C. M. and Johnston, I. A. (2018) 'Genomic tools and selective breeding in molluscs', *Frontiers in Genetics*, 9(JUL), p. 334494. doi: 10.3389/FGENE.2018.00253/BIBTEX.

Holsman, K. *et al.* (2018) *Climate change impacts, vulnerabilities and adaptations: North Pacific and Pacific Arctic marine fisheries*, *Impacts of climate change on fisheries and aquaculture: synthesis of current knowledge, adaptation and mitigation options*.

Houston, R. D. *et al.* (2008) 'Major Quantitative Trait Loci Affect Resistance to Infectious Pancreatic Necrosis in Atlantic Salmon (Salmo salar)', *Genetics*, 178(2), p. 1109. doi: 10.1534/GENETICS.107.082974.

Houston, R. D. *et al.* (2009) 'The susceptibility of Atlantic salmon fry to freshwater infectious pancreatic necrosis is largely explained by a major QTL', *Heredity 2010 105:3*, 105(3), pp. 318–327. doi: 10.1038/hdy.2009.171.

Houston, R. D. *et al.* (2014) 'Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar)', *BMC Genomics*, 15(1), p. 90. doi: 10.1186/1471-2164-15-90.

Houston, R. D. *et al.* (2020) 'Harnessing genomics to fast-track genetic improvement in aquaculture', *Nature Reviews Genetics*. Nature Research, pp. 389–409. doi: 10.1038/s41576-020-0227-y.

Houston, R. D., Kriaridou, C. and Robledo, D. (2022) 'Animal board invited review: Widespread adoption of genetic technologies is key to sustainable expansion of global aquaculture', *animal*, 16(10), p. 100642. doi: 10.1016/J.ANIMAL.2022.100642.

Howard, R., Jarquin, D. and Crossa, J. (2022) 'Chapter 5 Overview of Genomic Prediction Methods and the Associated Assumptions on the Variance of Marker Effect, and on the Architecture of the Target Trait', in *Methods in Molecular Biology*, pp. 139–156. doi: 10.1007/978-1-0716-2205-6_5.

Hozé, C. *et al.* (2013) 'High-density marker imputation accuracy in sixteen French cattle breeds', *Genetics Selection Evolution*, 45(1), pp. 1–11. doi: 10.1186/1297-9686-45-33.

Hu, X. *et al.* (2018) 'Developments in Common Carp Culture and Selective Breeding of New Varieties', *Aquaculture in China: Success Stories and Modern Trends*, pp. 125–148. doi: 10.1002/9781119120759.CH2_3.

Hui, R. *et al.* (2020) 'Evaluating genotype imputation pipeline for ultra-low

coverage ancient genomes', *Scientific reports*, 10(1). doi: 10.1038/S41598-020-75387-W.

Hulst, A. D., de Jong, M. C. M. and Bijma, P. (2021) 'Why genetic selection to reduce the prevalence of infectious diseases is way more promising than currently believed', *Genetics*, 217(4). doi: 10.1093/GENETICS/IYAB024.

Iglesias, R. *et al.* (2001) 'Philasterides dicentrarchi (Ciliophora, Scuticociliatida) as the causative agent of scuticociliatosis in farmed turbot Scophthalmus maximus in Galicia (NW Spain)', *Diseases of Aquatic Organisms*, 46(1), pp. 47–55. doi: 10.3354/DAO046047.

Johnston, I. A. *et al.* (2024) 'Advancing fish breeding in aquaculture through genome functional annotation', *Aquaculture*, 583, p. 740589. doi: 10.1016/J.AQUACULTURE.2024.740589.

Johnston, J., Kistemaker, G. and Sullivan, P. G. (2011) *Comparison of Different Imputation Methods*, *Interbull Bulletin*. Available at: https://journal.interbull.org/index.php/ib/article/view/1186 (Accessed: 19 June 2021).

Joshi, R. *et al.* (2018) 'Development and validation of 58K SNP-array and high-density linkage map in Nile tilapia (O. niloticus)', *Frontiers in Genetics*, 9(OCT), p. 472. doi: 10.3389/FGENE.2018.00472/BIBTEX.

Joshi, R. *et al.* (2021) 'Bayesian genomic models boost prediction accuracy for survival to Streptococcus agalactiae infection in Nile tilapia (Oreochromus nilioticus)', *Genetics Selection Evolution*, 53(1), pp. 1–10. doi: 10.1186/S12711-021-00629-Y/FIGURES/3.

Kamatani, N. *et al.* (2004) 'Large-Scale Single-Nucleotide Polymorphism (SNP) and Haplotype Analyses, Using Dense SNP Maps, of 199 Drug-Related Genes in 752 Subjects: the Analysis of the Association between Uncommon SNPs within Haplotype Blocks and the Haplotypes Constructed with Haplotype-Tagging SNPs', *The American Journal of Human Genetics*, 75(2), pp. 190–203. doi: 10.1086/422853.

Kamprasert, N. *et al.* (2024) 'Short communication: Accuracy of whole-genome sequence imputation in Angus cattle using within-breed and multi breed reference populations', *animal*, 18(3), p. 101087. doi: 10.1016/J.ANIMAL.2024.101087.

Karaman, E. *et al.* (2016) 'An Upper Bound for Accuracy of Prediction Using GBLUP'. doi: 10.1371/journal.pone.0161054.

Kijas, J. *et al.* (2017) 'Diversity and linkage disequilibrium in farmed Tasmanian Atlantic salmon', *Animal Genetics*, 48(2), pp. 237–241. doi: 10.1111/age.12513.

Kizilkaya, K., Fernando, R. L. and Garrick, D. J. (2010) 'Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes', *Journal of animal science*, 88(2), pp. 544–551. doi: 10.2527/JAS.2009-2064.

Kjetså, M. H., Ødegård, J. and Meuwissen, T. H. E. (2020) 'Accuracy of genomic prediction of host resistance to salmon lice in Atlantic salmon (Salmo salar) using

imputed high-density genotypes', *Aquaculture*, 526, p. 735415. doi: 10.1016/j.aquaculture.2020.735415.

Kriaridou, C. *et al.* (2020) 'Genomic Prediction Using Low Density Marker Panels in Aquaculture: Performance Across Species, Traits, and Genotyping Platforms', *Frontiers in Genetics*, 11, pp. 1–8. doi: 10.3389/fgene.2020.00124.

Kumar, G., Engle, C. and Tucker, C. (2018) 'Factors Driving Aquaculture Technology Adoption', *Journal of the World Aquaculture Society*. Blackwell Publishing Inc., pp. 447–476. doi: 10.1111/jwas.12514.

Lamb, H. J., Nguyen, L. T., Copley, J. P., *et al.* (2023) 'Imputation strategies for genomic prediction using nanopore sequencing', *BMC Biology*, 21(1), pp. 1–18. doi: 10.1186/S12915-023-01782-0/FIGURES/9.

Lamb, H. J., Nguyen, L. T., Briody, T. E., *et al.* (2023) 'Skim-Nanopore sequencing for routine genomic evaluation and bacterial pathogen detection in cattle', *Animal Production Science*, 63(11), pp. 1074–1085. doi: 10.1071/AN22451.

Lashmar, S. F., Muchadeyi, F. C. and Visser, C. (2019) 'Genotype imputation as a cost-saving genomic strategy for South African Sanga cattle: A review', *South African Journal of Animal Sciences*, 49(2), pp. 263–280. doi: 10.4314/sajas.v49i2.7.

Leeds, T. D. *et al.* (2010) 'Response to selection for bacterial cold water disease resistance in rainbow trout', *Journal of animal science*, 88(6), pp. 1936–1946. doi: 10.2527/JAS.2009-2538.

Legarra, A. *et al.* (2008) 'Performance of Genomic Selection in Mice', *Genetics*, 180(1), pp. 611–618. doi: 10.1534/GENETICS.108.088575.

Legarra, A. *et al.* (2011) 'Improved Lasso for genomic selection', *Genetics Research*, 93(1), pp. 77–87. doi: 10.1017/S0016672310000534.

Legarra, A., Aguilar, I. and Misztal, I. (2009) 'A relationship matrix including full pedigree and genomic information', *Journal of dairy science*, 92(9), pp. 4656–4663. doi: 10.3168/JDS.2009-2061.

Legarra, A. and Ducrocq, V. (2012) 'Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction'. doi: 10.3168/jds.2011-4982.

Li, H. (2011) 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', *Bioinformatics*, 27(21), p. 2987. doi: 10.1093/BIOINFORMATICS/BTR509.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics (Oxford, England)*, 25(14), pp. 1754–1760. doi: 10.1093/BIOINFORMATICS/BTP324.

Li, J. *et al.* (2022) 'Genome-wide association studies for egg quality traits in White Leghorn layers using low-pass sequencing and SNP chip data', *Journal of Animal Breeding and Genetics*, 139(4), pp. 380–397. doi: 10.1111/JBG.12679.

Li, Q. *et al.* (2021) ‘Estimation of genetic parameters for carotenoid traits in Chinese mitten crab, Eriocheir sinensis, females’, *Aquaculture*, 532, p. 735990. doi: 10.1016/J.AQUACULTURE.2020.735990.

Li, Z. *et al.* (2019) ‘Integrating gene expression data into genomic prediction’, *Frontiers in Genetics*, 10(FEB), p. 430679. doi: 10.3389/FGENE.2019.00126/BIBTEX.

Lieke, T. *et al.* (2020) ‘Sustainable aquaculture requires environmental-friendly treatment strategies for fish diseases’, *Reviews in Aquaculture*, 12(2), pp. 943–965. doi: 10.1111/RAQ.12365.

Lillehammer, M. *et al.* (2020) ‘Genomic selection for white spot syndrome virus resistance in whiteleg shrimp boosts survival under an experimental challenge test’, *Scientific Reports 2020 10:1*, 10(1), pp. 1–13. doi: 10.1038/s41598-020-77580-3.

Liu, C. T. *et al.* (2019) ‘Revisit Population-based and Family-based Genotype Imputation’, *Scientific Reports 2019 9:1*, 9(1), pp. 1–9. doi: 10.1038/s41598-018-38469-4.

Liu, Y. *et al.* (2018) ‘Genomic Selection Using BayesCπ and GBLUP for Resistance Against Edwardsiella tarda in Japanese Flounder (Paralichthys olivaceus)’, *Marine Biotechnology*, 20(5), pp. 559–565. doi: 10.1007/s10126-018-9839-z.

Lloret-Villas, A., Pausch, H. and Leonard, A. S. (2023) ‘The size and composition of haplotype reference panels impact the accuracy of imputation from low-pass sequencing in cattle’, *Genetics, selection, evolution : GSE*, 55(1). doi: 10.1186/S12711-023-00809-Y.

Long, E. M. *et al.* (2022) ‘Genome-wide imputation using the practical haplotype graph in the heterozygous crop cassava’, *G3 Genes|Genomes|Genetics*, 12(1). doi: 10.1093/G3JOURNAL/JKAB383.

Longo, S. B. *et al.* (2019) ‘Aquaculture and the displacement of fisheries captures’, *Conservation Biology*, 33(4), pp. 832–841. doi: 10.1111/COBI.13295.

Luo, Y. *et al.* (2017) ‘Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7’, *Nature Genetics 2017 49:2*, 49(2), pp. 186–192. doi: 10.1038/ng.3761.

Lush, J. L. (1947) ‘Family Merit and Individual Merit as Bases for Selection. Part I on JSTOR’, *The American Naturalist*, 81(799), pp. 241–261. Available at: https://www.jstor.org/stable/2457881 (Accessed: 1 November 2023).

Ma, P. *et al.* (2013) ‘Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle’. doi: 10.3168/jds.2012-6316.

Ma, P. *et al.* (2019) ‘Use of a Bayesian model including QTL markers increases prediction reliability when test animals are distant from the reference population’, *Journal of Dairy Science*, 102(8), pp. 7237–7247. doi: 10.3168/JDS.2018-15815.

MacLeod, I. M. *et al.* (2016) ‘Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits’, *BMC*

Genomics, 17(1), pp. 1–21. doi: 10.1186/S12864-016-2443-6/TABLES/9.

MacLeod, I. M., Hayes, B. J. and Goddard, M. E. (2014) 'The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data', *Genetics*, 198(4), pp. 1671–1684. doi: 10.1534/GENETICS.114.168344.

MacLeod, M. J. *et al.* (2020) 'Quantifying greenhouse gas emissions from global aquaculture', *Scientific Reports 2020 10:1*, 10(1), pp. 1–8. doi: 10.1038/s41598-020-68231-8.

Mair, G. C. *et al.* (2023) 'A decadal outlook for global aquaculture', *Journal of the World Aquaculture Society*, 54(2), pp. 196–205. doi: 10.1111/JWAS.12977.

Manolio, T. A. *et al.* (2009) 'Finding the missing heritability of complex diseases', *Nature*. Nature Publishing Group, pp. 747–753. doi: 10.1038/nature08494.

Marchini, J. and Howie, B. (2010a) 'Genotype imputation for genome-wide association studies', *Nature Reviews Genetics*. Nature Publishing Group, pp. 499–511. doi: 10.1038/nrg2796.

Marchini, J. and Howie, B. (2010b) 'Genotype imputation for genome-wide association studies', *Nature Reviews Genetics 2010 11:7*, 11(7), pp. 499–511. doi: 10.1038/nrg2796.

Maroso, F. *et al.* (2018) 'Highly dense linkage maps from 31 full-sibling families of turbot (Scophthalmus maximus) provide insights into recombination patterns and chromosome rearrangements throughout a newly refined genome assembly', *DNA Research*, 25(4), pp. 439–450. doi: 10.1093/dnares/dsy015.

Martin, A. R. *et al.* (2021) 'Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations', *American journal of human genetics*, 108(4), pp. 656–668. doi: 10.1016/J.AJHG.2021.03.012.

Martínez, P. *et al.* (2021) 'A genome-wide association study, supported by a new chromosome-level genome assembly, suggests sox2 as a main driver of the undifferentiatiated ZZ/ZW sex determination of turbot (Scophthalmus maximus)', *Genomics*, 113(4), pp. 1705–1718. doi: 10.1016/J.YGENO.2021.04.007.

Maulu, S. *et al.* (2021) 'Fish Nutritional Value as an Approach to Children's Nutrition', *Frontiers in Nutrition*, 8, p. 780844. doi: 10.3389/FNUT.2021.780844.

McCarthy, S. *et al.* (2016) 'A reference panel of 64,976 haplotypes for genotype imputation', *Nature Genetics 2016 48:10*, 48(10), pp. 1279–1283. doi: 10.1038/ng.3643.

Merrick, L. F. and Carter, A. H. (2021) 'Comparison of genomic selection models for exploring predictive ability of complex traits in breeding programs', *The Plant Genome*, 14(3), p. e20158. doi: 10.1002/TPG2.20158.

Meuwissen, T. *et al.* (2022) 'Identification of Genomic Variants Causing Variation in Quantitative Traits: A Review', *Agriculture 2022, Vol. 12, Page 1713*, 12(10), p. 1713. doi: 10.3390/AGRICULTURE12101713.

Meuwissen, T. H. E., Hayes, B. J. and Goddard, M. E. (2001) 'Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps'. Available at:

https://academic.oup.com/genetics/article/157/4/1819/6048353 (Accessed: 6 November 2023).

Misztal, I., Legarra, A. and Aguilar, I. (2009) ‘Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information’, *Journal of dairy science*, 92(9), pp. 4648–4655. doi: 10.3168/JDS.2009-2064.

Moen, T. *et al.* (2009) ‘Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar): Population-level associations between markers and trait’, *BMC Genomics*, 10, pp. 1–14. doi: 10.1186/1471-2164-10-368.

Mollandin, F. *et al.* (2022a) ‘Accounting for overlapping annotations in genomic prediction models of complex traits’, *BMC Bioinformatics*, 23(1), pp. 1–22. doi: 10.1186/S12859-022-04914-5/FIGURES/7.

Mollandin, F. *et al.* (2022b) ‘Capitalizing on complex annotations in Bayesian genomic prediction for a backcross population of growing pigs.’, in *12th World Congress on Genetics Applied to Livestock Production*. Rotterdam, Netherlands. Available at: https://hal.inrae.fr/hal-03742045 (Accessed: 4 March 2024).

Morgante, F. *et al.* (2018) ‘Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals’, *Heredity 2018 120:6*, 120(6), pp. 500–514. doi: 10.1038/s41437-017-0043-0.

Morgante, F. *et al.* (2020) ‘Leveraging Multiple Layers of Data To Predict Drosophila Complex Traits’, *G3 Genes|Genomes|Genetics*, 10(12), pp. 4599–4613. doi: 10.1534/G3.120.401847.

Mrode, Raphael A. (2014) *Linear models for the prediction of animal breeding values*. Edited by R. A. Mrode. UK: CABI. doi: 10.1079/9781780643915.0000.

Nascimento-Schulze, J. C. *et al.* (2021) ‘Optimizing hatchery practices for genetic improvement of marine bivalves’, *Reviews in Aquaculture*, 13(4), pp. 2289–2304. doi: 10.1111/RAQ.12568.

Nayeri, S., Sargolzaei, M. and Tulpan, D. (2019) ‘A review of traditional and machine learning methods applied to animal breeding’, *Animal health research reviews*, 20(1), pp. 31–46. doi: 10.1017/S1466252319000148.

Naylor, R. L. *et al.* (2021) ‘A 20-year retrospective review of global aquaculture’, *Nature 2021 591:7851*, 591(7851), pp. 551–563. doi: 10.1038/s41586-021-03308-6.

Nejati-Javaremi, A., Smith, C. and Gibson, J. P. (1997) ‘Effect of Total Allelic Relationship on Accuracy of Evaluation and Response to Selection’, *J. Anim. Sci*, 75, pp. 1738–1745. Available at: https://academic.oup.com/jas/article/75/7/1738/4625037 (Accessed: 7 November 2023).

Nguyen, N. H., Phuthaworn, C. and Knibb, W. (2020) ‘Genomic prediction for disease resistance to Hepatopancreatic parvovirus and growth, carcass and quality traits in Banana shrimp Fenneropenaeus merguiensis’, *Genomics*, 112(2), pp. 2021–2027. doi: 10.1016/J.YGENO.2019.11.014.

Noble, T. H. *et al.* (2020) ‘Genetic parameters of Gill-associated virus infection

and body weight under commercial conditions in black tiger shrimp, Penaeus monodon', *Aquaculture*, 528, p. 735580. doi: 10.1016/J.AQUACULTURE.2020.735580.

Norris, A. (2017) 'Application of genomics in salmon aquaculture breeding programs by Ashie Norris: Who knows where the genomic revolution will lead us?', *Marine Genomics*, 36, pp. 13–15. doi: 10.1016/J.MARGEN.2017.11.013.

Nosková, A. *et al.* (2021) 'Characterization of a haplotype-reference panel for genotyping by low-pass sequencing in Swiss Large White pigs', *BMC Genomics*, 22(1), pp. 1–14. doi: 10.1186/S12864-021-07610-5/FIGURES/4.

Ober, U. *et al.* (2015) 'Accounting for Genetic Architecture Improves Sequence Based Genomic Prediction for a Drosophila Fitness Trait', *PLOS ONE*, 10(5), p. e0126880. doi: 10.1371/JOURNAL.PONE.0126880.

Ødegård, J. *et al.* (2014) 'Genomic prediction in an admixed population of Atlantic salmon (Salmo salar)', *Frontiers in Genetics*, 5(NOV), p. 117388. doi: 10.3389/FGENE.2014.00402/BIBTEX.

Palaiokostas, C. *et al.* (2018) 'Accuracy of genomic evaluations of juvenile growth rate in common carp (Cyprinus carpio) using genotyping by sequencing', *Frontiers in Genetics*, 9(MAR), p. 82. doi: 10.3389/fgene.2018.00082.

Palaiokostas, C. *et al.* (2019) 'Optimizing genomic prediction of host resistance to Koi herpesvirus disease in carp', *Frontiers in Genetics*, 10(JUN). doi: 10.3389/fgene.2019.00543.

Palaiokostas, C. (2021) 'Predicting for disease resistance in aquaculture species using machine learning models', *Aquaculture Reports*, 20, p. 100660. doi: 10.1016/J.AQREP.2021.100660.

Pan, Z. *et al.* (2021) 'Pig genome functional annotation enhances the biological interpretation of complex traits and human disease', *Nature Communications 2021 12:1*, 12(1), pp. 1–15. doi: 10.1038/s41467-021-26153-7.

Park, T. and Casella, G. (2008) 'The Bayesian Lasso', *Journal of the American Statistical Association*, 103(482), pp. 681–686. doi: 10.1198/016214508000000337.

Pasaniuc, B. *et al.* (2012) 'Extremely low-coverage sequencing and imputation increases power for genome-wide association studies', *Nature Genetics 2012 44:6*, 44(6), pp. 631–635. doi: 10.1038/ng.2283.

Pausch, H. *et al.* (2017) 'Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle', *Genetics Selection Evolution*, 49(1), pp. 1–14. doi: 10.1186/s12711-017-0301-x.

Peñaloza, C. *et al.* (2020) 'Development and Validation of an Open Access SNP Array for Nile Tilapia (Oreochromis niloticus)', *G3 Genes|Genomes|Genetics*, 10(8), pp. 2777–2785. doi: 10.1534/G3.120.401343.

Phocas, F. (2022) 'Genotyping, the Usefulness of Imputation to Increase SNP Density, and Imputation Methods and Tools', pp. 113–138. doi: 10.1007/978-1-0716-2205-6_4.

Piazzon, M. C., Leiro, J. and Lamas, J. (2013) ‘Fish immunity to scuticociliate parasites’, *Developmental & Comparative Immunology*, 41(2), pp. 248–256. doi: 10.1016/J.DCI.2013.05.022.

Pong-Wong, R. and Woolliams, J. A. (2014) ‘Bayes U: A Genomic Prediction Method Based on the Horseshoe Prior’, in *10th world congress of genetics applied to livestock production*. doi: 10.13140/2.1.4852.6723.

Pooley, C. M. *et al.* (2020) ‘Estimating individuals’ genetic and non-genetic effects underlying infectious disease transmission from temporal epidemic data’, *PLOS Computational Biology*, 16(12), p. e1008447. doi: 10.1371/JOURNAL.PCBI.1008447.

Prowse-Wilkins, C. P. *et al.* (2021) ‘Putative Causal Variants Are Enriched in Annotated Functional Regions From Six Bovine Tissues’, *Frontiers in Genetics*, 12, p. 664379. doi: 10.3389/FGENE.2021.664379/BIBTEX.

Purcell, S. *et al.* (2007) ‘PLINK: A tool set for whole-genome association and population-based linkage analyses’, *American Journal of Human Genetics*, 81(3), pp. 559–575. doi: 10.1086/519795.

Regan, T. *et al.* (2021) ‘Genetic improvement technologies to support the sustainable growth of UK aquaculture’, *Reviews in Aquaculture*. John Wiley and Sons Inc. doi: 10.1111/raq.12553.

Ren, D. *et al.* (2021) ‘Efficient weighting methods for genomic best linear-unbiased prediction (BLUP) adapted to the genetic architectures of quantitative traits’, *Heredity*, 126(2), p. 320. doi: 10.1038/S41437-020-00372-Y.

Robinson, G. K. (1991) ‘That BLUP is a Good Thing: The Estimation of Random Effects’, *Statistical Science*, 6(1), pp. 15–51. Available at: https://www.jstor.org/stable/2245695?seq=1 (Accessed: 3 November 2023).

Robledo, D., Palaiokostas, C., *et al.* (2018) ‘Applications of genotyping by sequencing in aquaculture breeding and genetics’, *Reviews in Aquaculture*, 10(3), pp. 670–682. doi: 10.1111/RAQ.12193.

Robledo, D., Matika, O., *et al.* (2018) ‘Genome-wide association and genomic selection for resistance to amoebic gill disease in Atlantic salmon’, *G3: Genes, Genomes, Genetics*, 8(4), pp. 1195–1203. doi: 10.1534/g3.118.200075.

Rohde, P. D. *et al.* (2021) ‘Prediction of complex phenotypes using the Drosophila melanogaster metabolome’, *Heredity*, 126(5), p. 717. doi: 10.1038/S41437-021-00404-1.

Rubinacci, S. *et al.* (2021) ‘Efficient phasing and imputation of low-coverage sequencing data using large reference panels’, *Nature Genetics 2021 53:1*, 53(1), pp. 120–126. doi: 10.1038/s41588-020-00756-0.

Rubinacci, S. *et al.* (2023) ‘Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes’, *Nature Genetics 2023 55:7*, 55(7), pp. 1088–1090. doi: 10.1038/s41588-023-01438-3.

Sargolzaei, M., Chesnais, J. P. and Schenkel, F. S. (2010) *Accuracy of a family-based genotype imputation algorithm. Open Industry Session.*

Sargolzaei, M., Chesnais, J. P. and Schenkel, F. S. (2014) 'A new approach for efficient genotype imputation using information from relatives', *BMC Genomics*, 15(1), pp. 1–12. doi: 10.1186/1471-2164-15-478.

Schaeffer, L. (1991) 'C. R. Henderson: Contributions to Predicting Genetic Merit', *Journal of Dairy Science*, 74, pp. 4052–4066. doi: 10.3168/jds.S0022-0302(91)78601-3.

Shen, Y. and Yue, G. (2019) 'Current status of research on aquaculture genetics and genomics-information from ISGA 2018', *Aquaculture and Fisheries*, 4(2), pp. 43–47. doi: 10.1016/J.AAF.2018.11.001.

Smits, M. *et al.* (2020) 'Potential for Genetic Improvement of Resistance to Perkinsus olseni in the Manila Clam, Ruditapes philippinarum, Using DNA Parentage Assignment and Mass Spawning', *Frontiers in Veterinary Science*, 7, p. 767. doi: 10.3389/FVETS.2020.579840/BIBTEX.

Soledad Peñaloza Navarro, C. (2017) *Characterization of genome-wide deviations from Mendelian inheritance in bivalve species*.

Sonesson, A. K., Meuwissen, T. H. E. and Goddard, M. E. (2010) 'The use of communal rearing of families and DNA pooling in aquaculture genomic selection schemes', *Genetics Selection Evolution*, 42(1), pp. 1–9. doi: 10.1186/1297-9686-42-41/TABLES/4.

Sonesson, A. K. and Ødegård, J. (2016) 'Mating structures for genomic selection breeding programs in aquaculture', *Genetics Selection Evolution*, 48(1), pp. 1–7. doi: 10.1186/S12711-016-0224-Y/TABLES/4.

Song, H. *et al.* (2022) 'Estimation of genetic parameters for growth and egg related traits in Russian sturgeon (Acipenser gueldenstaedtii)', *Aquaculture*, 546, p. 737299. doi: 10.1016/J.AQUACULTURE.2021.737299.

Song, H. *et al.* (2023) 'Genomic selection and its research progress in aquaculture breeding', *Reviews in Aquaculture*, 15(1), pp. 274–291. doi: 10.1111/RAQ.12716.

Song, H. and Hu, H. (2022a) 'Strategies to improve the accuracy and reduce costs of genomic prediction in aquaculture species', *Evolutionary Applications*, 15(4), pp. 578–590. doi: 10.1111/EVA.13262.

Song, H. and Hu, H. (2022b) 'Strategies to improve the accuracy and reduce costs of genomic prediction in aquaculture species', *Evolutionary Applications*, 15(4), pp. 578–590. doi: 10.1111/EVA.13262.

Song, K. (2020) 'Genomic Landscape of Mutational Biases in the Pacific Oyster Crassostrea gigas', *Genome Biology and Evolution*, 12(11), p. 1943. doi: 10.1093/GBE/EVAA160.

Spiliopoulou, A. *et al.* (2017) 'GeneImp: Fast Imputation to Large Reference Panels Using Genotype Likelihoods from Ultralow Coverage Sequencing', *Genetics*, 206(1), pp. 91–104. doi: 10.1534/GENETICS.117.200063.

Stentiford, G. D. *et al.* (2020) *Sustainable aquaculture through the One Health lens*, *Nature Food*. doi: 10.1038/s43016-020-0127-5.

Sui, J. *et al.* (2016) ‘Genetic parameters and response to selection for harvest body weight of pacific white shrimp, Litopenaeus vannamei’, *Aquaculture Research*, 47(9), pp. 2795–2803. doi: 10.1111/ARE.12729.

Tan, K., Zhang, H. and Zheng, H. (2020) ‘Selective breeding of edible bivalves and its implication of global climate change’, *Reviews in Aquaculture*, 12(4), pp. 2559–2572. doi: 10.1111/RAQ.12458.

Taslima, K. *et al.* (2016) ‘DNA sampling from mucus in the Nile tilapia, Oreochromis niloticus: minimally invasive sampling for aquaculture-related genetics research’, *Aquaculture Research*, 47(12), pp. 4032–4037. doi: 10.1111/are.12809.

Teletchea, F. (2021) ‘Fish domestication in aquaculture: 10 unanswered questions’, *Animal Frontiers*, 11(3), pp. 87–91. doi: 10.1093/AF/VFAB012.

Teng, J. *et al.* (2022) ‘Assessment of the performance of different imputation methods for low-coverage sequencing in Holstein cattle’, *Journal of Dairy Science*, 105(4), pp. 3355–3366. doi: 10.3168/JDS.2021-21360.

Thomas, M. *et al.* (2021) ‘When more is more: taking advantage of species diversity to move towards sustainable aquaculture’, *Biological Reviews*, 96(2), pp. 767–784. doi: 10.1111/BRV.12677.

Tonner, P. D., Pressman, A. and Ross, D. (2022) ‘Interpretable modeling of genotype–phenotype landscapes with state-of-the-art predictive power’, *Proceedings of the National Academy of Sciences of the United States of America*, 119(26), p. e2114021119. doi: 10.1073/PNAS.2114021119/SUPPL_FILE/PNAS.2114021119.SAPP.PDF.

Triay, C. *et al.* (2024) ‘Fast and accurate imputation of genotypes from noisy low-coverage sequencing data in bi-parental populations’, *bioRxiv*, p. 2024.03.13.584787. doi: 10.1101/2024.03.13.584787.

Tsai, H.-Y. *et al.* (2016) ‘Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations’, *Genetics Selection Evolution*, 48, p. 47. doi: 10.1186/s12711-016-0226-9.

Tsai, H. Y. *et al.* (2015) ‘Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array’, *BMC Genomics*, 16(1), pp. 1–9. doi: 10.1186/s12864-015-2117-9.

Tsai, H. Y. *et al.* (2017) ‘Genotype imputation to improve the cost-efficiency of genomic selection in farmed Atlantic salmon’, *G3: Genes, Genomes, Genetics*, 7(4), pp. 1377–1383. doi: 10.1534/g3.117.040717.

Tsairidou, S. (2019) *CVrep: Package to perform multiple cross-validation iterations to calculate genomic prediction accuracy*. Available at: https://github.com/SmaragdaT/CVrep (Accessed: 11 August 2024).

Tsairidou, S. *et al.* (2020) ‘Optimizing low-cost genotyping and imputation strategies for genomic selection in Atlantic salmon’, *G3: Genes, Genomes, Genetics*, 10(2), pp. 581–590. doi: 10.1534/g3.119.400800.

Uffelmann, E. *et al.* (2021) ‘Genome-wide association studies’, *Nature Reviews Methods Primers 2021 1:1*, 1(1), pp. 1–21. doi: 10.1038/s43586-021-00056-9.

Vallejo, R. L. *et al.* (2021) 'The accuracy of genomic predictions for bacterial cold water disease resistance remains higher than the pedigree-based model one generation after model training in a commercial rainbow trout breeding population', *Aquaculture*, 545. doi: 10.1016/J.AQUACULTURE.2021.737164.

Vandeputte, M. *et al.* (2009) 'Response to domestication and selection for growth in the European sea bass (Dicentrarchus labrax) in separate and mixed tanks', *Aquaculture*, 286(1–2), pp. 20–27. doi: 10.1016/J.AQUACULTURE.2008.09.008.

Vanraden, P. M. *et al.* (2011) 'Genomic evaluations with many more genotypes', *Genetics, Selection, Evolution : GSE*, 43(1), p. 10. doi: 10.1186/1297-9686-43-10.

VanRaden, P. M. (2008) 'Efficient Methods to Compute Genomic Predictions', *Journal of Dairy Science*, 91(11), pp. 4414–4423. doi: 10.3168/JDS.2007-0980.

VanRaden, P. M. *et al.* (2009) 'Invited review: Reliability of genomic predictions for North American Holstein bulls', *Journal of Dairy Science*, 92(1), pp. 16–24. doi: 10.3168/jds.2008-1514.

VanRaden, P. M. *et al.* (2013) 'Genomic imputation and evaluation using high-density Holstein genotypes', *Journal of Dairy Science*, 96(1), pp. 668–678. doi: 10.3168/JDS.2012-5702.

VanRaden, P. M. *et al.* (2017) 'Selecting sequence variants to improve genomic predictions for dairy cattle', *Genetics Selection Evolution*, 49(1), pp. 1–12. doi: 10.1186/S12711-017-0307-4/FIGURES/4.

Ventura, R. V. *et al.* (2016) 'Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population', *Genetics Selection Evolution*, 48(1), pp. 1–20. doi: 10.1186/S12711-016-0244-7/FIGURES/8.

Wang, H. *et al.* (2014) 'Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens'. doi: 10.3389/fgene.2014.00134.

Wang, Y. *et al.* (2016) 'Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle', *Springer Science Reviews*, 4(2), pp. 79–98. doi: 10.1007/s40362-017-0041-x.

Wang, Z. *et al.* (2022) 'PHARP: a pig haplotype reference panel for genotype imputation', *Scientific Reports 2022 12:1*, 12(1), pp. 1–11. doi: 10.1038/s41598-022-15851-x.

Wasik, K. *et al.* (2021) 'Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics', *BMC Genomics*, 22(1), pp. 1–7. doi: 10.1186/S12864-021-07508-2/FIGURES/2.

Weigel, K. A. *et al.* (2009) 'Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers', *Journal of Dairy Science*, 92(10), pp. 5248–5257. doi: 10.3168/JDS.2009-2092.

Whalen, A. *et al.* (2018) 'Assessment of the performance of hidden Markov models for imputation in animal breeding', *Genetics Selection Evolution*, 50(1), pp. 1–10. doi: 10.1186/S12711-018-0416-8/TABLES/2.

Whalen, A., Gorjanc, G. and Hickey, J. M. (2019) 'Family-specific genotype arrays increase the accuracy of pedigree-based imputation at very low marker densities', *Genetics Selection Evolution*, 51(1), pp. 1–9. doi: 10.1186/S12711-019-0478-2/FIGURES/4.

Whalen, A. and Hickey, J. M. (2020) 'AlphaImpute2: Fast and accurate pedigree and population based imputation for hundreds of thousands of individuals in livestock populations', *bioRxiv*, p. 2020.09.16.299677. doi: 10.1101/2020.09.16.299677.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. doi: 10.1007/978-3-319-24277-4_9.

Wilson, J. R. and Archer, B. (2010) 'Diversification pays: Economic perspectives on investment in diversified aquaculture', *Finfish Aquaculture Diversification*, pp. 514–530. doi: 10.1079/9781845934941.0514.

Wolc, A. and Dekkers, J. C. M. (2022) 'Application of Bayesian genomic prediction methods to genome-wide association analyses', *Genetics Selection Evolution 2022 54:1*, 54(1), pp. 1–12. doi: 10.1186/S12711-022-00724-8.

Wragg, D. *et al.* (2024) 'A cautionary tale of low-pass sequencing and imputation with respect to haplotype accuracy', *Genetics Selection Evolution*, 56(1), pp. 1–19. doi: 10.1186/S12711-024-00875-W/FIGURES/9.

Wu, X. *et al.* (2020) 'A unified local objective function for optimally selecting SNPs on arrays for agricultural genomics applications', *Animal Genetics*, 51(2), pp. 306–310. doi: 10.1111/age.12916.

Xiang, R. *et al.* (2021) 'Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations', *Nature Communications 2021 12:1*, 12(1), pp. 1–13. doi: 10.1038/s41467-021-21001-0.

Yáñez, J. M. *et al.* (2020) 'High-Throughput Single Nucleotide Polymorphism (SNP) Discovery and Validation Through Whole-Genome Resequencing in Nile Tilapia (Oreochromis niloticus)', *Marine Biotechnology*, 22(1), pp. 109–117. doi: 10.1007/S10126-019-09935-5/FIGURES/3.

Yáñez, J. M. *et al.* (2023a) 'Genome-wide association and genomic selection in aquaculture', *Reviews in Aquaculture*, 15(2), pp. 645–675. doi: 10.1111/RAQ.12750.

Yáñez, J. M. *et al.* (2023b) 'Genome-wide association and genomic selection in aquaculture', *Reviews in Aquaculture*, 15(2), pp. 645–675. doi: 10.1111/RAQ.12750.

Yáñez, J. M., Joshi, R. and Yoshida, G. M. (2020) 'Genomics to accelerate genetic improvement in tilapia', *Animal Genetics*. Blackwell Publishing Ltd, pp. 658–674. doi: 10.1111/age.12989.

Yáñez, J. M., Newman, S. and Houston, R. D. (2015) 'Genomics in aquaculture to better understand species biology and accelerate genetic progress', *Frontiers in Genetics*, 6(APR), p. 137804. doi: 10.3389/FGENE.2015.00128/BIBTEX.

Yang, J. *et al.* (2011) 'GCTA: A Tool for Genome-wide Complex Trait Analysis',

American Journal of Human Genetics, 88(1), p. 76. doi: 10.1016/J.AJHG.2010.11.011.

Yang, R. *et al.* (2021) 'Accelerated deciphering of the genetic architecture of agricultural economic traits in pigs using a low-coverage whole-genome sequencing strategy', *GigaScience*, 10(7). doi: 10.1093/GIGASCIENCE/GIAB048.

Yang, W. *et al.* (2020) 'Animal-ImputeDB: a comprehensive database with multiple animal reference panels for genotype imputation', *Nucleic acids research*, 48(D1), pp. D659–D667. doi: 10.1093/NAR/GKZ854.

Yang, Z. *et al.* (2024) 'Market integration and market leadership: Evidence for cyprinoid markets in China', *Aquaculture*, 578, p. 740010. doi: 10.1016/J.AQUACULTURE.2023.740010.

Yao, X. *et al.* (2024) 'PheSeq, a Bayesian deep learning model to enhance and interpret the gene-disease association studies', *Genome Medicine*, 16(1), pp. 1–26. doi: 10.1186/S13073-024-01330-7/FIGURES/8.

Ye, S. *et al.* (2018) 'Imputation from SNP chip to sequence: A case study in a Chinese indigenous chicken population', *Journal of Animal Science and Biotechnology*, 9(1), pp. 1–12. doi: 10.1186/s40104-018-0241-5.

Yoshida, G. M. *et al.* (2018) 'Accuracy of genotype imputation and genomic predictions in a two-generation farmed Atlantic salmon population using high-density and low-density SNP panels', *Aquaculture*, 491, pp. 147–154. doi: 10.1016/j.aquaculture.2018.03.004.

Yoshida, G. M. *et al.* (2019) 'Genome-wide association study and cost-efficient genomic predictions for growth and fillet yield in Nile tilapia (Oreochromis niloticus)', *G3: Genes, Genomes, Genetics*, 9(8), pp. 2597–2607. doi: 10.1534/g3.119.400116.

Yoshida, G. M. and Yáñez, J. M. (2021) 'Increased accuracy of genomic predictions for growth under chronic thermal stress in rainbow trout by prioritizing variants from GWAS using imputed sequence data', *Evolutionary applications*, 15(4), pp. 537–552. doi: 10.1111/EVA.13240.

Yoshida, G. M. and Yáñez, J. M. (2022) 'Increased accuracy of genomic predictions for growth under chronic thermal stress in rainbow trout by prioritizing variants from GWAS using imputed sequence data', *Evolutionary Applications*, 15(4), pp. 537–552. doi: 10.1111/EVA.13240.

Yoshida, G. M., Yáñez, J. M. and De Ciencias, F. (2021) 'Increased accuracy of genomic predictions for growth under chronic thermal stress in rainbow trout by prioritizing variants from GWAS using imputed sequence data', *Evolutionary Applications*, 00, pp. 1–16. doi: 10.1111/eva.13240.

Zeng, Q. *et al.* (2022) 'Aquaculture Molecular Breeding Platform (AMBP): a comprehensive web server for genotype imputation and genetic analysis in aquaculture', *Nucleic Acids Research*, 50(W1), pp. W66–W74. doi: 10.1093/NAR/GKAC424.

Zenger, K. R. *et al.* (2019) 'Genomic selection in aquaculture: application,

limitations and opportunities with special reference to marine shrimp and pearl oysters', *Frontiers in Genetics*, 9, p. 693. doi: 10.3389/fgene.2018.00693.

Zhai, S. *et al.* (2021) 'Estimation of genetic parameters for resistance to Vibrio alginolyticus infection in the Pacific oyster (Crassostrea gigas)', *Aquaculture*, 538, p. 736545. doi: 10.1016/J.AQUACULTURE.2021.736545.

Zhang, G. *et al.* (2012) 'The oyster genome reveals stress adaptation and complexity of shell formation', *Nature 2012 490:7418*, 490(7418), pp. 49–54. doi: 10.1038/nature11413.

Zhang, W. *et al.* (2021) 'Evaluation for the effect of low-coverage sequencing on genomic selection in large yellow croaker', *Aquaculture*, 534. doi: 10.1016/J.AQUACULTURE.2020.736323.

Zhang, X. *et al.* (2016) 'Weighting Strategies for Single-Step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and GWAS', *Frontiers in genetics*, 7(AUG). doi: 10.3389/FGENE.2016.00151.

Zhang, X. *et al.* (2023) 'Penaeid Shrimp Chromosome Studies Entering the Post-Genomic Era', *Genes*, 14(11). doi: 10.3390/GENES14112050.

Zhang, Z. *et al.* (2010) 'Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix', *PloS one*, 5(9), pp. 1–8. doi: 10.1371/JOURNAL.PONE.0012648.

Zhang, Z. *et al.* (2015) 'Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix', *G3 (Bethesda, Md.)*, 5(4), pp. 615–627. doi: 10.1534/G3.114.016261.

Zhang, Z. and Druet, T. (2010) 'Marker imputation with low-density marker panels in Dutch Holstein cattle', *Journal of Dairy Science*, 93(11), pp. 5487–5494. doi: 10.3168/jds.2010-3501.

Zhang, Z., Zhang, Q. and Ding, X. D. (2011) 'Advances in genomic selection in domestic animals', *Chinese Science Bulletin*, 56(25), pp. 2655–2663. doi: 10.1007/S11434-011-4632-7/METRICS.

Zhao, C. *et al.* (2021) 'Towards a Cost-Effective Implementation of Genomic Prediction Based on Low Coverage Whole Genome Sequencing in Dezhou Donkey', *Frontiers in Genetics*, 12, p. 728764. doi: 10.3389/FGENE.2021.728764/BIBTEX.

Zhong, X. *et al.* (2017) 'Estimates of Linkage Disequilibrium and Effective Population Size in Wild and Selected Populations of the Pacific Oyster Using Single-nucleotide Polymorphism Markers', *Journal of the World Aquaculture Society*, 48(5), pp. 791–801. doi: 10.1111/JWAS.12393.

Zhou, S. *et al.* (2020) 'Systems genetics of the Drosophila metabolome', *Genome Research*, 30(3), pp. 392–405. doi: 10.1101/GR.243030.118/-/DC1.

Zhu, X. *et al.* (2021) 'Genomic prediction of growth traits in scallops using convolutional neural networks', *Aquaculture*, 545, p. 737171. doi: 10.1016/J.AQUACULTURE.2021.737171.