UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

QUANTITATIVE GENETICS AND GENOMIC TOOLS APPLIED TO AQUACULTURE BREEDING

Baltasar Fernandes Garcia Neto

Zootecnista

2021

UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL

QUANTITATIVE GENETICS AND GENOMIC TOOLS APPLIED TO AQUACULTURE BREEDING

Baltasar Fernandes Garcia Neto
Orientador: Dr. Roberto Carvalheiro
Coorientadores: Dr. Hugo Horácio Montaldo
Dra. Laiza Helena de Souza Iung

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para obtenção do título de Doutor em Genética e Melhoramento Animal

2021

Garcia Neto, Baltasar Fernandes
216q Quantitative genetics and genomic tools applied to aquaculture breeding / Baltasar Fernandes Garcia Neto. -- Jaboticabal, 2021
102 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientador: Roberto Carvalheiro
Coorientador: Hugo Horácio Montaldo, Laiza Helena de Souza Iung

1. Uniformity of weight. 2. SNP panel. 3. Genotype imputation. 4. Genomics. 5. Aquaculture. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: QUANTITATIVE GENETICS AND GENOMIC TOOLS APPLIED TO AQUACULTURE BREEDING

AUTOR: BALTASAR FERNANDES GARCIA NETO
ORIENTADOR: ROBERTO CARVALHEIRO
COORIENTADORA: LAIZA HELENA DE SOUZA IUNG
COORIENTADOR: HUGO HORÁCIO MONTALDO VALDENEGRO

Aprovado como parte das exigências para obtenção do Título de Doutor em GENÉTICA E MELHORAMENTO ANIMAL, pela Comissão Examinadora:

Pesquisador Dr. ROBERTO CARVALHEIRO (Participaçao Virtual)
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal

Prof. Dr. DANISIO PRADO MUNARI (Participaçao Virtual)
Departamento de Engenharia e Ciências Exatas (DECEx) / FCAV / Unesp - Jaboticabal

Prof. Dr. RAFAEL VILHENA REIS NETO (Participaçao Virtual)
Curso de Engenharia de Pesca / UNESP, Câmpus de Registro, Registro-SP

Prof. Dr. CARLOS ANTONIO LOPES DE OLIVEIRA (Participaçao Virtual)
Centro de Ciências Agrarias / Universidade Estadual de Maringá/PR

PhD. JOSÉ MANUEL YAÑEZ LOPEZ (Participaçao Virtual)
Departamento de Medicina Veterinária Preventiva / Universidad del Chile (Santiago)

Jaboticabal, 04 de novembro de 2021

DADOS CURRICULARES DO AUTOR

Baltasar Fernandes Garcia Neto nasceu em Jaboticabal, São Paulo, em 7 de outubro de 1991. Em 2010, iniciou o curso de graduação em Zootecnia na Universidade Estadual Paulista - UNESP em Jaboticabal. Desde o início do curso, Baltasar atuou na área de Aquicultura, mais especificamente na criação de camarões. Participou de diversos estudos no setor de Carcinicultura do CAUNESP e foi bolsista de pesquisa científica concedida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Em 2013, recebeu uma bolsa do programa Ciência sem Fronteiras concedida pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) para estudar por um ano na Irlanda, na National University of Ireland Galway - NUIG, onde conseguiu aprimorar suas habilidades em inglês e cursar algumas disciplinas. Em 2015, formou-se em Zootecnia pela UNESP, tendo como trabalho de conclusão de curso uma revisão bibliográfica sobre estimativas de herdabilidade em camarões, sob orientação do Dr. Roberto Carvalheiro. Em 2016, ele iniciou seu mestrado na UNESP, no programa de pós-graduação em genética e melhoramento animal, com bolsa da CAPES, sob orientação do Dr. Roberto Carvalheiro, terminando em 2018. O tema da dissertação foi a avaliação de diferentes métodos estatísticos para o mapeamento de QTLs usando simulações. Ainda em 2018, ingressou no curso de doutorado no mesmo programa de pós-graduação e sob a mesma orientação, com bolsa CAPES. Em 2020, foi contemplado com uma bolsa CAPES-PrInt que o permitiu realizar parte das atividades do doutorado na Universidade do Chile (Santiago, Chile) sob orientação do Dr. José Manuel Yánez.

Dedico

Aos meus pais, Baltasar e Silvana, à minha irmã, Carol, e à minha esposa, Rachel, pelo apoio e companheirismo durante essa jornada.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus por todas as bênçãos em minha vida.

Gostaria de agradecer ao Dr. Roberto Carvalheiro por todos os ensinamentos e conselhos. Agradeço por toda paciência, gentileza e por ser um exemplo de zootecnista: ético, profissional e acima de tudo humano.

Gostaria de agradecer a todos os membros das comissões examinadoras interna e externa: Dr. Danísio Prado Munari, Dr. Diogo Teruo Hashimoto, Dr. José Manuel Yánez e Dr. Carlos Antônio Lopes de Oliveira pelo precioso tempo e atenção dispensados com esta tese e pelo seu valor inestimável nas sugestões e contribuições.

Agradeço à Dra. Laiza Helena de Souza Iung por toda ajuda e paciência que além de coorientadora é uma excelente amiga.

Agradeço ao Dr. Hugo Horácio Montaldo, em nome da Universidad Nacional Autónoma de México (UNAM), que gentilmente nos forneceu os dados e contribuiu para o desenvolvimento do segundo capítulo desta tese.

Agradeço ao Dr. José Manuel Yánez, em nome da Universidad de Chile (UChile), que me recebeu de forma muito acolhedora em seu laboratório, forneceu os dados para o desenvolvimento dos capítulos 3 e 4 da presente tese, e também contribuiu ativamente no desenvolvimento destes capítulos.

Agradeço as empresas: Maricultura del Pacífico (México), OPUMARSA (Equador), Aquacorporación Internacional (Costa Rica) e AquaAmerica (Brasil) pelo fornecimento dos dados utilizados nesta tese.

Agradeço ao programa de pós-graduação em Genética e Melhoramento Animal e o curso de graduação em Zootecnia da FCAV- UNESP Jaboticabal. Obrigado por todo conhecimento que me foi apresentado desde 2010. Agradeço à Dra. Sandra Aidar de Queiroz, Dra. Lucia Galvão de Albuquerque e Dr. Danísio Prado Munari, professores em minha graduação e pós-graduação. Obrigado por todo o conhecimento passado e incentivo ao estudo do melhoramento genético animal.

Agradeço aos meus amigos do departamento de Zootecnia e companheiros da “salinha”: Laiza, Thaise, Daiane, Giovana, André Mauric, Andrés (Colômbia), Lúcio, Caio (Índio), Diego (Pumba), Rafael (Japa), Willian, Ivan, Thales, Isabella, Delvan,

Kétuly e Gabriel, pela ajuda em assuntos acadêmicos e também pelos momentos de confraternização nos cafés e churrascos.
Agradeço aos integrantes do “Laboratorio de genómica aplicada a la Acuicultura” na FAVET (UChile): Paulina, Grazyella, Carolina, Lucas, Rodrigo, Pablo, Sebastián, Tamara, David e Jousepth, por me receberem tão bem no grupo de pesquisa e me fazerem sentir em “casa” durante os 12 meses que estive no Chile.
O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

SUMMARY

GENÉTICA QUANTITATIVA E FERRAMENTAS GENÔMICAS APLICADAS AO MELHORAMENTO GENÉTICO NA AQUICULTURA

RESUMO – A procura por animais mais uniformes frente as adversidades ambientais e a aplicação de ferramentas genômicas são exemplos de novas estratégias para tornar o melhoramento genético em espécies aquícolas mais eficiente. No intuito de avaliar a efetividade destas estratégias, os objetivos do presente trabalho foram: i) estimar os componentes genéticos da uniformidade de peso a despesca (PD) e investigar se o PD e sua uniformidade (PD_u) podem afetar a sobrevivência (SOB) em camarões; ii) estimar o desequilíbrio de ligação (LD) em uma população de camarões cultivados usando um novo painel de 50k polimorfismos de nucleotídeos únicos (SNPs) e iii) avaliar a viabilidade da imputação ao nível de sequência em tilápia do Nilo e os impactos do tamanho e origem da população de referência na acurácia. Para i), foram utilizados 149.919 registros de PD e 164.023 de SOB de uma larvicultura de camarões mexicana. Os dados foram agrupados em três conjuntos de acordo com o tipo de produção: viveiros escavados com baixa densidade (S1), tanques de concreto com recirculação em alta densidade (S2) e juntando ambos conjuntos de dados (S1+S2). Um modelo linear generalizado hierárquico duplo foi aplicado para estimar os componentes de (co)variância de PD-PD_u e um modelo bivariado linear misto foi utilizado para estimar os componentes referentes a PD-SOB. As correlações genéticas (r_g) para estas características foram estimadas, além da correlação entre valores genéticos (VG) para PD_u-SOB. Para ii), 96 camarões (40 machos e 56 fêmeas) reprodutores originários de uma larvicultura do Equador foram genotipados utilizando um novo painel de 50k SNPs. Posteriormente, o LD foi estimado usando três controles de qualidade distintos com diferentes filtros de frequência de alelo menor: 0,1 (CQ1); 0,05 (CQ2) e 0,01 (CQ3). Para iii), amostras de DNA de 326 tilápias provenientes de três populações de origens distintas (P_A, P_B e P_C) foram extraídas e sequenciadas. Após o controle de qualidade dos dados de sequência foram obtidos 4,6 milhões de SNPs em comum para todas as populações. A imputação foi feita em quatro cenários distintos: dois tamanhos (10 ou 90% dos animais de cada população) e duas origens de referência (apenas duas populações diferentes ou todas as três populações). Os animais de validação tiveram seus genótipos ocultados mantendo somente 50k SNPs para a imputação a nível de sequência. Foi utilizado o software FImpute3 e a acurácia de imputação foi avaliada através da correlação entre o genótipo imputado e observado (r^2). No estudo de uniformidade, uma proporção significativa de variância genética foi detectada na variância residual de PD, revelando que existe a possibilidade de selecionar para uniformidade desta característica em camarões (coeficiente de variação genética variando de 17 a 35%). As r_g entre PD e SOB foram diferentes em sinal e magnitude com base no sistema de produção, sendo iguais a 0,36, -0,59 e -0,02 para S1, S2 e S1+S2, respectivamente. Os coeficientes de correlação entre o VG de PD_u e SOB foram significativos apenas para S1 (-0,32), sugerindo que a seleção para famílias mais uniformes também pode aumentar as taxas de sobrevivência em um ambiente de menor densidade. Para os dados genômicos de camarões, foram obtidos 34.425, 39.091 e 42.789 SNPs após CQ1, CQ2 e CQ3, respectivamente, exibindo alto grau de polimorfismo e validando a aplicação deste painel de SNPs para esta população de camarões cultivados. O LD decaiu rapidamente nos primeiros 30 KB de distância de 0,2 para 0,07 e depois diminuiu para 0,02 para distancias mais longas (>80 KB). Esses resultados sugerem

incorporação recente de animais de diferentes populações ou linhagens neste grupo de reprodutores e que a seleção genômica e estudos de associação amplo do genoma são viáveis em camarões usando este painel de SNP como ferramenta. Para a imputação, no geral, o r^2 por animal mostrou resultados intermediários variando de 0,37 a 0,56 para P_A e 0,43 a 0,58 para P_B e Pc. Entretanto, os resultados mostraram que foi possível imputar de 50k a aproximadamente 680k com alta acurácia usando dados de sequência de tilápia. Foi observado um aumento de 31,5% para P_A e 24,6% para P_B e P_C no r^2, quando 90% dos animais da mesma população foram usados como referência em comparação a 10%. Não houve diferenças significativas para r^2 entre os cenários que usaram 90% dos animais da mesma população e usaram animais das três populações como referência, mostrando que a estratégia de usar informações de outra população para aumentar a população de referência teve pouco efeito na acurácia de imputação.

Palavras-chave: Camarão, Desequilíbrio de ligação, Imputação de genótipos, Tilápia, Uniformidade

QUANTITATIVE GENETICS AND GENOMIC TOOLS APPLIED TO AQUACULTURE BREEDING

ABSTRACT – The search for more uniform animals in the face of environmental adversities and the application of genomic tools are examples of new strategies to make genetic improvement in aquaculture species more efficient. In order to evaluate the effectivity of these strategies, the aims of the present study were: i) to estimate the genetic components of uniformity of harvest weight (HW) and investigate whether the HW and its uniformity (HW_v) may affect the survival (SUR) in shrimp; ii) to estimate linkage disequilibrium (LD) in a population of farmed shrimp using a new 50k single-nucleotide polymorphisms (SNPs) panel and; iii) to assess the feasibility of genotype imputation to sequence level in Nile tilapia and the impacts of size and origin of population reference on accuracy of imputation. For i), 149,919 records of HW and 164,023 of SUR were obtained from a Mexican shrimp hatchery. The data were grouped into three sets according to the system of production: excavated ponds with low density (S1), concrete recirculation tanks with high density (S2) and joining both sets of data (S1+S2). A double hierarchical generalized linear model was applied to estimate the (co)variance components of HW-HW_v and a bivariate mixed linear model was used to estimate the components referring to HW-SUR. The genetic correlations (r_g) for these traits were estimated, in addition to the correlation between the estimated breeding values (EBV) for HW_v-SUR. For ii), 96 broodstock shrimp (40 males and 56 females) from a commercial hatchery from Ecuador were genotyped using a new 50k SNPs panel. Then, the LD was estimated using three distinct quality controls with different minor allele frequency filters: 0.1 (QC1); 0.05 (QC2) and 0.01 (QC3). For iii), DNA samples from 326 tilapia from three populations (P_A, P_B and P_C) were extracted and sequenced. After quality control of sequence data, 4.6 million of SNPs were obtained in common for all populations. The imputation was performed in four different scenarios: two sizes (10 or 90% of animals from each population) and two origins of population reference (only two different populations or all three populations). The validation animals had their genotypes masked keeping only 50k SNPs for imputation to the sequence level. The FImpute3 software was used and the imputation accuracy was evaluated through the correlation between the imputed and observed genotypes (r^2). In the uniformity study, a significant proportion of genetic variance was detected in the residual variance of HW, revealing that there is possibility to select for uniformity of this trait in shrimp (genetic coefficient of variation ranging between 17 and 35%). The genetic correlations between HW and SUR were different in sign and magnitude based on the production system, being equal to 0.36, -0.59 and -0.02 for S1, S2 and S1+S2, respectively. The correlation coefficients between the EBV of HW_v and SUR were significant only for S1 (-0.32), suggesting that selection for more uniform families may also increase survival rates in a lower density environment. For shrimp genomic data, 34,425, 39,091 and 42,789 SNPs were obtained after QC1, QC2 and QC3, respectively, showing high polymorphism and validating the application of this panel of SNPs to this farmed shrimp population. The LD declined rapidly in the first 30 KB of distance from 0.2 to 0.07 and then decreased to 0.02 for longer distances (>80 KB). These results suggest recent incorporation of animals from different populations or strains into this breeding population and that genomic selection and genome-wide association studies are feasible in shrimp using this SNP panel as tool. For the imputation using tilapia sequence data, in general, the r^2 per animal showed intermediate results ranging from 0.37 to 0.56 for P_A and 0.43 to 0.58 for P_B and P_C.

However, the results showed that it was possible to impute from 50k to approximately 680k with high accuracy. An increase in r^2 of 31.5% for P_A and 24.6% for P_B and P_C was observed, when 90% of animals from the same population were used as reference. There were no significant differences for r^2 between the scenarios that used 90% of animals from the same population and used animals from the three populations as a reference, showing that the strategy of using information from another population to increase the reference population had minor effect on the accuracy of imputation.

Keywords: Shrimp, Linkage Disequilibrium, Genotype Imputation, Tilapia, Uniformity

CHAPTER 1 – General Considerations

1.1 Introduction

Currently, aquaculture has played an important role in food production. In 2019, aquaculture was responsible by the production of approximately 120.1 million of tons representing an increase of 54% in comparison to 2009. Similar tendency was observed in the Brazilian sector, which approximately 600.3 thousands of tons were produced in 2019 (45.7% more than the production registered ten years ago) (FAO, 2021). Shrimp and tilapia farming were segments that represented large proportions of the total aquaculture production in Brazil in 2019, 9.1 and 53.9%, respectively (FAO, 2021).

The implementation of breeding programs is a strategy employed to improve the mean level of traits that present economic return. Traits related to growth (e.g., weight at harvest or specific age), survival (e.g., survival in specific production systems and after disease challenge), reproduction (e.g., age at sexual maturation and number of eggs) and meat quality (e.g., fillet color and yield) are groups of traits often contemplated in the aquaculture breeding programs (Gjedrem and Baranski, 2009). Although breeding programs focused on improving the mean level of these traits have been successful (Gjedrem and Rye, 2018), there is a new concern to decrease the phenotypic variability of such traits, i.e., to increase uniformity. In order to select for more uniform animals, the low residual (environmental) variance may be used as selection criterium, if this is a trait under genetic control (Mulder et al., 2008). Despite its importance to aquaculture industry and relevance to the effectivity of breeding programs, uniformity studies are scarce in this field.

Another strategy to increase the effectivity of breeding programs is the inclusion of genomic information. With the advent of new technologies, it is possible to extract DNA from animals, genotype this samples and include this as additional source of information. For breeding programs, the genomic information has two main applications: genomic selection (GS) and genome-wide association studies (GWAS). GS consists of estimating the genomic breeding values of selection candidates using a prediction equation that was previously calculated in a reference population (Meuwissen et al., 2001). GS requires a sufficient number (thousands) of single-nucleotide polymorphisms (SNPs) markers, allowing to capture genetic variability

through linkage disequilibrium (LD) with quantitative trait loci (QTL). On the other hand, GWAS may help to disentangle the genetic architecture of traits and detect genomic regions associated to genetic variation by using LD information between SNPs and QTLs. In this sense, when a SNP panel is developed, it is important to determine the level and extension of LD to evaluate its applicability in breeding programs.

The use of whole-genome sequencing (WGS) to obtain genotypic information is expected to accelerate the genetic gain in comparison to genotyping panels because it may include millions of SNPs with potential causal variants related to relevant phenotypes (Meuwissen and Goddard, 2010). Nevertheless, WGS faces an important drawback compared to SNP panels: it is much more expensive mainly when many animals are required. In this sense, the genotype imputation is a cost-effective strategy to obtain genotypes at sequence level. The imputation is performed using WGS information from key-animals as reference to predict the missing genotypes of target animals that were previously genotyped in a lower density. The size and origin of reference population are important aspects that should be evaluated to perform genotype imputation with high accuracy (Carvalheiro et al., 2014; Hickey et al., 2012). In spite of the increasing number of WGS studies in aquaculture species, none study was found evaluating how these factors may affect the final accuracy of imputation.

1.2 Literature Review

1.2.1 Uniformity

1.2.1.1 Genetics of uniformity in aquaculture species

The uniformity is a new concern in animal production mainly by three reasons: industry and consumer preferences, operational problems during the grow-out phase and phenotypic stability facing environmental changes (Sae-Lim et al., 2016). A more uniform product may ease the processing in industry and also aggregates value to the final product because it is often appreciated by the consumer market (Khaw et al., 2016). In addition to the direct economic return associated to uniformity, more homogeneous animals may reduce economic loss during the rearing phase. This effect is particularly noticed in aquaculture species in which large variation in size, for example, may stimulate competition for feed increasing mortality, encouraging cannibalistic behavior and reducing the animal welfare (Marjanovic et al., 2018). A final concept may be associated to uniformity in production animals, the “robustness”.

“Robustness”, also called phenotypic plasticity, means that in the advent of environmental changes the animal will not present variation in its phenotype, i.e., more uniform animals are expected to present lower phenotypic fluctuation when exposed to environmental variations (Elgersma et al., 2018).

The genotype by environment (GxE) interaction occurs when the performance of different individuals is affected by changes in the environmental conditions (Falconer and Mackay, 1996). Thinking about breeding programs, the GxE interaction may be a problem because individuals selected as superior in one environment may present different performance in another environment causing reranking of animals. The environmental factors that may cause GxE interaction may be grouped into macro-environmental and micro-environmental sensitivity. The macro-environmental factors are measurable and noticeable changes, such as different production systems, farm locations and stocking densities. On the other hand, micro-environmental sensitivity is related to non-measurable factors, such as pathogens, social interactions, and epigenetic effects (Sae-Lim et al., 2016). It is in the micro-environmental sensitivity that the uniformity may be evaluated, i.e., to evaluate the ability of each individual to keep its phenotype stable facing minor environmental disturbances (Mulder et al., 2008).

Aquaculture environments are very unstable and are subject to higher micro-environmental variation compared to other livestock species (e.g. pig and chicken) mainly because most of production systems involve constant interaction among individuals, for example, “raceways” in shrimp and net cages in tilapia farming (Castillo-Juárez et al., 2012; Khaw et al., 2016). Selection for more uniform animals is possible if the genetic component related to this trait is heritable. Several studies showed that, possibly, the residual variance allocates the genetic variance of uniformity (Hill and Mulder, 2010; Sorensen and Waagepetersen, 2003). Thus, in order to select for uniformity, a strategy would be to detect if there is genetic heterogeneity of environmental (residual) variance for determined trait (Mulder et al., 2006; SanCristobal-Gaudy et al., 2001).

1.2.1.2 Models to evaluate the uniformity

Several methods may be used to estimate the genetic component of the residual variance, such as the additive model (Hill and Zhang, 2004; Mulder et al., 2007), the exponential model (SanCristobal-Gaudy et al., 1998; Sorensen and

Waagepetersen, 2003) and the double hierarchical generalized model (DHGLM) (Mulder et al., 2009). Similarly to the exponential model, the DHGLM uses the natural logarithm of the residual variance in the exponential form $\ln(\hat{e}^2)$, but it is implemented in a two-step approach. In the first step, the following mean model is applied:

$$y = Xb + Za + e,$$

then, the residual variance model is applied in the second step:

$$\ln(\hat{e}^2) = Xb_V + Za_V + e_V,$$

where y and $\ln(\hat{e}^2)$ are the responsible variables for the mean and residual variance models, respectively, Xb (Xb_V), Za (Za_V), e(e_V) are fixed and random effects for the mean and residual variance models, respectively. In order to estimate the variance components in a single-step, Felleki et al. (2012) proposed an extension to the DHGLM in which a bivariate model for the mean and residual variance could be applied:

$$\begin{bmatrix} \mathrm{y} \\ \Psi \end{bmatrix} = \begin{bmatrix} \mathrm{X} & 0 \\ 0 & \mathrm{X_v} \end{bmatrix} \begin{bmatrix} \mathrm{b} \\ \mathrm{b_v} \end{bmatrix} + \begin{bmatrix} \mathrm{Z} & 0 \\ 0 & Z_V \end{bmatrix} \begin{bmatrix} \mathrm{a} \\ \mathrm{a_v} \end{bmatrix} + \begin{bmatrix} \mathrm{e} \\ \mathrm{e_v} \end{bmatrix},$$

where y is the mean trait and Ψ is the linearized response variable for the residual variance defined as:

$$\Psi_\mathrm{i} = \log\left(\hat{\sigma}^2_{\mathrm{e_i}}\right) + (\{[\hat{\mathrm{e}}^2_\mathrm{i}/(1 - \mathrm{h_i})] - \hat{\sigma}^2_{\mathrm{e_i}}\}/\hat{\sigma}^2_{\mathrm{e_i}}),$$

where $\hat{\sigma}^2_{\mathrm{e_i}}$ is the predicted residual variance for the *i*th observation obtained in the previous iteration, $\hat{\mathrm{e}}^2_\mathrm{i}$ is the squared residual from the mean model estimated for the *i*th observation, $\mathrm{h_i}$ is the leverage for the *i*th observation, which is the *i*th diagonal element of the hat matrix (Hoaglin and Welsch, 1978), $X(X_V)$ and $Z(Z_V)$, are incidence matrices, $b(b_V)$, $a(a_V)$, $e(e_V)$, are fixed and random effects for the mean and residual variance models, respectively. Other models may be applied to estimate the genetic components of the residual variance using, for example, Bayesian models (Sorensen and Waagepetersen, 2003) and models that incorporate the phenotypic coefficient of variation (Khaw et al., 2016). However, the most adequate method depends on the data structure available. It is important to highlight that uniformity requires multiple measures either at individual level (as repeated records) or within-family level. Repeated measures allow the estimation of standard deviation and coefficient variation which may be a direct measure of uniformity (Khaw et al., 2016). Most commonly, the aquaculture breeding programs are based on multiple families with a large number of full and half-sibs with a single observation, thus, the DHGLM is an

appropriate model to evaluate uniformity considering multiples observations at family level (Iung et al., 2019).

1.2.1.3 Estimation of uniformity genetic parameters

Two main genetic parameters may be used to interpret and compare the uniformity results: heritability of residual variance (h_V^2) and genetic coefficient of variation (GCV_E), also called evolvability (Mulder et al., 2007). The h_V^2 is related to the EBV accuracy of prediction for the residual variance and may be estimated as:

$$h_V^2 = \frac{\sigma^2_{a_{v,add}}}{\left(2\sigma_P^4 + 3\sigma^2_{a_{v,add}}\right)},$$

where $\sigma^2_{a_{v,add}}$ is the additive genetic variance estimated for the residual variance on the additive scale and σ_P^4 is the squared phenotypic variance. The $\sigma^2_{a_{v,add}}$ may be obtained as (Mulder et al., 2009):

$$\sigma^2_{a_{v,add}} = h^2_{res} 2\left(\bar{\sigma}^2_{\hat{e}}\right)^2,$$

where h^2_{res} is the heritability of the response variable used in the residual variance model, and $\bar{\sigma}^2_{\hat{e}}$ is the average residual variance in the mean model. The GCV_E indicates the potential response to selection for residual variance and may be calculated as:

$$GCV_E = \frac{\sigma_{a_{v,add}}}{\bar{\sigma}^2_{\hat{e}}},$$

where $\sigma_{a_{v,add}}$ is the standard deviation of the additive genetic variance for the residual variance on the additive scale.

Iung et al. (2019) reviewed uniformity studies in several livestock and aquaculture species. The authors found only six studies regarding this topic in aquaculture, all of them focused in fish: rainbow trout (Janhunen et al., 2012; Sae-Lim et al., 2015), Atlantic salmon (Sae-Lim et al., 2017; Sonesson et al., 2013) and Nile tilapia (Agha et al., 2018; Marjanovic et al., 2016). A low h_V^2 was found in most of studies implying that a high number of animals would be necessary to achieve high accuracy of selection for uniformity. They also found that compared to livestock species (sheep, pig, chicken and cattle), fish showed high GCV_E (0.36 on average) indicating that uniformity may be improved through selection in aquaculture species efficiently. More recently, other uniformity studies were also performed for other aquaculture species: shrimp (García-Ballesteros et al., 2021; Garcia et al., 2021b), oyster (Vu et al., 2021) and lumpfish (Sae-Lim et al., 2020). These studies also revealed potential to increase uniformity using selection.

1.2.2 Genomic tools applied to aquaculture breeding

1.2.2.1 Genomic resources and genotyping methods in aquaculture

The application of genomic tools for breeding purposes is more recent in aquaculture species than in other livestock species. Several important aquaculture species still have poor quality of reference genome and many more species have limited genomic resources available hardening the implementation of genomic tools (Houston et al., 2020). The lack of genomic resources and high-quality assembling goes beyond the more recent interest and investment issues. The genome assembling is much more challenging in aquatic species due to more complex genomic characteristics. The genome duplication events in salmonids (Lien et al., 2016) and the excess of heterozygosity loci in crustaceans (Zhang et al., 2019) are examples of intrinsic factors that difficult the assembling. Nevertheless, each year, more species have their genome sequenced increasing the knowledge of genomic aspects of important aquaculture species. You et al. (2020) found WGS projects for genome assembling in 52 aquatic species of economic relevance. The majority of these studies use preliminary versions of genome which have scaffolds instead of chromosomes. However, some species, such as Atlantic salmon and Nile tilapia, have already updated and more accurate genome versions which has allowed the development of several genomic studies varying from functional, such as high-density linkage maps and gene annotation (Lu and Luo, 2020) to breeding applications, such as GS and GWAS (You et al., 2020).

The genetic markers, predominantly SNPs, are the basic unity of genomic information for breeding applications. In order to genotype animals and discover markers, three main techniques have been employed in aquatic species: genotype by sequencing (GBS), SNP arrays and whole-genome sequencing (WGS). The GBS methods include a series of other techniques (e.g., RAD-seq, 2bRAD and ddRAD) which combine high-throughput sequencing and the construction of libraries to genotype specific sites at high-coverage. The main advantage of GBS is that a reference genome is not necessary to genotype and discover SNPs which is particularly interesting for emerging aquaculture species with poor genomic information (Robledo et al., 2018). However, due to complex preparations and pipelines, GBS may lead to genotyping errors compromising its repeatability across different datasets (Palti et al., 2015b). The SNP arrays are simpler platforms of pre-

defined SNPs that require previous genomic knowledge about the species. For this reason, there is a limited number of species with this tool available: Atlantic salmon (e.g., Yáñez et al., 2016), common carp (e.g., Xu et al., 2014), channel catfish (e.g., Liu et al., 2014), rainbow trout (e.g., Palti et al., 2015a), Pacific oyster (Gutierrez et al., 2018), Nile tilapia (e.g., Yáñez et al., 2020), sea bass (Griot et al., 2021), serrasalmid fish (Mastrochirico-Filho et al., 2021) and Pacific white shrimp (Garcia et al., 2021a). The SNP arrays have been largely used for GS and GWAS in livestock species and there is similar tendency for aquaculture. The WGS method is currently the most complete source of genomic information. This technology yields larges volume of data and millions of SNPs may be discovered from it. The main benefit of using WGS for breeding and population genetics purposes is the higher accuracy of information because the complete genome is sequenced several times. In spite of reducing costs in the last years, the WGS method is still very expensive in comparison to other genotyping methods. However, due to progressive advances in bioinformatics, and consequently in the reduction of costs, some authors predict that WGS will become prevalent in the future for the genotyping of production animals (Rubinacci et al., 2021; You et al., 2020).

1.2.2.2 Linkage disequilibrium

The linkage disequilibrium (LD) has a fundamental role in the application of genomic tools. For example, both GS and GWAS assumes that SNPs distributed over the genome are in LD with QTL affecting the interesting traits (Goddard and Hayes, 2009; Meuwissen et al., 2001). Looking at LD over the genome, it is also possible to know more about past events that affected the population (e.g., cross-breeding and geographical bottlenecks), and a more detailed view (LD at chromosome segments) reflects evolutionary forces that may have changed the population's genetic composition (e.g. artificial/natural selection and mutation) (Heuertz et al., 2006).

By definition, the LD is a non-random association of alleles at two or more loci (Falconer and Mackay, 1996). Several methods were proposed in order to measure LD. Currently, D' and r^2 are the most common methods applied in genomic studies. Considering a pair of loci with alleles *A* and *a* at locus one, and *B* and *b* at locus two, with allele frequencies P_A, P_a, P_B, and P_b, respectively, the resulting haplotype frequencies are P_{AB}, P_{Ab}, P_{aB}, and P_{ab}. We may calculate then:

$$D_{ab} = P_{AB} - P_A P_B,$$

where D_{AB} is the LD coefficient, P_{AB} is the frequency of gametes carrying the pair of alleles A and B at two loci, $P_A P_B$ are the product of the frequencies of alleles A and B, respectively. Using this coefficient, we may obtain D' (Lewontin, 1964):

$|D'| = \frac{(D_{ab})^2}{min(P_A P_b, P_a P_B)}$, for D_{ab} < 0;

$|D'| = \frac{(D_{ab})^2}{min(P_A P_B, P_a P_b)}$, for D_{ab} >0

In case of no recombination between these alleles (i.e., D'=1), there is complete LD. In the same way, we may obtain r^2 (Hill and Robertson, 1968):

$r^2 = \frac{(D_{ab})^2}{P_A P_a P_B P_b}$

Similarly, if r^2 is equal to 1, we may say there is perfect LD and, it is only possible, if the markers have not been separated by recombination and have the same allele frequency. The D' measure is highly affected by sample size, mainly for SNPs with low MAF, resulting in high levels of LD between loci when they are actually in equilibrium (Ardlie et al., 2002). Thus, r^2 is more recommended to measure LD between loci using SNP panels.

LD has been evaluated in different aquaculture species using distinct SNP platforms and densities, for example: Atlantic salmon (Barría et al., 2018); coho salmon (Barría et al., 2019); rainbow trout (Vallejo et al., 2018); Nile tilapia (Yoshida et al., 2019a) and Pacific oyster (Zhong et al., 2017). For shrimp, most of studies used low density SNP panels (up to 6.5K) (e.g., Jones et al., 2017) or GBS techniques (e.g., Wang et al., 2019) that may not represent with high accuracy the population LD for this species. A new 50K SNP array is commercially available for shrimp, but to the best of our knowledge, no LD studies were performed using this new tool. It is important to evaluate LD considering that several factors may affect it, such as allelic frequencies, recombination rates, population admixture, population bottlenecks, among others (Espigolan et al., 2013). Thus, LD is population-specific and its distinct patterns across different populations may compromise the repeatability of genomic studies. For this reason, the extent of LD should be evaluated before considering to implement GS and GWAS and it may validate the applicability of this novel tool to farmed populations.

1.2.2.3 Genotype imputation

The genotype imputation is a cost-effective strategy to maximize the number of animals with genotypes at higher densities reducing costs. In this strategy, a low proportion of animals is genotyped at high density (or sequenced) to be used as reference, then target animals that have genomic information at lower density may have their genotypes imputed at higher density. This approach is feasible using low, medium and high SNP densities and also using WGS.

Using information imputed to sequence level to genomic predictions is advantageous because it potentially includes causal mutations increasing the accuracy of estimated breeding values (Meuwissen and Goddard, 2010). In addition, the detection of QTL is also improved in GWAS due to the inclusion of rarer variants and greater genome coverage increasing the possibility of finding SNPs that are close and/or in LD with the QTL (MacLeod et al., 2016). However, the success of this strategy depends on how accurate was the imputation process.

Several factors may influence the accuracy of imputation, such as the number of SNPs present in the low and high density panel, method of imputation, reference population size and origin, minor allele frequency (MAF) of the imputed SNP, among others (Calus et al., 2014). The reference size and origin are more important for methods of imputation that use family information, such as in FImpute3 (Sargolzaei et al., 2014) and AlphaImpute software (Hickey et al., 2011). Basically, these methods search for haplotypes shared between reference and target animals to predict the missing genotypes. The haplotype construction is more accurate for closer relatives because they share longer haplotype blocks requiring less animals (Sargolzaei et al., 2014). In this way, if the reference animals are ancestors of target animals, for example, a lower number of animals is necessary to have high accuracy of imputation. On the other hand, more animals are needed to achieve the same accuracy level if the relationship between target and reference animals is low.

The genotype imputation has been successfully applied in aquaculture. For example, Yoshida et al. (2019) performed imputation from low density (0.5K, 1K and 3K) to medium density (50K) in Nile tilapia and reported high accuracy of imputation results (correlation between imputed and observed genotypes greater than 0.9). The authors also showed high accuracy in the genomic prediction for growth and fillet yield using the imputed data. However, genotype imputation studies using WGS is scarce

in aquaculture and we found only two studies using this type of data (Yoshida and Yáñez, 2021a, 2021b). Still, none study was found evaluating how reference size and origin may affect the accuracy of imputation.

1.3 Objectives

- to estimate the genetic component of residual variance for harvest weight and to investigate the effect of weight and its uniformity on the survival of Pacific white shrimp reared in different culture systems during the grow-out period.
- to characterize the genetic diversity, estimate the linkage disequilibrium and the effective population size present in a farmed population of Pacific white shrimp.
- to evaluate the feasibility of genotype imputation from approximately 50k SNPs to sequence level in Nile tilapia and to investigate the impacts of size and origin of reference population in the accuracy of imputation.

1.4 References

Agha, S., Mekkawy, W., Ibanez-Escriche, N., Lind, C.E., Kumar, J., Mandal, A., Benzie, J.A.H., Doeschl-Wilson, A., 2018. Breeding for robustness: investigating the genotype-by-environment interaction and micro-environmental sensitivity of Genetically Improved Farmed Tilapia (Oreochromis niloticus). Anim. Genet. 49, 421–427. https://doi.org/10.1111/age.12680

Ardlie, K.G., Kruglyak, L., Seielstad, M., 2002. Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. https://doi.org/10.1038/nrg777

Barría, A., Christensen, K.A., Yoshida, G., Jedlicki, A., Leong, J.S., Rondeau, E.B., Lhorente, J.P., Koop, B.F., Davidson, W.S., Yáñez, J.M., 2019. Whole Genome Linkage Disequilibrium and Effective Population Size in a Coho Salmon (Oncorhynchus kisutch) Breeding Population Using a High-Density SNP Array. Front. Genet. 10, 498. https://doi.org/10.3389/fgene.2019.00498

Barría, A., López, M.E., Yoshida, G., Carvalheiro, R., Lhorente, J.P., Yáñez, J.M., 2018. Population Genomic Structure and Genome-Wide Linkage Disequilibrium in Farmed Atlantic Salmon (Salmo salar L.) Using Dense SNP Genotypes. Front. Genet. 9, 649. https://doi.org/10.3389/fgene.2018.00649

Calus, M.P.L., Bouwman, A.C., Hickey, J.M., Veerkamp, R.F., Mulder, H.A., 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. animal 8, 1743–1753. https://doi.org/10.1017/S1751731114001803

Carvalheiro, R., Boison, S.A., Neves, H.H.R., Sargolzaei, M., Schenkel, F.S., Utsunomiya, Y.T., O'Brien, A.M.P., Sölkner, J., McEwan, J.C., Van Tassell, C.P., Sonstegard, T.S., Garcia, J.F., 2014. Accuracy of genotype imputation in Nelore cattle. Genet. Sel. Evol. 46, 1–11. https://doi.org/10.1186/s12711-014-0069-1

Castillo-Juárez, H., Montaldo, H.H., Campos-Montes, G.R., 2012. Genetic parameter estimates of the environmental variation for body weight at harvest size in a Pacific white shrimp breeding population, in: 4th International Conference on Quantitative Genetics: Understanding Variation in Complex Traits. Edinburgh, p. 167.

Elgersma, G.G., de Jong, G., van der Linde, R., Mulder, H.A., 2018. Fluctuations in milk yield are heritable and can be used as a resilience indicator to breed healthy cows. J. Dairy Sci. 101, 1240–1250. https://doi.org/10.3168/jds.2017-13270

Espigolan, R., Baldi, F., Boligon, A.A., Souza, F.R.P., Gordo, D.G.M., Tonussi, R.L., Cardoso, D.F., Oliveira, H.N., Tonhati, H., Sargolzaei, M., Schenkel, F.S., Carvalheiro, R., Ferro, J.A., Albuquerque, L.G., 2013. Study of whole genome linkage disequilibrium in Nellore cattle. BMC Genomics 14, 1–8. https://doi.org/10.1186/1471-2164-14-305

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to quantitative genetics., 4th ed, Introduction to quantitative genetics. Longmans Green, Harlow, Essex, UK.

FAO, 2021. FAO Fisheries & Aquaculture - Statistics [WWW Document]. URL http://www.fao.org/fishery/statistics/en (accessed 6.30.21).

Felleki, M., Lee, D., Lee, Y., Gilmour, A.R., Rönnegård, L., 2012. Estimation of breeding values for mean and dispersion, their variance and correlation using double hierarchical generalized linear models. Genet. Res. (Camb). 94, 307–317. https://doi.org/10.1017/S0016672312000766

García-Ballesteros, S., Villanueva, B., Fernández, J., Gutiérrez, J.P., Cervantes, I., 2021. Genetic parameters for uniformity of harvest weight in Pacific white shrimp (Litopenaeus vannamei). Genet. Sel. Evol. 2021 531 53, 1–9. https://doi.org/10.1186/S12711-021-00621-6

Garcia, B.F., Bonaguro, Á., Araya, C., Carvalheiro, R., Yáñez, J.M., 2021a. Application of a novel 50K SNP genotyping array to assess the genetic diversity and linkage disequilibrium in a farmed Pacific white shrimp (Litopenaeus vannamei) population. Aquac. Reports 20, 100691. https://doi.org/10.1016/J.AQREP.2021.100691

Garcia, B.F., Montaldo, H.H., Iung, L.H.S., Carvalheiro, R., 2021b. Effect of harvest weight and its uniformity on survival in Litopenaeus vannamei reared in different systems. Aquaculture 531, 735891. https://doi.org/10.1016/J.AQUACULTURE.2020.735891

Gjedrem, T., Baranski, M., 2009. Selective Breeding in Aquaculture: An Introduction, 1st ed, Reviews: Methods and Technologies in Fish Biology and Fisheries. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-90-481-2773-3

Gjedrem, T., Rye, M., 2018. Selection response in fish and shellfish: a review. Rev. Aquac. 1–12. https://doi.org/10.1111/raq.12154

Goddard, M.E., Hayes, B.J., 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10, 381–391.

https://doi.org/10.1038/nrg2575

Griot, R., Allal, F., Phocas, F., Brard-Fudulea, S., Morvezen, R., Bestin, A., Haffray, P., François, Y., Morin, T., Poncet, C., Vergnet, A., Cariou, S., Brunier, J., Bruant, J.S., Peyrou, B., Gagnaire, P.A., Vandeputte, M., 2021. Genome-wide association studies for resistance to viral nervous necrosis in three populations of European sea bass (Dicentrarchus labrax) using a novel 57k SNP array DlabChip. Aquaculture 530, 735930. https://doi.org/10.1016/J.AQUACULTURE.2020.735930

Gutierrez, A.P., Matika, O., Bean, T.P., Houston, R.D., 2018. Genomic Selection for Growth Traits in Pacific Oyster (Crassostrea gigas): Potential of Low-Density Marker Panels for Breeding Value Prediction. Front. Genet. 0, 391. https://doi.org/10.3389/FGENE.2018.00391

Heuertz, M., De Paoli, E., Källman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M., Gyllenstrand, N., 2006. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [Picea abies (L.) Karst]. Genetics 174, 2095–2105. https://doi.org/10.1534/genetics.106.065102

Hickey, J.M., Crossa, J., Babu, R., de los Campos, G., 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52, 654–663. https://doi.org/10.2135/cropsci2011.07.0358

Hickey, J.M., Kinghorn, B.P., Tier, B., Wilson, J.F., Dunstan, N., van der Werf, J.H., 2011. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet. Sel. Evol. 2011 431 43, 1–13. https://doi.org/10.1186/1297-9686-43-12

Hill, W.G., Mulder, H.A., 2010. Genetic analysis of environmental variation. Genet. Res. (Camb). 92, 381–395. https://doi.org/10.1017/S0016672310000546

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38, 226–231. https://doi.org/10.1007/BF01245622

Hill, W.G., Zhang, X.-S., 2004. Effects on phenotypic variability of directional selection arising through genetic differences in residual variability. Genet. Res. (Camb). 83, 121–132. https://doi.org/10.1017/S0016672304006640

Hoaglin, D.C., Welsch, R.E., 1978. The Hat Matrix in Regression and ANOVA. Am. Stat. 32, 17–22. https://doi.org/10.1080/00031305.1978.10479237

Houston, R.D., Bean, T.P., Macqueen, D.J., Gundappa, M.K., Jin, Y.H., Jenkins, T.L.,

Selly, S.L.C., Martin, S.A.M., Stevens, J.R., Santos, E.M., Davie, A., Robledo, D., 2020. Harnessing genomics to fast-track genetic improvement in aquaculture. Nat. Rev. Genet. 2020 217 21, 389–409. https://doi.org/10.1038/s41576-020-0227-y

Iung, L.H.S., Carvalheiro, R., Neves, H.H.R., Mulder, H.A., 2019. Genetics and genomics of uniformity and resilience in livestock and aquaculture species: A review. J. Anim. Breed. Genet. 1–18. https://doi.org/10.1111/jbg.12454

Janhunen, M., Kause, A., Vehviläinen, H., Järvisalo, O., 2012. Genetics of Microenvironmental Sensitivity of Body Weight in Rainbow Trout (Oncorhynchus mykiss) Selected for Improved Growth. PLoS One 7, e38766. https://doi.org/10.1371/JOURNAL.PONE.0038766

Jones, D.B., Jerry, D.R., Khatkar, M.S., Raadsma, H.W., Steen, H. Van Der, Prochaska, J., Forêt, S., Zenger, K.R., 2017. A comparative integrated gene-based linkage and locus ordering by linkage disequilibrium map for the Pacific white shrimp, Litopenaeus vannamei. Sci. Rep. 7, 1–16. https://doi.org/10.1038/s41598-017-10515-7

Khaw, H.L., Ponzoni, R.W., Yee, H.Y., Aziz, M.A. bin, Mulder, H.A., Marjanovic, J., Bijma, P., 2016. Genetic variance for uniformity of harvest weight in Nile tilapia (Oreochromis niloticus). Aquaculture 451, 113–120. https://doi.org/10.1016/j.aquaculture.2015.09.003

Lewontin, R.C., 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics 49, 49–67.

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., Hvidsten, T.R., Leong, J.S., Minkley, D.R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R.A., von Schalburg, K., Rondeau, E.B., Di Genova, A., Samy, J.K.A., Olav Vik, J., Vigeland, M.D., Caler, L., Grimholt, U., Jentoft, S., Inge Våge, D., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D.R., Yorke, J.A., Nederbragt, A.J., Tooming-Klunderud, A., Jakobsen, K.S., Jiang, X., Fan, D., Hu, Y., Liberles, D.A., Vidal, R., Iturra, P., Jones, S.J.M., Jonassen, I., Maass, A., Omholt, S.W., Davidson, W.S., 2016. The Atlantic salmon genome provides insights into rediploidization. Nat. 2016 5337602 533, 200–205. https://doi.org/10.1038/nature17164

Liu, S., Sun, L., Li, Y., Sun, F., Jiang, Y., Zhang, Y., Zhang, J., Feng, J., Kaltenboeck,

L., Kucuktas, H., Liu, Z., 2014. Development of the catfish 250K SNP array for genome-wide association studies. BMC Res. Notes 2014 71 7, 1–12. https://doi.org/10.1186/1756-0500-7-135

Lu, G., Luo, M., 2020. Genomes of major fishes in world fisheries and aquaculture: Status, application and perspective. Aquac. Fish. 5, 163–173. https://doi.org/10.1016/J.AAF.2020.05.004

MacLeod, I.M., Bowman, P.J., Vander Jagt, C.J., Haile-Mariam, M., Kemper, K.E., Chamberlain, A.J., Schrooten, C., Hayes, B.J., Goddard, M.E., 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17, 144. https://doi.org/10.1186/s12864-016-2443-6

Marjanovic, J., Mulder, H.A., Khaw, H.L., Bijma, P., 2016. Genetic parameters for uniformity of harvest weight and body size traits in the GIFT strain of Nile tilapia. Genet. Sel. Evol. 48, 1–10. https://doi.org/10.1186/s12711-016-0218-9

Marjanovic, J., Mulder, H.A., Rönnegård, L., Bijma, P., 2018. Modelling the co-evolution of indirect genetic effects and inherited variability. Heredity (Edinb). 631–647. https://doi.org/10.1038/s41437-018-0068-z

Mastrochirico-Filho, V.A., Ariede, R.B., Freitas, M. V., Borges, C.H.S., Lira, L.V.G., Mendes, N.J., Agudelo, J.F.G., Cáceres, P., Berrocal, M.H.M., Sucerquia, G.A.L., Porto-Foresti, F., Yáñez, J.M., Hashimoto, D.T., 2021. Development of a multi-species SNP array for serrasalmid fish Colossoma macropomum and Piaractus mesopotamicus. Sci. Reports 2021 111 11, 1–11. https://doi.org/10.1038/s41598-021-98885-x

Meuwissen, T., Goddard, M., 2010. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. Genetics 185, 623–631. https://doi.org/10.1534/GENETICS.110.116590

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819–1829.

Mulder, H.A., Bijma, P., Hill, W.G., 2008. Selection for uniformity in livestock by exploiting genetic heterogeneity of residual variance. Genet. Sel. Evol. 40, 37. https://doi.org/10.1186/1297-9686-40-1-37

Mulder, H.A., Bijma, P., Hill, W.G., 2007. Prediction of breeding values and selection responses with genetic heterogeneity of environmental variance. Genetics 175,

1895–1910. https://doi.org/10.1534/genetics.106.063743

Mulder, H.A., Hill, W.G., Vereijken, A., Veerkamp, R.F., 2009. Estimation of genetic variation in residual variance in female and male broiler chickens. animal 3, 1673–1680. https://doi.org/10.1017/S1751731109990668

Mulder, H.A., Veerkamp, R.F., Ducro, B.J., Van Arendonk, J.A.M., Bijma, P., 2006. Optimization of dairy cattle breeding programs for different environments with genotype by environment interaction. J. Dairy Sci. 89, 1740–1752. https://doi.org/10.3168/jds.S0022-0302(06)72242-1

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

Palti, Y., Vallejo, R.L., Gao, G., Liu, S., Hernandez, A.G., III, C.E.R., Wiens, G.D., 2015b. Detection and Validation of QTL Affecting Bacterial Cold Water Disease Resistance in Rainbow Trout Using Restriction-Site Associated DNA Sequencing. PLoS One 10, e0138435. https://doi.org/10.1371/JOURNAL.PONE.0138435

Robledo, D., Palaiokostas, C., Bargelloni, L., Martínez, P., Houston, R., 2018. Applications of genotyping by sequencing in aquaculture breeding and genetics. Rev. Aquac. 10, 670–682. https://doi.org/10.1111/raq.12193

Rubinacci, S., Ribeiro, D.M., Hofmeister, R.J., Delaneau, O., 2021. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat. Genet. 2021 531 53, 120–126. https://doi.org/10.1038/s41588-020-00756-0

Sae-Lim, P., Gjerde, B., Nielsen, H.M., Mulder, H., Kause, A., 2016. A review of genotype-by-environment interaction and micro-environmental sensitivity in aquaculture species. Rev. Aquac. 8, 369–393. https://doi.org/10.1111/raq.12098

Sae-Lim, P., Kause, A., Janhunen, M., Vehviläinen, H., Koskinen, H., Gjerde, B., Lillehammer, M., Mulder, H.A., 2015. Genetic (co)variance of rainbow trout (Oncorhynchus mykiss) body weight and its uniformity across production environments. Genet. Sel. Evol. 47, 1–10. https://doi.org/10.1186/s12711-015-0122-8

Sae-Lim, P., Kause, A., Lillehammer, M., Mulder, H.A., 2017. Estimation of breeding values for uniformity of growth in Atlantic salmon (Salmo salar) using pedigree relationships or single-step genomic evaluation. Genet. Sel. Evol. 49, 1–12.

https://doi.org/10.1186/s12711-017-0308-3

Sae-Lim, P., Khaw, H.L., Nielsen, H.M., Puvanendran, V., Hansen, Ø., Mortensen, A., 2020. Genetic variance for uniformity of body weight in lumpfish (Cyclopterus lumpus) used a double hierarchical generalized linear model. Aquaculture 514, 734515. https://doi.org/10.1016/J.AQUACULTURE.2019.734515

SanCristobal-Gaudy, M., Bodin, L., Elsen, J.M., Chevalet, C., 2001. Genetic components of litter size variability in sheep. Genet. Sel. Evol. 33, 249–271. https://doi.org/10.1186/1297-9686-33-3-249

SanCristobal-Gaudy, M., Elsen, J.-M., Bodin, L., Chevalet, C., 1998. Prediction of the response to a selection for canalisation of a continuous trait in animal breeding. Genet. Sel. Evol. 1998 305 30, 1–29. https://doi.org/10.1186/1297-9686-30-5-423

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

Sonesson, A.K., Ødegård, J., Rönnegård, L., 2013. Genetic heterogeneity of within-family variance of body weight in Atlantic salmon (Salmo salar). Genet. Sel. Evol. 45, 1–8. https://doi.org/10.1186/1297-9686-45-41

Sorensen, D., Waagepetersen, R., 2003. Normal linear models with genetically structured residual variance heterogeneity: A case study. Genet. Res. 82, 207–222. https://doi.org/10.1017/S0016672303006426

Vallejo, R.L., Silva, R.M.O., Evenhuis, J.P., Gao, G., Liu, S., Parsons, J.E., Martin, K.E., Wiens, G.D., Lourenco, D.A.L., Leeds, T.D., Palti, Y., 2018. Accurate genomic predictions for BCWD resistance in rainbow trout are achieved using low-density SNP panels: Evidence that long-range LD is a major contributing factor. J. Anim. Breed. Genet. 135, 263–274. https://doi.org/10.1111/jbg.12335

Vu, S. V., Gilmour, A.R., Nguyen, N.T.H., Dove, M., Vu, I. V., Le, T.S., Knibb, W., O'Connor, W., 2021. Does genetic correlation change across environments for harvest whole weight and its uniformity in the Portuguese oyster (Crassostrea angulata). Aquaculture 536, 736444. https://doi.org/10.1016/J.AQUACULTURE.2021.736444

Wang, Q., Yu, Y., Zhang, Q., Zhang, X., Yuan, J., Huang, H., Xiang, J., Li, F., 2019. A Novel Candidate Gene Associated With Body Weight in the Pacific White Shrimp Litopenaeus vannamei. Front. Genet. 10, 520.

https://doi.org/10.3389/fgene.2019.00520

Wood, S.N., 2006. On confidence intervals for generalized additive models based on penalized regression splines. Aust. N. Z. J. Stat. 48, 445–464. https://doi.org/10.1111/J.1467-842X.2006.00450.X

Xu, J., Zhao, Z., Zhang, X., Zheng, X., Li, J., Jiang, Y., Kuang, Y., Zhang, Y., Feng, J., Li, C., Yu, J., Li, Q., Zhu, Y., Liu, Y., Xu, P., Sun, X., 2014. Development and evaluation of the first high-throughput SNP array for common carp (Cyprinus carpio). BMC Genomics 2014 151 15, 1–10. https://doi.org/10.1186/1471-2164-15-307

Yáñez, J.M., Naswa, S., López, M.E., Bassini, L., Correa, K., Gilbey, J., Bernatchez, L., Norris, A., Neira, R., Lhorente, J.P., Schnable, P.S., Newman, S., Mileham, A., Deeb, N., Genova, A. Di, Maass, A., 2016. Genomewide single nucleotide polymorphism discovery in Atlantic salmon (Salmo salar): validation in wild and farmed American and European populations. Mol. Ecol. Resour. 16, 1002–1011. https://doi.org/10.1111/1755-0998.12503

Yáñez, J.M., Yoshida, G., Barria, A., Palma-Véjares, R., Travisany, D., Díaz, D., Cáceres, G., Cádiz, M.I., López, M.E., Lhorente, J.P., Jedlicki, A., Soto, J., Salas, D., Maass, A., 2020. High-Throughput Single Nucleotide Polymorphism (SNP) Discovery and Validation Through Whole-Genome Resequencing in Nile Tilapia (Oreochromis niloticus). Mar. Biotechnol. 22, 109–117. https://doi.org/10.1007/s10126-019-09935-5

Yoshida, G.M., Barria, A., Correa, K., Cáceres, G., Jedlicki, A., Cadiz, M.I., Lhorente, J.P., Yáñez, J.M., 2019a. Genome-Wide Patterns of Population Structure and Linkage Disequilibrium in Farmed Nile Tilapia (Oreochromis niloticus). Front. Genet. 10, 745. https://doi.org/10.3389/fgene.2019.00745

Yoshida, G.M., Lhorente, J.P., Correa, K., Soto, J., Salas, D., Yáñez, J.M., 2019b. Genome-wide association study and cost-efficient genomic predictions for growth and fillet yield in Nile tilapia (Oreochromis niloticus). G3 Genes, Genomes, Genet. 9, 2597–2607. https://doi.org/10.1534/g3.119.400116

Yoshida, G.M., Yáñez, J.M., 2021a. Increased accuracy of genomic predictions for growth under chronic thermal stress in rainbow trout by prioritizing variants from GWAS using imputed sequence data. Evol. Appl. 00, 1–16. https://doi.org/10.1111/eva.13240

Yoshida, G.M., Yáñez, J.M., 2021b. Multi-trait GWAS using imputed high-density genotypes from whole-genome sequencing identifies genes associated with body traits in Nile tilapia. BMC Genomics 22, 1–13. https://doi.org/10.1186/s12864-020-07341-z

You, X., Shan, X., Shi, Q., 2020. Research advances in the genomics and applications for molecular breeding of aquaculture animals. Aquaculture 526, 735357. https://doi.org/10.1016/J.AQUACULTURE.2020.735357

Zhang, Xiaojun, Yuan, J., Sun, Y., Li, S., Gao, Y., Yu, Y., Liu, C., Wang, Q., Lv, X., Zhang, Xiaoxi, Ma, K.Y., Wang, X., Lin, W., Wang, Long, Zhu, X., Zhang, C., Zhang, J., Jin, S., Yu, K., Kong, J., Xu, P., Chen, J., Zhang, H., Sorgeloos, P., Sagi, A., Alcivar-Warren, A., Liu, Z., Wang, Lei, Ruan, J., Chu, K.H., Liu, B., Li, F., Xiang, J., 2019. Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. Nat. Commun. 10, 1–14. https://doi.org/10.1038/s41467-018-08197-4

Zhong, X., Li, Q., Kong, L., Yu, H., 2017. Estimates of Linkage Disequilibrium and Effective Population Size in Wild and Selected Populations of the Pacific Oyster Using Single-nucleotide Polymorphism Markers. J. World Aquac. Soc. 48, 791–801. https://doi.org/10.1111/JWAS.12393@10.1111/(ISSN)1749-7345.HARAPAC_2018

Chapter 2 - Effect of harvest weight and its uniformity on survival in *Litopenaeus vannamei* reared in different systems[a]

ABSTRACT - A new goal in animal breeding programs is to improve not only trait means, but also to decrease the level of phenotypic variation. One way to achieve uniformity is through selection for reduced residual (environmental) variance, if this is a trait under genetic control. The objective of our study was to evaluate the opportunity of selection for uniformity of harvest weight (HW) and to investigate whether HW and its uniformity (HW_v) may affect survival (SUR) in shrimp reared in different production systems. A total of 149,919 and 164,023 records of HW and SUR, respectively, were obtained from a commercial shrimp hatchery and used in this study. Data was grouped into three sets based on the analyses: low density earthen ponds (S1), high density concrete recirculation tanks (S2) and S1+S2 for joint analyses. A sire-dam double hierarchical generalized linear model (DHGLM) was applied to estimate (co)variance components for HW-HW_v, while a bivariate mixed animal model was applied for HW-SUR, using the three datasets. Genetic correlations among these traits were then estimated, in addition to the correlation coefficient between estimated breeding values (EBV) for HW_v and SUR. The results revealed a significant proportion of genetic variance in the environmental variance of HW and, consequently, the opportunity to select for uniformity (genetic coefficient of variation ranging from 17 to 35%). The genetic correlations between HW and SUR were different in sign and magnitude based on the production system, being equal to 0.36, -0.59 and -0.02 for S1, S2 and S1+S2, respectively. Correlation coefficients between EBV for HW_v and SUR were significant only for S1 (-0.32), suggesting that selection for more uniform families may also increase survival rates in a less intensive environment. In conclusion, we found substantial genetic variation in the uniformity of harvest weight in *Litopenaeus vannamei* and this trait could be improved through selection. Harvest weight and its uniformity have a genetic correlation with survival affected by the production system and environmental condition.

Keywords: Genetic heterogeneity of residual variance, Harvest weight, Shrimp, Survival, Uniformity of production

[a]Article published in the “Aquaculture” doi: 10.1016/j.aquaculture.2020.735891

2.1 Introduction

Pacific white shrimp (*Litopenaeus vannamei*) is a species with economic importance in different parts of the globe with remarkable production in Asia– primarily in China, Indonesia and India– and in the Americas– especially in Ecuador, Mexico and Brazil (FAO, 2017). One strategy to increase shrimp productivity is the adoption of breeding programs that consider traits with economic interests such as harvest weight (Campos-Montes et al., 2013; Perez-Rostro and Ibarra, 2003) and survival rate (Moss et al., 2005; Ødegård and Olesen, 2011). Recently, breeding programs have considered improving not only the trait mean; attention has also been paid to the performance of individuals in multiple environments and approaches to reduce variation around mean levels. The latter topic may indicate stability of phenotypes, also called "robustness", and it is associated with the potential of individuals to keep their phenotypes stable when disturbed by stressor agents (Mulder et al., 2008; SanCristobal-Gaudy et al., 2001).

It is known that different genotypes may have distinct responses to environmental changes– the so-called genotype by environment (GxE) interaction (Falconer and Mackay, 1996; Lynch and Walsh, 1998)– and that phenotypic variation may be attributed to two main categories: 1) measurable or visible environmental changes, such as changes in production systems, temperature or stocking density, called macro-environmental sensitivity and, 2) non-measurable and animal-specific environmental factors that could encompass external factors such as pathogens and social interactions, as well as internal factors, such as homozygosity, epigenetic effects and mutations. Both factors are considered unknown and related to micro-environmental sensitivity (Mulder et al., 2013b). In shrimp and other aquaculture species, production systems may have substantial differences and are generally exposed to greater environmental interferences in comparison to livestock species (e.g., chickens and pigs), making it difficult to maintain a completely stable environment (Sae-Lim et al., 2013). In order to investigate genetic variation in macro-environmental sensitivity, GxE interaction effects have been studied in shrimp regarding different gradient environments, such as farm location and management (Campos-Montes et al., 2013, 2009), density (Ibarra and Famula, 2008; Tan et al., 2017), rearing structure (Nguyen et al., 2019), temperature (Li et al., 2015), presence of disease (Argue et al., 2002) and salinity level (Lu et al., 2017). The presence of GxE

interaction may help breeders decide if single or multiple breeding programs should be implemented to more efficiently select animals across environments (Agha et al., 2018).

The unknown nature of micro-environmental sensitivity may be measured by the uniformity of a trait. If individuals display substantial ability to maintain phenotype stability across minor environmental fluctuations, they are expected to have lower variation (higher uniformity) around group or family levels (Hill and Mulder, 2010). Uniformity is desired by consumers and farmers considering that more homogenous animals are associated with a final product of superior quality and ease of operational processing in industry resulting in greater profitability (Marjanovic et al., 2016). In aquaculture, for example, wide variability in size or weight may increase competition for food, stimulate cannibalistic behavior and reduce the survival rate, causing economic losses or requiring laborious intensive management to classify animals by size (Marjanovic et al., 2018). Increasing uniformity through selection is possible when there is genetic control of the environmental (residual) variance. In practical terms, uniformity may be investigated through the genetic heterogeneity of residual variance, i. e. the residual component of evaluation models is no longer considered to follow same distribution and variance; rather, genetic and environmental factors are assumed to have an effect on the residual (Mulder et al., 2007; Sorensen and Waagepetersen, 2003). Iung et al. (2019) reviewed uniformity studies in aquaculture species and found substantive proportions of genetic variance in the residual variance. These authors reported that, compared to livestock species, farmed aquatic animals have a higher potential to explore uniformity using selection due to the greater phenotypic variation, which may indicate high proportions of genetic variability for residual variance. However, these studies focus on fish (salmonids and tilapia) and only one study was found regarding this topic in shrimp (Castillo-Juárez et al., 2012).

The objective of this study was to estimate the genetic component of residual variance for harvest weight and to investigate the effect of weight and its uniformity on the survival of shrimp reared in different culture systems during the grow-out period.

2.2 Material and methods

2.2.1 Population background

The data used in this study was obtained from a Mexican commercial hatchery. The population structure and detailed rearing protocol are documented in Campos-Montes et al. (2013) and Castillo-Juárez et al. (2007). In brief, the family-based selection program was established in 1998 comprised of domesticated strains from Venezuela, Colombia, the United States and Ecuador and a natural population from Mexico, followed by mass selection for body weight. In the family-based selection program, the selected animals were separated by sex, tagged using numbered rings and then transferred to maturation tanks. Following an adaptation period, mature females were artificially inseminated in a 1M:2F (M: male; F: female) ratio and avoiding sib-mating.

The selection program from 1998 to 2002 was applied using two selection criteria: full-sib family phenotypic means of body weight (BW) at around 130 days of age and within-family information for the same trait. In the 2003-2007 period, the selection criteria changed to family estimated breeding values (EBV) and within-family information for BW at 130 days of age. Since 2008, a two-stage selection program has been implemented using EBV, first using BW at 28 days and subsequently BW and survival at 130 days of age as selection criteria.

2.2.2 Larvae and post-larvae management

After artificial insemination, females were moved to spawning tanks and had their eggs collected after 5-6 hours. The eggs were spawned in tanks under constant aeration conditions and then the number of nauplii were estimated separately and seeded into 200-L tanks with controlled water conditions of aeration, salinity, density and temperature. Special care was devoted to dividing families into two replicates, placing the family units in different zones of the greenhouse larvae room to reduce a possible confounding effect between additive genetic and common environment effects (Montaldo et al., 2013). The feed consisted of inert material with 40-50% protein and natural feed (*Chaetoceros spp.* microalgae, *Spirulina spp.* and *Artemia spp*).

After the final metamorphosis to post-larvae (at 28 days of age), the animals were seeded into cages placed inside ponds with controlled temperature and aeration. When they reached around 2 g of BW, samples were randomly drawn and the shrimps

were tagged using elastomers with combinations of different colours and tagging positions to identify families. After tagging, individuals of each family were randomly chosen and placed in two different production systems: low density earthen ponds (S1: System 1) and high density concrete recirculation tanks (S2: System 2). Thus, each family had individuals represented in both systems. In S1, ponds were approximately 0.2 ha, with a 1.4m water column and were seeded at a density of between 10 and 30 shrimp/m². Feeding consisted of a commercial diet containing 35-40% crude protein and the amount of feed was offered as 3% of the total biomass in the pond. The water exchange rate ranged from between 5 and 20% daily. For the recirculation system (S2), concrete tanks measured 16x4 m, with a 2m water column and a density of between 85 and 100 shrimp/m². A 35-40% crude protein commercial diet was also provided to these animals, but the quantity was equal to 6% of the total biomass. Constant aeration was provided and 4-5% of the water was exchanged daily. The number of animals per family in each system ranged from between 30 and 50, depending on the year.

2.2.3 Collection and edition of phenotypes

Shrimp were harvested at approximately 130 days, and each animal was identified, sexed, and weighted. Animals with deformities, undefined sex information and miscoding were discarded from the final dataset. Shrimp survival was also recorded and treated as a binary trait (0=dead; 1=alive) by using information from the number of animals seeded and harvested in each pond. From now on, harvest weight and survival will be called HW and SUR, respectively.

2.2.4 Data structure and statistical analysis

Data included in this study was recorded from multiple batches per year produced daily, ranging over an 11-year period of the breeding program (Table 1). A summary of the total number of animals, dams, sires and offspring may be observed in Table 2. Descriptive statistics (records distribution, mean, minimum, maximum and coefficient of variation of phenotypic records) for each production system were obtained and are shown in Table 3.

Normality was also tested for HW using the Skewness and Kurtosis measures (results not shown). All statistical analyses were performed in terms of two levels of information: a single system, combining information from S1 and S2 and considering HW as a single trait; or as separate systems, considering each trait in each system as

independent different trait. The DMU package (Jensen and Madsen, 2002) was used to perform all analysis through the AI-REML module.

Table 1. Number of batches and animals produced in each year of breeding program

Year	Batches	n[a]	Age (days)[b]
1	22	575 (482)	133 (7)
2	15	1,175 (1,121)	131 (22)
3	18	1,217 (835)	128 (5)
4	18	726 (357)	127 (5)
5	10	1,759 (956)	130 (3)
6	12	1,507 (702)	125 (4)
7	11	852 (315)	123 (3)
8	13	1,019 (503)	127 (4)
9	24	687 (269)	134 (7)
10	7	1,855 (1,058)	128 (2)
11	4	690 (344)	130 (1)

[a] Average number of animals harvested (standard deviation)
[b] Average age of animals harvested (standard deviation)

Table 2. Data structure, number of animals and families for each production system

Production System[a]	n[b]	Sire[c]	Dam[d]	Full-sib families[e]	Animals/family[f]
S1	105,340	698	945	966	111 (64)
S2	58,683	698	945	966	62 (29)
S1+S2	164,023	698	945	966	173 (62)

[a] S1: earthen ponds seeded at low density, S2: concrete recirculation tanks seeded at high density and, S1+S2: combination of records from both systems
[b] Number of animals (percentage of total)
[c] Number of sires
[d] Number of dams
[e] Number of full-sib families
[f] Average number of animals in each family (standard deviation)

Table 3. Data distribution and descriptive statistics for harvest weight (HW)* and survival (SUR)**

Trait	Production system[a]	Records[b]	Records/ family[c]	Min[d]	Max[e]	Mean	CV (%)[f]
HW (g)	S1	104,411 (69.64%)	80 (38)	3.31	31.92	16.98	26.19
	S2	45,508 (30.36%)	42 (24)	3.16	31.88	15.84	31.81
	S1+S2	149,919	114 (51)	3.16	31.92	16.64	28.04

SUR (%)	S1	105,340 (64.22%)	111 (64)	14.07	100	77.38	30.01
	S2	58,683 (35.78%)	62 (29)	0.09	100	74.74	45.20
	S1+S2	164,023	173 (70)	20.55	100	67.38	18.18

* Fewer number of records because dead animals had no weight.
** Statistics expressed in terms of family means
[a] S1: earthen ponds seeded at low density, S2: concrete recirculation tanks seeded at high density and, S1+S2: combination of records from both systems
[b] Number of records (percentage of total)
[c] Number of records per family (standard deviation)
[d] Minimum value recorded
[e] Maximum value recorded
[f] Coefficient of variation in percentage

2.2.5 Harvest weight and its uniformity

Double hierarchical generalized linear models (DHGLM) may be applied to detect the genetic heterogeneity of residual variance and to estimate the heritability of residual variance (h_v^2) and the genetic coefficient of variation of residual variance (GCV_v), also known as evolvability (Mulder et al., 2007; Rönnegård et al., 2010). In this model, a linear mixed model for the mean (HW) and a *gamma* generalized linear model for the residual variance of HW (HW_v) were fitted simultaneously using an algorithm with a log-link function to account for the non-normality of the residual variance. We used an extension of DHGLM developed by Felleki et al. (2012) that allows for fitting a bivariate sire-dam linear mixed model for HW and HW_v and estimating the genetic correlation between them. The DHGLM (Model 1) applied was:

$$\begin{bmatrix}\mathbf{y}\\ \boldsymbol{\Psi}\end{bmatrix}=\begin{bmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{X_v}\end{bmatrix}\begin{bmatrix}\mathbf{b}\\ \mathbf{b_v}\end{bmatrix}+\begin{bmatrix}(\mathbf{Z_S}+\mathbf{Z_D}) & \mathbf{0}\\ \mathbf{0} & (\mathbf{Z_S}+\mathbf{Z_D})_\mathbf{v}\end{bmatrix}\begin{bmatrix}\mathbf{u}\\ \mathbf{u_v}\end{bmatrix}+\begin{bmatrix}\mathbf{Q} & \mathbf{0}\\ \mathbf{0} & \mathbf{Q_v}\end{bmatrix}\begin{bmatrix}\mathbf{c}\\ \mathbf{c_v}\end{bmatrix}+\begin{bmatrix}\mathbf{e}\\ \mathbf{e_v}\end{bmatrix} \qquad [1]$$

where **y** and $\boldsymbol{\Psi}$ are the vectors of response variables for the mean (HW) and residual variance (HW_v) models, respectively; **b(b_v)** is the vector of fixed effects that included year-place (farm unit), pond/tank, sex and age (linear covariate), and **u(u_v)** is the vector of additive genetic effects of the sire and dam of each individual, assuming:

$$\begin{bmatrix}\mathbf{u}\\ \mathbf{u_v}\end{bmatrix}\sim N\left(0,\frac{1}{4}\mathbf{G}\otimes\mathbf{A}\right),$$

where **G** is a 2x2 sire-dam (co)variance matrix, and **A** is the numerator relationship matrix; **c(c_v)** is the vector of common to full-sibs effect, assuming:

$$\begin{bmatrix}\mathbf{c}\\ \mathbf{c_v}\end{bmatrix}\sim N(0,\mathbf{C}\otimes\mathbf{I}),$$

where **C** is a 2x2 (co)variance matrix of the common full-sib environmental effect and **I** an identity matrix. The residuals of **y(e)** and $\boldsymbol{\Psi(e_v)}$ were treated as independent and normally distributed as:

$$\begin{bmatrix}\mathbf{e}\\ \mathbf{e_v}\end{bmatrix} \sim N\left(0, \begin{bmatrix}\mathbf{W^{-1}\sigma_e^2} & \mathbf{0}\\ \mathbf{0} & \mathbf{W_v}^{-1}\boldsymbol{\sigma}^2_{\mathbf{e_v}}\end{bmatrix} \otimes \mathbf{I}\right),$$

where $\mathbf{W} = diag(\Psi^{-1})$ and $\mathbf{W_v} = diag\left(\frac{1-h_i}{2}\right)$ contain the individual residual variances and $\boldsymbol{\sigma}^2_{\mathbf{e}}(\boldsymbol{\sigma}^2_{\mathbf{e_v}})$ is a scaling variance expected to be 1. **X**, **(Z_S+Z_D)** and **Q** are the incidence matrices for each effect presented above. Vectors, matrices and scalar components indicated by the subscript **v** are related to the residual variance model. The linearized response variable $\boldsymbol{\Psi}$ was obtained as follows:

$$\boldsymbol{\Psi_i} = \mathbf{log}\left(\widehat{\boldsymbol{\sigma}}^2_{\mathbf{e_i}}\right) + (\{[\hat{\mathbf{e}}^2_{\mathbf{i}}/(\mathbf{1} - \mathbf{h_i})] - \widehat{\boldsymbol{\sigma}}^2_{\mathbf{e_i}}\}/\widehat{\boldsymbol{\sigma}}^2_{\mathbf{e_i}}),$$

where $\widehat{\boldsymbol{\sigma}}^2_{\mathbf{e_i}}$ is the predicted residual variance for the *i*th observation obtained in the previous iteration, $\hat{\mathbf{e}}^2_{\mathbf{i}}$ is the squared residual from the mean model estimated for the *i*th observation and $\mathbf{h_i}$ is the leverage for the *i*th observation, which is the *ith* diagonal element of the hat matrix (Hoaglin and Welsch, 1978). The model was iteratively performed, automatically updating $\boldsymbol{\Psi}$, **W** and $\mathbf{W_v}$ in each iteration until convergence running the DMU within a routine in R software (R Core Team, 2017). The convergence criterion was assumed to be achieved when the difference in the estimates of variance components between two successive interactions was less than 10^{-6} (Iung et al., 2017).

2.2.6 Harvest weight and survival

The following bivariate mixed animal model was fitted to estimate genetic parameters for HW and SUR (Model 2):

$$\begin{bmatrix}\mathbf{y_1}\\ \mathbf{y_2}\end{bmatrix} = \begin{bmatrix}\mathbf{X_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{X_2}\end{bmatrix}\begin{bmatrix}\mathbf{b_1}\\ \mathbf{b_2}\end{bmatrix} + \begin{bmatrix}\mathbf{Z_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{Z_2}\end{bmatrix}\begin{bmatrix}\mathbf{u_1}\\ \mathbf{u_2}\end{bmatrix} + \begin{bmatrix}\mathbf{Q_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{Q_2}\end{bmatrix}\begin{bmatrix}\mathbf{c_1}\\ \mathbf{c_2}\end{bmatrix} + \begin{bmatrix}\mathbf{e_1}\\ \mathbf{e_2}\end{bmatrix} \qquad [2]$$

where $\mathbf{y_1}$ and $\mathbf{y_2}$ are vectors of individual observations (continuous response for HW and binary for SUR, respectively); $\mathbf{b_1}$ is the vector of fixed effects as described in Model 1; $\mathbf{b_2}$ is the vector of fixed effects that includes only year, place (farm unit), and pond/tank, because sex and age were unknown for SUR; $\mathbf{u_1(u_2)}$ and $\mathbf{e_1(e_2)}$ are the vectors of additive genetic and residual random effects, respectively; $\mathbf{c_1}$ and $\mathbf{c_2}$ are the vectors of random effects common to full-sibs; $\mathbf{X_1}(\mathbf{X_2})$, $\mathbf{Z_1}(\mathbf{Z_2})$ and $\mathbf{Q_1}(\mathbf{Q_2})$ are the

incidence matrices related to their respective effects. For both traits, it was assumed that:

$$\begin{bmatrix}\mathbf{u_1}\\ \mathbf{u_2}\end{bmatrix} \sim N(0, \mathbf{G}\otimes\mathbf{A}),\ \begin{bmatrix}\mathbf{c_1}\\ \mathbf{c_2}\end{bmatrix} \sim N(0, \mathbf{C}\otimes\mathbf{I}) \text{ and } \begin{bmatrix}\mathbf{e_1}\\ \mathbf{e_2}\end{bmatrix} \sim N(0, \mathbf{R}\otimes\mathbf{I})$$

where **A**, **C** and **I** are the same matrices defined in Model 1; **G** and **R** are the (co)variance matrices of additive genetic and residual effects, respectively. For both models, the *c* effect was considered as the common environment from rearing of full-sibs until tagging.

2.2.7 Genetic parameter estimations

In the sire-dam DHGLM, the same estimated variance component was assumed for sire and dam resulting in one quarter of the additive genetic variance. Thus, the additive genetic variance for HW (σ_a^2) and its uniformity ($\sigma_{a_{v,exp}}^2$) were equal to ($4\sigma_u^2$) and ($4\sigma_{u_{v,exp}}^2$), respectively. Variance components estimated for uniformity (HW_v) were on the exponential scale and were transformed to an additive scale using conversion equations (Mulder et al., 2007; Sae-Lim et al., 2015). For the HW in the DHGLM, phenotypic variance was calculated as: $\sigma_p^2 = 2\sigma_u^2 + \sigma_c^2 + \sigma_e^2$; where σ_c^2 and σ_e^2 are the common to full-sibs and residual variances, respectively. Heritability (h^2) and common full-sib environmental effect (c^2) for the mean were obtained by $h^2 = \sigma_a^2/\sigma_p^2$ and $c^2 = \sigma_c^2/\sigma_p^2$, respectively. For uniformity, the same parameters were calculated by $h_v^2 = \frac{\sigma_{a_v}^2}{2\sigma_p^4+3(\sigma_a^2+\sigma_{c_v}^2)}$ and $c_v^2 = \frac{\sigma_{c_v}^2}{2\sigma_p^4+3(\sigma_a^2+\sigma_{c_v}^2)}$, respectively (Felleki and Lundeheim, 2013; Sae-Lim et al., 2015). The genetic coefficient of variation for uniformity (GCV_v) was calculated as: $GCV_v = \sigma_{a_v}/\sigma_E^2$, where $\sigma_E^2 = \sigma_e^2 - 2\sigma_u^2$. The h_v^2 and GCV_v are estimated to interpret and compare uniformity studies, the former is related to the accuracy of EBV for residual variance and the latter indicates the opportunity of selection response using the residual variance (Iung et al., 2019). Standard errors of both parameters were estimated using equations derived from Mulder et al. (2016). For the bivariate animal model (HW-SUR), h^2 and c^2 of each trait was estimated using the same equations for the mean in the DHGLM, except for σ_p^2, that now is equal to the sum of σ_a^2, σ_c^2 and σ_e^2.

Genetic correlations (r_g) between HW-HW_v and HW-SUR were directly obtained from DHGLM and the bi-trait animal model, respectively. Due to convergence problems, probably because of a hyper-parameterized model, a DHGLM including

HW, HW_v and SUR as response variables to estimate genetic correlation between them was not successful. Hence, correlation coefficients between EBV were performed. First, the accuracy of EBV (EBV_{acc}) was estimated for HW_v and SUR using: $EBV_{acc} = \sqrt{1 - \frac{PEV}{\sigma_a^2}}$, where PEV is the predicted error variance and σ_a^2 is the additive genetic variance of the trait (Meyer, 2004). Then, a correlation between EBV for HW_v and SUR was performed using only information from animals with EBV_{acc} greater than 0.5 for both traits. By means of this approach, we expected to approximate the real genetic correlation.

2.3. Results

2.3.1 Descriptive statistics

Descriptive statistics for records are shown in Table 3. The number of records was greater in S1 (low density earthen ponds) than in S2 (high density concrete recirculation tanks) as well as the average number of animals per family. All families had records for siblings in both environments. No significant differences in phenotype means were found between environments for HW (S1: 16.98g; S2: 15.84g) and SUR (S1: 77.38%; S2: 74.74%). However, a higher coefficient of variation for SUR was observed when calculated in terms of family means between systems (S1: 30.01%; S2: 45.20%). For HW, the coefficient of variation was more stable with only 5.6% more variation in S2 than in S1.

2.3.2 Genetic parameters for harvest weight, uniformity and survival

The variance components and genetic parameters for HW and HW_v are presented in Table 4. In general, the estimates of genetic parameters for HW and HW_v varied according to the production system. For HW, the lowest h^2 was observed for S1, followed by S2 and S1+S2 (0.15 ± 0.02, 0.24 ± 0.02 and 0.27 ± 0.01, respectively) and substantial c^2 effect was detected in all production systems (9.4, 9.5 and 7.2 % of the total phenotypic variance for S1, S2 and S1+S2, respectively).

Results obtained in S2 and S1+S2 were similar to the univariate models fitting residual variance as homogeneous (results not shown) while for S1, most variance components seemed to be underestimated. Although the estimates of h_v^2 were low for all production systems (0.01–0.03), GCV_v were considerably high (0.35 ± 0.03, 0.17 ± 0.05 and 0.28 ± 0.02, for S1, S2 and S1+S2, respectively). These results indicate that

using the residual variance of harvest weight as selection criteria may increase the uniformity of HW in shrimp.

Table 4. Estimates of variance components and genetic parameters (SE) for HW and its uniformity using DHGLM including information from different production systems

Trait[b]/Parameter[c]	Production System[a]		
	S1	S2	S1+S2
HW			
σ_p^2	6.19	8.19	6.12
σ_a^2	0.89 (0.18)	1.74 (0.24)	1.76 (0.18)
σ_c^2	0.58 (0.05)	0.78 (0.07)	0.44 (0.05)
σ_e^2	5.68 (0.04)	5.68 (0.04)	5.24 (0.02)
h^2	0.15 (0.02)	0.24 (0.02)	0.27 (0.01)
c^2	0.10 (0.01)	0.09 (0.01)	0.07 (0.01)
HW_v			
$\sigma_{a_{v,exp}}^2$	0.11 (0.02)	0.03 (0.02)	0.08 (0.02)
$\sigma_{c_v}^2$	0.11 (0.01)	0.12 (0.01)	0.04 (0.003)
GCV_v	0.35 (0.03)	0.17 (0.05)	0.28 (0.02)
h_v^2	0.03 (0.01)	0.01 (0.003)	0.02 (0.003)
c_v^2	0.003	0.002	0.001

[a] S1: earthen ponds seeded at low density, S2: concrete recirculation tanks seeded at high density and, S1+S2: combination of records from both systems
[b] HW: harvest weight, HW_v: uniformity for harvest weight
[c] σ_p^2: phenotypic, σ_a^2: additive genetic, σ_c^2: full-sib family and σ_e^2: residual variances, h^2: heritability for HW, c^2: effect common to full sibs for HW. $\sigma_{a_{v,exp}}^2$: additive genetic variance for HW_v in the exponential scale, $\sigma_{c_v}^2$: common environmental variance for HW_v, GCV_v: genetic coefficient of variation of residual variance, h_v^2: heritability for HW_v and c_v^2: common environmental effect for uniformity of HW_v

The variance components and genetic parameters using bivariate (HW - SUR) animal models are shown in Table 5. In contrast to the h^2 for HW obtained in DHGLM, similar estimates of h^2 across production systems were found with the bivariate model (0.13 ± 0.02, 0.17 ± 0.03 and 0.12 ± 0.02 for S1, S2 and S1+S2, respectively). For SUR, h^2 estimates were both low, ranging from 0.05 ± 0.01 to 0.07 ± 0.01, which were also similar to univariate SUR models (results not shown).

2.3.3 Genetic correlations among traits

Genetic correlations (r_g) between HW and HW_v were of low to moderate magnitude with a positive sign for production systems considered independently (S1 and S2) and negative for S1+S2 (Table 6). High standard errors were observed for these correlations. On the other hand, HW-SUR genetic correlations showed highly significant and opposite values when the production systems were considered

separately. Positive genetic correlations were found for S1 (r_g =0.36 ± 0.10) and negative correlations for S2 (r_g = -0.59 ± 0.10), implying that the environments evaluated may considerably change the trend and magnitude of a correlated response between HW and SUR. Correlation coefficients between EBV for HW_v and SUR were not significant for S2 and S1+S2 (0.18 ± 0.10 and -0.03± 0.002, respectively) but were highly significant for S1 (-0.32 ± 0.01), indicating that lower residual variance (more uniform individuals) may be genetically correlated to individuals with greater chances for survival.

Table 5. Estimates of variance components and genetic parameters (SE) for HW and SUR using bivariate animal models including information from different production systems

Trait[b]/Parameter[c]	Production System[a]		
	S1	S2	S1+S2
HW			
σ_p^2	5.64	7.02	6.13
σ_a^2	0.72 (0.13)	1.20 (0.21)	0.74 (0.13)
σ_c^2	0.56 (0.05)	0.74 (0.07)	0.57 (0.05)
σ_e^2	4.36 (0.71)	5.08 (0.11)	4.83 (0.07)
h^2	0.13 (0.02)	0.17 (0.03)	0.12 (0.02)
c^2	0.10 (0.01)	0.10 (0.01)	0.09 (0.01)
SUR			
σ_p^2	4.12	4.08	3.68
σ_a^2	0.28 (0.04)	0.30 (0.05)	0.18 (0.03)
σ_c^2	0.55 (0.05)	0.49 (0.05)	0.21 (0.02)
σ_e^2	3.29	3.29	3.29
h^2	0.07 (0.01)	0.07 (0.01)	0.05 (0.01)
c^2	0.13 (0.01)	0.12 (0.01)	0.06 (0.01)

[a] S1: earthen ponds seeded at low density, S2: concrete recirculation tanks seeded at high density and, S1+S2: combination of records from both systems
[b] HW: harvest weight, SUR: survival
[c] σ_p^2: phenotypic, σ_a^2: additive genetic, σ_c^2: full-sib family and σ_e^2: residual variances. h^2: Heritability, c^2: effect common to full sibs

Table 6. Genetic correlation estimates (SE) between harvest weight, uniformity and survival across different environments

Genetic correlation	Production System[a]		
(r_g)[b]	S1	S2	S1+S2
HW-HW_v	0.17 ± 0.13	0.34 ± 0.22	-0.23 ± 0.08
HW-SUR	0.36 ± 0.10	-0.59 ± 0.10	-0.02 ± 0.10
HW_v-SUR*	-0.32 ± 0.01	0.18 ± 0.10	-0.03 ± 0.002

[a] S1: earthen ponds seeded at low density, S2: concrete recirculation tanks seeded at high density and, S1+S2: combination of records from both systems
[b] HW: harvest weight, HW_v: uniformity for harvest weight, SUR: survival
* Simple correlation coefficient.

2.4. Discussion

2.4.1 Genetic parameters for harvest weight and uniformity

The h^2 estimates obtained for HW were of low to moderate magnitude and are consistent with other studies for growth related traits in *Litopenaeus vannamei* (ranging from 0.14 to 0.42; Gitterle et al., 2005b, 2005a; Ibarra and Famula, 2008; Kanchanachai et al., 2011; Li et al., 2015; Sui et al., 2015; Zhang et al., 2016). Low to high h^2 estimates (0.07 ± 0.03 to 0.52 ± 0.09) were reported for several growth traits in studies using data from the same breeding program (Caballero-Zamora et al., 2015, 2013; Campos-Montes et al., 2017, 2013, 2009; Castillo-Juárez et al., 2007; Hernández-Ruíz et al., 2020; Montaldo et al., 2013; Ríos-Pérez et al., 2015). The wide range of h^2 observed in these studies may be explained by several factors, such as: trait (morphometric measures and growth under specific conditions), number of families, statistical model and methodology and age of the animal at the time of measurement. In our study, h^2 estimates for HW were also affected by the source of information, i.e., the culture environment. Our results found a downward estimation of the additive genetic effect in S1, accounting for about half of the estimates in S2 and S1+S2. In contrast, single-trait animal models for the same trait and dataset produced very similar genetic parameters (e.g., h^2~0.28), irrespective of the production system (results not shown). It is likely that the sire-dam DHGLM produced this effect because it incorporates only information between sires and dams, rather than the whole family-based relationship applied in animal models. However, when a single observation per animal is available, animal models may produce biased variance components for residual variance models estimates because of the high dependence between EBV and residuals (Sae-Lim et al., 2017). For this reason, the most suitable model was the sire-dam. Additionally, it may be observed that most of parameters estimated in the residual part of the model in S1 are higher than in S2 and S1+S2 (lower section of Table 4). Perhaps a fraction of the genetic variance of HW, in S1, was captured by the DHGLM and directed to HW_v. Nevertheless, more studies of the genetic heterogeneity of residual variance for this data structure and species are necessary to clearly understand these results.

The GCV_v are consistent with other studies on aquaculture species ranging from 0.14 to 0.38 (Sae-Lim et al., 2017, 2015; Sonesson et al., 2013). GCV_v may be interpreted as the change in residual variance by one standard deviation of selection

response for genetic variance (Mulder et al., 2008). Thus, selection using residual variance for HW in shrimp may be used to increase its uniformity. However, the low estimates of h_v^2 suggest poor accuracy of selection for uniformity. Similarly low h_v^2 estimates were found in aquaculture and other livestock species as reviewed by Iung et al. (2019). Other environmental factors may be included in the total residual variance, making it difficult to disentangle the true residual variance (Houle, 1992). The accuracy of selection for uniformity could be improved by using sib-testing to increase the number of relatives (Sae-Lim et al., 2015), or, alternatively, by using genomic information (Mulder et al., 2013a). However, these options may imply costly operations or investments and should be carefully planned.

Production systems also have an effect on the GCV_v estimates. This variation could possibly be explained by the number of records per family. The lowest expected response to selection for uniformity was obtained in S2 in which, on average, families had 42 records. According to Hill and Mulder (2010), a reasonable number of records per family for a moderate heritability trait to estimate variance components for its uniformity is 85, which is close to the number of records found in S1 (80). Another possible explanation for the divergent estimates of GCV_v obtained across production systems is due to GxE interaction. Sae-Lim et al. (2015) found re-ranking in uniformity of body weight for rainbow trout between production and breeding environments despite the fact that no GxE interaction was found for the mean, i.e., the genetic configuration that underlies the mean and residual variance of a trait may be in different loci or interact distinctly. In our case, no GxE interaction was found for HW between S1 and S2 (r_g=0.88, results not shown). A very similar result was previously obtained by Campos-Montes et al. (2009) with the same database. A GxE interaction model for uniformity was attempted, but convergence was not achieved.

2.4.2 Genetic parameters for harvest weight and survival

The genetic parameters for HW and SUR remained more stable across production systems according to the results of the bivariate animal model. The h^2 estimates for HW ranged from 0.12 to 0.17 and for SUR from 0.05 to 0.07, both low. Similar results for SUR were found in other studies for *L. vannamei*: 0.02 ± 0.02 – 0.12 ± 0.06 (Gitterle et al., 2005a); 0.06 ± 0.02 – 0.11 ± 0.03 (Li et al., 2015) and 0.04 ± 0.01 for a study of the same population (Campos-Montes et al., 2013). Tan et al. (2017) reported greater heritability values for SUR in different environments (h^2>0.30)

but they did not include the common effect to full-sibs in the models which could overestimate variance components (Montaldo et al., 2013). In the present study, the common full-sib environmental effect represented up to 13% of the total phenotypic variance, highlighting the importance of this effect to obtain more accurate estimates. Several factors may affect the estimation of genetic parameters for survival in growth ponds, such as density (Ibarra and Famula, 2008; Tan et al., 2017), presence of diseases (Caballero-Zamora et al., 2015; Gitterle et al., 2005b; Ødegård et al., 2011) and temperature (Li et al., 2015). When production environments have substantial differences as in our study, these factors may be simultaneously present and make it difficult to identify which factor is responsible for the variation in the estimates (Nguyen et al., 2019). Thus, survival to specific factors rather than general survival may be preferred in aquaculture breeding programs (Castillo-Juárez et al., 2015; Yáñez et al., 2015).

2.4.3 Genetic correlations

The r_g between HW and HW_v were inaccurate (large standard errors) for S1 and S2, making it difficult to identify a clear additive genetic relationship between mean and residual variance levels. However, the sign and magnitude of correlations are consistent with the only available results for *L. vannamei* (Castillo-Juárez et al., 2012) with a r_g of 0.18 ± 0.14, which are approximately similar. Studies on other aquaculture species for untransformed HW also revealed similar tendency, ranging from 0.30 to 0.60 (Marjanovic et al., 2016; Sae-Lim et al., 2015), in which selection for growth traits may be genetically tied to decreased uniformity, i.e., to an increase in the coefficient of variation. Given that uniformity is also one of the breeding goals, mean and residual variance of a trait could be included in a selection index to achieve a more efficient response to selection for both traits simultaneously. There was an unexpected sign change in the correlation for S1+S2. Negative genetic correlations between mean and residual variance for weight were reported in several species such as pigs (Yang et al., 2011), chickens (Mulder et al., 2016), cattle (Iung et al., 2017; Neves et al., 2011) and aquaculture species (Agha et al., 2018; Sae-Lim et al., 2015). However, a factor in common among these studies is data transformation. Very often data transformation is performed to obtain normality that may potentially reduce the additive genetic variance of residual variance (Sonesson et al., 2013). In our study, no scale effect was observed in HW, thus only untransformed data were used.

The r_g between HW and SUR had different values and direction depending on the production system. Our results also suggest that different responses to selection may be obtained according to the production system. Some studies have shown high and positive correlations between growth and survival traits (0.44 – 0.98; Campos-Montes et al., 2013; Gitterle et al., 2005a; Hernández-Ruíz et al., 2020). In the present study, the correlation estimate was positive (0.36 ± 0.10) in S1, negative (-0.59 ± 0.10) in S2 and close to zero (-0.02 ± 0.10) in S1+S2. General survival is strongly influenced by uncontrolled and unknown environmental factors, as reflected in the low h^2 estimates for SUR. If a disease occurred in one environment in a given year, for example, it is perfectly reasonable that the other environment remained unaffected because S1 and S2 were in different geographical locations. Argue et al. (2002), for instance, identified negative genetic correlations between growth and survival traits in the presence of Taura syndrome virus (TSV). Another plausible explanation for divergent r_g between HW and SUR in both environments is related to stocking density. In S2, animals were stocked in higher densities, which could result in competition and/or aggressive behaviour in shrimp (Griffith and Wigglesworth, 1993; Zhang et al., 2008). Animals that have higher growth rates will require more environmental resources, i.e., growth may be a threshold factor to survive in a competitive environment (Ellen et al., 2008). To confirm this hypothesis, models using indirect genetic effects (social effects) could be applied but specific breeding designs are necessary to estimate social effects (Luan et al., 2015; Sae-Lim and Bijma, 2016), which was not the case for the dataset available to this study.

Production systems seemed to affect r_g estimation between HW_v and SUR. A negative r_g (-0.32 ± 0.01) was observed in the less intensive system suggesting that families with lower residual variance for weight, i.e., higher uniformity, may display an increase in survival rates (Figure 1). It is possible to observe that in S1, one of the families with the higher survival rate also displayed less variation in weight records compared to a family with a lower survival rate and higher mean weight. This result is consistent with recent studies in pigs (Mulder et al., 2015; Putz et al., 2019) and cattle (Elgersma et al., 2018; Poppe et al., 2020). All of them report an increase in resilience/health traits when selection for uniformity is performed. Given that disease outbreaks are a major concern in shrimp farming (Lightner, 2011; Thitamadee et al., 2016), the use of residual variance as a selection criterion is an option to improve

overall survival and resistance to stressor agents and to increase “robustness”. For S2, this correlation was positive and with a high standard deviation, making it difficult to identify a trend between survival and uniformity. The correlations between HW_v and SUR were calculated on the basis of EBVs with accuracy greater than 0.50, resulting in only 215 animals analysed in S2, which is substantially less than in S1 (9,811).

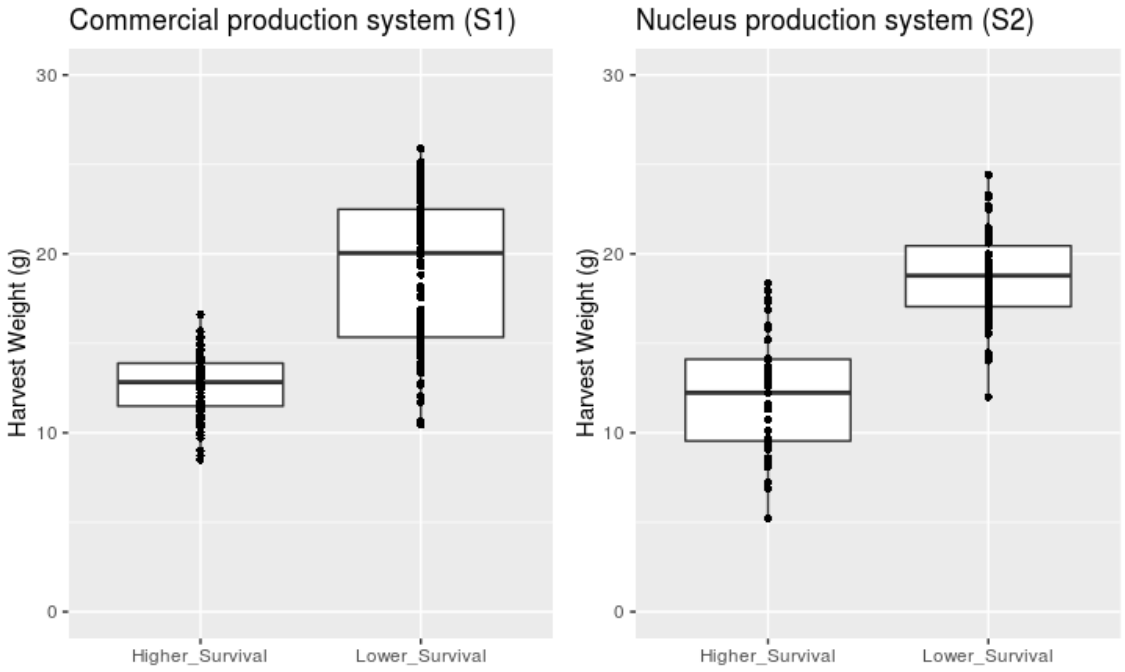


Figure 1. Boxplot of two contrasting families (high survival *vs* low survival) in both production environments. Each black dot represents a weight record (g) from an individual showing the level of variation around the family mean.

2.5 Conclusion

A substantial genetic proportion in the residual variance was found for harvest weight in *Litopenaeus vannamei*, suggesting an opportunity to increase uniformity through selection. As in other aquaculture species, low heritability and a high genetic coefficient of variation of residual variance were found. Production systems seemed to affect genetic correlations between weight, uniformity and survival. A negative genetic correlation between uniformity and survival, under low-density culture conditions, suggests that more uniform families have greater chances for survival during the grow-out phase in a commercial-like production system.

2.6 References

Agha, S., Mekkawy, W., Ibanez-Escriche, N., Lind, C.E., Kumar, J., Mandal, A., Benzie, J.A.H., Doeschl-Wilson, A., 2018. Breeding for robustness: investigating the genotype-by-environment interaction and micro-environmental sensitivity of Genetically Improved Farmed Tilapia (Oreochromis niloticus). Anim. Genet. 49, 421–427. https://doi.org/10.1111/age.12680

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

Caballero-Zamora, A., Cienfuegos-Rivas, E.G., Montaldo, H.H., Campos-Montes, G.R., Martínez-Ortega, A., Castillo-Juárez, H., 2013. Genetic parameters for spawning and growth traits in the Pacific white shrimp (Penaeus (Litopenaeus) vannamei). Aquac. Res. 46, 833–839. https://doi.org/10.1111/are.12235

Caballero-Zamora, A., Montaldo, H.H., Campos-Montes, G.R., Cienfuegos-Rivas, E.G., Martínez-Ortega, A., Castillo-Juárez, H., 2015. Genetic parameters for body weight and survival in the Pacific White Shrimp Penaeus (Litopenaeus) vannamei affected by a White Spot Syndrome Virus (WSSV) natural outbreak. Aquaculture 447, 102–107. https://doi.org/10.1016/J.AQUACULTURE.2014.08.028

Campos-Montes, G.R., Montaldo, H.H., Armenta-Córdova, M., Martínez-Ortega, A., Caballero-Zamora, A., Castillo-Juárez, H., 2017. Incorporation of tail weight and tail percentage at harvest size in selection programs for the Pacific white shrimp Penaeus (Litopenaeus) vannamei. Aquaculture 468, 293–296. https://doi.org/10.1016/j.aquaculture.2016.10.034

Campos-Montes, G.R., Montaldo, H.H., Martínez-Ortega, A., Castillo-Juárez, H., 2009. Genotype by environment interaction effects for body weight at 130 days of age in the Pacific white shrimp [Penaeus (Litopenaeus) vannamei]. Vet. Mex. 40, 255–267.

Campos-Montes, G.R., Montaldo, H.H., Martínez-Ortega, A., Jiménez, A.M., Castillo-Juárez, H., 2013. Genetic parameters for growth and survival traits in Pacific white shrimp Penaeus (Litopenaeus) vannamei from a nucleus population undergoing a two-stage selection program. Aquac. Int. 21, 299–310. https://doi.org/10.1007/s10499-012-9553-1

Castillo-Juárez, H., Campos-Montes, G.R., Caballero-Zamora, A., Montaldo, H.H., 2015. Genetic improvement of Pacific white shrimp [Penaeus (Litopenaeus)

vannamei]: perspectives for genomic selection. Front. Genet. 06, 1–4. https://doi.org/10.3389/fgene.2015.00093

Castillo-Juárez, H., Casares, J.C.Q., Campos-Montes, G., Villela, C.C., Ortega, A.M., Montaldo, H.H., 2007. Heritability for body weight at harvest size in the Pacific white shrimp, Penaeus (Litopenaeus) vannamei, from a multi-environment experiment using univariate and multivariate animal models. Aquaculture 273, 42–49. https://doi.org/10.1016/j.aquaculture.2007.09.023

Castillo-Juárez, H., Montaldo, H.H., Campos-Montes, G.R., 2012. Genetic parameter estimates of the environmental variation for body weight at harvest size in a Pacific white shrimp breeding population, in: 4th International Conference on Quantitative Genetics: Understanding Variation in Complex Traits. Edinburgh, p. 167.

Elgersma, G.G., de Jong, G., van der Linde, R., Mulder, H.A., 2018. Fluctuations in milk yield are heritable and can be used as a resilience indicator to breed healthy cows. J. Dairy Sci. 101, 1240–1250. https://doi.org/10.3168/jds.2017-13270

Ellen, E.D., Visscher, J., Van Arendonk, J.A.M., Bijma, P., 2008. Survival of laying hens: Genetic parameters for direct and associative effects in three purebred layer lines. Poult. Sci. 87, 233–239. https://doi.org/10.3382/ps.2007-00374

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to quantitative genetics., 4th ed, Introduction to quantitative genetics. Longmans Green, Harlow, Essex, UK.

FAO, 2017. FAO Fisheries & Aquaculture - Statistics [WWW Document]. URL http://www.fao.org/fishery/statistics/en (accessed 12.30.19).

Felleki, M., Lee, D., Lee, Y., Gilmour, A.R., Rönnegård, L., 2012. Estimation of breeding values for mean and dispersion, their variance and correlation using double hierarchical generalized linear models. Genet. Res. (Camb). 94, 307–317. https://doi.org/10.1017/S0016672312000766

Felleki, M., Lundeheim, N., 2013. Genetic control of residual variance for teat number in pigs, in: Proc. Assoc. Advmt. Anim. Breed. Genet. pp. 538–541.

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

Gitterle, T., Salte, R., Gjerde, B., Cock, J., Johansen, H., Salazar, M., Lozano, C., Rye, M., 2005b. Genetic (co)variation in resistance to White Spot Syndrome Virus (WSSV)

and harvest weight in Penaeus (Litopenaeus) vannamei. Aquaculture 246, 139–149. https://doi.org/10.1016/j.aquaculture.2005.02.011

Griffith, D.R.W., Wigglesworth, J.M., 1993. Growth rhythms in the shrimp Penaeus vannamei and P. schmitti. Mar. Biol. 115, 295–299. https://doi.org/10.1007/BF00346347

Hernández-Ruíz, H., Montaldo, H.H., Bustos-Martínez, J., Campos-Montes, G.R., Castillo-Juárez, H., 2020. Heritability and genetic correlations for infectious hypodermal and hematopoietic necrosis virus load, body weight at harvest, and survival rate in Pacific white shrimp (*Litopenaeus vannamei*). J. World Aquac. Soc. 51, 312–323. https://doi.org/10.1111/jwas.12664

Hill, W.G., Mulder, H.A., 2010. Genetic analysis of environmental variation. Genet. Res. (Camb). 92, 381–395. https://doi.org/10.1017/S0016672310000546

Hoaglin, D.C., Welsch, R.E., 1978. The Hat Matrix in Regression and ANOVA. Am. Stat. 32, 17–22. https://doi.org/10.1080/00031305.1978.10479237

Houle, D., 1992. Comparing evolvability and variability of quantitative traits. Genetics 130, 195–204.

Ibarra, A.M., Famula, T.R., 2008. Genotype by environment interaction for adult body weights of shrimp Penaeus vannamei when grown at low and high densities. Genet. Sel. Evol. 40, 541. https://doi.org/10.1186/1297-9686-40-5-541

Iung, L.H.S., Carvalheiro, R., Neves, H.H.R., Mulder, H.A., 2019. Genetics and genomics of uniformity and resilience in livestock and aquaculture species: A review. J. Anim. Breed. Genet. 1–18. https://doi.org/10.1111/jbg.12454

Iung, L.H.S., Neves, H.H.R., Mulder, H.A., Carvalheiro, R., 2017. Genetic control of residual variance of yearling weight in nellore beef cattle. J. Anim. Sci. 95, 1425–1433. https://doi.org/10.2527/jas2016.1326

Jensen, J., Madsen, P., 2002. DMU - a package for analyzing multivariate mixed models.

Kanchanachai, Y., Poompuang, S., Koonawootrittriron, S., Uraiwan, S., 2011. Estimating Genetic Parameters for Weight and Body Size of Pacific White Shrimp (Litopenaeus vannamei) by Restricted Maximum Likelihood Method. Kasetsart J. - Nat. Sci. 45, 1047–1057.

Li, W., Luan, S., Luo, K., Sui, J., Xu, X., Tan, J., Kong, J., 2015. Genetic parameters and genotype by environment interaction for cold tolerance, body weight and survival

of the Pacific white shrimp Penaeus vannamei at different temperatures. Aquaculture 441, 8–15.https://doi.org/10.1016/j.aquaculture.2015.02.013

Lightner, D. V., 2011. Virus diseases of farmed shrimp in the Western Hemisphere (the Americas): A review. J. Invertebr. Pathol. 106, 110–130. https://doi.org/10.1016/j.jip.2010.09.012

Lu, X., Luan, S., Cao, B., Meng, X., Sui, J., Dai, P., Luo, K., Shi, X., Hao, D., Han, G., Kong, J., 2017. Estimation of genetic parameters and genotype-by-environment interactions related to acute ammonia stress in Pacific white shrimp (Litopenaeus vannamei) juveniles at two different salinity levels. PLoS One 12, e0173835. https://doi.org/10.1371/journal.pone.0173835

Luan, S., Luo, K., Chai, Z., Cao, B., Meng, X., Lu, X., Liu, N., Xu, S., Kong, J., 2015. An analysis of indirect genetic effects on adult body weight of the Pacific white shrimp Litopenaeus vannamei at low rearing density. Genet. Sel. Evol. 47, 1–8. https://doi.org/10.1186/s12711-015-0164-y

Lynch, M., Walsh, B., 1998. Genetics and analysis of quantitative traits, 1st ed. Sinauer Associates, Sunderland, MA.

Marjanovic, J., Mulder, H.A., Khaw, H.L., Bijma, P., 2016. Genetic parameters for uniformity of harvest weight and body size traits in the GIFT strain of Nile tilapia. Genet. Sel. Evol. 48, 1–10. https://doi.org/10.1186/s12711-016-0218-9

Marjanovic, J., Mulder, H.A., Rönnegård, L., Bijma, P., 2018. Modelling the co-evolution of indirect genetic effects and inherited variability. Heredity (Edinb). 631–647. https://doi.org/10.1038/s41437-018-0068-z

Meyer, K., 2004. Scope for a random regression model in genetic evaluation of beef cattle for growth. Livest. Prod. Sci. 86, 69–83. https://doi.org/10.1016/S0301-6226(03)00142-8

Montaldo, H.H., Castillo-Juárez, H., Campos-Montes, G., Pérez-Enciso, M., 2013. Effect of the data family structure, tank replication and the statistical model, on the estimation of genetic parameters for body weight at 28 days of age in the Pacific white shrimp (Penaeus (Litopenaeus) vannamei Boone, 1931). Aquac. Res. 44, n/a-n/a. https://doi.org/10.1111/j.1365-2109.2012.03176.x

Moss, S.M., Doyle, R.W., Lightner, D. V, 2005. Breeding shrimp for disease resistance: challenges and opportunities for improvement. Dis. Asian Aquac. 5 Proc. fifth Symp. Dis. Asian Aquac. 24-28 Novemb. 2002, Queensland, Aust. 15–393.

Mulder, H.A., Bijma, P., Hill, W.G., 2008. Selection for uniformity in livestock by exploiting genetic heterogeneity of residual variance. Genet. Sel. Evol. 40, 37. https://doi.org/10.1186/1297-9686-40-1-37

Mulder, H.A., Bijma, P., Hill, W.G., 2007. Prediction of breeding values and selection responses with genetic heterogeneity of environmental variance. Genetics 175, 1895–1910. https://doi.org/10.1534/genetics.106.063743

Mulder, H.A., Crump, R.E., Calus, M.P.L., Veerkamp, R.F., 2013a. Unraveling the genetic architecture of environmental variance of somatic cell score using high-density single nucleotide polymorphism and cow data from experimental farms. J. Dairy Sci. 96, 7306–7317. https://doi.org/10.3168/jds.2013-6818

Mulder, H.A., Hill, W.G., Knol, E.F., 2015. Heritable environmental variance causes nonlinear relationships between traits: Application to birth weight and stillbirth of pigs. Genetics 199, 1255–1269. https://doi.org/10.1534/genetics.114.173070

Mulder, H.A., Rönnegård, L., Fikse, W.F., Veerkamp, R.F., Strandberg, E., 2013b. Estimation of genetic variance for macro- and micro-environmental sensitivity using double hierarchical generalized linear models. Genet. Sel. Evol. 45, 1–14. https://doi.org/10.1186/1297-9686-45-23

Mulder, H.A., Visscher, J., Fablet, J., 2016. Estimating the purebred–crossbred genetic correlation for uniformity of eggshell color in laying hens. Genet. Sel. Evol. 48, 39. https://doi.org/10.1186/s12711-016-0212-2

Neves, H.H.R., Carvalheiro, R., Roso, V.M., Queiroz, S.A., 2011. Genetic variability of residual variance of production traits in Nellore beef cattle. Livest. Sci. 142, 164–169. https://doi.org/10.1016/j.livsci.2011.07.010

Nguyen, N.H., Ninh, N.H., Hung, N.H., 2019. Evaluation of two genetic lines of Pacific White leg shrimp Liptopenaeus vannamei selected in tank and pond environments. Aquaculture 734522. https://doi.org/10.1016/j.aquaculture.2019.734522

Ødegård, J., Gitterle, T., Madsen, P., Meuwissen, T.H., Yazdi, M.H., Gjerde, B., Pulgarin, C., Rye, M., 2011. Quantitative genetics of taura syndrome resistance in pacific white shrimp (penaeus vannamei): a cure model approach. Genet. Sel. Evol. 43, 14. https://doi.org/10.1186/1297-9686-43-14

Ødegård, J., Olesen, I., 2011. Comparison of testing designs for genetic evaluation of social effects in aquaculture species. Aquaculture 317, 74–78. https://doi.org/10.1016/J.AQUACULTURE.2011.04.016

Perez-Rostro, C.I., Ibarra, A.M., 2003. Heritabilities and genetic correlations of size traits at harvest size in sexually dimorphic Pacific white shrimp (Litopenaeus vannamei) grown in two environments. Aquac. Res. 34, 1079–1085. https://doi.org/10.1046/j.1365-2109.2003.00913.x

Poppe, M., Veerkamp, R.F., van Pelt, M.L., Mulder, H.A., 2020. Exploration of variance, autocorrelation, and skewness of deviations from lactation curves as resilience indicators for breeding. J. Dairy Sci. 103, 1667–1684. https://doi.org/10.3168/jds.2019-17290

Putz, A.M., Harding, J.C.S., Dyck, M.K., Fortin, F., Plastow, G.S., Dekkers, J.C.M., 2019. Novel Resilience Phenotypes Using Feed Intake Data From a Natural Disease Challenge Model in Wean-to-Finish Pigs. Front. Genet. 9, 660. https://doi.org/10.3389/fgene.2018.00660

R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ríos-Pérez, L., Campos-Montes, G.R., Martínez-Ortega, A., Castillo-Juárez, H., Montaldo, H.H., 2015. Inbreeding Effects on Body Weight at Harvest Size and Grow-out Survival Rate in a Genetic Selected Population of Pacific White Shrimp Penaeus (Litopenaeus) vannamei. J. World Aquac. Soc. 46, 53–60. https://doi.org/10.1111/jwas.12169

Rönnegård, L., Felleki, M., Fikse, F., Mulder, H.A., Strandberg, E., 2010. Genetic heterogeneity of residual variance - Estimation of variance components using double hierarchical generalized linear models. Genet. Sel. Evol. 42, 1–10. https://doi.org/10.1186/1297-9686-42-8

Sae-Lim, P., Bijma, P., 2016. Comparison of designs for estimating genetic parameters and obtaining response to selection for social interaction traits in aquaculture. Aquaculture 451, 330–339. doi.or10.1016/j.aquaculture.2015.09.017

Sae-Lim, P., Kause, A., Janhunen, M., Vehviläinen, H., Koskinen, H., Gjerde, B., Lillehammer, M., Mulder, H.A., 2015. Genetic (co)variance of rainbow trout (Oncorhynchus mykiss) body weight and its uniformity across production environments. Genet. Sel. Evol. 47, 1–10. https://doi.org/10.1186/s12711-015-0122-8

Sae-Lim, P., Kause, A., Lillehammer, M., Mulder, H.A., 2017. Estimation of breeding values for uniformity of growth in Atlantic salmon (Salmo salar) using pedigree relationships or single-step genomic evaluation. Genet. Sel. Evol. 49, 1–12.

https://doi.org/10.1186/s12711-017-0308-3

Sae-Lim, P., Kause, A., Mulder, H.A., Martin, K.E., Barfoot, A.J., Parsons, J.E., Davidson, J., Rexroad, C.E., Van Arendonk, J.A.M., Komen, H., 2013. Genotype-by-environment interaction of growth traits in rainbow trout (Oncorhynchus mykiss): A continental scale study. J. Anim. Sci. 91, 5572–5581. https://doi.org/10.2527/jas.2012-5949

SanCristobal-Gaudy, M., Bodin, L., Elsen, J.M., Chevalet, C., 2001. Genetic components of litter size variability in sheep. Genet. Sel. Evol. 33, 249–271. https://doi.org/10.1186/1297-9686-33-3-249

Sonesson, A.K., Ødegård, J., Rönnegård, L., 2013. Genetic heterogeneity of within-family variance of body weight in Atlantic salmon (Salmo salar). Genet. Sel. Evol. 45, 1–8. https://doi.org/10.1186/1297-9686-45-41

Sorensen, D., Waagepetersen, R., 2003. Normal linear models with genetically structured residual variance heterogeneity: A case study. Genet. Res. 82, 207–222. https://doi.org/10.1017/S0016672303006426

Sui, J., Luan, S., Luo, K., Meng, X., Lu, X., Cao, B., Li, W., Chai, Z., Liu, N., Xu, S., Kong, J., 2015. Genetic parameters and response to selection for harvest body weight of pacific white shrimp, Litopenaeus vannamei. Aquac. Res. 47, 2795–2803. https://doi.org/10.1111/are.12729

Tan, J., Luan, S., Luo, K., Guan, J., Li, W., Sui, J., Guo, Z., Xu, S., Kong, J., 2017. Heritability and genotype by environment interactions for growth and survival in Litopenaeus vannamei at low and high densities. Aquac. Res. 48, 1430–1438. https://doi.org/10.1111/are.12978

Thitamadee, S., Prachumwat, A., Srisala, J., Jaroenlak, P., Salachan, P.V., Sritunyalucksana, K., Flegel, T.W., Itsathitphaisarn, O., 2016. Review of current disease threats for cultivated penaeid shrimp in Asia. Aquaculture. https://doi.org/10.1016/j.aquaculture.2015.10.028

Yáñez, J.M., Newman, S., Houston, R.D., 2015. Genomics in aquaculture to better understand species biology and accelerate genetic progress. Front. Genet. 6, 1–3. https://doi.org/10.3389/fgene.2015.00128

Yang, Y., Christensen, O.F., Sorensen, D., 2011. Analysis of a genetically structured variance heterogeneity model using the Box-Cox transformation. Genet. Res. (Camb). 93, 33–46. https://doi.org/10.1017/S0016672310000418

Zhang, J., Cao, F., Liu, J., Yuan, R., Hu, Z., 2016. Genetic Parameters for Growth and Hypoxic Tolerance Traits in Pacific White Shrimp Litopenaeus vannamei at Different Ages. N. Am. J. Aquac. 79, 75–83. https://doi.org/10.1080/15222055.2016.1194923

Zhang, P., Zhang, X., Li, J., Meng, Q., 2008. Observation of behavior in Fenneropenaeus chinensis and Litopenaeus vannamei postlarvae. J. Fish. China 32, 223–228.

Chapter 3 - Application of a novel 50K SNP genotyping array to assess the genetic diversity and linkage disequilibrium in a farmed Pacific white shrimp (*Litopenaeus vannamei*) population[a]

ABSTRACT - The Pacific white shrimp (*Litopenaeus vannamei*) is the most farmed shrimp species globally. The inclusion of genomic information became a reality in shrimp breeding and it is expected to accelerate the genetic gain over time. The decay of linkage disequilibrium (LD) between single nucleotide polymorphisms (SNPs) is an important measure to evaluate the feasibility of implementing genomic selection. The aim of this study was to evaluate the use of a novel 50K SNP array tool to characterize the genomic diversity, LD and effective population size (N_e) in a farmed shrimp population. A total of 96 animals (40 sires and 56 dams) were genotyped using the novel Illumina AquaArray HD (50K) vannamei®. Quality control (QC) of genomic data was performed and three different minor allele frequency (MAF) exclusion thresholds were applied: < 0.10 (QC1), < 0.05 (QC2) and < 0.01 (QC3). After QC, 34,425, 39,091 and 42,789 SNPs were retained for QC1, QC2 and QC3, respectively, validating the high informativeness of this SNP array to this particular shrimp breeding population. The population showed a considerable high overall heterozygosity in comparison to other aquaculture species meaning that genetic diversity is stable despite selection. The principal component analysis revealed three genetically distant groups with the first two principal components explaining 27.7% of total variation. LD decayed rapidly in the first 30Kb of distance between markers from 0.20 to 0.07 and then decreased to 0.02 in the long-range distance. These results suggest a relatively recent incorporation of animals from different populations in the broodstock. N_e size reduced from 7,871 to 301 animals in 827 generations for QC1, 8,253 to 305 animals for QC2 and 8,957 to 315 animals for QC3, both in 899 generations. Contemporary N_e was close to 86 for all QCs. The level of LD estimated suggests that genomic selection is feasible in shrimp by using this SNP array. For GWAS, the rapid LD decay indicates that a higher proportion of SNPs would be necessary to accurately map quantitative trait loci (QTLs).

Keywords: Shrimp, SNP array, Linkage disequilibrium, Effective population size

[a]Article published in the “Aquaculture Reports” doi: 10.1016/j.aqrep.2021.100691

3.1 Introduction

The Pacific white shrimp (*Litopenaeus vannamei*) is the most produced shrimp world-wide, accounting for approximately 83% of all shrimp and prawn farmed in 2018 (FAO, 2020). In the Americas, *L. vannamei* farming is present mainly in Brazil, Mexico and Ecuador accounting for more than 80% of all shrimp produced in this continent in 2018 (FAO, 2020). A strategy to increase overall production in shrimp and other aquaculture species is the implementation of breeding programs (Gjedrem, 2005). Several studies showed expressive improvement (around 5-15% by generation) in economic relevant traits in shrimp through selection, such as, growth (Argue et al., 2002; Campos-Montes et al., 2013) and resistance to diseases (Hernández-Ruíz et al., 2020; Trang et al., 2019).

As in other aquaculture species, most of breeding programs in shrimp farming are based on recording phenotypic and pedigree information to estimate breeding values (Neira, 2010). With the advent of genomic technologies, genomic information has also been included in breeding programs paving the way for which has been called genomic selection (GS) (Meuwissen et al., 2001). GS is feasible when a sufficient number (thousands) of single nucleotide polymorphism (SNP) markers, distributed all over the genome, are available, allowing to capture genetic variability through linkage disequilibrium (LD) with quantitative trait loci (QTL) (Goddard, 2009).

The use of genomic information may allow higher accuracy in estimating the genetic merit (i.e., genomic estimated breeding values) in comparison to traditional pedigree methods, thus accelerating the rate of genetic gain (Correa et al., 2017; Ødegård et al., 2014; Vallejo et al., 2017; Yoshida et al., 2019b, 2018). GS has special importance for traits that are difficult and/or expensive to measure (e.g., resistance to pathogens and product quality traits), because it explores the relationship between phenotyped and selection candidates more accurately (Calus and Veerkamp, 2011). Another application of genomic information is related to genome-wide association studies (GWAS). This approach may help to unravel the genetic architecture of traits and detect genomic regions associated to genetic variation by using LD information between markers and QTLs. Besides, the LD pattern may be used to understand historical

processes and evolutionary forces that affected the population over time (e.g., artificial/natural selection, bottlenecks, cross-breeding) (Heuertz et al., 2006). Thus, the level and extension of LD may help to determine the potential of a specific SNP panel to predict SNP effects in the implementation of GS as well as mapping of QTLs in GWAS and also provide important information about population genetic diversity (Johnson et al., 2010).

Different methods were proposed to evaluate LD in populations, among them, r^2 is the most common measure (Pritchard and Przeworski, 2001). By definition, the LD is the non-independent association between two loci, resulting in frequency of occurrence distinct that the expected by chance. In case of no recombination between these alleles, there is complete LD. LD in farmed populations has been subject of study in livestock as, for example, in dairy and beef cattle (Bohmanova et al., 2010; Espigolan et al., 2013); pigs (Ai et al., 2013; Amaral et al., 2008; Du et al., 2007), sheep (Al-Mamun et al., 2015; Prieur et al., 2017), chicken (Andreescu et al., 2007; Fu et al., 2015) and goats (Berihulay et al., 2019; Brito et al., 2015; Mdladla et al., 2016). LD has also been investigated in aquaculture species as, for example, in Atlantic salmon (Barría et al., 2018; Gutierrez et al., 2015; Hayes et al., 2006; Kijas et al., 2017); coho salmon (Barría et al., 2019); rainbow trout (Rexroad and Vallejo, 2009; Vallejo et al., 2018); Nile tilapia (Yoshida et al., 2019a) and Pacific oyster (Zhong et al., 2017).

For shrimp, Du et al. (2010) and Ciobanu et al. (2010) reported the first results of LD estimation using 1,344 and 1,221 putative SNPs, respectively. Jones et al. (2017) used the first 6.4K SNP array for shrimp to estimate the mean r^2 of adjacent SNP pairs as 0.15 for distances lower than 1 cM. More recently, Wang et al. (2019), using the 2b-RAD method for discovering and genotyping SNPs in 200 shrimps, estimated LD decay for about 2.6 million of paired SNPs. A novel 50K SNP array is commercially available for shrimp breeding. Up to date, there are no studies aimed at evaluating the applicability of this tool for the incorporation of genomic information in shrimp breeding populations. In this study, we present the first application of a 50K SNP to characterize the genetic diversity, estimate the linkage disequilibrium and the effective population size of a breeding population of Pacific white shrimp from Ecuador.

3.2 Material and methods

3.2.1 Population and samples

Animals were obtained from the shrimp breeding population owned by Opúsculo del Mar S.A. (Santa Elena, Ecuador). The population was established in 1999, in which a mass selection scheme has been implemented since then to improve harvest weight for about 20 generations. An average of 140 families are created on each generation by using hierarchical matings (1:2-3 male:female ratio) and the generation interval spans 12 months. A considerable increase in growth rate has been achieved in the population across 20 generations. For instance, the average body weight of females at the age of six months in 1999 and 2020 was 29 g and 60 g, respectively. No pedigree information is available for this breeding program. With the objective to evaluate a SNP genotyping tool to incorporate genomics to assist selection, a total of 96 animals (40 sires and 56 females) were randomly sampled in January 2020 and used in this study. Due to unknown information about the families, the relatedness among animals was assessed using genomic relationship through the G matrix (Strandén and Garrick, 2009). The average relatedness was close to 0 (-1.03^{-10} SD = 0.08), with a minimum of -0.09 and maximum of 0.53 (negative relatedness estimate means that the individuals are less related than the average). It is important to mention that these samples have no genetic links with the samples used to develop the 50K SNP array.

3.2.2 Genotyping, quality control and genetic diversity

Pleopods of the 96 animals were sampled and used for genomic DNA extraction. Further, genotyping was performed using the novel Illumina AquaArray HD (50K) vannamei® panel developed by Neogen® (Nebraska, USA) and commercially available through the Center for Aquaculture Technologies (San Diego, USA). This panel uses the assembly ASM378908v1 (GenBank accession GCF_003789085.1; Zhang et al., 2019) of *L. vannamei* as a reference to discovery SNPs.

Quality control of genotypes consisted of four exclusion criteria: call-rate < 0.8 for SNPs and samples, Hardy-Weinberg equilibrium (p-value < 10^{-5}) and minor allele frequency (MAF) with different threshold values, forming three quality control sets (QC): < 0.10 (QC1), < 0.05 (QC2) and < 0.01 (QC3). By using

different levels of MAF, we investigated the effect of including less frequent alleles in the LD estimation.

The genetic diversity of the population was evaluated by calculating the observed (H_0) and expected (H_E) mean heterozygosity for each QC and using principal component analysis (PCA) on the genomic relationship matrix. The observed heterozygosity for each SNP ($H_{0(SNP)}$) was obtained as: $H_{0(SNP)} = \frac{N_{AB}}{(N_{AA}+N_{AB}+N_{BB})}$, where N_{AA}, N_{AB}, N_{BB} are the number of individuals carrying AA, AB and BB alleles for each SNP, respectively. The expected heterozygosity ($H_{E(SNP)}$) for each SNP was estimated as: $H_{E(SNP)} = 2pq$, where p is the frequency of A allele and q is the frequency of B allele (obtained simply as: $1 - p$), respectively. The H_0 and H_E values may be compared to other populations to indicate the level of genetic diversity genome-wide. If H_0 in a breeding population is lower in comparison to a wild population, for instance, there is loss of genetic diversity in the first population and inbreeding may be a cause of such poor genetic variability. On the other hand, a higher value of H_0 indicates high genetic variability due to introduction of meta-populations, for example. A PCA was also performed based on the variance-standardized relationship matrix (Yang et al., 2011) which may be calculated as:

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

where A_{jk} is the genetic relationship between individuals j and k, N is the number of SNPs after quality control, x_{ij} and x_{ik} are the numbers of copies of the reference allele for the *i*th SNP of the *j*th individual and *k*th individual, respectively, and p is the frequency of the reference allele. The building of the genetic relationship allows to investigate the possible existence of subpopulations by plotting the two principal components that explained higher proportions of variation in this matrix. The quality control and genetic diversity parameters described above were evaluated using the PLINK 1.9 software (Purcell et al., 2007).

3.2.3 LD estimation

Considering two biallelic loci, one with alleles a and A and other with alleles b and B, the LD was estimated using the r^2 formula (Hill and Robertson, 1968):

$r^2(p_A, p_B, p_{AB}) = \frac{(p_{AB} - p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)}$, where p_{AB} is the frequency of haplotypes with the allele A at locus 1 and allele B at locus 2, and p_A and p_B are the frequencies of alleles A and B, respectively. The parameters: --ld-window-kb 10000, --ld-window 99999 and --ld-window-r2 set to zero were used in PLINK 1.9 to estimate r^2 between all SNPs in each contig for windows of 10Mb maximum size. The minimum and maximum distances between two adjacent SNPs in the LD analysis were 2.21 and 10,000Kb, respectively. Each pair-wise r^2 was included in a determined bin based on the physical distance between syntenic SNPs to generate the LD decay plot.

3.2.4 Effective population size (N_e)

The historical effective population size (N_e) was estimated using the SNeP v1.1 tool (Barbato et al., 2015). SNeP applies a formula developed by Corbin et al. (2012):

$$N_{et} = \frac{1}{(4f(c_t))}\left(\frac{1}{E[r^2_{adj}|c_t]} - \alpha\right)$$

where N_{et} is the effective population size t generations ago, c_t is the recombination rate for a specific physical distance between SNPs, r^2_{adj} is the LD adjusted for sample size ($r^2_{adj} = r^2 - (1/sample\ size)$) and α is the adjustment for mutation rate. The c_t component is inferred using mapping functions that translate physical distance into linkage distance. Firstly, the recombination rate between two SNPs is inferred using an approximation: $\delta = kd$, where δ is the physical distance in Mb, k is a constant equal to 10^{-8} and d is the linkage distance in centiMorgan (cM). This approximation is not valid for large values of d due to higher probabilities of recombination and non-linearity between map distance and the recombination rate. To overcome this, we applied a mapping function to modify the c_t developed by Sved (1971):

$$c_t = \frac{1 - \left(\frac{d}{2}\right)}{(1-d)^2}$$

where c_t is the recombination rate between two SNPs and d is the linkage distance previously estimated. We considered $\alpha = 2$, meaning that mutation does occur. This is a theoretical value that assumes mutations occurring at different time points on the same chromosome which has been previously derived by Ohta

and Kimura (1971). The N_e was estimated between SNPs within a distance of 0 to 5Mb and the number and size of each bin was 30 and 50Kb, respectively. Due to the low number of SNPs per contig, the mean N_e was expressed using harmonic means calculated for each bin (Alvarenga et al., 2018). In order to estimate the contemporary N_e, we used the software NeEstimator v2.01 (Do et al., 2014). First, the genotype files in PLINK format were transformed to GENEPOP format (Rousset, 2008) using the R software (R Core Team, 2017) through the radiator package (Gosselin et al., 2020). Then, the LD method was chosen assuming non-random mating and a critical value for removing rare alleles of 0.05 to obtain the contemporary N_e considering each quality control.

3.3 Results

3.3.1 SNP array validation and quality control

The initial genotype dataset had 50,811 SNPs from 96 animals. After quality control 6 animals were removed due to low call-rate (<0.80). The number of SNPs that were not present in at least 80% of animals was only 3,570 showing the high quality of genotyping of samples using the novel 50K array (Table 1). It is possible to notice that MAF was the filter that eliminated most SNPs (11,075, 7,374 and 3,664, for QC1, QC2 and QC3, respectively). The final numbers of SNPs after each quality control were 34,425, 39,091 and 42,789 for QC1, QC2 and QC3, respectively.

Table 1. Summary of number of markers included after different quality controls (QC) performed in the 50K SNP genotypes in a farmed shrimp population.

QC[1]	Call-rate[2]		MAF[3]	HWE[4]	Heterozygosity[5]	
	Samples	SNPs (%)*	SNPs (%)	SNPs (%)	H_O	H_E
QC1	90	47,241 (93.0)	36,166 (71.2)	35,425 (69.2)	0.4041	0.3994
QC2	90	47,241 (93.0)	39,867 (78.5)	39,091 (76.9)	0.3783	0.3744
QC3	90	47,241 (93.0)	43,577 (85.8)	42,789 (84.2)	0.3496	0.3461

[1] Quality control applied (differences only in MAF filter)
[2] Number of SNPs and samples included after the exclusion criterion call-rate < 0.8
[3] Number and proportion of SNPs included by minor allele frequency (MAF) threshold: QC1 (MAF > 0.10), QC2 (MAF > 0.05), and QC3 (MAF > 0.01)
[4] Number and proportion of SNPs in Hardy–Weinberg Equilibrium (HWE) ($p > 10^{-5}$)
[5] Heterozygosity observed (H_O) and expected (H_E)
* All percentages are in comparison to the initial number of SNPs (50,811)

There was small variation in MAF levels regarding the quality controls (Fig. 1). Comparing the two most contrasting scenarios (QC1 *vs* QC3), the proportion of SNPs loci ranged from 1.85 to 4.44% considering all MAF classes. The mean MAF was 0.31 ± 0.12 for QC1, 0.28 ± 0.13 for QC2 and 0.26 ± 0.14 for QC3. In addition, a large proportion of SNPs have MAF > 0.2, showing high level of polymorphism present in the studied population and captured by the SNP array. The number of SNPs that were eliminated by more than one criterion is shown in the supplementary material (Appendix A).

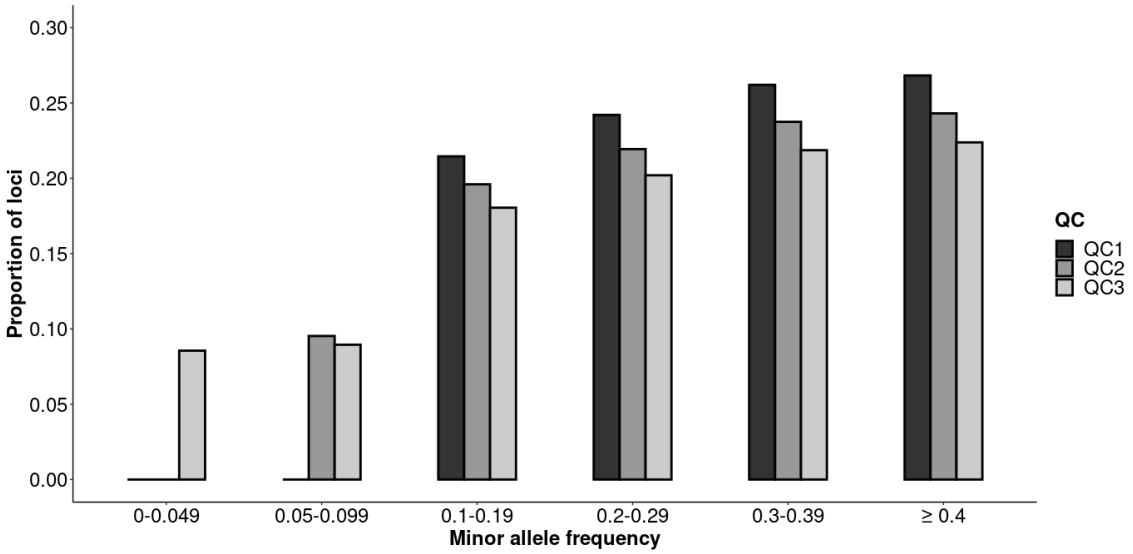


Figure 1. Proportion of SNPs in different minor allele frequency (MAF) intervals after distinct quality controls (QC) using three MAF exclusion criteria thresholds: QC1 (< 0.10), QC2 (< 0.05) and QC3 (< 0.01).

3.3.2 Genetic diversity

Observed and expected heterozygosity values are showed in Table 1. The observed and expected values were very similar for all QCs and the overall heterozygosity seems to be at an acceptable level despite selection for 20 generations. As expected, a reduction in the H_E is noticed as MAF decreases, i.e., the proportion of homozygous loci increases when SNPs with lower MAF are included. This reduction in the heterozygosity was equal to 6.4 and 13.5% for QC2 and QC3 in comparison to QC1, respectively. For the PCA, all QCs produced similar results, thus, we presented results of QC1 only. The two most important principal components explained 27.7% of the genomic relationship variation among animals in this population (Fig. 2). Three clusters may be

observed in the PCA: a larger group of 75 animals (top-left), a small group of 7 females (bottom-centre) and a last small group with 8 animals of both sexes (top-right). The first principal component differentiates the three clusters (14.4% of variation) and the second component differentiates the female group from the other two groups (13.3% of variation).

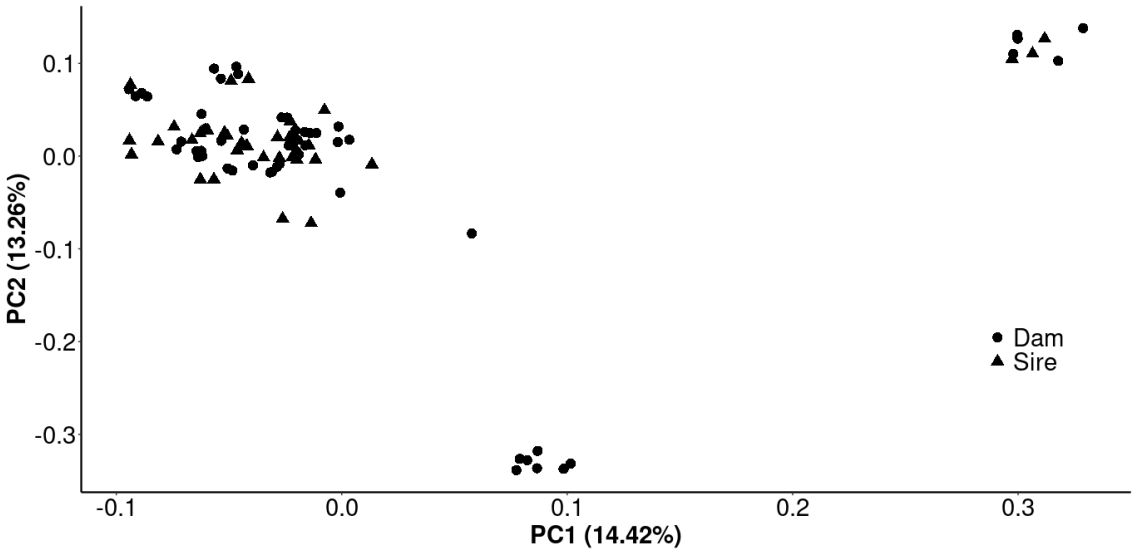


Figure 2. Genetic structure of a farmed shrimp population based on principal component analysis of the genomic relationship matrix (G). Sires and dams are identified by different shapes. Only SNPs with minor allele frequency ≥ 0.10 were considered to compute G.

3.3.3 LD and N_e estimation

The estimates of LD for each contig are shown in Table 2. Overall, LD levels varied among contigs ranging from 0.21 to 0.26, considering all adjacent SNPs in each contig and quality controls. The MAF had minor effect on LD estimations within each contig. The LD decay (Fig. 3) revealed a rapid decrease of r^2 as the physical distance between markers increased. The most significant drop was observed in the first 30Kb of distance and then the r^2 level slightly decreased to values close to 0.02. Regarding the different QCs, the inclusion of SNPs with less frequent alleles (QC2 and QC3) seemed to decrease the r^2 in the

Table 2. Estimated average linkage disequilibrium (r^2 ± standard deviation), number of SNPs and mean distance between adjacent SNPs (Dist. ± standard deviation in Mb), for a farmed shrimp population. Values are expressed for each contig after distinct quality controls (QC) using different minor allele frequency (MAF) exclusion criteria thresholds: QC1 (< 0.10), QC2 (< 0.05) and QC3 (< 0.01).

Contig	QC1			QC2			QC3		
	r^2 ± SD	SNPs	Dist. ± SD	r^2 ± SD	SNPs	Dist. ± SD	r^2 ± SD	SNPs	Dist. ± SD
1	0.021 ± 0.035	1401	4.88 ± 2.87	0.021 ± 0.035	1508	4.88 ± 2.88	0.020 ± 0.034	1660	4.88 ± 2.88
2	0.021 ± 0.034	1333	4.88 ± 2.88	0.021 ± 0.033	1441	4.88 ± 2.88	0.020 ± 0.032	1557	4.88 ± 2.88
3	0.021 ± 0.033	1173	4.86 ± 2.88	0.021 ± 0.033	1309	4.85 ± 2.88	0.020 ± 0.032	1415	4.85 ± 2.88
4	0.021 ± 0.035	1090	4.85 ± 2.88	0.021 ± 0.034	1198	4.85 ± 2.88	0.021 ± 0.033	1296	4.85 ± 2.88
5	0.021 ± 0.032	1028	4.83 ± 2.87	0.021 ± 0.031	1124	4.84 ± 2.87	0.020 ± 0.031	1212	4.84 ± 2.87
6	0.022 ± 0.033	980	4.82 ± 2.88	0.021 ± 0.033	1080	4.82 ± 2.88	0.021 ± 0.032	1169	4.82 ± 2.88
7	0.022 ± 0.033	948	4.83 ± 2.88	0.021 ± 0.033	1042	4.83 ± 2.88	0.021 ± 0.033	1130	4.83 ± 2.88
8	0.021 ± 0.034	945	4.80 ± 2.88	0.021 ± 0.033	1025	4.80 ± 2.88	0.020 ± 0.032	1093	4.81 ± 2.88
9	0.022± 0.033	898	4.81 ± 2.87	0.022 ± 0.033	977	4.81 ± 2.87	0.021 ± 0.032	1065	4.81 ± 2.87
10	0.023 ± 0.035	894	4.80 ± 2.88	0.022 ± 0.034	973	4.80 ± 2.88	0.022 ± 0.034	1064	4.80 ± 2.88
11	0.022 ± 0.036	854	4.80 ± 2.88	0.022 ± 0.035	952	4.81 ± 2.88	0.021 ± 0.034	1022	4.80 ± 2.88
12	0.023 ± 0.040	843	4.82 ± 2.88	0.023 ± 0.039	926	4.81 ± 2.88	0.022 ± 0.037	1013	4.80 ± 2.88
13	0.022 ± 0.034	854	4.80 ± 2.87	0.021 ± 0.034	944	4.80 ± 2.88	0.021 ± 0.033	1021	4.80 ± 2.88
14	0.021 ± 0.034	813	4.80 ± 2.88	0.021 ± 0.033	909	4.79 ± 2.87	0.020 ± 0.032	988	4.80 ± 2.87
15	0.022 ± 0.038	765	4.77 ± 2.87	0.022 ± 0.036	842	4.77 ± 2.87	0.021 ± 0.035	925	4.76 ± 2.87
16	0.025 ± 0.041	705	4.75 ± 2.88	0.024 ± 0.039	784	4.75 ± 2.88	0.023 ± 0.038	856	4.75 ± 2.88
17	0.022 ± 0.032	768	4.77 ± 2.87	0.021 ± 0.032	853	4.76 ± 2.87	0.021 ± 0.031	911	4.76 ± 2.87
18	0.022 ± 0.034	700	4.72 ± 2.87	0.021 ± 0.034	772	4.72 ± 2.87	0.021 ± 0.033	835	4.72 ± 2.87
19	0.021 ± 0.033	736	4.76 ± 2.86	0.021 ± 0.033	810	4.75 ± 2.87	0.021 ± 0.032	865	4.75 ± 2.87
20	0.022 ± 0.035	650	4.74 ± 2.88	0.022 ± 0.034	715	4.72 ± 2.88	0.021 ± 0.033	785	4.72 ± 2.87

Table 2. Continued.

21	0.022 ± 0.035	630	4.66 ± 2.87	0.021 ± 0.034	693	4.69 ± 2.86	0.021 ± 0.033	750	4.70 ± 2.87
22	0.023 ± 0.039	637	4.71 ± 2.86	0.022 ± 0.038	694	4.71 ± 2.86	0.021 ± 0.036	763	4.71 ± 2.86
23	0.021 ± 0.038	633	4.72 ± 2.87	0.021 ± 0.036	699	4.71 ± 2.87	0.020 ± 0.035	765	4.71 ± 2.87
24	0.021 ± 0.032	623	4.66 ± 2.86	0.021 ± 0.032	673	4.67 ± 2.86	0.020 ± 0.031	720	4.65 ± 2.86
25	0.021 ± 0.034	597	4.69 ± 2.87	0.021 ± 0.033	654	4.69 ± 2.86	0.021 ± 0.033	702	4.68 ± 2.86
26	0.022 ± 0.037	587	4.71 ± 2.87	0.021 ± 0.036	650	4.72 ± 2.87	0.021 ± 0.035	713	4.71 ± 2.87
27	0.023 ± 0.037	547	4.65 ± 2.86	0.023 ± 0.037	594	4.65 ± 2.86	0.022 ± 0.036	633	4.66 ± 2.86
28	0.022 ± 0.033	484	4.56 ± 2.84	0.022 ± 0.033	528	4.56 ± 2.84	0.021 ± 0.032	573	4.55 ± 2.84
29	0.020 ± 0.033	419	4.52 ± 2.84	0.020 ± 0.032	446	4.52 ± 2.84	0.020 ± 0.032	496	4.52 ± 2.85
30	0.021 ± 0.030	434	4.53 ± 2.84	0.020 ± 0.030	482	4.50 ± 2.84	0.020 ± 0.029	524	4.53 ± 2.85
31	0.020 ± 0.031	424	4.52 ± 2.84	0.020 ± 0.030	468	4.52 ± 2.83	0.019 ± 0.030	504	4.51 ± 2.83
32	0.020 ± 0.031	381	4.43 ± 2.83	0.020 ±0 .031	429	4.43 ± 2.83	0.019 ± 0.031	467	4.44 ± 2.83
33	0.021 ± 0.031	364	4.37 ± 2.82	0.020 ± 0.030	400	4.38 ± 2.81	0.020 ± 0.030	437	4.39 ± 2.81
34	0.026 ± 0.046	360	4.39 ± 2.81	0.025 ± 0.044	400	4.39 ± 2.81	0.024 ± 0.043	426	4.38 ± 2.81
35	0.021 ± 0.032	339	4.26 ± 2.79	0.020 ± 0.031	363	4.26 ± 2.78	0.020 ± 0.030	393	4.27 ± 2.78
36	0.021 ± 0.035	313	4.22 ± 2.76	0.021 ± 0.034	352	4.24 ± 2.78	0.020 ± 0.033	379	4.23 ± 2.77
37	0.021 ± 0.032	314	4.30 ± 2.80	0.021 ± 0.032	341	4.29 ± 2.79	0.020 ± 0.031	369	4.30 ± 2.79
38	0.022 ± 0.035	328	4.32 ± 2.80	0.021 ± 0.034	353	4.33 ± 2.80	0.020 ± 0.033	388	4.32 ± 2.79
39	0.020 ± 0.033	315	4.27 ± 2.80	0.020 ± 0.032	342	4.28 ± 2.79	0.019 ± 0.031	368	4.27 ± 2.78
40	0.023 ± 0.036	302	4.24 ± 2.79	0.022 ± 0.035	338	4.25 ± 2.80	0.021 ± 0.034	368	4.25 ± 2.79
41	0.021 ± 0.035	270	4.01 ± 2.70	0.021 ± 0.034	294	4.03 ± 2.72	0.020 ± 0.033	318	4.02 ± 2.72
42	0.022 ± 0.041	156	2.57 ± 1.80	0.021 ± 0.039	169	2.58 ± 1.81	0.020 ± 0.037	186	2.60 ± 1.83
43	0.023 ± 0.040	72	1.22 ± 0.85	0.023 ± 0.039	80	1.22 ± 0.86	0.022 ± 0.037	85	1.21 ± 0.84
44	0.026 ± 0.046	60	1.16 ± 0.82	0.025 ± 0.045	65	1.14 ± 0.80	0.024 ± 0.043	69	1.12 ± 0.78

short-range distances, compared to QC1, but all had similar pattern of LD decay in the long-range distances.

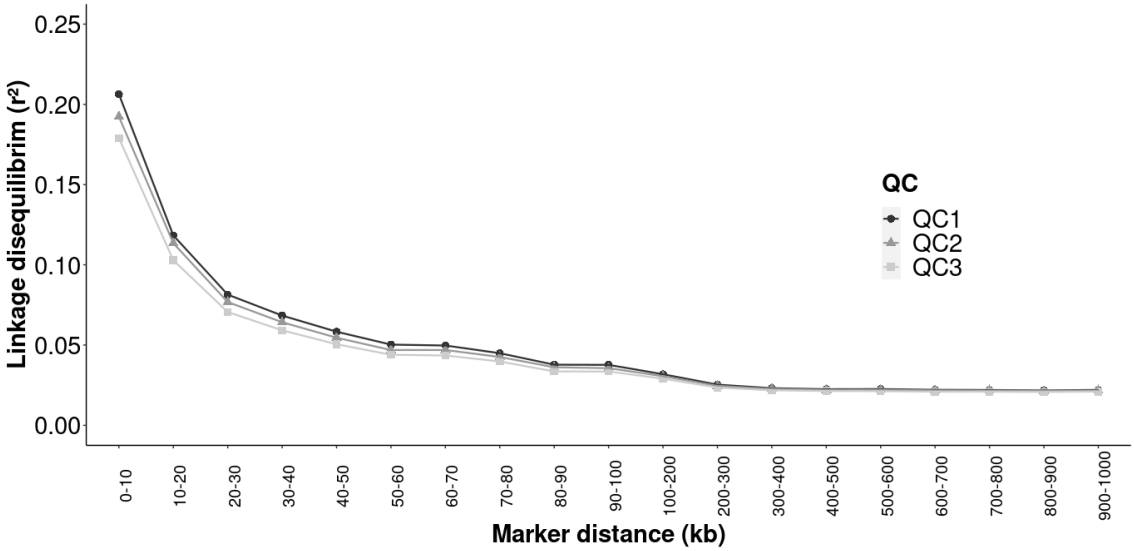


Figure 3. Estimated average linkage disequilibrium (r^2) over different classes of distance between syntenic markers, for a farmed shrimp population, using different quality controls (QC) with three minor allele frequency exclusion criteria thresholds: QC1 (< 0.10), QC2 (< 0.05) and QC3 (< 0.01).

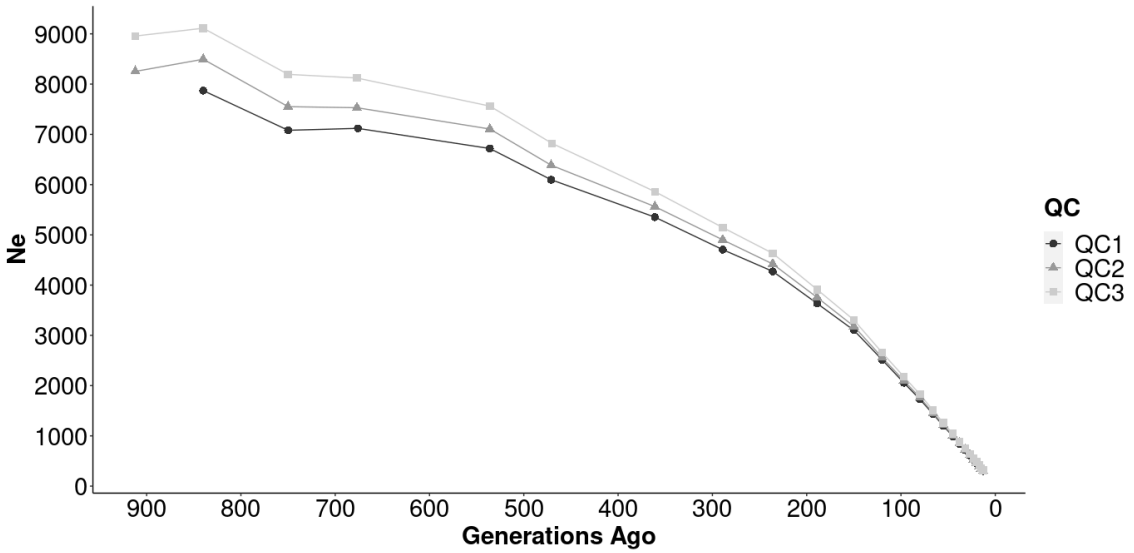


Figure 4. Effective population size (N_e) estimated in a farmed shrimp population using different quality controls (QC) with three minor allele frequency exclusion criteria thresholds: QC1 (< 0.10), QC2 (< 0.05) and QC3 (< 0.01).

The historic N_e using the LD method showed a decrease in size of 7,871 to 301 animals in 827 generations for QC1, 8,253 to 305 animals in 899 generations for QC2 and 8,957 to 315 animals in 899 generations for QC3 (Fig.

4). The most significant decrease was observed in the last 300 generations with approximately 62% of total reduction in this period. The contemporary N_e estimated was very similar among all quality controls 85.0, 86.6 and 86.7 for QC1, QC2 and QC3, respectively.

3.4 Discussion

3.4.1 Genetic diversity

This study shows the first application of a 50K SNP genotyping array to a farmed shrimp population. The panel presents 50,811 SNPs, which after quality control yielded 69.2, 76.9 and 84.2% out of the total SNPs for QC1, QC2 and QC3, respectively. Furthermore, the proportion of SNPs with MAF > 0.2 was close to 70% with a mean value of 0.3 for all QCs. These results show the high adequacy of this novel tool for this particular farmed shrimp population from Ecuador.

The evaluation of genetic diversity is important because it may help to assess the genetic variability of broodstock, genetic “bottlenecks”, level of inbreeding and show perspectives for genomic selection (Yáñez et al., 2015). A possible effect of mass selection is the reduction of overall heterozygosity through generations. Our results showed that overall heterozygosity is at a satisfactory level and its reduction overtime was likely slow in this population. Despite genetic diversity has been studied in shrimp using different types of genetic markers (see Benzie, 2000 for a review), few studies were found using SNPs (Ciobanu et al., 2010; Lu et al., 2018; Perez-Enriquez et al., 2018) and none study was found characterizing a farmed population using dense genomic information becoming difficult to compare our results with previous reports in literature. Indeed, the comparison of the results obtained with wild *L. vannamei* or other commercial populations would be more adequate to clearly visualize the effect of selection on genetic diversity. However, expected heterozygosity found in the present study (0.35-0.40) is in the upper level to what has been found in other farmed aquaculture species such as, Tilapia (0.28-0.36) (Yáñez et al., 2020; Yoshida et al., 2019a), sea trout (0.27) (Drywa et al., 2013) and Atlantic salmon (0.22-0.38) (Kijas et al., 2017).

A possible explanation for the considerable overall heterozygosity may be the high genetic variability present in the population. Likely, there was inclusion of animals from different strains in this broodstock recently. This assumption may be verified by the PCA results which display three distinct groups. We expect that these groups represent sub-populations rather than families because hybridization with other wild and farmed populations could not be discarded. Although inbreeding is not a direct measure of genetic diversity, the directional selection applied in breeding programs may generate diversity loss and inbreeding depression (i.e., inbred individuals have reduced average performance in comparison to non-inbred) (De Donato et al., 2005). In this population, the recent introduction of animals of different populations or strains allowed a more flexible mating system by broadening options of selection, reducing the chances of reproduction between relatives and controlling inbreeding.

Another aspect that reinforces the consistence of heterozygosity results obtained is associated to the lack of genetic links between the population studied and the population used to discover and build the SNP array. In order to select SNPs to include in the array, a small base population is genome-sequenced and the SNPs are chosen following mainly quality, spacing and frequency as criteria (Malomane et al., 2018). The non-random selection of animals as discovery population and the pre-filter applied on SNPs may generate the ascertainment bias, i.e., the observed allelic frequencies are different than those expected for a random sample (Lachance and Tishkoff, 2013). As a result, measures of genetic variability, such as heterozygosity, in the discovery population may be overestimated (Rogers and Jorde, 1996). Thus, we expect that these parameters were accurately estimated. The inclusion of SNPs by softening the MAF filter (QC2 and QC3) decreased the overall heterozygosity. This result is expected because SNPs with less frequent alleles will increase the proportion of homozygote loci.

3.4.2 LD and N_e

LD provides important information about previous events that may have changed the population to what is observable on its current status. Our results presented variable r^2 estimations across all 44 contigs (0.21-0.26) with high

variation within each contig. This variation most likely happened due to different recombination rates at specific genomic regions (hotspots and coldspots) as a possible result of genetic drift and different genomic structures (Arias et al., 2009; Saunders et al., 2005). Several factors may affect LD, such as admixture, founder effect, inbreeding, mutation and selection (Gaut and Long, 2003). Accuracy of LD estimation depends on demographic and biological factors, as mentioned before, but also on sample size and relationship among animals. Bohmanova et al. (2010) suggested that a satisfactory number of animals to estimate LD is 55 using the r^2 measure. Gutierrez et al. (2015) stated that LD may be overestimated when the animals under study are highly related. The present study used 90 animals and, as explained in the methods section, there is low relationship level among them. Thus, we expect that LD was estimated accurately.

LD was estimated in other shrimp populations (Du et al., 2010; Jones et al., 2017; Wang et al., 2019). However, a direct comparison is inappropriate due to differences in the quality control of genotypes, method of SNP discovery, SNP density and intrinsic characteristics of each population. Most studies of LD in shrimp used low density SNPs arrays (<7K) which may not represent properly the whole LD considering that the genome of *L. vannamei* has approximately 2,600Mb (Zhang et al., 2019). In addition, the SNP array used in this study offers higher accuracy regarding the SNP-calling process and even distribution of markers across the genome in comparison to genotyping-by-sequencing methods applied in studies that used denser SNPs (Robledo et al., 2018).

The LD decay pattern was different than the one reported by Wang et al. (2019) for a *L. vannamei* population cultivated in China. It is possible to observe that LD at short-range was weaker in our study (~0.2) in comparison to Wang's study (~0.4) and that long-range LD was low (< 0.05) in both studies. The recent admixture of distinct populations is suspected to decrease LD at short-range (Ødegård et al., 2014) also known as admixture linkage disequilibrium (Pfaff et al., 2001). We believe that the level of admixture found within the Ecuadorian population studied here was absent in the Chinese population studied by Wang et al. (2019) generating the difference in the short-range LD. The *L. vannamei* is naturally found in the Pacific coast from Mexico to Northern Peru (Holthuis, 1980) while it was commercially introduced in Asia as an exotic species in 1996 (Biao

and Kaijin, 2007; Liao and Chien, 2011). Due to its natural occurrence, the inclusion of animals of different genetic basis or wild strains is recurrent in the Americas (Moss et al., 2012). This “admixture effect” was also described in other species, such as Nile tilapia (Yoshida et al., 2019a) and Salmon (Barría et al., 2018).

Some studies also suggested that the persistency of LD levels through distance (long-range LD) is influenced by admixture as well (Barría et al., 2018; Ødegård et al., 2014; Vallejo et al., 2018). However, the induced LD at long-range was not observed in this population. As suggested by Yoshida et al. (2019a), other biological and demographical factors also play important roles in the LD extension, such as recombination events and effective population size. The MAF filter had minor effect on LD estimation overall. The inclusion of SNPs with less frequent alleles (QC2 and QC3) decreased slightly r^2 at short-range distances. This probably happened because higher MAFs tend to overestimate the level of LD (Espigolan et al., 2013).

The historical N_e decreased considerably overtime with a more pronounced drop in the last 300 generations with a reduction of about 5,000 to 300 in the N_e. However, these results should be interpreted with caution. According to Corbin et al. (2012) the estimation of N_e, in the most recent generations, involves high rates of recombination which may compromise the N_e prediction model derived by Hayes et al. (2003). In this method, N_e is estimated assuming a relationship between the length of chromosomes segments and number of generations fitting better to lower rates of recombination. Moreover, Santiago et al. (2020) using simulation showed that coalescent methods (such as the one we applied in this study) may produce biased estimations of N_e when populations suffer abrupt reductions or expansions, especially in recent past generations. This likely has occurred in this shrimp population as suggested by the PCA results. This bias could also explain the sharply reduction of N_e size in the more distant past (300 generations ago) rather than when admixture likely occurred (~20 generations ago). Nevertheless, this method is a useful alternative to extract demographic information about unknown events that affected a population and has been widely used for N_e estimation in humans (Tenesa et al.,

2007), livestock (Makina et al., 2015) and aquaculture species (Barría et al., 2019, 2018; Yoshida et al., 2019a).

The MAF had major effect in the long-term N_e and QC3 presented higher N_e estimates at early generations in comparison to QC1 and QC2. The N_e is estimated using LD information and the different LD levels in each quality control as well as distinct recombination rates likely caused distinct patterns of early N_e. The contemporary N_e was close to 86 showing that N_e reduced 214 in the last 13 years (assuming one year of generation interval). The contemporary N_e observed is similar to the reported in other shrimp populations using microsatellites markers ranging from 50 to 173 (Cruz et al., 2004; De Donato et al., 2005; Ren et al., 2018; Vela Avitúa et al., 2013). The number estimated in this population is greater than 50 which was recommended by Ponzoni et al. (2010) to keep inbreeding rate limit close to 1% in breeding programs. Thus, our results showed that despite mass selection applied in this population, genetic variability and effective population size are at acceptable levels most likely as a result of recent incorporation of different populations.

3.4.3 Practical applications

The results of LD obtained here are relevant for the prospect of genomic tools applied to shrimp breeding. The extent of LD may affect GS and GWAS basically in two ways: for GS, it has implications on the accuracy of genomic estimated breeding values, and for GWAS, on the mapping power and resolution. For GS, the predictive ability of QTL effects may be totally dependent of LD when QTLs are not represented in the panel (Kizilkaya et al., 2010). Meuwissen (2009), suggested that, for unrelated individuals, high accuracy of genomic breeding values may be obtained using $2N_eL$ number of animals and $10N_eL$ number of markers, where N_e is the effective population size and L is the length of genome in Morgans. For this study, assuming N_e equals to 85 and that the genome size of *L. vannamei* has approximately 45.325 Morgans (Jones et al., 2017), 7,706 training records and 38,527 SNPs would be necessary to achieve accuracy of 0.88 – 0.93. Although, these formulas are empirical and were obtained using simulation of small populations, the marker density required is close to the obtained after QC2 and QC3 showing the feasibility of GS using this SNP array.

The level of LD observed in the present study shows opportunity to conduct GWAS using this SNP array. However, the QTL mapping power depends on other factors rather than SNP density, such as genetic architecture, heritability of the trait, sample size and statistical method which should be evaluated before performing association analysis.

3.5 Conclusion

The 50K SNP array panel applied in this study showed a high proportion of polymorphic loci being a suitable tool to estimate genetic variability, LD and N_e in a farmed shrimp population. The heterozygosity level and PCA results showed recent incorporation of animals from different populations or strains in this broodstock. This study revealed the potential use of a novel SNP array technology to GS and GWAS in shrimp breeding.

3.6 References

Ai, H., Huang, L., Ren, J., 2013. Genetic Diversity, Linkage Disequilibrium and Selection Signatures in Chinese and Western Pigs Revealed by Genome-Wide SNP Markers. PLoS One 8, e56001. https://doi.org/10.1371/journal.pone.0056001

Al-Mamun, H.A., A Clark, S., Kwan, P., Gondro, C., 2015. Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep. Genet. Sel. Evol. 47, 90. https://doi.org/10.1186/s12711-015-0169-6

Alvarenga, A.B., Rovadoscki, G.A., Petrini, J., Coutinho, L.L., Morota, G., Spangler, M.L., Pinto, L.F.B., Carvalho, G.G.P., Mourão, G.B., 2018. Linkage disequilibrium in Brazilian Santa Inês breed, Ovis aries. Sci. Rep. 8, 1–11. https://doi.org/10.1038/s41598-018-27259-7

Amaral, A.J., Megens, H.J., Crooijmans, R.P.M.A., Heuven, H.C.M., Groenen, M.A.M., 2008. Linkage disequilibrium decay and haplotype block structure in the pig. Genetics 179, 569–579. https://doi.org/10.1534/genetics.107.084277

Andreescu, C., Avendano, S., Brown, S.R., Hassen, A., Lamont, S.J., Dekkers, J.C.M., 2007. Linkage disequilibrium in related breeding lines of chickens. Genetics 177, 2161–2169. https://doi.org/10.1534/genetics.107.082206

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

Arias, J.A., Keehan, M., Fisher, P., Coppieters, W., Spelman, R., 2009. A high density linkage map of the bovine genome. BMC Genet. 10, 1–12. https://doi.org/10.1186/1471-2156-10-18

Barbato, M., Orozco-terWengel, P., Tapio, M., Bruford, M.W., 2015. SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. Front. Genet. 6, 109. https://doi.org/10.3389/fgene.2015.00109

Barría, A., Christensen, K.A., Yoshida, G., Jedlicki, A., Leong, J.S., Rondeau, E.B., Lhorente, J.P., Koop, B.F., Davidson, W.S., Yáñez, J.M., 2019. Whole Genome Linkage Disequilibrium and Effective Population Size in a Coho Salmon (Oncorhynchus kisutch) Breeding Population Using a High-Density SNP Array. Front. Genet. 10, 498. https://doi.org/10.3389/fgene.2019.00498

Barría, A., López, M.E., Yoshida, G., Carvalheiro, R., Lhorente, J.P., Yáñez, J.M., 2018. Population Genomic Structure and Genome-Wide Linkage Disequilibrium in Farmed Atlantic Salmon (Salmo salar L.) Using Dense SNP Genotypes. Front. Genet. 9, 649. https://doi.org/10.3389/fgene.2018.00649

Benzie, J.A.H., 2000. Population genetic structure in penaeid prawns. Aquac. Res. 31, 95–119. https://doi.org/10.1046/j.1365-2109.2000.00412.x

Berihulay, H., Islam, R., Jiang, L., Ma, Y., 2019. Genome-Wide Linkage Disequilibrium and the Extent of Effective Population Sizes in Six Chinese Goat Populations Using a 50K Single Nucleotide Polymorphism Panel. Animals 9, 350. https://doi.org/10.3390/ani9060350

Biao, X., Kaijin, Y., 2007. Shrimp farming in China: Operating characteristics, environmental impact and perspectives. Ocean Coast. Manag. 50, 538–550. https://doi.org/10.1016/j.ocecoaman.2007.02.006

Bohmanova, J., Sargolzaei, M., Schenkel, F.S., 2010. Characteristics of linkage disequilibrium in North American Holsteins. BMC Genomics 11, 1–11. https://doi.org/10.1186/1471-2164-11-421

Brito, L.F., Jafarikia, M., Grossi, D.A., Kijas, J.W., Porto-Neto, L.R., Ventura, R. V., Salgorzaei, M., Schenkel, F.S., 2015. Characterization of linkage disequilibrium, consistency of gametic phase and admixture in Australian and Canadian goats. BMC Genet. 16, 1–15. https://doi.org/10.1186/s12863-015-0220-1

Calus, M.P., Veerkamp, R.F., 2011. Accuracy of multi-trait genomic selection using different methods. Genet. Sel. Evol. 43, 26. https://doi.org/10.1186/1297-9686-43-26

Campos-Montes, G.R., Montaldo, H.H., Martínez-Ortega, A., Jiménez, A.M., Castillo-Juárez, H., 2013. Genetic parameters for growth and survival traits in Pacific white shrimp Penaeus (Litopenaeus) vannamei from a nucleus population undergoing a two-stage selection program. Aquac. Int. 21, 299–310. https://doi.org/10.1007/s10499-012-9553-1

Ciobanu, D.C., Bastiaansen, J.W.M., Magrin, J., Rocha, J.L., Jiang, D.-H., Yu, N., Geiger, B., Deeb, N., Rocha, D., Gong, H., Kinghorn, B.P., Plastow, G.S., van der Steen, H.A.M., Mileham, A.J., 2010. A major SNP resource for dissection of phenotypic and genetic variation in Pacific white shrimp (*Litopenaeus vannamei*

). Anim. Genet. 41, 39–47. https://doi.org/10.1111/j.1365-2052.2009.01961.x

Corbin, L.J., Liu, A.Y.H., Bishop, S.C., Woolliams, J.A., 2012. Estimation of historical effective population size using linkage disequilibria with marker data. J. Anim. Breed. Genet. 129, 257–270. https://doi.org/10.1111/j.1439-0388.2012.01003.x

Correa, K., Bangera, R., Figueroa, R., Lhorente, J.P., Yáñez, J.M., 2017. The use of genomic information increases the accuracy of breeding value predictions for sea louse (Caligus rogercresseyi) resistance in Atlantic salmon (Salmo salar). Genet. Sel. Evol. 49, 1–5. https://doi.org/10.1186/s12711-017-0291-8

Cruz, P., Ibarra, A.M., Mejia-Ruiz, H., Gaffney, P.M., Pérez-Enríquez, R., 2004. Genetic variability assessed by microsatellites in a breeding program of pacific white shrimp (Litopenaeus vannamei). Mar. Biotechnol. 6, 157–164. https://doi.org/10.1007/s10126-003-0017-5

De Donato, M., Manrique, R., Ramirez, R., Mayer, L., Howell, C., 2005. Mass selection and inbreeding effects on a cultivated strain of Penaeus (Litopenaeus) vannamei in Venezuela, in: Aquaculture. Elsevier, pp. 159–167. https://doi.org/10.1016/j.aquaculture.2005.02.005

Do, C., Waples, R.S., Peel, D., Macbeth, G.M., Tillett, B.J., Ovenden, J.R., 2014. NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Mol. Ecol. Resour. 14, 209–214. https://doi.org/10.1111/1755-0998.12157

Drywa, A., Poćwierz-Kotus, A., Was, A., Dobosz, S., Kent, M.P., Lien, S., Bernaś, R., Wenne, R., 2013. Genotyping of two populations of Southern Baltic Sea trout Salmo trutta m. trutta using an Atlantic salmon derived SNP-array. Mar. Genomics 9, 25–32. https://doi.org/10.1016/j.margen.2012.08.001

Du, F.X., Clutter, A.C., Lohuis, M.M., 2007. Characterizing linkage disequilibrium in pig populations. Int. J. Biol. Sci. https://doi.org/10.7150/ijbs.3.166

Du, Z.-Q., Ciobanu, D.C., Onteru, S.K., Gorbach, D., Mileham, A.J., Jaramillo, G., Rothschild, M.F., 2010. A gene-based SNP linkage map for pacific white shrimp, Litopenaeus vannamei. Anim. Genet. 41, 286–294. https://doi.org/10.1111/j.1365-2052.2009.02002.x

Espigolan, R., Baldi, F., Boligon, A.A., Souza, F.R.P., Gordo, D.G.M., Tonussi, R.L., Cardoso, D.F., Oliveira, H.N., Tonhati, H., Sargolzaei, M., Schenkel, F.S.,

Carvalheiro, R., Ferro, J.A., Albuquerque, L.G., 2013. Study of whole genome linkage disequilibrium in Nellore cattle. BMC Genomics 14, 1–8. https://doi.org/10.1186/1471-2164-14-305

FAO, 2020. FAO Fisheries & Aquaculture - Statistics [WWW Document]. URL http://www.fao.org/fishery/statistics/en (accessed 12.30.19).

Fu, W., Dekkers, J.C.M., Lee, W.R., Abasht, B., 2015. Linkage disequilibrium in crossbred and pure line chickens. Genet. Sel. Evol. 47, 11. https://doi.org/10.1186/s12711-015-0098-4

Gaut, B.S., Long, A.D., 2003. The lowdown on linkage disequilibrium. Plant Cell. https://doi.org/10.1105/tpc.150730

Gjedrem, T., 2005. Breeding Programs. Springer, Dordrecht, The Netherlands.

Goddard, M., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136, 245–257. https://doi.org/10.1007/s10709-008-9308-0

Gosselin, T., Lamothe, M., Devloo-Delva, F., Grewe, P., 2020. radiator: RADseq Data Exploration, Manipulation and Visualization using R. https://doi.org/10.5281/zenodo.3687060

Gutierrez, A.P., Yáñez, J.M., Fukui, S., Swift, B., Davidson, W.S., 2015. Genome-Wide Association Study (GWAS) for Growth Rate and Age at Sexual Maturation in Atlantic Salmon (Salmo salar). PLoS One 10, e0119730. https://doi.org/10.1371/journal.pone.0119730

Hayes, B.J., Gjuvsland, A., Omholt, S., 2006. Power of QTL mapping experiments in commercial Atlantic salmon populations, exploiting linkage and linkage disequilibrium and effect of limited recombination in males. Heredity (Edinb). 97, 19–26. https://doi.org/10.1038/sj.hdy.6800827

Hayes, B.J., Visscher, P.M., McPartlan, H.C., Goddard, M.E., 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 13, 635–643. https://doi.org/10.1101/gr.387103

Hernández-Ruíz, H., Montaldo, H.H., Bustos-Martínez, J., Campos-Montes, G.R., Castillo-Juárez, H., 2020. Heritability and genetic correlations for infectious hypodermal and hematopoietic necrosis virus load, body weight at harvest, and survival rate in Pacific white shrimp (*Litopenaeus vannamei*). J. World Aquac. Soc. 51, 312–323. https://doi.org/10.1111/jwas.12664

Heuertz, M., De Paoli, E., Källman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M., Gyllenstrand, N., 2006. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [Picea abies (L.) Karst]. Genetics 174, 2095–2105. https://doi.org/10.1534/genetics.106.065102

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38, 226–231. https://doi.org/10.1007/BF01245622

Holthuis, L.B., 1980. FAO species catalogue: Shrimps and prawns of the world., 1st ed. Food and Agriculture Organizationof the United Nations, Rome, Italy.

Johnson, R.C., Nelson, G.W., Troyer, J.L., Lautenberger, J.A., Kessing, B.D., Winkler, C.A., O'Brien, S.J., 2010. Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics 11, 1–6. https://doi.org/10.1186/1471-2164-11-724

Jones, D.B., Jerry, D.R., Khatkar, M.S., Raadsma, H.W., Steen, H. Van Der, Prochaska, J., Forêt, S., Zenger, K.R., 2017. A comparative integrated gene-based linkage and locus ordering by linkage disequilibrium map for the Pacific white shrimp, Litopenaeus vannamei. Sci. Rep. 7, 1–16. https://doi.org/10.1038/s41598-017-10515-7

Kijas, J., Elliot, N., Kube, P., Evans, B., Botwright, N., King, H., Primmer, C.R., Verbyla, K., 2017. Diversity and linkage disequilibrium in farmed Tasmanian Atlantic salmon. Anim. Genet. 48, 237–241. https://doi.org/10.1111/age.12513

Kizilkaya, K., Fernando, R.L., Garrick, D.J., 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes1. J. Anim. Sci. 88, 544–551. https://doi.org/10.2527/jas.2009-2064

Lachance, J., Tishkoff, S.A., 2013. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays 35, 780–786. https://doi.org/10.1002/bies.201300014

Liao, I.C., Chien, Y.-H., 2011. The Pacific White Shrimp, Litopenaeus vannamei, in Asia: The World's Most Widely Cultured Alien Crustacean, in: In the Wrong Place - Alien Marine Crustaceans: Distribution, Biology and Impacts. Springer Netherlands, pp. 489–519. https://doi.org/10.1007/978-94-007-0591-3_17

Lu, X., Kong, J., Meng, X., Cao, B., Luo, K., Dai, P., Luan, S., 2018. Identification

of SNP markers associated with tolerance to ammonia toxicity by selective genotyping from de novo assembled transcriptome in Litopenaeus vannamei. Fish Shellfish Immunol. 73, 158–166. https://doi.org/10.1016/j.fsi.2017.12.005

Makina, S.O., Taylor, J.F., van Marle-Köster, E., Muchadeyi, F.C., Makgahlela, M.L., MacNeil, M.D., Maiwashe, A., 2015. Extent of Linkage Disequilibrium and Effective Population Size in Four South African Sanga Cattle Breeds. Front. Genet. 6, 337. https://doi.org/10.3389/fgene.2015.00337

Malomane, D.K., Reimer, C., Weigend, S., Weigend, A., Sharifi, A.R., Simianer, H., 2018. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. BMC Genomics 19, 22. https://doi.org/10.1186/s12864-017-4416-9

Mdladla, K., Dzomba, E.F., Huson, H.J., Muchadeyi, F.C., 2016. Population genomic structure and linkage disequilibrium analysis of South African goat breeds using genome-wide SNP data. Anim. Genet. 47, 471–482. https://doi.org/10.1111/age.12442

Meuwissen, T.H., 2009. Accuracy of breeding values of “unrelated” individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41, 1–9. https://doi.org/10.1186/1297-9686-41-35

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819–1829.

Moss, S.M., Moss, D.R., Arce, S.M., Lightner, D. V., Lotz, J.M., 2012. The role of selective breeding and biosecurity in the prevention of disease in penaeid shrimp aquaculture. J. Invertebr. Pathol. https://doi.org/10.1016/j.jip.2012.01.013

Neira, R., 2010. Breeding in Aquaculture Species: Genetic Improvement Programs in Developing Countries, in: 9th World Congress on Genetics Applied to Livestock Production. Leipzig, Germany, p. 8.

Ødegård, J., Moen, T., Santi, N., Korsvoll, S.A., Kjøglum, S., Meuwissen, T.H.E., 2014. Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). Front. Genet. 5, 402. https://doi.org/10.3389/fgene.2014.00402

Ohta, T., Kimura, M., 1971. LINKAGE DISEQUILIBRIUM BETWEEN TWO SEGREGATING NUCLEOTIDE SITES UNDER THE STEADY FLUX OF MUTATIONS IN A FINITE POPULATION. Genetics 68.

Perez-Enriquez, R., Robledo, D., Houston, R.D., Llera-Herrera, R., 2018. SNP markers for the genetic characterization of Mexican shrimp broodstocks. Genomics 110, 423–429. https://doi.org/10.1016/j.ygeno.2018.10.001

Pfaff, C.L., Parra, E.J., Bonilla, C., Hiester, K., McKeigue, P.M., Kamboh, M.I., Hutchinson, R.G., Ferrell, R.E., Boerwinkle, E., Shriver, M.D., 2001. Population structure in admixed populations: Effect of admixture dynamics on the pattern of linkage disequilibrium. Am. J. Hum. Genet. 68, 198–207. https://doi.org/10.1086/316935

Ponzoni, R.W., Khaw, H.L., Nguyen, N.H., Hamzah, A., 2010. Inbreeding and effective population size in the Malaysian nucleus of the GIFT strain of Nile tilapia (Oreochromis niloticus). Aquaculture 302, 42–48. https://doi.org/10.1016/j.aquaculture.2010.02.009

Prieur, V., Clarke, S.M., Brito, L.F., McEwan, J.C., Lee, M.A., Brauning, R., Dodds, K.G., Auvray, B., 2017. Estimation of linkage disequilibrium and effective population size in New Zealand sheep using three different methods to create genetic maps. BMC Genet. 18, 1–19. https://doi.org/10.1186/s12863-017-0534-2

Pritchard, J.K., Przeworski, M., 2001. Linkage disequilibrium in humans: Models and data. Am. J. Hum. Genet. https://doi.org/10.1086/321275

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

R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ren, S., Mather, P.B., Tang, B., Hurwood, D.A., 2018. Levels of genetic diversity and inferred origins of Penaeus vannamei culture resources in China: Implications for the production of a broad synthetic base population for genetic improvement. Aquaculture 491, 221–231. https://doi.org/10.1016/j.aquaculture.2018.03.036

Rexroad, C.E., Vallejo, R.L., 2009. Estimates of linkage disequilibrium and effective population size in rainbow trout. BMC Genet. 10, 1–12.

https://doi.org/10.1186/1471-2156-10-83

Robledo, D., Palaiokostas, C., Bargelloni, L., Martínez, P., Houston, R., 2018. Applications of genotyping by sequencing in aquaculture breeding and genetics. Rev. Aquac. 10, 670–682. https://doi.org/10.1111/raq.12193

Rogers, A.R., Jorde, L.B., 1996. Ascertainment bias in estimates of average heterozygosity. Am. J. Hum. Genet. 58, 1033–1041.

Rousset, F., 2008. GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. Mol. Ecol. Resour. 8, 103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Santiago, E., Novo, I., Pardiñas, A.F., Saura, M., Wang, J., Caballero, A., 2020. Recent Demographic History Inferred by High-Resolution Analysis of Linkage Disequilibrium. Mol. Biol. Evol. 37, 3642–3653. https://doi.org/10.1093/molbev/msaa169

Saunders, M.A., Slatkin, M., Garner, C., Hammer, M.F., Nachman, M.W., 2005. The extent of linkage disequilibrium caused by selection on G6PD in humans. Genetics 171, 1219–1229. https://doi.org/10.1534/genetics.105.048140

Strandén, I., Garrick, D.J., 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92, 2971–2975. https://doi.org/10.3168/JDS.2008-1929

Sved, J.A., 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. 2, 125–141. https://doi.org/10.1016/0040-5809(71)90011-6

Tenesa, A., Navarro, P., Hayes, B.J., Duffy, D.L., Clarke, G.M., Goddard, M.E., Visscher, P.M., 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17, 520–526. https://doi.org/10.1101/gr.6023607

Trang, T.T., Hung, N.H., Ninh, N.H., Knibb, W., Nguyen, N.H., 2019. Genetic Variation in Disease Resistance Against White Spot Syndrome Virus (WSSV) in Liptopenaeus vannamei. Front. Genet. 10, 264. https://doi.org/10.3389/fgene.2019.00264

Vallejo, R.L., Leeds, T.D., Gao, G., Parsons, J.E., Martin, K.E., Evenhuis, J.P., Fragomeni, B.O., Wiens, G.D., Palti, Y., 2017. Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease

resistance compared to a traditional pedigree-based model in rainbow trout aquaculture. Genet. Sel. Evol. 49, 1–13. https://doi.org/10.1186/s12711-017-0293-6

Vallejo, R.L., Silva, R.M.O., Evenhuis, J.P., Gao, G., Liu, S., Parsons, J.E., Martin, K.E., Wiens, G.D., Lourenco, D.A.L., Leeds, T.D., Palti, Y., 2018. Accurate genomic predictions for BCWD resistance in rainbow trout are achieved using low-density SNP panels: Evidence that long-range LD is a major contributing factor. J. Anim. Breed. Genet. 135, 263–274. https://doi.org/10.1111/jbg.12335

Vela Avitúa, S., Montaldo, H.H., Márquez Valdelamar, L., Campos Montes, G.R., Castillo Juárez, H., 2013. Decline of genetic variability in a captive population of Pacific white shrimp Penaeus (Litopenaeus) vannamei using microsatellite and pedigree information. Electron. J. Biotechnol. 16, 9–9. https://doi.org/10.2225/vol16-issue4-fulltext-11

Wang, Q., Yu, Y., Zhang, Q., Zhang, X., Yuan, J., Huang, H., Xiang, J., Li, F., 2019. A Novel Candidate Gene Associated With Body Weight in the Pacific White Shrimp Litopenaeus vannamei. Front. Genet. 10, 520. https://doi.org/10.3389/fgene.2019.00520

Yáñez, J.M., Newman, S., Houston, R.D., 2015. Genomics in aquaculture to better understand species biology and accelerate genetic progress. Front. Genet. 6, 1–3. https://doi.org/10.3389/fgene.2015.00128

Yáñez, J.M., Yoshida, G., Barria, A., Palma-Véjares, R., Travisany, D., Díaz, D., Cáceres, G., Cádiz, M.I., López, M.E., Lhorente, J.P., Jedlicki, A., Soto, J., Salas, D., Maass, A., 2020. High-Throughput Single Nucleotide Polymorphism (SNP) Discovery and Validation Through Whole-Genome Resequencing in Nile Tilapia (Oreochromis niloticus). Mar. Biotechnol. 22, 109–117. https://doi.org/10.1007/s10126-019-09935-5

Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., 2011. GCTA: A tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yoshida, G.M., Bangera, R., Carvalheiro, R., Correa, K., Figueroa, R., Lhorente, J.P., Yáñez, J.M., 2018. Genomic prediction accuracy for resistance against Piscirickettsia salmonis in farmed rainbow trout. G3 Genes, Genomes, Genet. 8,

719–726. https://doi.org/10.1534/g3.117.300499

Yoshida, G.M., Barria, A., Correa, K., Cáceres, G., Jedlicki, A., Cadiz, M.I., Lhorente, J.P., Yáñez, J.M., 2019a. Genome-Wide Patterns of Population Structure and Linkage Disequilibrium in Farmed Nile Tilapia (Oreochromis niloticus). Front. Genet. 10, 745. https://doi.org/10.3389/fgene.2019.00745

Yoshida, G.M., Carvalheiro, R., Rodríguez, F.H., Lhorente, J.P., Yáñez, J.M., 2019b. Single-step genomic evaluation improves accuracy of breeding value predictions for resistance to infectious pancreatic necrosis virus in rainbow trout. Genomics 111, 127–132. https://doi.org/10.1016/j.ygeno.2018.01.008

Zhang, Xiaojun, Yuan, J., Sun, Y., Li, S., Gao, Y., Yu, Y., Liu, C., Wang, Q., Lv, X., Zhang, Xiaoxi, Ma, K.Y., Wang, X., Lin, W., Wang, Long, Zhu, X., Zhang, C., Zhang, J., Jin, S., Yu, K., Kong, J., Xu, P., Chen, J., Zhang, H., Sorgeloos, P., Sagi, A., Alcivar-Warren, A., Liu, Z., Wang, Lei, Ruan, J., Chu, K.H., Liu, B., Li, F., Xiang, J., 2019. Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. Nat. Commun. 10, 1–14. https://doi.org/10.1038/s41467-018-08197-4

Zhong, X., Li, Q., Kong, L., Yu, H., 2017. Estimates of Linkage Disequilibrium and Effective Population Size in Wild and Selected Populations of the Pacific Oyster Using Single-nucleotide Polymorphism Markers. J. World Aquac. Soc. 48, 791–801. https://doi.org/10.1111/JWAS.12393@10.1111

Chapter 4 - Accuracy of genotype imputation to sequence level using different populations of Nile tilapia

ABSTRACT - A cost-effective strategy to obtain genomic information at sequence level is to sequence part of population and perform genotype imputation from a lower density to sequence for the remaining animals. The aims of this study were to evaluate the feasibility of genotype imputation from medium density to sequence level in Nile tilapia and to investigate the impacts of size and origin of reference population in the accuracy of imputation. DNA samples were extracted from fin-clip samples of 326 animals from 3 different populations (P_A, P_B and P_C). After sequencing, alignment and quality control, approximately 4.6 million of single-nucleotide polymorphisms (SNPs) were retained in common to all populations as the final imputation density. Four scenarios were evaluated for each population: two reference sizes (10 or 90% of animals of each population) and two reference origins (two different populations only or all three populations). The animals in the validation set had part of their genotypes masked keeping only 49,216 SNPs and the accuracy of imputation was assessed using the correlation between the imputed and observed genotypes ($\boldsymbol{R^2}$). Imputation was carried out using FImpute3 software. Overall, the $\boldsymbol{R^2}$ per animal showed intermediate results ranging from 0.37 to 0.56 for P_A, and 0.43 to 0.58 for P_B and P_C. The results showed an increase in the $\boldsymbol{R^2}$ of 31.5% for P_A and 24.6% for P_B and P_C, when 90% animals from same population were used as reference in comparison to only 10%. Using all three populations as reference yielded the best results in terms of number of SNPs imputed with accuracy greater than 0.8, being on average equal to 676,233, 666,559 and 592,187 for P_A, P_B and P_C, respectively. Considering only these top accuracy SNPs (accuracy>0.8), the average imputation accuracy of samples was equal to 0.95 for P_A and 0.92 for P_B and P_C, considering scenarios that included more animals as reference (90% of same population as reference, two and three populations). The $\boldsymbol{R^2}$ for scenarios that used 90% of animals from the same population and used animals from the three population as reference were very similar showing that the strategy of using information from other population to increase the reference population had minor effect on accuracy of imputation. In conclusion, it was feasible to impute from 50k to approximately

680k with high accuracy using tilapia sequence data. We also expect that the use of more animals from these populations or animals from ascending lines as reference could help in the imputation process to obtain millions of imputed SNPs with high accuracy.

Keywords: Nile tilapia, whole-genome sequencing, genotype imputation

4.1 Introduction

Nile tilapia (*Oreochromis niloticus*) is one of the most farmed freshwater fish in the world reaching approximately 4.6 megatons (MT) in 2019 (FAO, 2021). The four main Nile tilapia producer countries in 2019 were: China (1.23 MT), Indonesia (1.18 MT), Egypt (1.08 MT) and Brazil (0.32 MT) altogether representing more than 83% of total production in this year (FAO, 2021). A milestone in the increasing of tilapia production was the establishment of the GIFT (Genetic Improvement of Farmed Tilapias) project in the end of the 1980s (Eknath et al., 1993). The GIFT breeding program was a strategy to spread and popularize the tilapia around the globe that showed excellent accumulated genetic gain for selection to growth performance (85% of increase in comparison to a base population in five years) (Eknath and Hulata, 2009). Due to its success, currently, Nile tilapia is produced in more than 87 countries and it is the aquaculture species that has the highest number of breeding programs registered in developing countries (19) (Neira, 2010; Yáñez et al., 2020a).

As in other aquaculture species, most of Nile tilapia breeding programs are based on pedigree and phenotype information to estimate breeding values and select animals. However, recent advances in molecular biotechnology and quantitative analysis have allowed the incorporation of genomic information in aquaculture breeding programs as well (Yáñez et al., 2015). When genotyping of animals is performed, it is possible to estimate the genomic breeding values and select the candidates accurately using between and within familiar genetic information, the so-called genomic selection (GS) (Meuwissen et al., 2001). Recent studies have shown GS benefits over pedigree-based methods in tilapia, mainly for pathogen resistance (Joshi et al., 2021a, 2021b, 2020; Lu et al., 2020; Sukhavachana et al., 2020) and growth traits (Yoshida et al., 2019b). Another application for genomic information is the mapping of quantitative trait loci (QTL) through genome-wide association studies (GWAS). GWAS identify statistical associations between phenotypes and markers (e.g., single-nucleotide polymorphisms or SNPs) in order to detect signals of genetic variation over the genome (Goddard and Hayes, 2009). GWAS have also been performed in tilapia trying to understand the genetic architecture of interest traits and detect QTLs

associated to them (Cáceres et al., 2019; Sukhavachana et al., 2020; Yoshida et al., 2019b; Yoshida and Yáñez, 2021a; Yu et al., 2021).

The density of SNPs in which the animals are genotyped affects the accuracy of genomic prediction and precision of QTL mapping (Daetwyler et al., 2014; Pérez-Enciso et al., 2015). It is expected that genotypes obtained through whole-genome sequencing (WGS) (millions of SNPs) will offer more accurate results in comparison to SNP panels (thousands of SNPs) (MacLeod et al., 2016). Due to the high costs associated to WGS technology, the genotype imputation is a cost-effective strategy to obtain sequence information for several animals. This strategy uses a small number of sequenced animals as reference to infer and impute nongenotyped markers of target animals that were genotyped using lower density SNP panels. Several factors may influence the accuracy of genotype imputation such as, number of SNPs in the lower density panel, number of animals used as reference, relatedness between reference and target populations, method and software of imputation, marker frequency and precision of genotyping calling during the WGS process (Carvalheiro et al., 2014; Fernandes Júnior et al., 2021; Hayes et al., 2012; Hickey et al., 2012a). Although genotype imputation has been successfully applied for aquaculture species using different SNP panel densities (Dufflocq et al., 2019; Tsai et al., 2017; Tsairidou et al., 2020; Yoshida et al., 2018) and WGS data (Yoshida and Yáñez, 2021a, 2021b), none imputation study was found evaluating population reference size and origin as factors that may influence genotype imputation accuracy in aquaculture species using WGS data.

The objectives of this study were to evaluate the technical feasibility of genotype imputation from approximately 50k SNPs to sequence level in Nile tilapia and to investigate the impacts of size and origin of reference population in the accuracy of imputation.

4.2 Material and methods

4.2.1 Origin of animals

The animals used in this study belong to three different Nile tilapia breeding populations. The first population (P_A) is from Brazil and belong to AquaAmerica corporation, where animals had been evaluated in cages using

daily weight gain as selection criteria for three generations. The P_A founders are GIFT originally imported from Malaysia in 2005. The second population (P_B) is from Costa Rica and belong to Aquacorporación Internacional, where animals had been evaluated in ponds using body weight as selection criterion for seven generations. The founders of P_B are a mixture of different populations from Israel, Singapore, Taiwan and Thailand that gave origin to the GIFT strain. The last population (P_C) is from the same breeding company and country of P_B which had been selected for eight generations. However, the founders of P_C are a mixture of populations from the eighth generation of GIFT and two African strains that originated GIFT. More detailed information about the origin of these populations is available in Yáñez et al. (2020).

4.2.2 Whole-genome sequencing and quality control of genotypes

DNA of 326 animals (57 P_A, 126 P_B and 143 P_C) were obtained from fin-clip samples using a Wizard Genomic DNA purification kit (Promega). The sequencing of samples was performed using Illumina HiSeq 2500 equipment (Illumina, USA). The assembly ASM185804v2 (GenBank accession GCF_001858045.1) (Conte et al., 2017) of *O. niloticus* was used as a reference genome for alignment of the raw sequencing reads for each sample. The GATK v3.5.0 software (McKenna et al., 2010) was implemented to discover the variants and generate the raw SNP file in the variant call format (VCF). A more detailed description of the variant discovery step is fully described in Cáceres et al. (2019) and Yáñez et al. (2020). Due to a new version of Nile tilapia genome (O_niloticus_UMD_NMBU, GCA_001858045.3) (Conte et al., 2019), probes of 200 base pairs were extracted and used to reallocate the SNPs according to the new positions.

An initial quality control of genotypes was applied using the VCFtools software (Danecek et al., 2011) by removing: insertions/deletions, non-biallelic SNPs, genotypes with quality score (phred) < 15, mitochondrial SNPs and SNPs assigned to other than the 23 chromosomes. After that, the filtered VCF file was transformed into binary PLINK format. Then, a second round of QC was applied, but at this time, separated by population. Using the PLINK v1.9 software (Purcell et al., 2007), the following excluding criteria were applied: Hardy-Weinberg equilibrium (HWE) p-value < $1x10^{-8}$, minor allele frequency (MAF) < 0.01 and

call-rate of SNPs < 0.80. A final filter was applied to keep only the SNPs that were in common among the three populations.

4.2.3 Imputation scenarios

We evaluated the impact of two main factors in the accuracy of genotype imputation: size and origin of reference population. The size of reference population was tested within each population exclusively (i.e., reference and target animals were from the same population). Two reference size scenarios were evaluated: 10_{REF} and 90_{REF}, corresponding to 10 or 90% of animals of each population, respectively. For example, $P_A\_10_{REF}$ and $P_A\_90_{REF}$ scenarios represent 6 and 51 animals used as reference, respectively (10 and 90% of 57 P_A animals). The origin of reference population was assessed in two other scenarios: using only animals from other populations as reference or using animals from all populations as reference. For example, $P_A\_P_{BC}$ represents imputation of P_A animals using only P_B and P_C animals as reference (6 P_A animals as target and 126 P_B + 143 P_C animals as reference) while $P_A\_P_{ABC}$ represents imputation of P_A animals using P_A, P_B and P_C animals as reference (6 P_A animals as target and 51 P_A + 126 P_B + 143 P_C animals as reference). The four scenarios were applied for the three populations totalizing 12 imputation scenarios which are summarized in Table 1.

Table 1. Imputation scenarios and number of animals used as reference (Ref) and validation (Val) in each cross-validation group.

Scenario[a]	Ref	Val	Scenario[b]	Ref	Val	Scenario[c]	Ref	Val
$P_A\_10_{REF}$	6	51	$P_B\_10_{REF}$	13	113	$P_C\_10_{REF}$	14	129
$P_A\_90_{REF}$	51	6	$P_B\_90_{REF}$	113	13	$P_C\_90_{REF}$	129	14
$P_A\_P_{BC}$	269	6	$P_B\_P_{AC}$	200	13	$P_C\_P_{AB}$	183	14
$P_A\_P_{ABC}$	320	6	$P_B\_P_{ABC}$	313	13	$P_C\_P_{ABC}$	312	14

[a] Imputation scenarios using Population A (P_A) as target: $P_A\_10_{REF}$ and $P_A\_90_{REF}$ scenarios represent 10 or 90% of P_A animals used as reference, respectively. $P_A\_P_{BC}$ and $P_A\_P_{ABC}$ represent imputation of P_A animals using only P_B and P_C or using P_A, P_B and P_C animals as reference, respectively.

[b] Imputation scenarios using Population B (P_B) as target: $P_B\_10_{REF}$ and $P_B\_90_{REF}$ scenarios represent 10 or 90% of P_B animals used as reference, respectively. $P_B\_P_{AC}$ and $P_B\_P_{ABC}$ represent imputation of P_B animals using only P_A and P_C or using P_A, P_B and P_C animals as reference, respectively.

[c] Imputation scenarios using Population C (P_C) as target: $P_C\_10_{REF}$ and $P_C\_90_{REF}$ scenarios represent 10 or 90% of P_C animals used as reference, respectively. $P_C\_P_{AB}$

and P_C_P_{ABC} represent imputation of P_C animals using only P_A and P_B or using P_A, P_B and P_C animals as reference, respectively.

4.2.4 Genotype imputation and accuracy

The imputation was performed using the FImpute v3 software (Sargolzaei et al., 2014) by diving each target population randomly in 10 groups. The animals in the validation set had their genotypes masked keeping only 49,216 SNPs. These validation SNPs were the same obtained in a previous study to develop a 50K SNP array to Nile tilapia (Yáñez et al., 2020). Animals used as reference had information at sequence level and target animals had genotypes at medium density (50k). Finally, the target animals had genotypes imputed to sequence level using the reference scenarios described previously. To assess the accuracy of imputation, we used the square of the Pearson correlation coefficient between observed and imputed genotypes (R^2). The R may be estimated as (Huang et al., 2009):

$$R = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(g_i - \overline{g})}{\sqrt{\sum_{i'1}^{n}(x_i - \overline{x})^2 \sum_{i'1}^{n}(g_i - \overline{g})^2}}$$

where for each SNP in which the true genotypes were masked, the possible genotypes were coded as 0, 1, or 2, representing the possible number of the minor allele for this SNP in the target population; x_i is the imputed genotype for individual *i*, and $\overline{x}$ denotes the mean value of the imputed genotypes across all individuals. Similarly, g_i and $\overline{g}$ express the correspondent variables for the observed genotypes and *n* is the number of individuals in the validation group. The R^2 was calculated by SNP and averaged by individual. We also evaluated the imputation efficiency according to different MAF levels by plotting R^2 and the allelic imputation error rate ($AIER$) over different MAF ranges. The $AIER$ in percentage was estimated as (Zhang and Druet, 2010):

$$AIER\ (\%) = \frac{1}{2 * N}\sum_{j=1}^{N} |O(n_{ij}^1) - I(n_{ij}^1)| * 100$$

where N is the number of markers, $O(n_{ij}^1)$ and $I(n_{ij}^1)$ are the observed genotypes and imputed number of alleles "1" for individual *i* at marker *j* respectively, and $2 * N$ is the final number of imputed alleles. The $AIER$ was estimated for each SNP imputed. Both R^2 and $AIER$ statistics may be used to measure correctness of imputation. R^2 is a direct measure of accuracy of imputation and $AIER$ is a

proportion of alleles incorrectly imputed. Due to the high dependence of $AIER$ on the population allelic frequencies (Calus et al., 2014), we decided to evaluate the MAF effect on this metric, but we used R^2 as the main comparison criteria for accuracy of imputation.

4.3 Results

4.3.1 WGS and quality control

After WGS and posterior alignment of the 326 samples, 76.3 ± 64.6 million reads were mapped to a single position in the *O. niloticus* genome. The mean coverage per sample was 8.7 ± 8.9*x*, ranging from 2.1 to 65.7*x*. The discovery step yielded a total of 38,454,943 variants that were posteriorly filtered to remove redundant SNPs and re-positioned using the latest genome version resulting in 36,399,577 SNPs. Quality control results of WGS genotypes are available in Table 2.

Table 2. Initial and population-specific genotype quality control of whole-genome sequencing (WGS) tilapia data.

Initial quality control[a]	Before	After	
All populations	36,399,577	24,664,286	
Population-specific quality control	HWE[b]	MAF[c]	Call-Rate[d]
P_A	24,651,013	13,862,830	8,942,268
P_B	24,077,033	15,332,113	10,475,788
P_C	23,766,617	14,924,264	10,048,571
Common SNPs[e]	4,631,757		

[a] Initial quality control applied to all populations by excluding: insertions/deletions, non-biallelic SNPs, genotypes with quality score (phred) < 15, mitochondrial SNPs and SNPs assigned to other than the 23 chromosomes
[b] Number of SNPs in Hardy–Weinberg Equilibrium (HWE) ($p > 10^{-8}$)
[c] Number of SNPs included by minor allele frequency (MAF) threshold (MAF > 0.01)
[d] Number of SNPs included after the exclusion criterion call-rate > 0.8
[e] Final number of SNPs that are in common among the three populations

After the initial quality control on raw WGS data (exclusion of insertions/deletions, non-biallelic SNPs, genotypes phred < 15, SNPs not assigned to one of the 23 chromosomes), 11,735,291 SNPs (32.2%) were removed. For the population-specific quality control, the filter that proportionally most excluded SNPs was MAF (29.6, 24.0 and 24.3%) followed by call-rate (13.5,

13.3 and 13.4%) and HWE (0.04, 1.6 and 2.5%) for P_A, P_B and P_C, respectively. The final quality control yielded 4,631,757 SNPs that were in common among all populations and was used as the final imputation density.

4.3.2 Accuracy of genotype imputation

Overall, genotype imputation per animal presented low to moderate accuracy (Table 3).

Table 3. Summary statistics of correlation between imputed and observed genotypes (R^2) for individuals, number and mean R^2 of SNPs with accuracy greater than 0.8 in each scenario.

Scenario[a]	Mean ± SD[b]	Max[c]	Min[d]	SNPs>0.8 ± SD[e]	Mean>0.8 ± SD[f]
P_A_10_{REF}	0.37 ± 0.04	0.52	0.25	24,791 ± 7,582	0.86 ± 0.06
P_A_90_{REF}	0.54 ± 0.07	0.71	0.35	585,283 ± 112,914	0.95 ± 0.06
P_A_P_{BC}	0.47 ± 0.06	0.64	0.36	546,775 ± 119,290	0.95 ± 0.06
P_A_P_{ABC}	0.56 ± 0.07	0.78	0.39	676,233 ± 142,291	0.95 ± 0.06
P_B_10_{REF}	0.43 ± 0.05	0.64	0.23	47,857 ± 4,190	0.85 ± 0.04
P_B_90_{REF}	0.57 ± 0.07	0.72	0.29	595,759 ± 48,323	0.92 ± 0.07
P_B_P_{AC}	0.45 ± 0.06	0.59	0.22	448,678 ± 34,039	0.92 ± 0.07
P_B_P_{ABC}	0.58 ± 0.08	0.76	0.27	666,559 ± 52,648	0.92 ± 0.07
P_C_10_{REF}	0.43 ± 0.05	0.63	0.30	30,650 ± 2,267	0.85 ± 0.05
P_C_90_{REF}	0.57 ± 0.07	0.76	0.38	543,525 ± 73,811	0.92 ± 0.07
P_C_P_{AB}	0.45 ± 0.06	0.60	0.28	386,706 ± 57,739	0.91 ± 0.07
P_C_P_{ABC}	0.58 ± 0.07	0.80	0.37	592,187 ± 89,663	0.92 ± 0.07

[a] Scenarios of imputation: first acronyms are the target populations (before "_"), 10_{REF} and 90_{REF} represent 10 or 90% of same population used as reference and second acronyms are the reference populations (after "_")
[b] Mean ± standard deviation of correlation between imputed and observed genotypes (R^2) for individuals
[c] Maximum correlation between imputed and observed genotypes (R^2) for individuals
[d] Minimum correlation between imputed and observed genotypes (R^2) for individuals
[e] Mean ± standard deviation of number of SNPs with correlation between imputed and observed genotypes (R^2) greater than 0.8
[f] Mean ± standard deviation of correlation between imputed and observed genotypes (R^2) only using SNPs with R^2 greater than 0.8

The scenario with the lowest mean individual correlation between imputed and observed genotypes (R^2) was P_A_10_{REF} (0.37 ± 0.04) and the highest were P_B_P_{ABC} and P_C_P_{ABC} (0.58 ± 0.08 and 0.58 ± 0.07, respectively). The minimum and maximum individual accuracies were achieved in the P_B_P_{AC} and P_C_P_{ABC} scenarios, 0.22 and 0.80, respectively. In spite of moderate accuracy values at individual level, some scenarios showed a high proportion of genotypes imputed with accuracy greater than 0.8 (SNPs>0.8).

For instance, considering scenarios using 90% of animals from same population as reference (90_{REF}) and using two or three populations as reference (P_{BC}, P_{AC}, P_{AB} and P_{ABC}), the number of SNPs with accuracy greater than 0.8 ranged between 386,706 ± 57,739 (P_C_P_{AB}) and 676,233 ± 142,291 (P_A_P_{ABC}) (Figure 1).

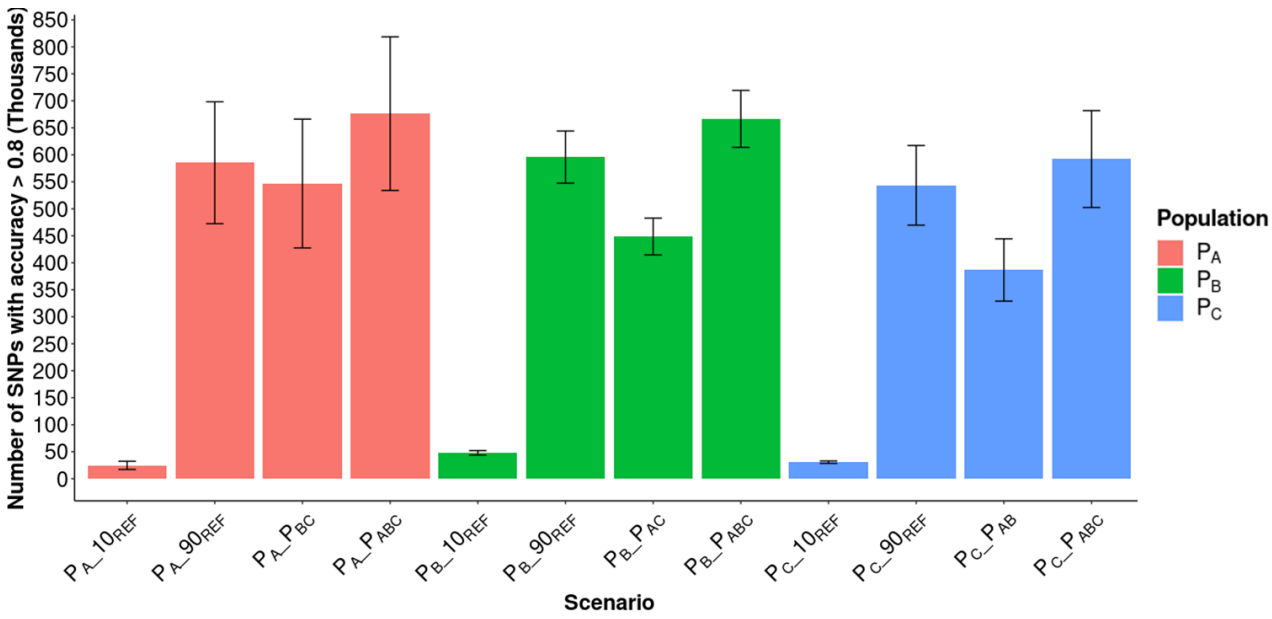


Figure 1. Mean and standard deviation of number of SNPs with correlation between imputed and observed genotypes (R^2) greater than 0.8 for all cross-validation groups.

As a result, the average accuracy of imputation per individual increased, for example, from 0.45 ± 0.06 to 0.91 ± 0.07 and from 0.56 ± 0.07 to 0.95 ± 0.06 for P_C_P_{AB} and P_A_P_{ABC}, respectively, when using only the SNPs>0.8 (Table 3).

4.3.2.1 Effect of reference population size and origin

There was a clear difference in the accuracy of imputation between scenarios that used 10 and 90% of same population as reference. The increase in the R^2 was equal to 0.17 for P_A and 0.14 for P_B and P_C when 90% of animals were used as reference compared with 10%. The same trend was observed in the number of SNPs imputed with accuracy greater than 0.8 in which 24,791, 47,857 and 30,650 SNPs>0.8 were obtained in 10_{REF} scenarios, and 585,283, 595,759 and 543,525 in 90_{REF}, for P_A, P_B and P_C, respectively. Regarding the reference population origin, using all three populations as reference (P_{ABC})

resulted in greater accuracy of imputation in comparison with two population scenarios (P_{BC}, P_{AC}, P_{AB}). This difference was equal to 0.09 for P_A and 0.13 for P_B and P_C when animals from all populations were used as reference. A positive effect of using two different populations as reference was observed in the number of SNPs imputed with accuracy above 0.8. Assuming that only two different populations had WGS data, it was possible to obtain 546,775, 448,678 and 386,706 SNPs imputed with accuracy greater than 0.8 for P_A, P_B and P_C, respectively (Figure 1).

4.3.2.2 Effect of MAF

The average individual accuracy of imputation (R^2) and allelic imputation error rate ($AIER$) were calculated for different MAF classes (0-0.049; 0.050-0.099; 0.100-0.149; 0.150-0.199; 0.199-0.249; 0.250-0.299; 0.300-0.349; 0.350-0.399; 0.400-0.450) in Figure 2 and Figure 3, respectively. It is possible to observe that, for most of scenarios, accuracy of imputation was greater for higher MAF levels and this effect was more pronounced for scenarios with a reduced number of reference animals (10_{REF}). In contrast, imputation of SNPs with higher MAF levels tended to increase the $AIER$ and, such effect is softened by increasing the reference size (90_{REF}) and using all populations (P_{ABC}) as reference.

4.3.2.3 Effect of sequencing coverage

An implicit factor that may affect the accuracy of imputation is the sequencing coverage. Due to the relatively low average coverage (8.7*x*) obtained in this study, we computed the accuracy of imputation according to each individual coverage (Figure 4). These results showed a strong and positive relationship between the average sequencing coverage and accuracy of imputation mainly in the first 10*x* coverage spectrum. This effect was more pronounced in the 90_{REF} and _P_{ABC} scenarios and, after 10*x*, the accuracy of imputation seemed to stabilize.

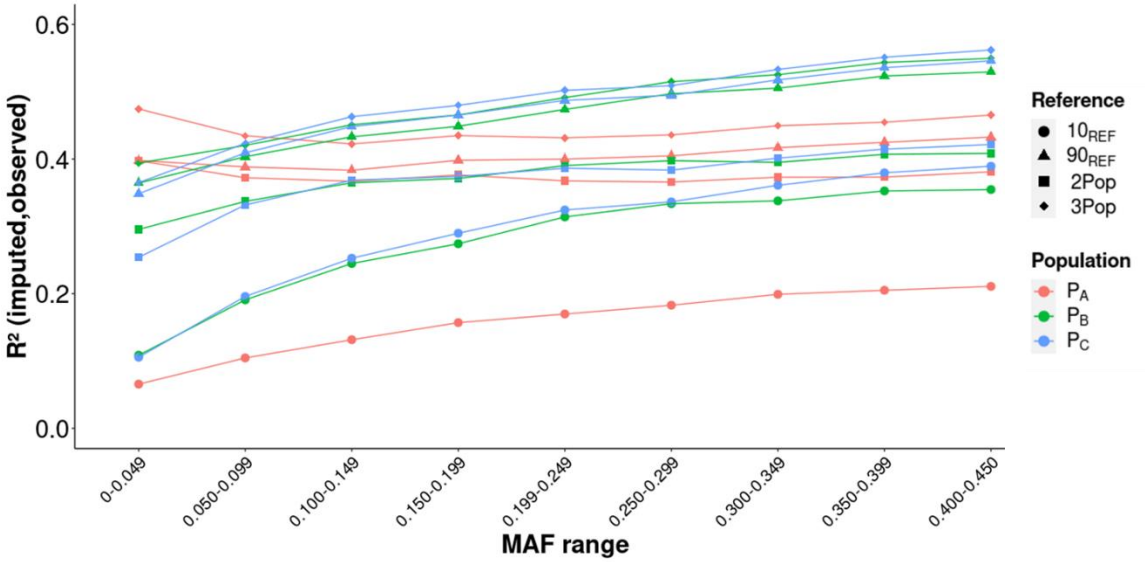


Figure 2. Correlation between imputed and observed genotypes (R^2) by minor allele frequency (MAF) range. (10_{REF}: 10% of same population individuals as reference; 90_{REF}: 90% of same population individuals as reference; 2Pop: two other populations as reference only; 3Pop: all three populations as reference).

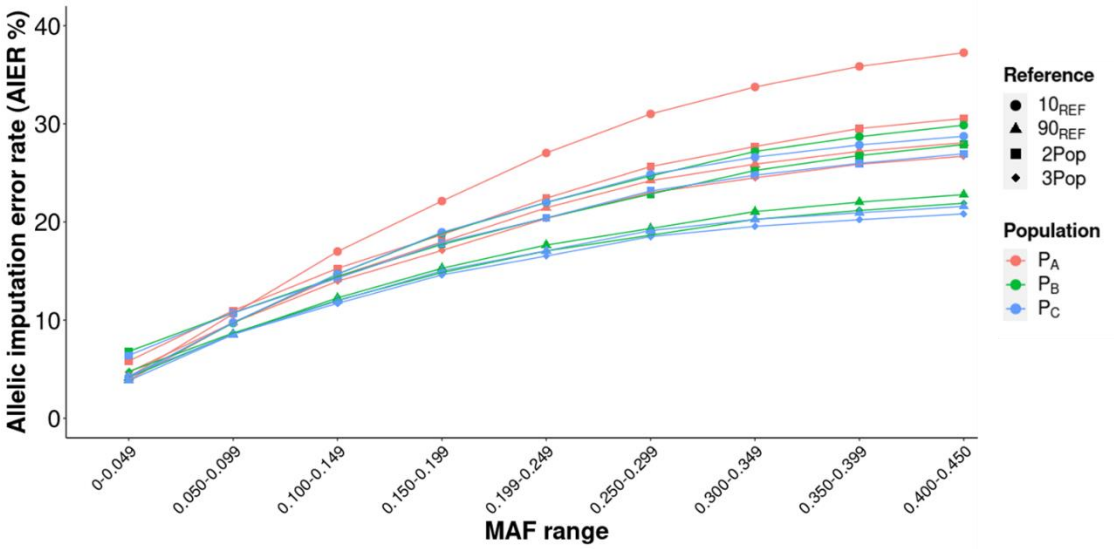


Figure 3. Allelic imputation error rate ($AIER$) by minor allele frequency (MAF) range. (10_{REF}: 10% of same population individuals as reference; 90_{REF}: 90% of same population individuals as reference; 2Pop: two other populations as reference only; 3Pop: all three populations as reference).

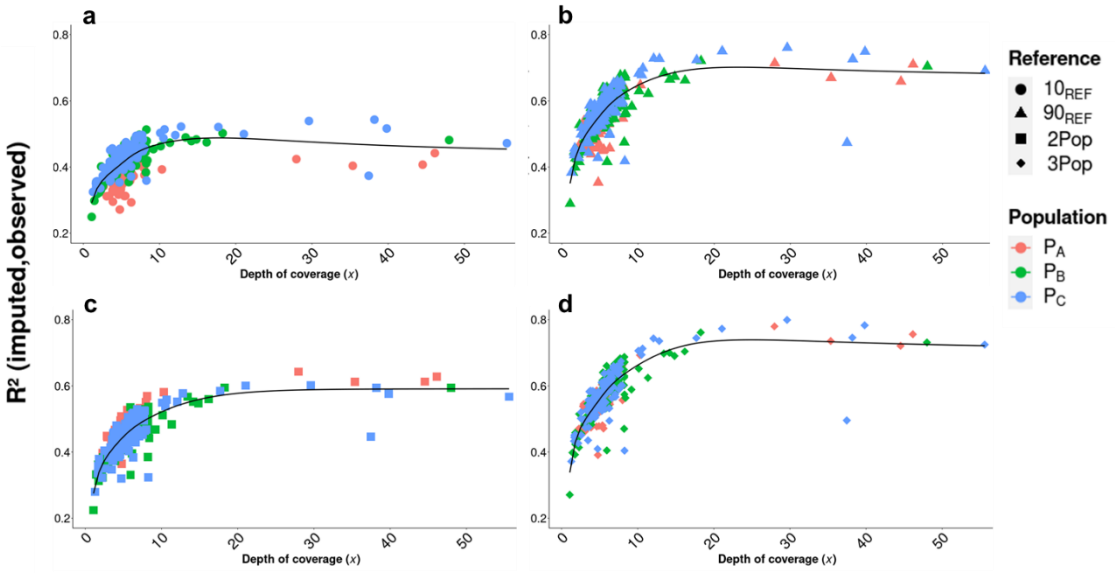


Figure 4. Correlation between imputed and observed genotypes (R^2) as a function of sequencing coverage depth for each individual. **a**: 10% of same population individuals as reference; **b**: 90% of same population individuals as reference; **c**: two other populations as reference only; **d**: all three populations as reference. (The solid line represents a generalized additive model (GAM) fitted by a sum of smooth functions of the covariates plus a conventional parametric component of the linear predictor, see Wood (2006) for more details).

4.4 Discussion

The permanent reduction of costs to obtain genotypes in livestock and aquaculture species have increased the interest in the use of WGS data. In addition to the low-cost projection in a near future, WGS data offers more accurate results in comparison to high density SNP arrays for genomic prediction and GWAS mainly when the causal mutations are expected to present low MAF (Druet et al., 2013). Currently, genotype imputation has a key role to obtain WGS data with reduced costs on a large scale. In this study, we evaluated two important factors that may affect accuracy of imputation: size and origin of reference population. To the best of our knowledge, both effects have not been investigated in genotype imputation to WGS in aquaculture species.

We found low to moderate mean individual accuracy of imputation (0.37 to 0.58) for the evaluated scenarios. These estimates are lower than the accuracy of imputation to the sequence level reported for other species, for example, 0.78 to 0.95 in dairy cattle, (Binsbergen et al., 2014; Pausch et al., 2017), 0.94 to 0.95 in beef cattle (Fernandes Júnior et al., 2021) and 0.87 to 0.97 for sheep

(Bolormaa et al., 2019), considering different reference sizes and the individual R^2 as accuracy of imputation. The main challenge in the imputation process from medium density to sequence level is the high number of genotypes to be imputed. Some studies showed that larger differences between the density of SNPs from the reference and imputation panels may decrease the accuracy of imputation (Carvalheiro et al., 2014; Hickey et al., 2012a). This was a potential problem in our study because we aimed to impute from approximately 50k to 4.6 million of SNPs (approximately 92 times more). The imputation studies using WGS mentioned above used higher density SNP panels (500-777k) or stepwise imputation (i.e., sequential imputation from medium density to high density before imputation to sequence level), meaning lower distance between reference and target densities. In order to reduce the distance between densities, we evaluated the use of a high-density SNP panel for tilapia. However, at this moment, the development of a more expensive genotyping chip would be an unrealistic scenario for this species and we opted to use only the 50k platform as validation.

There was an increase in the accuracy of imputation of 31.5% for P_A and 24.6% for P_B and P_C, when more animals from the same population were used as reference (i.e., 10_{REF} vs 90_{REF}). In livestock, the success of genotype imputation is highly affected by the relationship between reference and target animals. If the target animals are immediate ancestors of the reference animals, for example, a low number of reference animals is enough to obtain high accuracy of imputation (Hayes et al., 2012). On the other hand, more distant animals would require a higher number of animals as reference to impute with the same accuracy. This happens because most of imputation software (including FImpute3) seek haplotypes shared between reference and target animals to predict the missing genotypes. The haplotype construction is more accurate for closer relatives because they share longer haplotype blocks requiring less animals (Sargolzaei et al., 2014). In the present study, there is weak genetic ties between animals of each population. The mean genomic relationship within population was 0.13 ± 0.12, 0.06 ± 0.09 and 0.05 ± 0.09 for P_A, P_B and P_C, respectively. These values were calculated through a genomic relationship matrix (VanRaden, 2008) using the WGS data (Appendix B). Thus, the shift of 10 to 90% of animals used as reference had a major impact in the accuracy of imputation.

A strategy to increase the number of animals used as reference would be the addition of WGS data available from other populations. For P_A, the accuracy increased from 0.37 in the lowest reference size scenario to 0.47 using only the other two populations as reference (i.e., $P_A\_10_{REF}$ vs $P_A\_P_{BC}$), while for P_B and P_C, this gain in accuracy was marginal (from 0.43 to 0.45 for both populations) comparing the analogous scenarios. Most likely, this benefit was higher for the P_A because only six animals were used as reference in the 10_{REF} scenario while for PB and PC this number was equal to 13 and 14, respectively.

Despite a previous study determined that these are three highly diverse populations (Yoshida et al., 2019a), we expected that the common GIFT origin could help in the imputation. However, the recent admixture and selection applied to these populations in different geographical locations may have difficulted the construction and traceability of haplotypes. Bouwman and Veerkamp (2014) also obtained moderate imputation accuracy imputing to sequence level and using a multi-breed cattle population as reference. These authors also observed that the persistency of linkage disequilibrium phase between markers in the low-density panel may limit the advantage of using different populations as reference. Similarly, this may also have affected the performance of imputation in scenarios that included animals from different populations in this study. Nevertheless, the final number of SNPs imputed with accuracy higher than 0.8 (SNPs>0.8) was much higher when two different populations were used as reference (21 times greater for P_A, 8.4 for P_B and 11.6 for P_C) in comparison to 10_{REF} scenarios. These results reinforce the idea of Bouwman and Veerkamp (2014) that a “reference bank” with WGS data from different populations or breeds may be an advantageous strategy for imputation, mainly when a low number of sequenced animals is available.

The best results of accuracy of imputation were obtained when animals from all populations were used as reference. The benefit of using P_{ABC} as reference was more evident when compared to 10_{REF} scenarios (51.3% higher accuracy in comparison to $P_A\_10_{REF}$ and 34.9% higher in comparison to both $P_B\_10_{REF}$ and $P_C\_10_{REF}$) and scenarios with two populations used as reference (19.1% higher accuracy in comparison to P_{BC} and 28.9% higher in comparison to both P_{AC} and P_{AB}). On the other hand, these differences were marginals in

comparison to 90_{REF} scenarios (3.7% for P_A and 1.7% for both P_B and P_C). The same tendency was observed for the SNPs>0.8. For example, a higher mean proportion of SNPs>0.8 was observed for P_{ABC} in comparison to scenarios that used only other two populations as reference (23.7, 48.6 and 53.1% more SNPs>0.8 for P_{ABC} in comparison to P_{BC}, P_{AC} and P_{AB}, respectively). On average, there was also more SNPs>0.8 for P_{ABC} in comparison to 90_{REF} scenarios, however these differences were not significant due to high standard deviations. These results showed that the use of more related animals as reference was more important than adding more animals from other populations to the imputation. Regardless the differences between scenarios and populations, the average accuracy of imputation for individuals using only SNPs>0.8 could be increased from 0.37-0.58 to 0.86-0.95 maximizing the number of known genotypes from 50k to approximately 680k.

An advantage of imputing to sequence level is the possibility of inclusion of rare variants in the subsequent analysis. We observed that for most of scenarios the accuracy improved as the MAF increased. This result was expected because SNPs with lower MAF in the target population have few haplotypes, containing these rarer SNPs, in common with the reference animals hardening the imputation (Hickey et al., 2012a). This effect is intensified when the reference size is lower which also explains the poor accuracy of imputation in 10_{REF} scenarios mainly for P_A. The accuracy of imputation by MAF class followed the results of individual accuracy of imputation: the inclusion of animals from the other 2 populations increased the accuracy but the P_{ABC} and 90_{REF} presented the best results with no significant difference between them. Interestingly, the accuracy of imputation was higher in $P_A\_90_{REF}$, $P_A\_P_{BC}$ and $P_A\_P_{ABC}$ for SNPs imputed with low MAF (0-0.049). A possible explanation for this result is the fact that the low number of validation animals (6) in P_A are overrepresenting the SNPs in common with lower MAF in the others populations assisting the haplotype construction. It is worthy to mention that FImpute3 allows the inclusion of pedigree information which could improve substantially the accuracy of imputation of rare variants (Fernandes Júnior et al., 2021; Ma et al., 2013). Unfortunately, this information was not available in our study.

An alternative metric used to evaluate the correctness of accuracy of imputation is the allelic imputation error rate ($AIER$). A similar trend was found in the $AIER$ curves regarding the size and origin of reference population. There was a reduction in $AIER$ in scenarios that included more animals as reference specially when these reference animals had individuals from the same target population. However, it is possible to observe an opposite relationship between MAF and the $AIER$ in comparison to the accuracy of imputation (R^2), i.e., imputation of SNPs with higher MAF increases the $AIER$. This antagonistic effect caused by different MAF levels between both metrics may be explained by how they are calculated. For lower MAF SNPs, the missing genotypes are most likely homozygous for the more frequent allele resulting in very low $AIER$ (Calus et al., 2014). On the other hand, the R^2 is only a Pearson's correlation between imputed and observed genotypes not being influenced by this effect. For this reason, the accuracy of imputation measured as correlation (R^2) is the most common metric applied in imputation studies.

Another aspect that possibly affected the accuracy of imputation in this study was the sequencing depth or coverage. The coverage is related to the average number of times that a locus is sequenced which influences the precision of "genotype calling" process in WGS. Our results showed that most of individuals with mean coverage greater than 10*x* had the highest accuracy of imputation levels. The mean individual coverage considering all animals was 8.7*x* suggesting that a deeper sequencing could improve the performance of imputation. The sequencing coverage is also linked to costs in WGS because the higher the average coverage, the higher the sequencing costs. In order to reduce these costs, VanRaden et al. (2015) reported a trade-off between coverage and number of sequenced animals for accuracy of imputation, i.e., it is possible to decrease WGS costs without loss of accuracy of imputation by sequencing several animals with low coverage. With the tendency of reduction for WGS costs, the strategy of low-coverage sequencing of several animals to perform imputation is a promising alternative for genomic studies (Rubinacci et al., 2021).

Two additional factors may also affect imputation studies: the position of the imputed SNP along the chromosome and the method or software of imputation. An unequal accuracy of imputation along chromosomes was reported

in different species (Bolormaa et al., 2019; Liu et al., 2015; Pausch et al., 2017). Several reasons may cause this effect, e.g., level of heterozygosity, density of SNPs in determined regions and segmental duplications (Daetwyler et al., 2014; Fernandes Júnior et al., 2021). For tilapia, specifically, we think that it may be associated to assembling errors. Although we used the latest version of genome assembling, we noticed in some chromosomes, regions with systematic low accuracy of imputation (Appendix C). It is expected that future improvements on tilapia genome assembling may also increase accuracy of imputation in these regions. Several software have been applied for imputing genotypes in livestock and aquaculture species, for instance, AlphaImpute (Hickey et al., 2012b), Beagle (Browning and Browning, 2008), Minimac (Howie et al., 2012) and FImpute (Sargolzaei et al., 2014). In addition to FImpute, we tested Minimac as the imputation software in some scenarios. However, we opted for FImpute because it was more time-efficient and no significant differences in accuracy were observed in the scenarios evaluated (results not shown).

4.5 Conclusion

It was possible to impute from 50k to approximately 680k SNPs using tilapia WGS data with expected accuracy between 0.92 and 0.95, considering only SNPs imputed with accuracy greater than 0.8. The use of more animals from same population as reference had higher impact in the accuracy of imputation than animals from different populations. However, a higher proportion of SNPs imputed with high accuracy was obtained when animals from all populations were used as reference.

4.6 References

Binsbergen, R. van, Bink, M.C., Calus, M.P., Eeuwijk, F.A. van, Hayes, B.J., Hulsegge, I., Veerkamp, R.F., 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 2014 461 46, 1–13. https://doi.org/10.1186/1297-9686-46-41

Bolormaa, S., Chamberlain, A.J., Khansefid, M., Stothard, P., Swan, A.A., Mason, B., Prowse-Wilkins, C.P., Duijvesteijn, N., Moghaddar, N., Werf, J.H. van der, Daetwyler, H.D., MacLeod, I.M., 2019. Accuracy of imputation to whole-genome sequence in sheep. Genet. Sel. Evol. 2019 511 51, 1–17. https://doi.org/10.1186/S12711-018-0443-5

Bouwman, A.C., Veerkamp, R.F., 2014. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genet. 2014 151 15, 1–9. https://doi.org/10.1186/S12863-014-0105-8

Browning, B.L., Browning, S.R., 2008. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84, 210–223. https://doi.org/10.1016/j.ajhg.2009.01.005

Cáceres, G., López, M.E., Cádiz, M.I., Yoshida, G.M., Jedlicki, A., Palma-Véjares, R., Travisany, D., Díaz-Domínguez, D., Maass, A., Lhorente, J.P., Soto, J., Salas, D., Yáñez, J.M., 2019. Fine Mapping Using Whole-Genome Sequencing Confirms Anti-Müllerian Hormone as a Major Gene for Sex Determination in Farmed Nile Tilapia (Oreochromis niloticus L.). G3 Genes, Genomes, Genet. 9, 3213–3223. https://doi.org/10.1534/g3.119.400297

Calus, M.P.L., Bouwman, A.C., Hickey, J.M., Veerkamp, R.F., Mulder, H.A., 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. animal 8, 1743–1753. https://doi.org/10.1017/S1751731114001803

Carvalheiro, R., Boison, S.A., Neves, H.H.R., Sargolzaei, M., Schenkel, F.S., Utsunomiya, Y.T., O'Brien, A.M.P., Sölkner, J., McEwan, J.C., Van Tassell, C.P., Sonstegard, T.S., Garcia, J.F., 2014. Accuracy of genotype imputation in Nelore cattle. Genet. Sel. Evol. 46, 1–11. https://doi.org/10.1186/s12711-014-0069-1

Conte, M.A., Gammerdinger, W.J., Bartie, K.L., Penman, D.J., Kocher, T.D.,

2017. A high quality assembly of the Nile Tilapia (Oreochromis niloticus) genome reveals the structure of two sex determination regions. BMC Genomics 18, 1–19. https://doi.org/10.1186/s12864-017-3723-5

Conte, M.A., Joshi, R., Moore, E.C., Nandamuri, S.P., Gammerdinger, W.J., Roberts, R.B., Carleton, K.L., Lien, S., Kocher, T.D., 2019. Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes. Gigascience 8, 1–20. https://doi.org/10.1093/gigascience/giz030

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brøndum, R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C., Esquerré, D., Bouchez, O., Rossignol, M.N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P.J., Coote, D., Chamberlain, A.J., Anderson, C., Vantassell, C.P., Hulsegge, I., Goddard, M.E., Guldbrandtsen, B., Lund, M.S., Veerkamp, R.F., Boichard, D.A., Fries, R., Hayes, B.J., 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46, 858–865. https://doi.org/10.1038/ng.3034

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., Group, 1000 Genomes Project Analysis, 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. https://doi.org/10.1093/BIOINFORMATICS/BTR330

Druet, T., Macleod, I.M., Hayes, B.J., 2013. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Hered. 2014 1121 112, 39–47. https://doi.org/10.1038/hdy.2013.13

Dufflocq, P., Pérez-Enciso, M., Lhorente, J.P., Yáñez, J.M., 2019. Accuracy of genomic predictions using different imputation error rates in aquaculture breeding programs: A simulation study. Aquaculture 503, 225–230. https://doi.org/10.1016/j.aquaculture.2018.12.061

Eknath, A.E., Hulata, G., 2009. Use and exchange of genetic resources of Nile tilapia (Oreochromis niloticus). Rev. Aquac. 1, 197–213. https://doi.org/10.1111/j.1753-5131.2009.01017.x

Eknath, A.E., Tayamen, M.M., Palada-de Vera, M.S., Danting, J.C., Reyes, R.A., Dionisio, E.E., Capili, J.B., Bolivar, H.L., Abella, T.A., Circa, A. V., Bentsen, H.B.,

Gjerde, B., Gjedrem, T., Pullin, R.S.V., 1993. Genetic improvement of farmed tilapias: the growth performance of eight strains of Oreochromis niloticus tested in different farm environments. Aquaculture 111, 171–188. https://doi.org/10.1016/0044-8486(93)90035-W

FAO, 2021. FAO Fisheries & Aquaculture - Statistics [WWW Document]. URL http://www.fao.org/fishery/statistics/en (accessed 6.30.21).

Fernandes Júnior, G.A., Carvalheiro, R., Oliveira, H.N., Sargolzaei, M., Costilla, R., Ventura, R. V., Fonseca, L.F.S., Neves, H.H.R., Hayes, B.J., de Albuquerque, L.G., 2021. Imputation accuracy to whole-genome sequence in Nellore cattle. Genet. Sel. Evol. 53, 27. https://doi.org/10.1186/s12711-021-00622-5

Goddard, M.E., Hayes, B.J., 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10, 381–391. https://doi.org/10.1038/nrg2575

Hayes, B.J., Bowman, P.J., Daetwyler, H.D., Kijas, J.W., Van Der Werf, J.H.J., 2012. Accuracy of genotype imputation in sheep breeds. Anim. Genet. 43, 72–80. https://doi.org/10.1111/j.1365-2052.2011.02208.x

Hickey, J.M., Crossa, J., Babu, R., de los Campos, G., 2012a. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52, 654–663. https://doi.org/10.2135/cropsci2011.07.0358

Hickey, J.M., Kinghorn, B.P., Tier, B., van der Werf, J.H., Cleveland, M.A., 2012b. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. 2012 441 44, 1–11. https://doi.org/10.1186/1297-9686-44-9

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., Abecasis, G.R., 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 2012 448 44, 955–959. https://doi.org/10.1038/ng.2354

Huang, L., Li, Y., Singleton, A.B., Hardy, J.A., Abecasis, G., Rosenberg, N.A., Scheet, P., 2009. Genotype-Imputation Accuracy across Worldwide Human Populations. Am. J. Hum. Genet. 84, 235–250. https://doi.org/10.1016/J.AJHG.2009.01.013

Joshi, R., Almeida, D.B., da Costa, A.R., Skaarud, A., de Pádua Pereira, U., Knutsen, T.M., Moen, T., Alvarez, A.T., 2021a. Genomic selection for resistance

to Francisellosis in commercial Nile tilapia population: Genetic and genomic parameters, correlation with growth rate and predictive ability. Aquaculture 537, 736515. https://doi.org/10.1016/j.aquaculture.2021.736515

Joshi, R., Skaarud, A., Alvarez, A.T., Moen, T., Ødegård, J., 2021b. Bayesian genomic models boost prediction accuracy for survival to Streptococcus agalactiae infection in Nile tilapia (Oreochromus nilioticus). Genet. Sel. Evol. 53, 37. https://doi.org/10.1186/s12711-021-00629-y

Joshi, R., Skaarud, A., de Vera, M., Alvarez, A.T., Ødegård, J., 2020. Genomic prediction for commercial traits using univariate and multivariate approaches in Nile tilapia (Oreochromis niloticus). Aquaculture 516, 734641. https://doi.org/10.1016/j.aquaculture.2019.734641

Liu, Q., Cirulli, E.T., Han, Y., Yao, S., Liu, S., Zhu, Q., 2015. Systematic assessment of imputation performance using the 1000 Genomes reference panels. Brief. Bioinform. 16, 549–562. https://doi.org/10.1093/BIB/BBU035

Lu, S., Zhu, J., Du, X., Sun, S., Meng, L., Liu, S., Fan, G., Wang, J., Chen, S., 2020. Genomic selection for resistance to Streptococcus agalactiae in GIFT strain of Oreochromis niloticus by GBLUP, wGBLUP, and BayesCπ. Aquaculture 523, 735212. https://doi.org/10.1016/j.aquaculture.2020.735212

Ma, P., Brøndum, R.F., Zhang, Q., Lund, M.S., Su, G., 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. J. Dairy Sci. 96, 4666–4677. https://doi.org/10.3168/JDS.2012-6316

MacLeod, I.M., Bowman, P.J., Vander Jagt, C.J., Haile-Mariam, M., Kemper, K.E., Chamberlain, A.J., Schrooten, C., Hayes, B.J., Goddard, M.E., 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17, 144. https://doi.org/10.1186/s12864-016-2443-6

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. https://doi.org/10.1101/gr.107524.110

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total

Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819–1829.

Neira, R., 2010. Breeding in Aquaculture Species: Genetic Improvement Programs in Developing Countries, in: 9th World Congress on Genetics Applied to Livestock Production. Leipzig, Germany, p. 8.

Pausch, H., MacLeod, I.M., Fries, R., Emmerling, R., Bowman, P.J., Daetwyler, H.D., Goddard, M.E., 2017. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genet. Sel. Evol. 2017 491 49, 1–14. https://doi.org/10.1186/S12711-017-0301-X

Pérez-Enciso, M., Rincón, J.C., Legarra, A., 2015. Sequence- vs. chip-assisted genomic selection: Accurate biological information is advised. Genet. Sel. Evol. 47, 1–14. https://doi.org/10.1186/s12711-015-0117-5

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

Rubinacci, S., Ribeiro, D.M., Hofmeister, R.J., Delaneau, O., 2021. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat. Genet. 2021 531 53, 120–126. https://doi.org/10.1038/s41588-020-00756-0

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

Sukhavachana, S., Tongyoo, P., Massault, C., McMillan, N., Leungnaruemitchai, A., Poompuang, S., 2020. Genome-wide association study and genomic prediction for resistance against Streptococcus agalactiae in hybrid red tilapia (Oreochromis spp.). Aquaculture 525, 735297. https://doi.org/10.1016/j.aquaculture.2020.735297

Tsai, H.Y., Matika, O., Edwards, S.M.K., Antolín-Sánchez, R., Hamilton, A., Guy, D.R., Tinch, A.E., Gharbi, K., Stear, M.J., Taggart, J.B., Bron, J.E., Hickey, J.M., Houston, R.D., 2017. Genotype imputation to improve the cost-efficiency of genomic selection in farmed Atlantic salmon. G3 Genes, Genomes, Genet. 7, 1377–1383. https://doi.org/10.1534/g3.117.040717

Tsairidou, S., Hamilton, A., Robledo, D., Bron, J.E., Houston, R.D., 2020. Optimizing low-cost genotyping and imputation strategies for genomic selection in atlantic salmon. G3 Genes, Genomes, Genet. 10, 581–590. https://doi.org/10.1534/g3.119.400800

VanRaden, P.M., 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91, 4414–4423. https://doi.org/10.3168/JDS.2007-0980

VanRaden, P.M., Sun, C., O'Connell, J.R., 2015. Fast imputation using medium or low-coverage sequence data. BMC Genet. 2015 161 16, 1–12. https://doi.org/10.1186/S12863-015-0243-7

Wood, S.N., 2006. On confidence intervals for generalized additive models based on penalized regression splines. Aust. N. Z. J. Stat. 48, 445–464. https://doi.org/10.1111/J.1467-842X.2006.00450.X

Yáñez, J.M., Joshi, R., Yoshida, G.M., 2020a. Genomics to accelerate genetic improvement in tilapia. Anim. Genet. 51, 658–674. https://doi.org/10.1111/AGE.12989

Yáñez, J.M., Newman, S., Houston, R.D., 2015. Genomics in aquaculture to better understand species biology and accelerate genetic progress. Front. Genet. 6, 1–3. https://doi.org/10.3389/fgene.2015.00128

Yáñez, J.M., Yoshida, G., Barria, A., Palma-Véjares, R., Travisany, D., Díaz, D., Cáceres, G., Cádiz, M.I., López, M.E., Lhorente, J.P., Jedlicki, A., Soto, J., Salas, D., Maass, A., 2020b. High-Throughput Single Nucleotide Polymorphism (SNP) Discovery and Validation Through Whole-Genome Resequencing in Nile Tilapia (Oreochromis niloticus). Mar. Biotechnol. 22, 109–117. https://doi.org/10.1007/s10126-019-09935-5

Yoshida, G.M., Barria, A., Correa, K., Cáceres, G., Jedlicki, A., Cadiz, M.I., Lhorente, J.P., Yáñez, J.M., 2019a. Genome-Wide Patterns of Population Structure and Linkage Disequilibrium in Farmed Nile Tilapia (Oreochromis niloticus). Front. Genet. 10, 745. https://doi.org/10.3389/fgene.2019.00745

Yoshida, G.M., Carvalheiro, R., Lhorente, J.P., Correa, K., Figueroa, R., Houston, R.D., Yáñez, J.M., 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, 147–154. https://doi.org/10.1016/j.aquaculture.2018.03.004

Yoshida, G.M., Lhorente, J.P., Correa, K., Soto, J., Salas, D., Yáñez, J.M., 2019b. Genome-wide association study and cost-efficient genomic predictions for growth and fillet yield in Nile tilapia (Oreochromis niloticus). G3 Genes, Genomes, Genet. 9, 2597–2607. https://doi.org/10.1534/g3.119.400116

Yoshida, G.M., Yáñez, J.M., 2021a. Multi-trait GWAS using imputed high-density genotypes from whole-genome sequencing identifies genes associated with body traits in Nile tilapia. BMC Genomics 22, 1–13. https://doi.org/10.1186/s12864-020-07341-z

Yoshida, G.M., Yáñez, J.M., 2021b. Increased accuracy of genomic predictions for growth under chronic thermal stress in rainbow trout by prioritizing variants from GWAS using imputed sequence data. Evol. Appl. 00, 1–16. https://doi.org/10.1111/eva.13240

Yu, X., Megens, H.-J., Mengistu, S.B., Bastiaansen, J.W.M., Mulder, H.A., Benzie, J.A.H., Groenen, M.A.M., Komen, H., 2021. Genome-wide association analysis of adaptation to oxygen stress in Nile tilapia (Oreochromis niloticus). BMC Genomics 22, 426. https://doi.org/10.1186/s12864-021-07486-5

Zhang, Z., Druet, T., 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. J. Dairy Sci. 93, 5487–5494. https://doi.org/10.3168/JDS.2010-3501

CHAPTER 5 – Final Considerations

In the present thesis, we tried to explore three important topics related to new perspectives of quantitative genetics and genomic tools applications in aquaculture breeding. The first topic was linked to uniformity of harvest weight in shrimp. Our results showed that there is a large proportion of genetic variance in the residual variance for harvest weight in this shrimp species, suggesting opportunity to increase uniformity through selection. However, there is low heritability for uniformity indicating that a large number of animals is necessary to select animals for this trait with high accuracy. In order to increase accuracy of selection for uniformity, sib-testing or the inclusion of genomic information may be interesting alternatives. We also observed a negative correlation between estimated breeding values (EBV) for uniformity of harvest weight and survival, when the animals are reared in low density conditions. It may indicate that most likely families that are more uniform may also present higher survival rates, i.e., these families may be more "robust" in this condition. "Robustness" may be a desirable trait in shrimp farming given the constant outbreak of diseases and climate change issues that affect this industry.

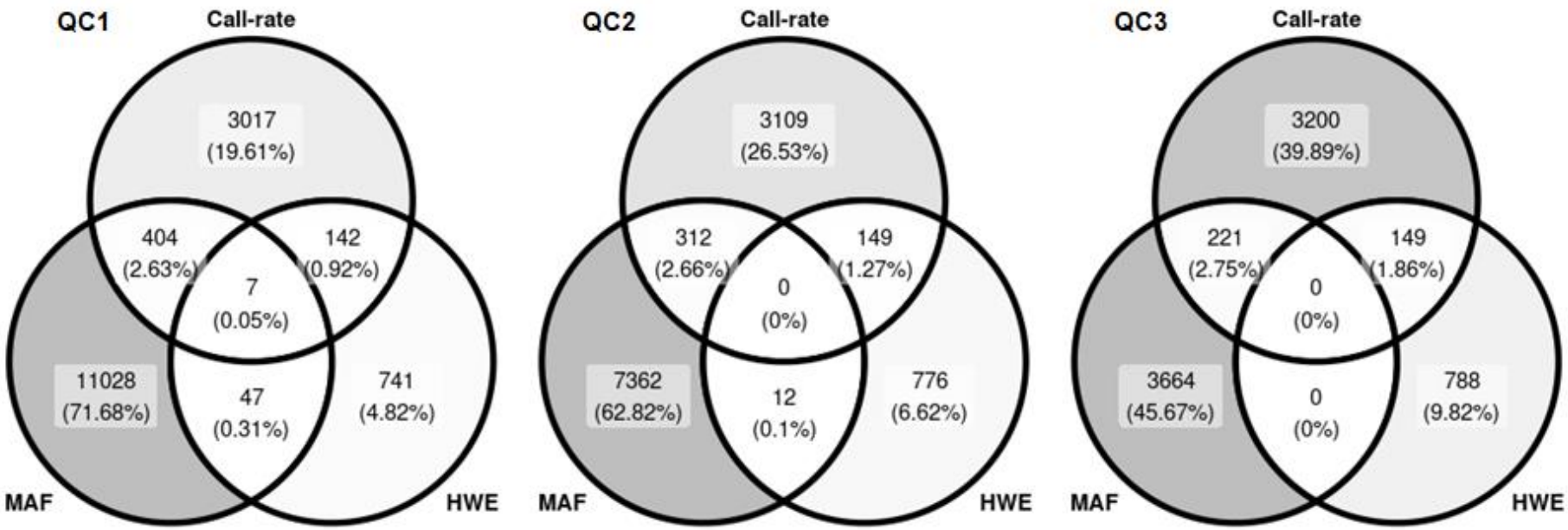
Currently, there is a clear expansion in the knowledge of genomic resources of aquaculture species. Each year, more genomic studies and tools are developed to assist in the aquaculture breeding programs. One example was showed in the second topic of this thesis in which we evaluated the application of the first 50k SNP panel commercially available for shrimp. Our results showed a high proportion of polymorphic loci validating this tool to this farmed shrimp population. We also evaluated population parameters associated to genetic variability, linkage disequilibrium and effective population size. These results converge to suggest recent incorporation of animals from different populations or strains in this broodstock. Finally, the level of linkage disequilibrium (LD) extent captured by the panel revealed its potential use to genomic selection (GS) and genome-wide association studies (GWAS) in shrimp breeding. For GS, as shown in other species, this level of LD and the number of SNPs seem to be satisfactory to estimate with high accuracy the genomic EBV of animals. This is particularly important for traits that are difficult or expensive to be measured, such as disease resistance and meat quality. Nevertheless, care must be taken in the application

of GWAS. Spurious associations between QTL and SNPs may be obtained due to the rapid linkage disequilibrium decay at short distance, mainly for polygenic traits, requiring a high number of animals and probably a denser SNP panel to improve QTL detection.

Finally, the last topic was related to genotype imputation using whole-genome sequence data (WGS). This type of data is very informative and may boost the application of GS and GWAS. However, it is expensive to obtain mainly considering many thousands of individuals. In this sense, we evaluated the genotype imputation as a cost-effective strategy to obtain WGS in Nile tilapia. Using only SNPs with high imputation accuracy (r^2>0.8) we were able to impute from 50k to approximately 680k SNPs. If we only consider this high-accuracy SNPs as the final density, the expected accuracy of imputation would be high (between 0.92 and 0.95). We also observed that a “WGS bank” with animals from different populations may help in the imputation. However, a larger impact on accuracy is expected when target and reference animals share higher kinship levels. Thus, we do recommend more animals or animals from ascending lines as reference to obtain millions of SNPs imputed with high accuracy.

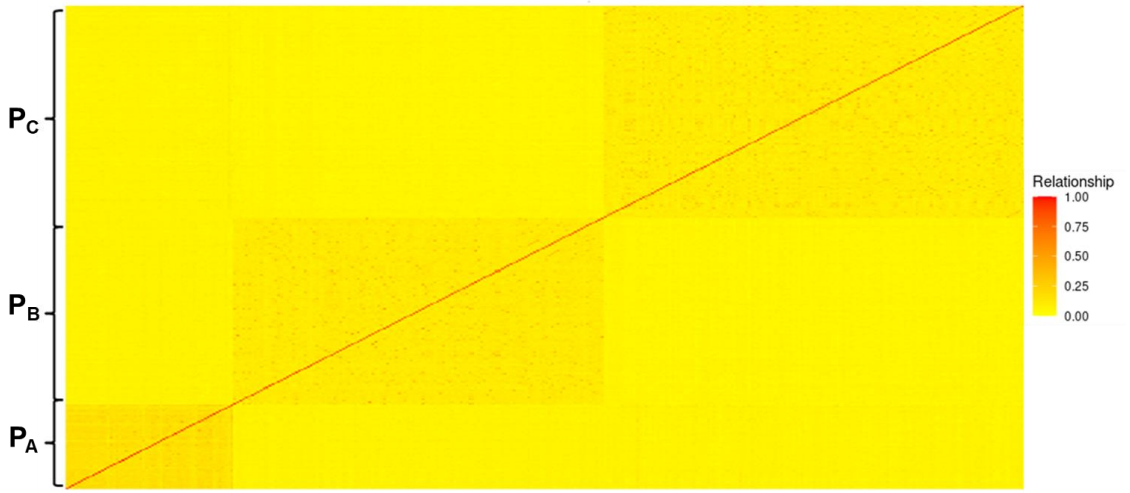
APPENDICES

Appendix A



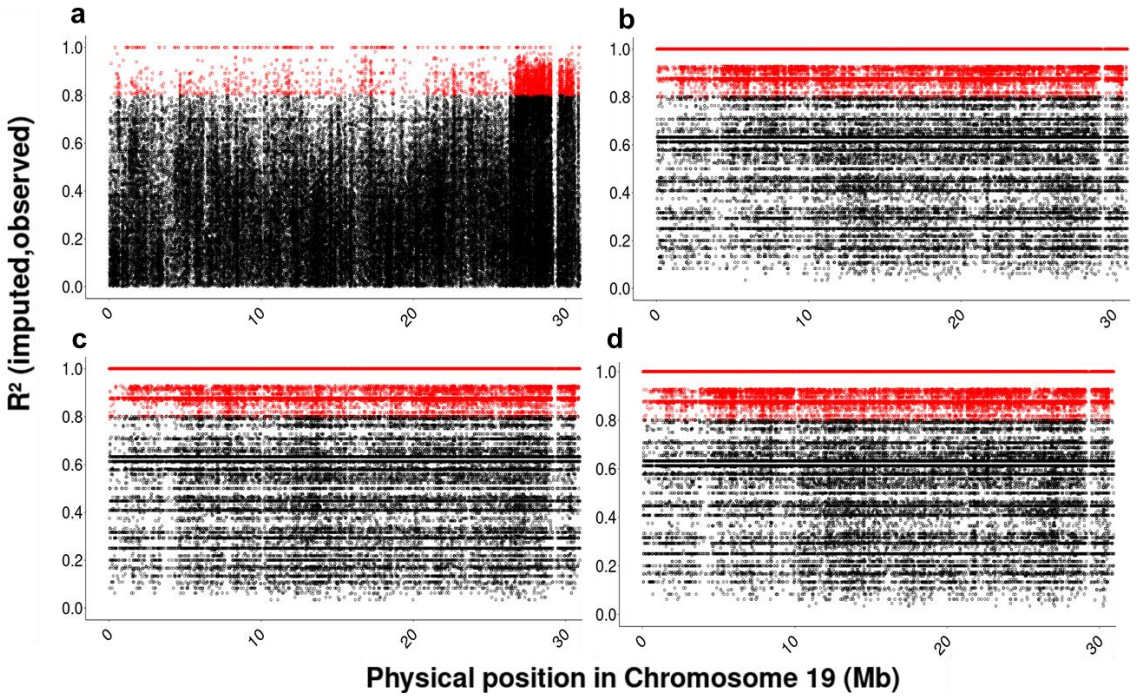
Appendix A. Venn diagrams of quality control (QC) of SNPs for a farmed shrimp population genotyped with a 50k array using three minor allele frequency (MAF) exclusion criteria thresholds: QC1 (< 0.10), QC2 (< 0.05) and QC3 (< 0.01). Call-rate and Hardy-Weinberg equilibrium (HWE) exclusion filters were the same for all QCs, < 0.8 and p-value < 10^{-5}, respectively.

Appendix B



Appendix B. Genomic relationship matrix among the tilapias studied from three different populations: 57 P_A, 126 P_B, 143 P_C.

Appendix C



Appendix C. Example of correlation between imputed and observed genotypes (R^2) as a function of physical position in chromosome 19 for P_A. **a**: 10% of same population individuals as reference; **b**: 90% of same population individuals as reference; **c**: two other populations as reference only; **d**: all three populations as reference. (Red dots are SNPs with accuracy greater than 0.8 considering the 10 validation groups).