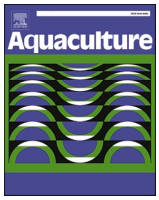
Aquaculture 621 (2026) 743971

Contents lists available at ScienceDirect

Aquaculture

journal homepage: www.elsevier.com/locate/aquaculture



Calculating genomic breeding values for growth using DNA pools within commercially reared black tiger shrimp (*Penaeus monodon*)

Gopala K. Guddanti [a,b,d,*], Cecile Massault [a,d], David B. Jones [a,c,d], Dean R. Jerry [a,c,d,e], Kyall R. Zenger [a,b,c,d]

[a] ARC ITRH for Supercharging Tropical Aquaculture through Genetic Solutions, College of Science and Engineering, James Cook University, Townsville 4811, Queensland, Australia
[b] Food Agility CRC Ltd., 175 Pitt St., Sydney 2000, NSW, Australia
[c] ARC ITRH for Advanced Prawn Breeding, College of Science and Engineering, James Cook University, Townsville 4811, Australia
[d] Centre for Sustainable Tropical Fisheries and Aquaculture, College of Science and Engineering, James Cook University, Townsville 4811, Queensland, Australia
[e] Tropical Futures Institute, James Cook University, Singapore

ARTICLE INFO

Keywords:
Aquaculture
Black tiger shrimp
DNA pool genotyping
Genomic prediction

ABSTRACT

Genomic estimated breeding values (gEBVs) are routinely used in genomic selection; however, their practical application in commercial shrimp aquaculture is impeded by high genotyping costs and unequal family contributions. DNA pooling offers a cost-efficient alternative to generate reliable gEBVs through hybrid genomic relationship matrixes (h-GRMs) and ranked phenotype groups. Despite its conceptual advantages, this approach has not been tested on commercial shrimp data. We evaluated pooled genotype data to predict genomic estimated breeding values (gEBVs) for body weight in *Penaeus monodon*. Reconstructed pools (RPs) of 2, 5, 10, 15, 20, and 25 individuals ranked by body weight across the population were generated *in-silico* using individual genotypes across 5097 DArTCap SNPs. In addition, the *in-silico* predictions from pool size (PS) 10 was tested using physical DNA pools (PDP) genotyped using a custom Axiom 70 k SNP array. Across the RP and PDP pools, gEBVs were estimated using a GBLUP animal model examined for three scenarios. First, parental gEBVs predicted from RPs showed high accuracy with small pool size (0.92 ± 0.005, PS2) and continuously declined as pool size increased (0.82 ± 0.07, PS25). Second, sibling gEBV prediction across ponds displayed high accuracy with small pool sizes (0.94 ± 0.15 for PS2) and reduced sharply in larger pool sizes (0.58 ± 0.62 for PS25). Within the top 20% of siblings, the gEBV accuracy ranged from 0.67 ± 0.57 (PS2) to 0.48 ± 1.59 (PS25). Together, these outcomes indicated PS10 represents an optimum balance between cost and accuracy. In the third scenario, gEBV accuracy was tested using PDP-PS10, and a higher prediction accuracy was obtained in a pond with moderate genetic diversity (0.53 ± 0.01) than in one with low diversity (0.43 ± 0.01). Pooling DNA from 10 ranked individuals would reduce genotyping costs up to eightfold; however, prediction accuracy decreased by 27%–30% when compared to individual genotyping. Although pooling reduces accuracy, the implementation of a structured pool design and moderate genetic diversity in reference populations can enable reliable gEBV estimation. These findings validate DNA pooling as a practical and cost-effective tool for genomic selection in commercial breeding programs for *P. monodon*.

1. Introduction

Shrimp farming is vital to the global aquaculture sector and serves as a crucial source of protein and socio-economic activity. However, its productivity is often constrained by factors such as inconsistent growth rate, high mortality rate, and susceptibility to a variety of pathogens. Advanced selective breeding has improved the productivity of farmed animals by identifying superior progeny that can be selected as parents to produce subsequent generations (Gjedrem and Rye, 2018). The application of selection enhances production traits such as growth, survival, and disease resistance (Gjedrem, 2005; Gjedrem and Robinson, 2014). However, several challenges limit the implementation of

* Corresponding author at: Centre for Sustainable Tropical Fisheries and Aquaculture, College of Science and Engineering, James Cook University, Townsville 4811, Australia.
E-mail address: gopalakrishna.guddanti@my.jcu.edu.au (G.K. Guddanti).

https://doi.org/10.1016/j.aquaculture.2026.743971
Received 14 October 2025; Received in revised form 5 March 2026; Accepted 1 April 2026
Available online 7 April 2026

advanced methods for selective breeding programs across different animal production systems, primarily because of the need to synchronise with commercial production, maintain the pedigree traceability of animals, and collect accurate phenotype measurements. Furthermore, small-scale production systems face considerable obstacles in adopting selection methods because investment costs often exceed their capabilities (Fugeray-Scarbel et al., 2021).

In early traditional family-based aquaculture breeding programs, each family was physically separated and reared in individual tanks; thus, superior animals were selected and tracked during production. However, identifying discrete family performance is impeded by the complexities of environmental variability, which influence the precise estimation of individual genetic merit, as well as the additional infrastructure and personnel expenses incurred while rearing families separately. The advent of genotyping techniques has provided a viable solution for tracing individuals to their respective family representatives, and farms can now deduce the parental and, thus, family origin of individuals (through parentage analysis) within a system in which all families are reared together (Massault et al., 2021; Vandeputte and Haffray, 2014). However, owing to survival differences, unequal family contributions are often observed in communal rearing systems (Foote et al., 2019). Such unequal family contributions necessitate the genotyping of many individuals to ensure the sufficient representation of all families. Consequently, accurately identifying individual family members using genotype information and the corresponding performance records requires substantially larger sample sizes (Domingos et al., 2014).

In modern selective breeding, such as that based on genomic selection methodologies, predictive models are formulated within a training cohort using phenotypic and genotypic data. The reliability of genomic prediction within the training dataset is assessed by forecasting the trait performance in a designated test population (Goddard et al., 2009; Hayes et al., 2009). The most insightful application of genomic selection in aquatic species is the identification of the best-performing individuals using highly accurate estimations of genetic merit within a combined family rearing system (Henshall et al., 2014; Song and Hu, 2021; Sui et al., 2020). However, in species that undergo mass spawning, a substantial expense is incurred by the large number of individuals that must be genotyped to accurately characterise family structures. This is a particular concern in species where the return on investment through genotyping of many individuals is low because of their production value and market demand (Wang et al., 2020). This highlights the need for cost-effective strategies to balance breeding accuracy and financial sustainability (Khatkar, 2017; Zenger et al., 2019).

DNA pooling is an alternative method that aims to reduce genotyping costs. This technique is based on pooling DNA derived from a number of individuals before genotyping. Once the pool is genotyped, the average allele frequency at each locus is calculated, and the likely proportion of individuals in each pool is computationally estimated. When pools are paired with phenotypic records (i.e., body weight ranks in the population), the overall genetic merit of individuals from the pool can also be predicted, subsequently lowering genotyping costs for establishing genetic breeding values. To implement this workflow, researchers have applied the hybrid genomic relationship matrix (h-GRM) to estimate the genetic relationships among individuals derived from pooled genotypes of progeny samples (Bell et al., 2017; Reverter et al., 2016).

The use of pooled DNA and allele frequency estimates has been validated as a cost-effective approach for estimating genomic breeding values (gEBVs) in terrestrial and aquatic species. In terrestrial species, simulation experiments in cattle have shown that genomic prediction accuracy improves when pools are constructed based on phenotype ranking rather than randomly (Alexandre et al., 2019). Similarly, pooling strategies applied to predict Dag scores in sheep indicated the broader applicability of the approach across traits (Bell et al., 2017). In aquatic species, pooled DNA has been used to estimate family performance (Kinghorn et al., 2010; Sonesson et al., 2010), identify family contributions, and estimate growth trait heritability in pooled shrimp samples (Henshall et al., 2014; Khalilisamani et al., 2022). In Atlantic salmon (*Salmo salar*), DNA pooling has been used to rank family performance for survival against salmonid alphavirus (Dagnachew et al., 2022). These findings highlight that pooling strategies can serve as cost-effective alternatives to individual genotyping, particularly when large populations are evaluated.

Conversely, a consistent limitation across these studies is the sensitivity of prediction accuracy to pool size and composition (Aldridge et al., 2022; Baller et al., 2020). Pools containing more than 10–15 individuals have a significantly lower accuracy in both cattle and sheep (Alexandre et al., 2019; Bell et al., 2017; Reverter et al., 2014). Similar trends were observed in pooled DNA derived from related individuals or correlated traits, where an increase in pool size led to a decline in sire gEBV prediction accuracy owing to the added complexity arising from contributions by multiple sires (Alexandre et al., 2020; Vargas Jurado et al., 2021). In sheep, the predictive accuracies for Dag scores have been shown to decline progressively with larger pool sizes (Bell et al., 2017). Evidence suggests that pools containing more than 10 individuals often introduce phenotypic and genetic variability, which can reduce model precision (Reverter et al., 2014). Thus, while pooling provides a viable strategy for reducing costs, its predictive power also depends on the pool size and the composition of families within pools.

The successful application of DNA pooling methods for aquaculture species represents a major advancement in the industry but faces many challenges. This is particularly evident for mass-spawning species, where families are communally reared from early developmental stages (i.e., larval stages). Simulation studies can provide theoretical expectations of accuracy; however, in real-world applications, substantial variation occurs owing to biological and practical commercial constraints and can be observed from one rearing environment to another (i.e., between tanks or ponds). Thus, it is important to use on-farm datasets to calculate the associated trait correlations and heritability before evaluating the merit of DNA pooling. This helps formulate selection strategies that closely align with industry production scenarios. Furthermore, commercial datasets can reveal pragmatic obstacles and different family structures, which may aid in the more effective customisation of genomic selection initiatives.

To address these barriers, this study evaluated DNA pooling using a farm-derived dataset of the body weight of *P. monodon* to estimate gEBVs for genomic selection. Individual shrimp phenotypic and genotypic records were obtained and combined into pools comprising different numbers of individuals, referred to as the reconstructed pools (RPs). To create these RPs, individuals were ranked by weight in descending order and pooled in sliding brackets across the distribution based on the Pool Sizes (PS). The gEBV prediction accuracies were then calculated across different PS using pool criteria (i.e., ranked phenotypes). In addition, Physical DNA Pools (PDP) were created and genotyped at PS10 to allow for validation of the RP methods. The prediction accuracy of body weight was assessed for three testing scenarios. First, the optimal pool size was explored from parental gEBV predictions (p-gEBV), as the prediction accuracy was influenced by the number of pools and their varying genetic diversity across multiple commercial ponds. Second, the effects of sibling gEBV (s-gEBV) prediction accuracy were tested randomly, as well as in a scenario in which the top 20% of the heaviest shrimp were selected before evaluation. Finally, actual physical DNA pool gEBVs (PDP-gEBVs) were compared with gEBV predictions based on RPs (RP-gEBV) to validate the gEBV prediction accuracy of the optimal pool size.

2. Materials & methods

2.1. Cohort production

As part of a broader research initiative, a commercial cohort of *P. monodon* progeny was produced from 55 males and 65 females (Foote

et al., 2019). At the post-larval stage, progenies were transferred to commercial grow-out ponds at a density of 40 animals/m^2 and were cultured for 140 days. At harvest, 5274 individuals from seven production ponds (IDs #149, 150, 152, 155, 156, 157, and 160) were weighed and sampled for genotyping, as described below. The resulting individual records ranged from 604 to 894 across all seven ponds (Table 1).

2.2. Phenotypes and reconstructed pool design

Individual body weight records collected from production farms were used to create reconstructed pools (RPs) across the phenotypic rank distribution of each pond. Weight records were arranged in descending order for each pond distribution. Different pool sizes (PS) were established across the distribution per pond as follows: animals were pooled in multiples of two (PS2), five (PS5), ten (PS10), 15 (PS15), 20 (PS20), and 25 (PS25). Any remaining individuals who did not form a complete pool for the respective PS were excluded from the analysis. As the pool size increased, the number of pools per pond decreased (Table 1). The body weight of each PS was calculated as the mean body weight of the individuals within each pool. A statistical normality test was performed for body weight distribution using the Shapiro-Wilk test. In addition, non-parametric tests were conducted using the Wilcoxon method for ponds that failed the normality test.

2.3. DNA extraction, genotyping and parentage analysis for cohort samples

Pleopod samples were collected from 5274 progeny across the ponds and all broodstock ($N = 120$) and preserved in 70% ethanol for subsequent DNA extraction. Total nucleic acids (TNA) were extracted using the KingFisher Flex 96 kit. The TNA was normalised to 25 ng/μl, as described in Foote et al. (2019), before being genotyped with Diversity Array Technology using custom DArTCap methods described by Guppy et al. (2020). The resulting SNP genotype dataset was filtered by silencing any genotype calls made from fewer than seven reads and then removing any SNPs with call rates of <0.80 or minor allele frequency (MAF) of <0.02 (Vu et al., 2020). Parentage analysis was conducted using CERVUS 3.0 (Kalinowski et al., 2007), with all parent samples included, using 5097 single nucleotide polymorphisms (SNPs). LOD scores between the parent and parent-offspring trios were used to assign parentage.

2.3.1. Calculating allele frequencies from reconstructed pools

To ensure the quality and suitability of the genotype data for genomic relationship estimation and model analyses, additional filtering and processing steps were applied to SNP datasets. In addition to the SNP filtering described in Section 2.3, DArTCap SNPs with a minor allele frequency (MAF) below 0.1 were removed. Genotype information across individuals was then recoded in PLINK 1.9 as AA = 0, AB = 1, and BB = 2, with missing data denoted as NA (Chang et al., 2015). Missing genotypes were imputed using Wright's method in the SNPReady package (Granato et al., 2018). Allele frequencies for individual progeny, sires, and dams were estimated from the second allele (B allele) to include both individual and pooled genotypes in the same analysis. Individual genotypes converted into allele frequencies and, therefore, become 0, 0.5, and 1 for homozygotes, heterozygotes, and alternative homozygotes, respectively (Bell et al., 2017). After genotype recoding and imputation, individual genotypes (0,1,2) were first converted into B-allele frequencies (0, 0.5, 1). These individual frequencies were then collated into reconstructed pools (RP-AF) by grouping according to pool size, summing their B-allele counts at each SNP, and dividing by the total number of observed allele calls. The denominator was adjusted at each SNP to exclude missing genotypes, ensuring that each pool frequency reflected the proportion of observed diploid counts rather than a fixed sample size. Parental genotypes were processed in the same way and converted to B-allele frequencies before being appended to the reconstructed offspring pools. This produced a single allele-frequency matrix representing progeny pools and parental samples on a common scale,

$$p = f(B) = \frac{2nBB + 2nBb}{2N}$$

where ***nBB*** and ***nBb*** represent the number of B alleles in each genotype, and ***N*** is the total number per pool and SNP, respectively. The ***f(B)*** matrix was applied to the M, P, and Z matrices using the same formula as VanRaden (2008), M = f(B)-0.5 × 2, for the P matrix (mean of f(B)-0.5) × 2, and the parents were appended to the pooled progenies before Z matrix calculation.

For ponds 152, 157, and 160, the number of pools was limited to 24 pools for PS25 and 30 for PS20, owing to the lower number of individual records for each pond. This small number of pools caused GRM diagnostics showed near-singularity issues, convergence failure during REML and consequently prediction errors in estimating the gEBV. To counteract this, an additional 300 individual genotype records were simulated for these ponds using the pedSimulate R package (Nilforooshan, 2022). By augmenting the data, we ensured sufficient genetic variance for prediction while recognising that simulated individuals ensure reliable gEBV estimation across all pool sizes. Genotypes were simulated by applying Mendelian sampling principles and sampled probabilistically based on respective pond allele frequencies and preserved within-pond family size distributions. The simulated genotypes were appended random iteratively to the existing genotype matrix. Phenotypic values for simulated individuals followed normal distribution parameterised by the observed pond-level phenotypic mean and variance. The empirical and simulated records were merged prior to construction of the genomic relationship matrix. This augmentation improved matrix conditioning and ensured stable convergence of genomic evaluations.

Table 1
The number of individual records and the number of pools with their respective pool sizes (PS: number of individuals in a pool) across the seven ponds.

Pond ID	Number of individual records	PS2	PS5	PS10	PS15	PS20	PS25
149	812	406	162	81	54	40	32
150	894	447	179	89	59	44	36
152	604	302	121	60	40	30	24
155	878	439	175	88	58	44	35
156	845	422	169	84	56	42	34
157	616	308	123	61	41	31	24
160	625	312	125	62	41	31	24
Total	5274	2636	1054	525	349	262	209

2.4. Physical DNA pools (PDP), genotyping and allele frequency calculation

2.4.1. DNA pool preparation

To assess any potential biases in calculating the allele frequency in RPs, individual DNA samples within PS10 were physically pooled in equimolar amounts, and the estimated allele frequencies (PDP-AF) were compared with RP-AF. A pool size of 10 was selected for evaluation as it displayed an intermediate balance between loss of prediction accuracy and the number of individuals included within RPs (See results Section 3.2.2). Two ponds were selected for this DNA pool experiment: pond 150, which consisted of a large number of families ($N = 38$; as an indicator of moderate diversity) and showed higher heritability for body weight ($h^2 = 0.29$; based on individual records), and pond 152, which had a smaller number of families ($N = 28$; low diversity, $h^2 = 0.19$).

To build physical DNA pools, we used 460 and 720 DNA samples from ponds 152 and 150, respectively. Each sample was quantified using NanoDrop™ (Thermo Scientific ™), resulting in an average DNA

concentration across all samples of 209.2 (± SD 189.15) ng/μl. The DNA samples were then ranked in descending order based on body weight to match the design of the RPs. To create PS10 DNA pools, equimolar amounts of DNA were collected from every group of 10 individuals in this ranked order and pooled independently three times to produce triplicate pools for each group. This procedure ensured that each replicate pool reliably depicted an equal concentration of individuals within each pool of 10 and enabled the testing of technical errors in pipetting through the replicate pools. This resulted in 72 pools (in triplicate) for pond 150 and 46 pools (in triplicate) for pond 152 (Supplementary Table 1).

2.4.2. DNA pool genotyping and filtering

Different genotyping approaches were used for the RP and PDP datasets to enhance the accuracy of the pooled allele frequency estimation and enable cross-platform genotype comparisons. Previously, individuals in the RP dataset were genotyped using DArTCap. To determine the pooled allele frequencies from PDP with higher accuracy and reduce SNP dropouts, 494 PDP samples were genotyped using a custom-made 70 K Axiom *P. monodon* array (Thermo Fisher Scientific). To ensure accurate cross-platform genotype comparisons, individual DNA samples that were previously genotyped on the DArTCap panel were genotyped with PDPs on the Axiom array. For this purpose, a subset of 140 individuals was selected, including 10 individuals from each of the 11 pools in pond 150 and 10 individuals from each of the three pools in pond 152 (Supplementary Table 1). Beyond validating genotypes and allele frequencies, these 140 individual samples facilitated the integration of genotypes from both datasets using common SNPs.

Quality control and genotype calling of the Axiom data were performed using the *Axiom Analysis Suite* (v5.3.0.45). The best-practice workflow from ThermoFisher Scientific was applied with slight adjustments to the threshold level (Table 3). Individual samples were filtered, retaining only samples with quality controls that assessed the performance of the array hybridisation pattern of signal intensity (DISH QC) > 0.82 and sample call rate > 0.97. The average call rate of the samples that passed these thresholds was 98.5. A total of 68,148 SNPs were consistently amplified using the Axiom platform, yielding a conversion rate of 97.3%. Additional filtering of SNPs using an SNP call rate > 0.97, Fisher's linear discriminant (FLD) > 3.6, and MAF > 0.01 resulted in 31,003 polymorphic SNPs that were used for downstream analysis (Table 2). In addition, the raw X and Y signal intensities were extracted for individual and pooled samples, as these were required to calculate the allele frequencies of the pooled samples. For computational purposes, the X intensities were rescaled between 0 and 2 (Peiris et al., 2011).

2.4.3. DArTCap and axiom genotype merge and validation

To enable direct comparison between RP and PDP samples, the allele-frequency information generated from DArTCap and Axiom platforms was merged into a common analytical framework. Allele frequencies of DNA pools (PDP-AF) were estimated from Axiom array genotyping based on the individual genotype calls of the samples represented within the pools and the X and Y intensities for the pools themselves (Peiris et al., 2011). Alignment was required because *in-silico* pools used DArTCap data, while physical pools were genotyped on the Axiom platform. Pooled allele frequencies ensured the h-GRM operated on a common SNP scale, enabling valid performance assessment based on overlapping markers and consistent calling thresholds. A set of 4656 SNPs was selected and identified as common between the Axiom and DArTCap platforms based on their common physical base pair positions within the reference genome (Uengwetwanit et al., 2021). All SNP positions were identified by mapping the associated sequences to the *P. monodon* genome using bwa and custom scripts (Li and Durbin, 2009). As stated in Section 2.4.2, the QC filter method resulted in the Axiom dataset containing 31,003 SNPs, 140 individual samples, and 354 pooled samples, producing a genotype concordance rate of 92.32%. Four individuals and seven SNPs that showed conflicts between the two genotyping platforms were removed before the final merge of the common genotypes. Axiom genotypes were prioritized over DArTCap genotypes (if both were present) because of their much lower genotype error rates. However, when comparing the merged genotype dataset across all individuals throughout the ponds, many SNPs showed high rate of missing genotype rate (> 35%). Therefore, a subset comprising 2818 SNPs was identified as having low missingness within the merged Axiom and DArTCap genotype datasets and was used in ongoing analyses of both the individual and pooled samples.

2.4.4. Physical DNA pool (PDP) allele frequencies

For the PDP samples, 'X' and 'Y' intensity values were used to calculate allele frequencies following the Pn^3 method described by Peiris et al. (2011) as follows:

$$\hat{p}_{n3} = \begin{cases} \frac{1}{2}\left(\frac{\hat{p}_k - \bar{\hat{p}}_{k_{BB}}}{\frac{1}{2} - \bar{\hat{p}}_{k_{BB}}}\right), & \hat{p}_k \le \frac{1}{2}, \\ 1 - \left(\frac{1}{2}\left(\frac{\hat{p}_k - \hat{p}_{k_{AA}}}{\frac{1}{2} - \hat{p}_{k_{AA}}}\right)\right), & \hat{p}_k > \frac{1}{2}, \end{cases}$$

$$\hat{k} = \bar{X}/\bar{Y},$$

$$\hat{p}_k = X/(X + \hat{k}Y)$$

where $\hat{p}_k$ is the heterozygote-corrected frequency estimation; k (the heterozygote correction factor) is the ratio of the X and Y intensities of heterozygotes. $\bar{X}$ and $\bar{Y}$ are the means of X and Y of known heterozygous (AB) individuals for each SNP, which were used to calculate the heterozygote-corrected frequency estimation ($\hat{p}_k$), and the heterozygote correction factor (k) for each SNP. The average values of $\hat{p}_k$ for each SNP with *pk_AA* and *pk_BB* were estimated as the frequency of the homozygous genotypes AA and BB for each SNP and calculated from the observed individual samples with the AA and BB genotypes. The initial X and Y intensities from a pool of individuals were directly compared with the average allele A and B frequencies of the pools, with the error rate being inversely proportional to the number of individuals in each pool. Therefore, by considering the error rate associated with varying numbers of pools and pool sizes, the $\hat{p}_{n3}$ method can identify the corrected allele A intensity and subsequently estimate the pooled allele frequency, directly correlating it to the actual allele A frequency. Allele frequencies for the physical DNA pool (PDP-AF) were independently estimated from the X and Y intensities for the three replicate sets. The similarity between the three replicate pooled samples was assessed using the root mean square deviation (RMSD) method before selecting the

Table 2
SNP filtering workflow, parameters and threshold levels as per the Best Practises recommendations (ThermoFisher Scientific).

Probe set QC parameters	Threshold levels	Number of SNPs retained
All	–	68,148
PolyHighResolution and NoMinorHom		33,172
Best and Recommended	1	33,064
Call Rate	>97	32,984
Minor Allele Frequency	>0.001	32,975
Fisher's Linear Discriminant (FLD)	>3.6	32,969
Homozygous FLD	>8.5	32,965
Homozygous Ratio Offset	>0.02	32,965
Minor Allele Count	>2	31,003

most reliable replicate set for downstream comparisons. The resulting genotype accuracies were measured using the Pearson correlation coefficient, and PDP-AF was compared to RP-AF.

2.5. Genetic parameter calculation of individual and pooled samples

2.5.1. Calculation of genomic relationship matrices (GRMs)

The relationship between pooled and individual samples was estimated using genomic relationship coefficient matrices (GRM). For individual samples, the genomic relationship was calculated using the R package *Gaston*, using VanRaden's method, whereby the observed mean allele frequencies were divided by a scaled numerator relationship matrix (VanRaden, 2008).

$$G = \frac{zz'}{2\Sigma p_i(1-p_i)}$$

Where $\boldsymbol{G}$ is the genomic relationship matrix, $\mathbf{z}$ is the incidence matrix of the markers, and $\boldsymbol{p_i}$ is the observed MAF of all genotyped individuals.

In addition to the GRM based on individual genotypes, a hybrid GRM (h-GRM) containing the parental genotypes (86 individually genotyped dams and sires), progeny included within the RPs was created. The h-GRM comprises three distinct blocks of relationship coefficients: 1) within parents (individual genotypes), 2) between RP samples and parents, and 3) within RP samples. The h-GRM was used to predict the weights of the parents using the RPs. The B-allele frequencies of progeny pools were used to represent the allele frequency of the second allele at each locus, allowing for a more nuanced understanding of genetic variation within the pooled samples. This estimation is crucial for linking mean phenotypic data to mean B-allele frequencies (Bell et al., 2017).

2.5.2. Estimation of breeding values

Genomic best linear unbiased prediction (GBLUP) was performed using ASReml-R version 4.2 (Butler et al., 2023) for variance component estimation and the breeding values prediction. Parents (p-gEBV) and siblings (s-gEBV) genomic breeding values were obtained from the pooled data using a univariate animal model.

$$y = X\beta + Zu + e$$

where $\boldsymbol{y}$ is the phenotype (body weight), $\boldsymbol{X}$ is the mean of the pooled phenotype, β is a fixed effect vector, $\boldsymbol{Z}$ is the incidence matrix associated with the random effect, $\boldsymbol{u}$ is the vector of random additive genetic effects and $\boldsymbol{e}$ is the residual effect. Additive genetic effects were assumed to follow $\boldsymbol{u \sim N(0, G\sigma_u^2)}$, where $\boldsymbol{\sigma_u^2}$ represents the additive genomic variance, and $\boldsymbol{G}$ is the GRM; $\boldsymbol{e \sim N\,(0, I\sigma_e^2)}$, where $\boldsymbol{\sigma_e^2}$ is the residual variance, and $\boldsymbol{I}$ is the identity matrix. Variance components were estimated by restricted maximum likelihood. The GRM was constructed using the VanRaden (2008) method I formulation, in which the centred genotype matrix was scaled by $2\sum pi(1-pi)$. Both sex and pond were tested as fixed effects; however, sex had no significant effect. Pond was included as a fixed effect in the animal mixed model to account for systematic differences in environmental conditions. This specification ensured that environmental heterogeneity arising from variation in pond management, feeding, and water parameters was absorbed by the fixed-pond term rather than contributing to the additive genetic variance. Model assumptions were evaluated by inspecting heterogeneity of residual variances across ponds and no major violations were detected.

The heritability (h^2) of body weight was estimated from the proportion of phenotypic variance attributed to the additive genetic variance.

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

where $\boldsymbol{\sigma_a^2}$ is the additive genetic variance and $\boldsymbol{\sigma_e^2}$ is the variance of the residuals.

2.5.3. Calculating prediction accuracy

The accuracy of the gEBVs prediction was assessed using the 10-fold cross-validation method with five replicates. In each fold of the replica, 10% of the population was randomly masked, and the remaining records were trained to predict masked individuals (Goddard et al., 2011). The accuracy of the predicted gEBVs was assessed by calculating the Pearson correlation coefficient between the observed phenotype and predicted gEBVs of the masked individuals, divided by the square root of heritability. The overall accuracy was estimated using ten-fold cross-validation as the average across five replicates. The accuracy of the genomic breeding values was estimated as follows:

$$Acc = \frac{Corr(y, p)}{\sqrt{h^2}}$$

where $\boldsymbol{Corr\ (y,\ p)}$ is the correlation between $\boldsymbol{y}$ (observed, individual gEBVs) and $\boldsymbol{p}$ (predictive, pooled gEBVs), divided by the square root of heritability to scale the correlation between estimated breeding values and observed phenotypes, providing a more accurate representation of the genetic component of predictability and understanding genetic merit can be predicted, independent of environmental or non-heritable influences (Jerry et al., 2022; Song et al., 2022).

2.6. Accuracy testing scenarios

The accuracy of gEBV predictions obtained from individual genotype data and pooled allele frequencies was assessed across the three scenarios as described below (Fig. 1). First, the optimum pool size was investigated across multiple ponds containing varied family diversity by predicting the parents' weights using RPs (Scenario 1). Second, sibling prediction was evaluated for model performance from two perspectives: one addressed two-stage selection by selecting a random 10% across the phenotype rank distribution, and the other selected a random 10% within the top 20% of siblings as test datasets (Scenario 2). Finally, PDP was performed (for PS10, as inferred from RP analysis) to enable a direct comparison of genetic parameters calculated from RP and PDP genotypes (Scenario 3).

2.6.1. Scenario 1: optimum pool size determination using reconstructed pools

This scenario aims to identify an optimal balance between pool size and family diversity across multiple ponds to minimise the loss of prediction accuracy, which, in turn, will provide insights into practical DNA pooling applications in pond-based breeding programs. Within this scenario (S1), the prediction accuracies were tested across different PS using RPs. First, changes in gEBV prediction accuracy were measured by comparing p-gEBVs in different pool sizes with individual gEBV accuracies at the pond-specific level. The optimal threshold for the pool size was measured as a weighted score determined by assigning the cost variable and prediction accuracy of the respective pool size.

2.6.2. Scenario 2: phenotype prediction of masked siblings using reconstructed pools

In this prediction test scenario, the test population was optimised by targeting individuals with no phenotypes as selection candidates for the next breeding cycle. This approach was followed for two types of target siblings, referred to as Scenarios 2a and 2b. For both scenarios, the training pools were structured based on the phenotype rank. In Scenario 2a (S2a), phenotypes of random siblings across the phenotype distribution were masked and their gEBVs (s-gEBVs) were predicted using pooled data from siblings. We tested the gEBV accuracy of the masked candidates for each pond and across all ponds.

To evaluate the pooling strategy for broodstock replacement, we tested the concept in Scenario 2b (S2b) by simulating a two-stage selection strategy. This approach involves using genotyped and phenotyped siblings to predict the gEBV for un-phenotyped siblings within the

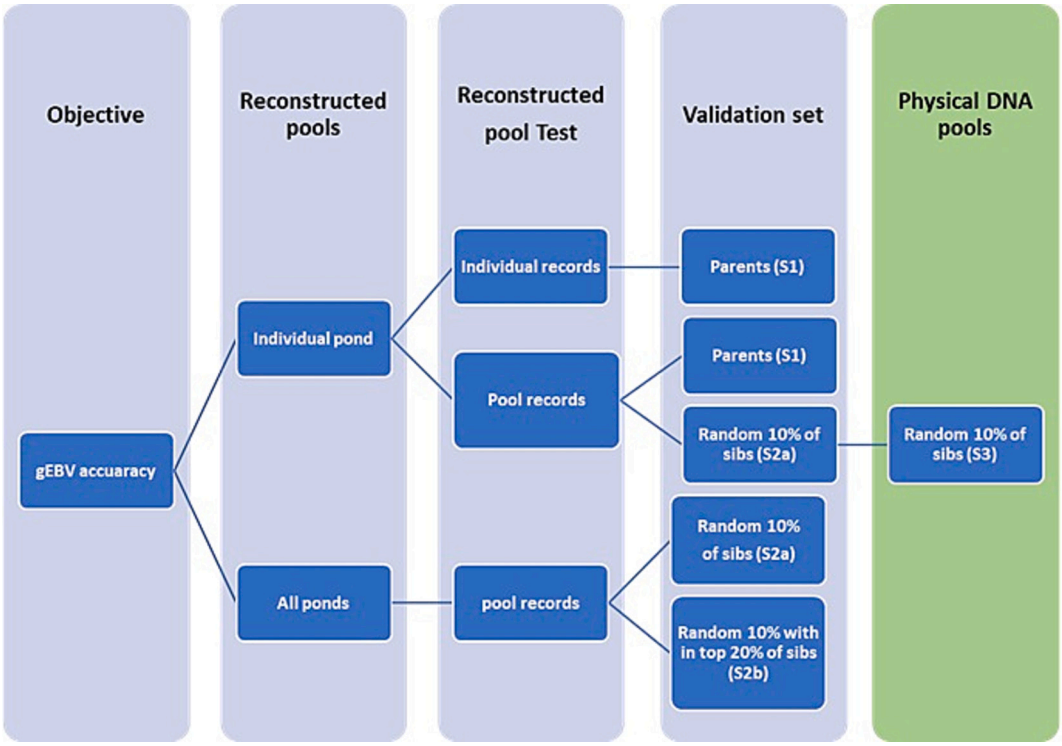


Fig. 1. Overview of datasets used for training and testing in gEBV prediction, highlighting three testing scenarios for gEBV accuracy between reconstructed pools (RP) and physical DNA pools (PDP).

top 20% of candidates whose phenotypes were masked. This approach reflects real-world breeding programs, where direct phenotyping of nucleus stock is often not feasible.

2.6.2.1. Scenario 2a: prediction of randomly masked siblings' phenotype across the distribution. To test the accuracy of the sibling gEBV prediction, 10% of the siblings across the phenotypic distribution were randomly masked to form the test set. The remaining siblings were re-ordered according to their phenotype rank and pooled at various pool sizes. The gEBVs of the masked individuals were then predicted using the training data from their pooled siblings, and these accuracies were compared with those of the re-estimated gEBVs. This sibling validation was performed in individual ponds, where the number of pools was limited by the sample size (for example, PS10 had 60 pools, and PS25 had 24 pools). We obtained an inadequate number of pools because of the limited number of records in individual ponds, which led to the model being unable to converge for large pool sizes (> PS20). Therefore, a combined all-ponds dataset was used, resulting in an increased number of pools (for example, PS10, 474 pools; PS25, 190 pools), while body weight was normalised within each pond to account for and nullify the multiple pond effects.

2.6.2.2. Scenario 2b: prediction of randomly masked siblings' phenotype within top-rank individuals. In this scenario, the test individuals were chosen to predict the top-performing individuals without requiring phenotypic data. The s-gEBV prediction was assessed by masking 10% of individuals randomly within the top 20% of the phenotype distribution. The accuracy of the gEBVs was evaluated by comparing the predicted and actual gEBVs. We then identified the masked individual gEBVs as those remaining within the top 20% after the prediction. Due to limited pond data, samples from multiple ponds were combined with standardised body weights, as explained in Section 2.6.2.1.

2.6.3. Scenario 3: physical DNA pool gEBV and prediction accuracy

The selection of PS10 for PDP validation was based on explicit decision criteria that PS10 preserved prediction accuracy within 15% of the individual-genotype, while achieving a substantial reduction in per-sample genotyping cost, and that it remained below variability thresholds that rendered larger pools unreliable. The convergence of independent indicators such as accuracy, cost per unit accuracy, and weighted scoring identified PS10 as the least-compromised option. These metrics indicated for prioritising PS10 in physical validation experiments.

The Physical DNA pool (PDP) methodology facilitates enhanced allele frequency estimation directly from pooled DNA samples. We assessed PDP performance based on the genetic diversity within the pools. For that, two distinct diversity ponds were used, as described in Section 2.4.1, and a common number of SNPs ($n = 2818$) was used for gEBV prediction (PDP-gEBV). Random sibling gEBV and 10-fold cross-validation were performed. The accuracies of PDP-gEBV and RP-gEBV were compared between the two ponds. The RP data were standardised, with a corresponding number of pools to facilitate optimised computational analysis comparable to PDP (Scenario 3, S3). Two-way ANOVA and statistical assumption tests were performed to test the differences in accuracy between RP-gEBV and PDP-gEBV.

3. Results

3.1. Summary statistics and distributions within empirical data

3.1.1. Phenotype distribution

The body weight distributions within and between the different commercial ponds are shown in Fig. 2. The mean body weight of the sampled ponds was 14.8 g, with an average standard deviation of 3.7 g. The observed coefficient of variation in females ranged from 22.9% to 33.2%, compared with 21.2% to 31.0% in males. The Shapiro-Wilk normality assessment indicated that ponds 149 and 150 deviated from a normal distribution ($P < 0.05$), which showed differences in the body weight distribution of males and females. In contrast, the remaining ponds followed a normal distribution ($P > 0.05$), with no differences

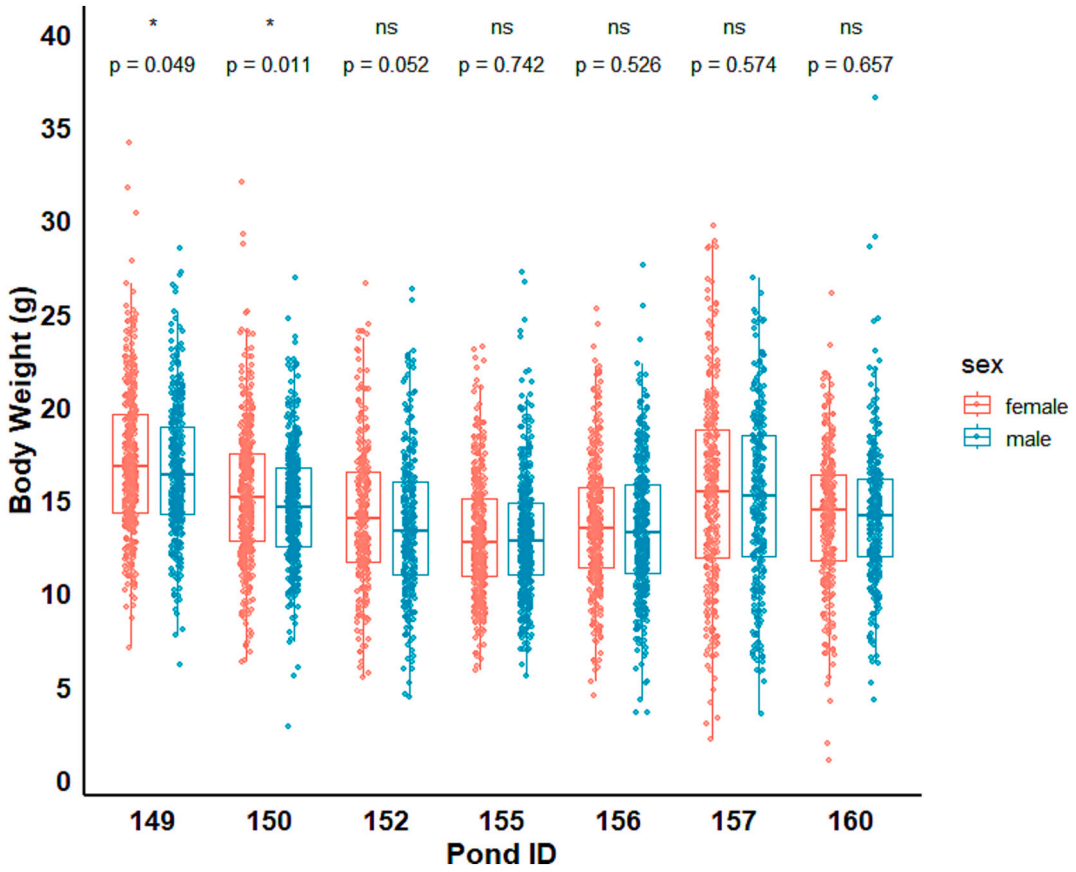


Fig. 2. Body weight distribution at harvest across the commercial ponds. Data points are distinguished by colour to represent female (red) and male (blue) individuals. Boxplots show the spread of body weights in each pond. Statistically significant *P*-values for differences between male and female body weights within each pond are indicated, with non-significant comparisons labelled as 'ns'. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

detected between the body weights of males and females. Given the non-normal distribution in ponds 149 and 150, sex differences in body weight were assessed using the Wilcoxon rank-sum test. Significant differences between males and females were detected in pond 149 ($P < 0.049$) and pond 150 ($P < 0.01$). However, the corresponding effect sizes were small in both ponds ($r < 0.3$). This suggests that the frequency of higher rank values has a small influence on the body weight differences between males and females.

3.1.2. *Family distributions and heritabilities across ponds within empirical data*

The number of full-sib families per pond ranged from 13 to 37 (Table 3). Because pools were constructed for each pond, the chances of numerous family contributions within the pool increased with the pool size (Table 3). This pattern was more pronounced in ponds with lower family diversity, indicating an uneven family distribution throughout the phenotypic distribution. Individual records from the DArTCap SNP dataset were used to estimate h^2 for body weight across each pond, with values ranging from a minimum of 0.19 in pond 152 to a maximum of 0.3 in pond 156. This variation in h^2 estimates across ponds is likely due to the differing compositions of stocked families and the overall genetic diversity observed within each pond (Table 4). Pond IDs 152, 157, and 160 exhibited lower family diversity, ranging from 0.19 to 0.23, whereas ponds 149, 150, 155, and 156 showed higher diversity (from 0.26 to 0.3).

The family distributions of PS are shown in Table 3. The number of families and their representation in each pond were relatively consistent across the distribution of pools, particularly for ponds 149, 150, 155, and 156. This indicates higher family diversity, with multiple families being evenly represented. In contrast, ponds 152, 157, and 160 displayed more uneven family distributions, with pond 152 showing a steep increase in family dominance at certain points, whereas ponds 157

Table 3
Summary of the number of families and heritability estimates across ponds: mean number of families represented within the resulting pool sizes (PS: 2, 5, 10, 15, 20, and 25).

Pond ID	Number of Full-sib families	Number of Half-sibs families	PS2	PS5	PS10	PS15	PS20	PS25
149	36	16	1.9	4.9	8.1	11.1	13.7	15.8
150	38	18	2.0	4.5	8.2	11.4	13.9	16.3
152	28	6	1.8	3.6	5.7	7.5	8.8	10.0
155	37	19	2.0	4.6	8.5	11.8	14.6	16.8
156	33	16	1.9	4.4	7.9	10.8	13.0	15.4
157	21	6	1.8	3.7	5.9	7.3	7.3	9.9
160	13	4	1.9	3.9	6.1	7.4	8.5	9.2
Total	**206**	**85**	**13.3**	**29.6**	**50.4**	**67.3**	**79.8**	**93.4**

Table 4
Genetic variance components, heritability, and genomic diversity indices summarised for each pond.

Pond ID	h^2 ± SE	Vp ± SE	Va ± SE	Ve ± SE	Hz	F
149	0.29 ± 0.06	12.62 ± 0.67	3.32 ± 0.68	9.30 ± 0.63	0.26	0.08
150	0.29 ± 0.06	13.79 ± 0.85	3.04 ± 0.69	11.30 ± 0.77	0.29	0.07
152	0.19 ± 0.06	13.44 ± 0.81	2.02 ± 0.65	11.42 ± 0.76	0.19	0.12
155	0.26 ± 0.05	12.56 ± 0.65	3.22 ± 0.72	9.34 ± 0.65	0.28	0.07
156	0.30 ± 0.06	22.55 ± 1.47	6.33 ± 1.72	16.22 ± 1.23	0.26	0.07
157	0.23 ± 0.07	21.76 ± 1.51	5.55 ± 1.65	16.22 ± 1.21	0.24	0.12
160	0.22 ± 0.11	21.98 ± 1.55	5.76 ± 1.73	16.22 ± 1.24	0.23	0.07

and 160 exhibited irregular patterns, reflecting skewed family representation.

The mean family body weight varied across ponds, with a CV ranging from 12 to 20%, with the overlap of full and half-sib families. Ponds 149, 150, and 155 had high family diversity and a relatively uniform body weight distribution across families, with fewer extreme variations in body weights (Supplementary Fig. 1). In contrast, pond IDs 152, 157, and 160 showed irregular distributions, with very few families outperforming the others. For example, pond 152 included families with narrow or broad distributions (e.g. families 17 and 26 or family 22, respectively). Similarly, pond 160 displayed smaller family counts with highly uneven weight distributions, suggesting unequal growth rates between families.

3.2. Scenario 1: optimum pool size validation

3.2.1. Reconstructed pool allele frequencies

The RP-AF for each pool size was estimated using a customised R script based on the respective mean of the individual genotypes for each DArTCap SNP. The Pearson correlation coefficient (r) was compared between genomic relationships based on the RP data and the mean GRM value of the individual relationship coefficient values for each pool category (Fig. 3). For example, PS10 includes ten individuals and their pairwise mean GRM values. The average relationship between the corresponding pooled individuals was used to obtain equivalent values. These were then compared with the GRM values based on the PS10 allele frequencies. High correlations were observed, with an r-value of >0.99 ($P < 0.05$) between the mean individual coefficient values and the values derived from RP-AF. Notably, a larger PS observed reduction in relatedness values between pools. In addition, the GRM values decreased as the pool size increased. For example, for PS15 and greater, the GRM values were more distinct, whereas pools from the PS2 and PS5 categories had a wider range of coefficient values. This suggests that as the pool size increases, the RP data average out individual variations, leading to less variability in GRM estimates (based on empirical pools) while maintaining a high correlation with GRM estimates from individual values. The relationship coefficient values reflect the differences in family diversity. The lower pairwise relatedness between pools, ranging from −0.09 to 0.20, indicated a more diverse family composition across pools. In contrast, high within-pool relatedness, ranging from 0.20 to 0.45, suggested that individuals within the same pool were more closely related, likely due to fewer contributing families (Supplementary Fig. 3).

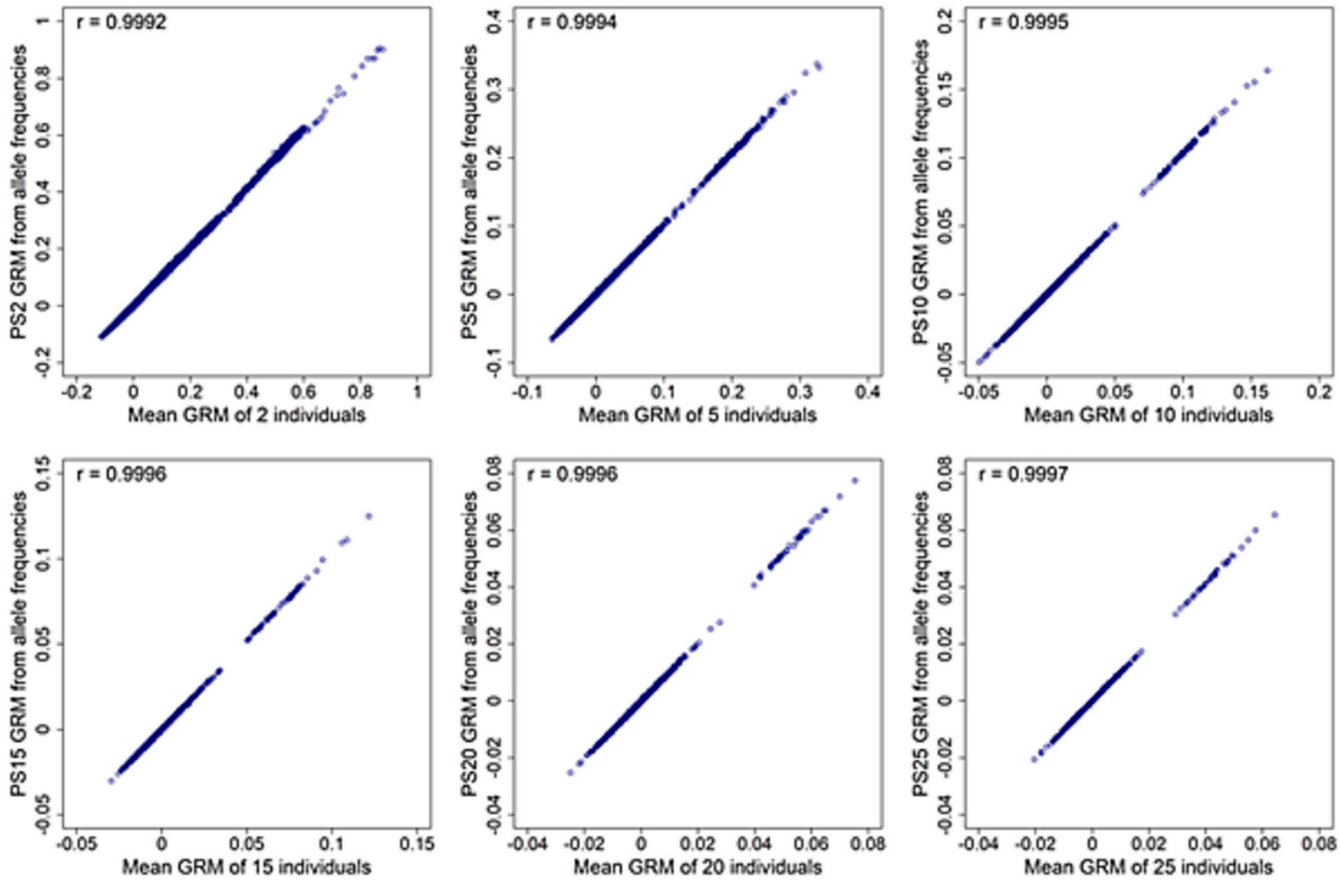


Fig. 3. Comparison of the relationship coefficient values between the mean GRM of individuals and B-allele frequency estimated for the respective pool sizes (PS).

3.2.2. gEBV estimation and accuracy of reconstructed pools

In scenario S1, the p-gEBV prediction accuracies were compared between gEBVs predicted from individual records and RPs. The variations in accuracy among the different PSs are shown in Fig. 4A. A prominent decreasing trend in accuracy loss was observed across the different ponds as a higher PS was generated. The accuracy ranged from 0.86 for PS2 to 0.57 for PS25. In the phenotype rank pools, smaller pool sizes (< PS10) were observed with high prediction accuracies. Remarkably, there was a significant difference ($P < 0.05$) in prediction accuracy between ponds for each PS. The accuracy significantly decreased below 0.6 ($P < 0.01$) for PS 25, particularly in ponds 152 and 160.

A statistical approach using the Weighted Sum Model (WSM) was used to determine the optimal pool size based on the prediction accuracy and test cost reduction. Cost reduction is the cost savings for each pool size, and it is estimated to reduce the number of test individuals. Each PS is ranked on a weighted score, calculated by weighting the prediction accuracy at 60% and the cost reduction at 40% (Supplementary Table 2). Accuracy was given the higher weight because loss of reliability directly compromises selection response and long-term genetic gain, whereas cost savings, although operationally relevant, are a secondary constraint rather than the primary breeding objective. PS5 had the highest weighted score of 0.74, whereas PS10 appeared to be a balanced choice with moderate accuracy and cost reduction.

Individual genotyping produced the highest prediction accuracy (0.68), with a rapid decline beyond PS10; PS10 yielded 0.59 accuracy (13% lower than individual genotyping), whereas PS15 and PS25 fell to 0.55 and 0.49, respectively. When expressed as cost per 1% accuracy, PS10 delivered a 30–35% improvement relative to individual genotyping, Pool size 10 save costs by 90% and is four times more efficient than PS5 when considering cost per 1% increase in accuracy. Although PS15 is marginally less expensive, the associated decline in prediction accuracy indicates that PS10 represents the more optimal choice. (Fig. 4B). From the comparisons of p-gEBV accuracies based on individual data versus RP and weighted scores, we determined that PS10 was the optimal pool size ($r > 0.50$) with individual data, while reductions in prediction accuracy (accuracy ≥0.75) and lower costs were incurred from the total test cost.

3.3. Scenario 2: sibling predictions

The prediction of the sibling genomic breeding value (s-gEBV) was evaluated using the GBLUP model across different PS. Validation was performed using two datasets: individual pond data (S2a), where pools were confined to a single pond, and a combined dataset across all ponds; in S2b, prediction accuracy was assessed for target individuals within the top 20% of the phenotypic distribution.

3.3.1. Scenario 2a: random sibling prediction

3.3.1.1. Individual ponds. The s-gEBV prediction accuracy across ponds exhibited distinct trends with increasing PS numbers (Fig. 5A). Pool sizes PS2 and PS5 consistently resulted in higher accuracy, but the rate of accuracy decline varied between ponds as the PS increased. We observed that average additive genetic variance greatly varied from 5.4 in PS2 to 2.2 for PS25, while average phenotypic variance increased from 9.8 in PS2 to 22.9 for PS25. Ponds 149, 150, and 155 exhibited relatively similar prediction accuracy trends, with a gradual and consistent decline as PS increased. In these ponds, the accuracy remained above 0.40, even for the largest pool size (PS25). In contrast, ponds 152, 156, and 157 exhibited more pronounced accuracy loss, particularly for PS20 and PS25. At PS25, the simulated hybrid data for ponds 152, 157, and 160 exhibited substantially greater variability than the corresponding observed values (Fig. 5A). Pond 152 experienced a rapid decline from 0.65 in PS2 to 0.35 in PS25. Similarly, ponds 156 and 157 exhibited higher variability, as indicated by the larger error at PS20 and PS25. These trends suggest that the accuracy was more sensitive to the pooling effects of these ponds, likely because of the lower sample size and reduced family diversity. Pond 160 also exhibited a steep decline, with an accuracy dropping below 0.30 at PS25, further highlighting the impact of small sample sizes and skewed family frequency distributions.

3.3.1.2. All ponds. In the test scenario (S2a), a random s-gEBV was predicted using all pond data, and the model was cross-validated. The target 10% of sib weights were randomly masked and predicted their breeding values from the entire cohort sample (combining all ponds) using different PS across the phenotype rank distribution (Fig. 5A). The

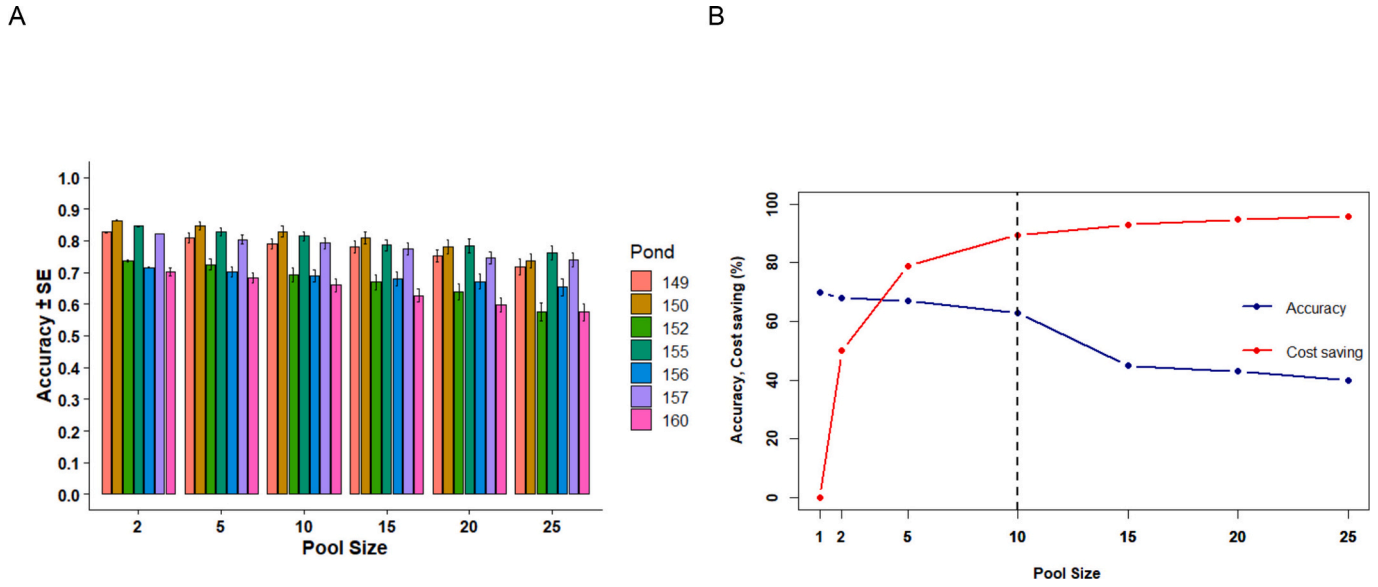


Fig. 4. A). Prediction accuracy of parental genomic breeding values (p-gEBV) in different pool sizes (PS). The X-axis represents the range of PS. The Y-axis represents the accuracy between p-gEBV derived from RP-AF and individual genotype data, while the coloured bars denote specific ponds, illustrating the variation in prediction across multiple ponds. B) Comparison between the percentage of genotyping cost and the percentage of genomic estimated breeding accuracy attained in each pool size (individual genotyping used as base line). The dashed line indicated for pool size that retains prediction accuracy at low marginal cost erosion.

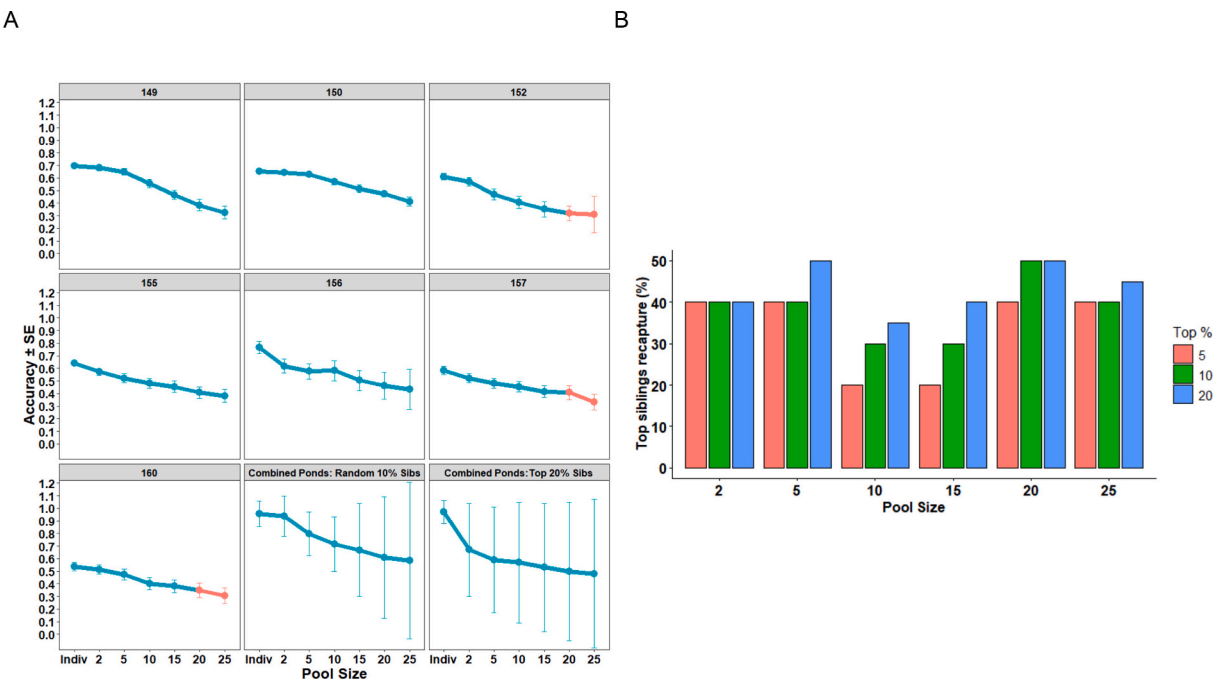


Fig. 5. A). Change in prediction accuracy of siblings randomly predicted from pooled siblings in across phenotype rank distribution (test scenario, S2a) in individual the ponds. (The blue line indicates the original data, and the red lines indicate the pools supplemented with simulated data). Prediction accuracy of random 10% and top 20% siblings compared to the increasing number of pools of 2 to pools of 25 individuals when all ponds were combined (test scenario S2b) **B)** Frequency of recapturing high-ranking sibling gEBV (s-gEBVs) when predicting from different pool sizes (PS). Different colours represent the top 5, top 10, and top 20 quantiles of the masked test data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

s-gEBV prediction accuracy was higher at 0.94 for PS2, and the accuracy further decreased to 0.79 for PS5 and 0.58 for PS25. The accuracy decreased from 0.9 to 0.5, and a corresponding increase in the standard error from 0.3 to 0.7 was observed for the pool size from PS2 to PS25.

3.3.2. *Scenario 2b: top 20% sibling prediction*

The re-prediction analysis confirmed the model's efficacy in identifying top rank masked s-gEBVs. The top 20% of the sibling gEBV prediction accuracies using the RP results are listed in Fig. 5A. The prediction accuracy varied by 0.67 in PS2 and 0.48 in PS25. Higher standard errors were observed between different pool sizes, with moderate to low prediction accuracy, primarily because of the limited ability of the prediction model to capture higher-ranked individuals within the training population. Fig. 5B illustrates the re-identification of top-ranked s-gEBVs and the distribution of predicted values for masked individuals. The analysis revealed that about 40% of masked individuals maintained their gEBV ranks within the top 5%, 10%, and 20% of the phenotype-based rank distribution when 10% of the top 20% siblings were masked and re-evaluated across various pool sizes.

3.4. *Scenario 3: physical DNA pooling*

3.4.1. *Accuracies of allele frequency estimation from physical DNA pools*

The PDP-AF estimated from the Axiom array generated X and Y intensities that were compared against the individual representatives of 14 pools (Supplementary Fig. 4**)**. This process ensured the accuracy and reliability of the method used for allele frequency estimation. The regression trend showed a positive correlation ($r = 0.96$), indicating that the pooled method accurately estimated the allele frequencies of individuals within the pool. There were noticeable large deviations in the extreme allele frequencies in the pooled-based estimation owing to the different relationships between individuals and pools.

Allele frequencies from the X and Y intensities of the PDP samples were estimated using three technical replicates for each sample. The average concordance of allele frequency estimates between these replicates was high (99.4%), indicating high repeatability across replicate sets within these datasets (Supplementary Fig. 5). However, for two pools (pool 2 and pool 8) across the distribution of pond 150, one of the replicates had higher root mean square deviation (RMSD) values than the other two pools within the respective triplicates (RMSD of 0.07 compared to an average RMSD of 0.04). Because this variation was only slight, the data from replicates 1 and 3 were deemed consistent and were used for further downstream analysis. Consensus DNA pool allele frequencies were calculated as the average of the three replicates for each pool.

To evaluate the accuracy of the GRM estimation, the coefficient values calculated from the PDP-AF were compared with the GRM values calculated from the RP-AF using Pearson correlation coefficients. Positive correlations were observed ($r = 0.87$ and 0.83, $P < 0.05$) for both the moderate- and low-diversity ponds (ponds 150 and 152, respectively; Fig. 6A and B).

3.4.2. *gEBV prediction accuracy of physical DNA pools*

A comparison of the gEBV prediction accuracies between RP and PDP for the ponds differed with high and low genetic diversity is summarised in Table 5. The difference between the RP and PDP was more pronounced for the low genetic diversity pond. A two-way ANOVA indicated that within the moderate-genetic diversity pond, the mean accuracy of RP was 11% higher (significant, $P < 0.04$) than that of PDP, with accuracy difference of 0.06. The accuracy was significantly lower in the low-diversity pond for both RP and PDP. Conversely, within the low-diversity pond, the difference was higher 23% at significant $P < 0.02$. Within the RP, mean accuracy differed by 0.03 (5%) between moderate and low genetic diversity ponds. Within PDP, this difference was larger, at 0.09 (18%; $P < 0.03$). This indicates that predictions in the low-genetic diversity pond varied between the two allele frequency estimation workflows, with RP-AF derived from DArTCap genotypes and PDP-AFs estimated from Axiom data.

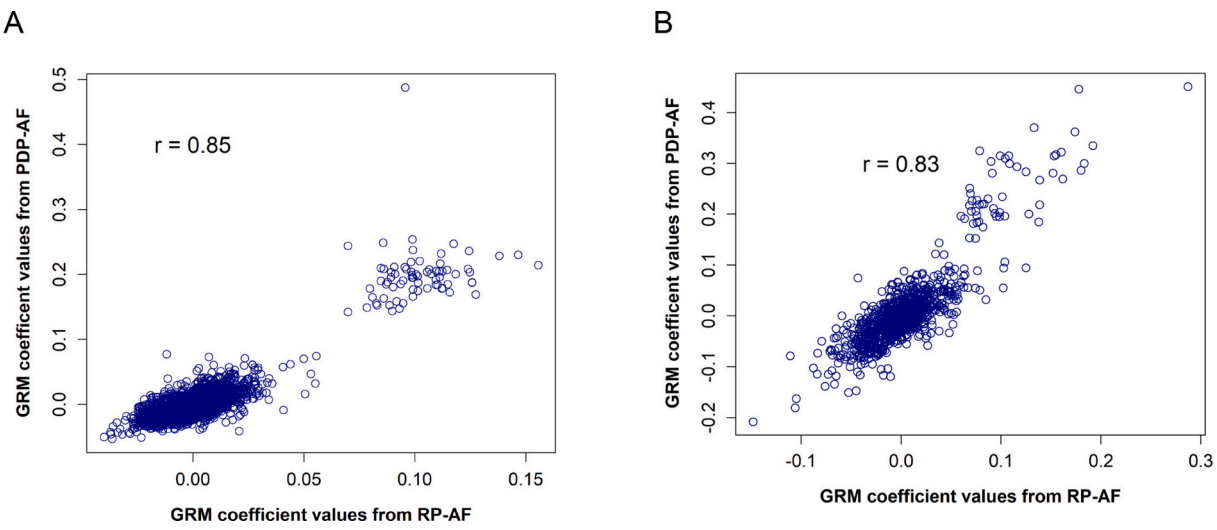


Fig. 6. Accuracy of the genomic relationship coefficients calculated from RP-AF and PDP-AF for a pool size of 10. (A) Pond 150: a large, diverse group of individuals in pools. (B) Pond 152: Most of the individuals had close relatives.

Table 5
Mean gEBV prediction accuracies for siblings using RP (DArTCap) and PDP (Axiom array). All accuracies were based on the pool size of 10 individuals.

Type	RP gEBV accuracy ± SE	PDP gEBV accuracy ± SE	Differences between RP and PDP
Moderate diversity (Pond 150)	0.591 ± 0.05	0.528 ± 0.013	0.06 (11%)
Low diversity (Pond 152)	0.563 ± 0.09	0.434 ± 0.01	0.13 (23%)
Difference between ponds	0.03 (5%)	0.09, (18%)	

4. Discussion

Genomic prediction using DNA pooling methodologies was performed to mitigate the requirement for individual genotyping in the genomic selection application of a black tiger shrimp breeding program. An evaluation of various pool sizes revealed that PS5 had a minimal impact on predictive accuracy, while a PS10 weighted score of 0.73 indicated the optimal balance between cost efficiency and preservation of accuracy.

4.1. Pooling strategies

The application of genomic selection in aquatic species requires significant capital investment for a large number of individuals to be genotyped, compared to terrestrial animal production (Kriaridou et al., 2020; Lhorente et al., 2019). However, the integration of pooling strategies into genomic selection workflows is emerging as a cost-effective approach for predicting individual performance (Dagnachew et al., 2022; Henshall et al., 2014; Kriaridou et al., 2020; Sui et al., 2020; Vargas Jurado et al., 2021). This study investigated the potential of applying a cost-effective DNA pooling method to predict the gEBVs of body weight in *P. monodon* using commercial, on-farm phenotypic, and genotypic data across different pool sizes (PS2, PS5, PS10, PS15, PS20, and PS25). Across ponds, prediction accuracy was primarily driven by family diversity and representation. Low-diversity ponds (152, 157 and 160) exhibited greater variability in gEBVs as pool size increased, reflecting skewed family contributions across the phenotypic distribution. In contrast, high-diversity ponds (149 and 150) maintained stable family frequencies even at smaller pool sizes, indicating that balanced genetic representation stabilises allele frequency estimation and prediction.

In scenario S1, phenotypic rank pools improved p-gEBV prediction accuracy with PS10 resulting in the highest accuracy (> 0.8). In this study, p-gEBV prediction accuracies consistently exceeded s-gEBV accuracy, which declined from 0.70 at PS2 to 0.35 at PS25. The higher parent accuracy reflects direct genetic contribution, whereas sibling prediction is diluted by environmental noise and reduced genetic representation within larger pools. A similar pattern of discrepancies between parents and siblings has been observed in cattle genomic predictions using phenotype based rank pooled genotype (Alexandre et al., 2019; Alexandre et al., 2020; Bell et al., 2017). Conversely, pools that are predominantly characterised by one or two families showed reduced accuracy due to limited genetic diversity and overfitting family specific effects. Such patterns were evident in ponds 152 and 157, whereby ponds 149, 150, and 155, comprised diverse family compositions resulted in improved accuracy. These results verified that when the number of families contributing to a pool is limited, the genomic prediction model becomes skewed towards specific family effects.

4.2. Factors influencing gEBV prediction accuracy using DNA pools

The accuracy of gEBV prediction for pools is influenced by several factors, such as genetic diversity within pools, the number of pools, genetic architecture of the trait, and a strong relationship between genomic training and the test population (Dekkers, 2004; Nadeau et al., 2023; Omeka et al., 2024). In commercial aquaculture populations, variations in prediction accuracy can be attributed to differences in mean phenotypes among families (Khalilisamani et al., 2021). In scenarios S2a and S2b, prediction accuracy influenced from relatedness between training and test data sets. We observed relatively stable prediction accuracy in S2a (individual ponds) this attributed to the inclusion of individuals across the population distribution preserved phenotypic and genetic variance, as a result, genomic breeding values were more stable and less sensitive to random sampling error. In contrast, the top-ranked subset (S2b) observed unstable prediction accuracy. This is due to reduction in additive genetic variance and decreased the covariance between marker effects and phenotypes meant that prediction within the top-rank subset more relied on stochastic within-family variation rather than between-family variation, consequently larger standard error of prediction accuracy. These scenario results indicate that maintaining broad family inclusion across the performance distribution is essential for robust genomic prediction.

The PDP-gEBV accuracy is influenced by the number of pools used to

train the model (Henshall et al., 2012). Our results demonstrate that RP datasets with varying sample sizes and family contributions affect gEBV prediction accuracy. In the case of the pond sample size with $n = 800$ records and over 35 families (ponds 149, 150, and 155), the contributions achieved gEBV-prediction accuracy of 0.55 for PS10. In contrast, in low-diversity ponds with fewer (⟨30) families (ponds 152,157, and 160), gEBV accuracies were less than 0.5. This finding suggests that diverse genetic representations within pooled samples contribute to more reliable gEBV predictions.

The influence of pool size was further explained by changes in variance structure. In our dataset, large pool sizes displayed increase of phenotypic variance more rapidly (i.e. from 9.8 in PS2 to 22.8 for PS25) than the additive genetic variance (i.e. 5.4 in PS2 to 2.2 in PS25), reducing the proportion of genetic variance captured by the prediction model. This inflation of variability reduces confidence in individual ranking, particularly in ponds with limited family diversity. In addition, when pools were ranked by phenotype, the predictions from larger pools became more inconsistent because averaging across many individuals results in larger standard deviations. These results were consistent with livestock pooling studies, which found that accuracy was impacted by high phenotypic variation within the pools (Alexandre et al., 2020; Baller et al., 2020).

A key finding was the wide variation in the standard error (SE) of gEBV prediction accuracy across pool sizes, with a consistent upward trend as the pool size increased in S2b. The higher SE was primarily due to the greater variability introduced when more individuals were pooled. In this study, we observed that an increased number of pools, when combined with multiple ponds, resulted in a standard error (SE) of 0.3 for PS2, with PS25 reaching a SE more than 1 (e.g., ±1.59 for PS25). Larger pools introduced greater heterogeneity in family contribution and allele frequency estimation, particularly where the number of pools was reduced. Consequently, accuracy in breeding value prediction declined, leading to increased variability and reduced reliability of genomic predictions. (Dagnachew et al., 2022). This effect was most pronounced in PS15 and larger pools with limited family diversity in the training set. Although pooling reduces genotyping costs, excessively large pools compromise ranking confidence. In applied breeding programmes, optimising pool size, standardising DNA contributions, and incorporating pool replication are essential to balance cost efficiency against predictive reliability.

4.3. Impact of SNP markers

The number of SNPs used in this study varied from pond to pond, and the heritability estimates varied according to the average MAF per pond. The call rate threshold levels were imposed before imputation, and an average of 3100 SNPs were imputed (Table 2). A higher number of missing genotypes reduced the predictive correlation (Supplementary Data Fig. 1). This supports studies where the imputation of missing genotypes improves the performance of genomic data analysis (Clouard et al., 2022). We found that increasing the number of SNPs beyond 3000 had a minimal impact on heritability estimation and correlations of predicted gEBV (Supplementary Fig. 2). This finding aligns with the observation that the number of SNPs influences genomic prediction accuracy in the training population (Dagnachew et al., 2022; Khalilisamani et al., 2022). Notably, prediction accuracy improves when marker effects are associated with quantitative traits (Sonesson et al., 2010) and incorporating trait-associated markers can further enhance this accuracy (Dagnachew et al., 2022). Therefore, within moderately dense markers, prediction accuracy depends more on SNP informativeness than on marker density.

A considerable proportion of missing genotype calls and monomorphic SNP markers was observed in the DArTCap genotype data. As a result, platform-specific differences in genotype completeness and marker informativeness were evident when comparing Axiom and DArTCap data for S3. Under these conditions, the correlation between GRM estimates remained high ($r = 0.8$; Fig. 6A and B), indicating only a minor reduction in concordance between platforms. This suggests that the remaining SNPs retained sufficient genetic relationship coverage for genomic prediction, with no substantial loss of prediction accuracy. Nevertheless, the prediction accuracy for S3 s-gEBV remained below 0.6. This likely reflects compounding errors from DNA pool standardisation and the effective informative SNP effects on the trait. However, identifying the specific factors that lead to individual variations within pooled samples is a significant challenge. Hence, these findings underline the importance of quality control in genotyping workflows, especially when integrating different genotype platforms. Furthermore, the quality and quantity of DNA are essential determinants for the formulation of PDP experiments. Therefore, for practical applications, careful consideration of pool design is vital to mitigate potential quality constraints.

Calculating allele frequencies from individual genotype data led to higher accuracy than calculating allele frequencies from the physical DNA pool. DNA pooling can introduce inaccuracies in allele frequency estimates, which may originate from various sources, including discrepancies between genotyping platforms (e.g. arrays versus sequencing) and the number of individuals incorporated into each pool. Several reports have highlighted the variability in genotyping precision between arrays and sequencing methodologies, with sequencing typically providing superior accuracy in identifying low-frequency variants, while being more vulnerable to biases due to sequencing depth and coverage (Dagnachew et al., 2022; Keele et al., 2021). Conversely, arrays yield more consistent and reproducible genotype calls but may overlook rare alleles because of their dependence on predefined SNP panels. Such disparities can result in bias in allele frequency estimations when pooling strategies are used.

4.4. Limitations of uniform family representation

In the current study, the accuracy was influenced owing to the lack of a uniform number of families and an equal number of pools. However, such homogeneous data may be impractical for intensive commercial farming operations. Moreover, a potential limitation is typical in the early domestication of wild *P. monodon,* which faces reproductive challenges that impede the production of a large number of families in a short time frame.

4.5. Recommendations

From an economic perspective, DNA pooling offers a cost-efficient alternative; however, our findings show a clear trade-off between genotyping expenditure and prediction accuracy: increasing the pool size reduce prediction accuracy and increase operational complexity, even though they lower costs and require fewer pools; while smaller pools maintain accuracy with less cost savings. Assuming the laboratory inputs required to obtain SNP information for 1000 individuals, two genotyping scenarios were evaluated: individual genotyping and pooled genotyping. Individual genotyping requires one reaction per animal, so 1000 animals require 1000 reactions, whereas pooling in groups of 10 reduces this to 100 reactions, lowering the SNP genotyping cost by ~90%. This study estimating breeding value using DNA pools demonstrated that PS 10 provided the most favourable balance between accuracy and reduced number of genotyping individuals, resulting in substantial cost savings with only moderate accuracy loss relative to individual genotyping and greater cost-efficiency than PS5. Although PS15 reduced costs slightly further, prediction accuracy declined sharply, therefore PS10 is optimal choice for large population.

DNA pooling can be economically advantageous under specific conditions in commercial breeding programmes. Lower prediction accuracy can reduce the rate of genetic gain by decreasing the precision of selection decisions, particularly skewed distribution and traits with low to moderate heritability. However, pooling may remain beneficial when genotyping budgets constrain the size of training populations, as pooling

allows more individuals to be evaluated and can increase selection intensity. Pooling may also support the expansion of training reference populations and facilitate genomic auditing of large breeding cohorts, where pooled allele frequency estimates provide a cost-effective approach for monitoring genetic diversity and population structure. In such situations, pooling enables the evaluation of larger populations, which may partially offset reductions in prediction accuracy, particularly when combined with targeted individual genotyping of top performing candidates. Pooling should therefore be considered a program-specific optimisation strategy, and breeding programmes should assess whether pooling-enabled increases in selection intensity compensate for the associated reduction in prediction accuracy in order to maintain or improve genetic gain per unit cost.

In commercial shrimp breeding, DNA pooling methodologies identify high-performing siblings through a structured approach. First, diverse full-sib families must be identified within the breeding cohort to ensure genetic variability and selection precision. Individuals are then aggregated based on pond performance indicators, including body weight, survival rate, disease resistance, feed conversion efficiency, and environmental adaptability. The gEBVs are estimated from individual DNA, and pools are formed with optimal size for cost-effectiveness and accuracy. Pool sizes $\leq$10 recommended while scope of increasing pool numbers, as larger pool size (>10) resulted in reduce predictive accuracy. Robust statistical prediction methods must evaluate family contribution variations and genetic correlation biases to refine selection. These frameworks enable shrimp breeding programs to use DNA pooling for improved broodstock selection while minimizing costs and maintaining genetic gains.

5. Conclusion

In conclusion, the efficiency of DNA pooling was tested to predict the genetic merit of black tiger shrimp growth. We performed phenotype rank distribution pooling strategies for parents' and siblings' body weight predictions with skewed family distributions under commercial conditions. We demonstrated that the PDP genotyping approach provides an optimum pool size, reliable on-farm application, and cost-effective means of predicting the phenotype of parents and siblings in a commercial cohort. A low-cost DNA pool genotyping method (over 3000 SNPs) for predicting breeding value was used, and its accuracy was maintained at the loss of accuracy compared to individual genotyping for 5000 individuals. Despite the higher standard errors associated with genetic diversity and varying numbers of family frequencies within the pools, we suggest that uniform families and mean family phenotype variations should be considered when designing pools. Moreover, findings from the PDP analysis require validation in larger and more biologically diverse datasets to rigorously evaluate the potential influence of genetic diversity on prediction accuracy. Overall, better prediction of breeding value using a cost-effective method could be of potential value for most commercial breeding programs for aquaculture-farmed species.

CRediT authorship contribution statement

Gopala K. Guddanti: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Cecile Massault:** Writing – review & editing, Visualization, Supervision, Investigation, Conceptualization. **David B. Jones:** Writing – review & editing, Visualization, Supervision, Investigation, Conceptualization. **Dean R. Jerry:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Kyall R. Zenger:** Writing – review & editing, Visualization, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no competing financial interests or personal relationships that could influence the work reported in this study.

Acknowledgements

This research was supported by the James Cook University Postgraduate Research Scholarship (JCUPRS) and the Food Agility Pvt. Ltd. HDR top-up scholarship. The authors acknowledge the assistance of the ARC ITRH Research Hub Program (IH130200013) and in-kind with support from the ARC ITRH Supercharging Tropical Aquaculture through Genetic Solutions (IH210100014). The authors participated in the project design and conceptual development, data analyses, and editing of the manuscript. The authors thank Nick Wade for his support in acquiring samples and Emmanuelle Botté from Manuscribe for editorial support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aquaculture.2026.743971.

Data availability

Data will be made available on request.

References

Aldridge, M.N., Marjanovic, J., Henshall, J.M., de Klerk, B., Peeters, K., de Haas, Y., 2022. DNA pooling is a cost effective method of including commercial crossbred data in selection of purebreds. In: Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP), p. 781. https://doi.org/10.3920/978-90-8686-940-4_182.

Alexandre, P.A., Porto-Neto, L.R., Karaman, E., Lehnert, S.A., Reverter, A., 2019. Pooled genotyping strategies for the rapid construction of genomic reference populations1. J. Anim. Sci. 97 (12), 4761–4769. https://doi.org/10.1093/jas/skz344.

Alexandre, P.A., Reverter, A., Lehnert, S.A., Porto-Neto, L.R., Dominik, S., 2020. In silico validation of pooled genotyping strategies for genomic evaluation in Angus cattle. J. Anim. Sci. 98 (6). https://doi.org/10.1093/jas/skaa170.

Baller, J.L., Kachman, S.D., Kuehn, L.A., Spangler, M.L., 2020. Genomic prediction using pooled data in a single-step genomic best linear unbiased prediction framework. J. Anim. Sci. 98 (6). https://doi.org/10.1093/jas/skaa184.

Bell, A.M., Henshall, J.M., Porto-Neto, L.R., Dominik, S., McCulloch, R., Kijas, J., Lehnert, S.A., 2017. Estimating the genetic merit of sires by using pooled DNA from progeny of undetermined pedigree. Genet. Sel. Evol. 49 (1), 28. https://doi.org/10.1186/s12711-017-0303-8.

Butler, D.G., Cullis, B.R., Gilmour, A.R., Gogel, B.G., Thompson, R., 2023. ASReml-R Reference Manual Version 4.2. VSN International Ltd., Hemel Hempstead, HP2 4TP, UK. https://asreml.kb.vsni.co.uk/wp-content/uploads/sites/3/ASReml-R-Reference-Manual-4.2.pdf.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4 (1), s13742–015.

Clouard, C., Ausmees, K., Nettelblad, C., 2022. A joint use of pooling and imputation for genotyping SNPs. BMC Bioinform. 23 (1), 421. https://doi.org/10.1186/s12859-022-04974-7.

Dagnachew, B., Aslam, M.L., Hillestad, B., Meuwissen, T., Sonesson, A., 2022. Use of DNA pools of a reference population for genomic selection of a binary trait in Atlantic salmon. Front. Genet. 13, 896774. https://doi.org/10.3389/fgene.2022.896774.

Dekkers, J.C., 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J. Anim. Sci. 82 (suppl_13), E313–E328.

Domingos, J.A., Smith-Keune, C., Jerry, D.R., 2014. Fate of genetic diversity within and between generations and implications for DNA parentage analysis in selective breeding of mass spawners: a case study of commercially farmed barramundi, *Lates calcarifer*. Aquaculture 424-425, 174–182. https://doi.org/10.1016/j.aquaculture.2014.01.004.

Foote, A., Simma, D., Khatkar, M., Raadsma, H., Guppy, J., Coman, G., Giardina, E., Jerry, D., Zenger, K., Wade, N., 2019. Considerations for maintaining family diversity in commercially mass-spawned penaeid shrimp: a case study on *Penaeus monodon*. Front. Genet. 10, 1127. https://doi.org/10.3389/fgene.2019.01127.

Fugeray-Scarbel, A., Bastien, C., Dupont-Nivet, M., Lemarié, S., 2021. Why and how to switch to genomic selection: lessons from plant and animal breeding experience. Front. Genet. 12. https://doi.org/10.3389/fgene.2021.629737.

Gjedrem, T., 2005. Selection and Breeding Program in Aquaculture. Springer. https://doi.org/10.1007/1-4020-3342-7_12.

Gjedrem, T., Robinson, N., 2014. Advances by selective breeding for aquatic species: a review. Agric. Sci. 05 (12), 1152–1158. https://doi.org/10.4236/as.2014.512125.

Gjedrem, T., Rye, M., 2018. Selection response in fish and shellfish: a review. Rev. Aquac. 10 (1), 168–179. https://doi.org/10.1111/raq.12154.
Goddard, M.E., Wray, N.R., Verbyla, K., Visscher, P.M., 2009. Estimating effects and making predictions from genome-wide marker data. Stat. Sci. 24 (4), 517–529. https://doi.org/10.1214/09-sts306.
Goddard, M.E., Hayes, B.J., Meuwissen, T.H., 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128 (6), 409–421. https://doi.org/10.1111/j.1439-0388.2011.00964.x.
Granato, I.S.C., Galli, G., de Oliveira Couto, E.G., e Souza, M.B., Mendonça, L.F., Fritsche-Neto, R., 2018. SnpReady: a tool to assist breeders in genomic analysis. Mol. Breed. 38 (8). https://doi.org/10.1007/s11032-018-0844-8.
Guppy, J.L., Jones, D.B., Kjeldsen, S.R., Le Port, A., Khatkar, M.S., Wade, N.M., Zenger, K.R., 2020. Development and validation of a RAD-Seq target-capture based genotyping assay for routine application in advanced black tiger shrimp (Penaeus monodon) breeding programs. BMC genomics 21 (1), 541.
Hayes, B.J., Bowman, P.J., Chamberlain, A.J., Goddard, M.E., 2009. Invited review: genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92 (2), 433–443. https://doi.org/10.3168/jds.2008-1646.
Henshall, J.M., Hawken, R.J., Dominik, S., Barendse, W., 2012. Estimating the effect of SNP genotype on quantitative traits from pooled DNA samples. Genet. Sel. Evol. 44 (1), 12. https://doi.org/10.1186/1297-9686-44-12.
Henshall, J.M., Dierens, L., Sellars, M.J., 2014. Quantitative analysis of low-density SNP data for parentage assignment and estimation of family contributions to pooled samples. Genet. Sel. Evol. 46, 51. https://doi.org/10.1186/s12711-014-0051-y.
Jerry, D.R., Jones, D.B., Lillehammer, M., Massault, C., Loughnan, S., Cate, H.S., Harrison, P.J., Strugnell, J.M., Zenger, K.R., Robinson, N.A., 2022. Predicted strong genetic gains from the application of genomic selection to improve growth related traits in barramundi (*Lates calcarifer*). Aquaculture 549. https://doi.org/10.1016/j.aquaculture.2021.737761.
Kalinowski, S.T., Taper, M.L., Marshall, T.C., 2007. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol. Ecol. 16 (5), 1099–1106.
Keele, J., McDaneld, T., Lawrence, T., Jennings, J., Kuehn, L., 2021. Estimation of pool construction and technical error. Agriculture 11 (11). https://doi.org/10.3390/agriculture11111091.
Khalilisamani, N., Thomson, P.C., Raadsma, H.W., Khatkar, M.S., 2021. Impact of genotypic errors with equal and unequal family contribution on accuracy of genomic prediction in aquaculture using simulation. Sci. Rep. 11 (1), 18318. https://doi.org/10.1038/s41598-021-97873-5.
Khalilisamani, N., Thomson, P.C., Raadsma, H.W., Khatkar, M.S., 2022. Estimating heritability using family-pooled phenotypic and genotypic data: a simulation study applied to aquaculture. Heredity (Edinb) 128 (3), 178–186. https://doi.org/10.1038/s41437-022-00502-8.
Khatkar, M.S., 2017. Genomic selection in aquaculture breeding programs. In: Book: Bioinformatics in Aquaculture Breeding.
Kinghorn, B.P., Bastiaansen, J.W.M., Ciobanu, D.C., Van Der Steen, 2010. Quantitative genotyping to estimate genetic contributions to pooled samples and genetic merit of the contributing entities. Acta Agric. Scand. Sect. A 60 (1), 3–12.
Kriaridou, C., Tsairidou, S., Houston, R.D., Robledo, D., 2020. Genomic prediction using low density marker panels in aquaculture: performance across species, traits, and genotyping platforms. Front. Genet. 11, 124. https://doi.org/10.3389/fgene.2020.00124.
Lhorente, J.P., Araneda, M., Neira, R., Yáñez, J.M., 2019. Advances in genetic improvement for salmon and trout aquaculture: the Chilean situation and prospects. Rev. Aquac. 11 (2), 340–353. https://doi.org/10.1111/raq.12335.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25 (14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
Massault, C., Jones, D.B., Zenger, K.R., Strugnell, J.M., Barnard, R., Jerry, D.R., 2021. A SNP parentage assignment panel for the silver lipped pearl oyster (Pinctada maxima). Aquacult. Rep. 20. https://doi.org/10.1016/j.aqrep.2021.100687.
Nadeau, S., Beaulieu, J., Gezan, S.A., Perron, M., Bousquet, J., Lenz, P.R.N., 2023. Increasing genomic prediction accuracy for un-phenotyped full-sib families by modeling additive and dominance effects with large datasets in white spruce. Front. Plant Sci. 14, 1137834. https://doi.org/10.3389/fpls.2023.1137834.
Nilforooshan, M.A., 2022. PedSimulate – an R package for simulating pedigree, genetic merit, phenotype, and genotype data. Rev. Bras. Zootec. 51. https://doi.org/10.37496/rbz5120210131.
Omeka, W.K.M., Liyanage, D.S., Lee, S., Udayantha, H.M.V., Kim, G., Ganeshalingam, S., Jeong, T., Jones, D.B., Massault, C., Jerry, D.R., Lee, J., 2024. Genomic prediction model optimisation for growth traits of olive flounder (Paralichthys olivaceus). Aquacult. Rep. 36. https://doi.org/10.1016/j.aqrep.2024.102132.
Peiris, B.L., Ralph, J., Lamont, S.J., Dekkers, J.C., 2011. Predicting allele frequencies in DNA pools using high density SNP genotyping data. Anim. Genet. 42 (1), 113–116. https://doi.org/10.1111/j.1365-2052.2010.02077.x.
Reverter, A., Henshall, J.M., McCulloch, R., Sasazaki, S., Hawken, R., Lehnert, S.A., 2014. Numerical analysis of intensity signals resulting from genotyping pooled DNA samples in beef cattle and broiler chicken. J. Anim. Sci. 92 (5), 1874–1885. https://doi.org/10.2527/jas2013-7133.
Reverter, A., Porto-Neto, L.R., Fortes, M.R.S., McCulloch, R., Lyons, R.E., Moore, S., Nicol, D., Henshall, J., Lehnert, S.A., 2016. Genomic analyses of tropical beef cattle fertility based on genotyping pools of Brahman cows with unknown pedigree. J. Anim. Sci. 94 (10), 4096–4108. https://doi.org/10.2527/jas2016-0675.
Song, H., Hu, H., 2021. Strategies to improve the accuracy and reduce costs of genomic prediction in aquaculture species. Evol. Appl. https://doi.org/10.1111/eva.13262.
Sonesson, A.K., Meuwissen, T.H., Goddard, M.E., 2010. The use of communal rearing of families and DNA pooling in aquaculture genomic selection schemes. Genet. Sel. Evol. 42 (1), 41.
Song, H., Dong, T., Yan, X., Wang, W., Tian, Z., Sun, A., Dong, Y., Zhu, H., Hu, H., 2022. Improving the accuracy of genomic predictions for disease resistance traits in fish using a multiple-trait linear-threshold model. Aquaculture 554. https://doi.org/10.1016/j.aquaculture.2022.738163.
Sui, J., Luan, S., Dai, P., Fu, Q., Meng, X., Luo, K., Cao, B., Kong, J., 2020. High accuracy of pooled DNA genotyping by 2b-RAD sequencing in the Pacific white shrimp, *Litopenaeus vannamei*. PLoS One 15 (7), e0236343. https://doi.org/10.1371/journal.pone.0236343.
Uengwetwanit, T., Pootakham, W., Nookaew, I., Sonthirod, C., Angthong, P., Sittikankaew, K., Rungrassamee, W., Arayamethakorn, S., Wongsurawat, T., Jenjaroenpun, P., Sangsrakru, D., Leelatanawit, R., Khudet, J., Koehorst, J.J., Schaap, P.J., Martins Dos Santos, V., Tangy, F., Karoonuthaisiri, N., 2021. A chromosome-level assembly of the black tiger shrimp (*Penaeus monodon*) genome facilitates the identification of growth-associated genes. Mol. Ecol. Resour. 21 (5), 1620–1640. https://doi.org/10.1111/1755-0998.13357.
Vandeputte, M., Haffray, P., 2014. Parentage assignment with genomic markers: a major advance for understanding and exploiting genetic variation of quantitative traits in farmed aquatic animals. Front. Genet. 5, 432. https://doi.org/10.3389/fgene.2014.00432.
VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91 (11), 4414–4423. https://doi.org/10.3168/jds.2007-0980.
Vargas Jurado, N., Kuehn, L.A., Keele, J.W., Lewis, R.M., 2021. Accuracy of GEBV of sires based on pooled allele frequency of their progeny. G3 Genes Genomes Genet. 11 (11). https://doi.org/10.1093/g3journal/jkab231.
Vu, N.T.T., Zenger, K.R., Guppy, J.L., Sellars, M.J., Silva, C.N.S., Kjeldsen, S.R., Jerry, D. R., 2020. Fine-scale population structure and evidence for local adaptation in australian giant black tiger shrimp (*Penaeus monodon*) using SNP analysis. BMC Genomics 21 (1), 669. https://doi.org/10.1186/s12864-020-07084-x.
Wang, L., Janss, L.L., Madsen, P., Henshall, J., Huang, C.H., Marois, D., Alemu, S., Sorensen, A.C., Jensen, J., 2020. Effect of genomic selection and genotyping strategy on estimation of variance components in animal models using different relationship matrices. Genet. Sel. Evol. 52 (1), 31. https://doi.org/10.1186/s12711-020-00550-w.
Zenger, K.R., Khatkar, M.S., Jones, D.B., Khalilisamani, N., Jerry, D.R., Raadsma, H.W., 2019. Genomic selection in aquaculture: application, limitations and opportunities with special reference to marine shrimp and pearl oysters. Front. Genet. 9, 693. https://doi.org/10.3389/fgene.2018.00693.