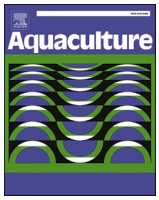
Aquaculture 613 (2026) 743314

Contents lists available at ScienceDirect

Aquaculture

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



Optimisation of genomic selection for harvest traits of Malabar red snapper (*Lutjanus malabaricus*)

Bing Liang [a,b,*], Dean R. Jerry [a,d,f], Vu Nguyen [a], Purushothaman Kathiresan [a,c,e], David B. Jones [d,f], Xueyan Shen [a], Joyce Koh [a,c], Celestine Terence [a], Maria G. Nayfa [a,c], Maura Carrai [a], Rachel Jia Wen Ho [a,c], Hazim Mohamed [a,c], Saraphina Dianne Rwei Qing Tneo [c], Grace Loo [c], Shubha Vij [a,c], Jose A. Domingos [a]

[a] *Tropical Futures Institute, James Cook University Singapore, 149 Sims Drive, 387380, Singapore*
[b] *Marine Aquaculture Centre, Singapore Food Agency, 52 Jurong Gateway Road, JEM Office Tower, #14-01, 608550, Singapore*
[c] *School of Applied Science, Republic Polytechnic, 9 Woodlands Avenue 9, 738964, Singapore*
[d] *ARC Research Hub for Supercharging Tropical Aquaculture through Genetic Solutions, James Cook University, Townsville, QLD 4810, Australia*
[e] *Department of Preclinical Sciences and Pathology, Faculty of Veterinary Medicine, Norwegian University of Life Sciences, Ås 1433, Norway*
[f] *Centre for Sustainable Tropical Fisheries and Aquaculture, College of Science and Engineering, James Cook University, Townsville, QLD, Australia*

ARTICLE INFO

Keywords:
Lutjanus malabaricus
Genomic selection
GBLUP
BayesR
KAML
GWAS

ABSTRACT

Selective breeding in aquaculture has traditionally relied on pedigree-based selection to improve economically important traits. However, these methods have limitations, including lower accuracy, pedigree errors, and an inability to fully capture and exploit within-family genetic variation. Genomic selection (GS) offers a more precise alternative by leveraging genome-wide single nucleotide polymorphisms (SNPs) to calculate genomic estimated breeding values (GEBVs). This study aimed to optimize GS for Malabar red snapper (*Lutjanus malabaricus*), a commercially valuable species in Southeast Asia. We evaluated different SNP densities, training population sizes, and three GS models: GBLUP (genomic best linear unbiased prediction), BayesR (a hierarchical Bayesian model), and KAML (kinship-adjusted multiple-loci linear mixed model). A total of 2547 snappers from three rearing sites were phenotyped and genotyped at harvest (~1.5 years of age) using a custom SNP array. Five traits—body weight (BW), total length (TL), body depth (BD), Fulton's condition factor (K), and body shape index (BSI)—were analysed using six SNP densities (500, 1000, 5000, 10000, 30,000, and 56,378 SNPs) selected randomly, or ranked based on genome-wide association study (GWAS) *p*-values on body weight. We also tested six training population sizes (80, 400, 800, 1200, 1600, and 2038 individuals). Prediction accuracy was assessed via five-fold cross-validation with 10 replicates. Results showed that prediction accuracy improved with larger training population sizes and higher SNP densities, plateauing at 1200 to 1600 training population size and 5000 SNPs for growth traits. GWAS-ranked SNP subsets significantly outperformed corresponding random subsets in predicting GEBVs for body weight when the training and test populations were from the same dataset. GBLUP and KAML achieved similar prediction accuracies for BW and TL (0.45–0.50) and outperformed BayesR (0.40–0.45). In contrast, BayesR performed better for the other traits, yielding the highest accuracies for BD (0.62), K (0.78) and BSI (0.71) using the full dataset. GEBVs from GBLUP and KAML were highly correlated across all traits ($\geq$0.95), whereas BayesR showed strong correlations with GBLUP and KAML for BW and BD (>0.93), moderate correlations for TL and BSI (0.54–0.74), and no significant correlation for K. These findings suggest that using 5000 SNPs and a training population of 1200–1600 individuals provide a cost-effective balance between accuracy and efficiency for growth traits when using the GBLUP model. This study offers key insights for genomic breeding programs in Malabar red snapper, contributing to the sustainability and productivity of aquaculture in Southeast Asia.

* Corresponding author at: Tropical Futures Institute, James Cook University Singapore, 149 Sims Drive, 387380, Singapore.
E-mail address: bing.liang@my.jcu.edu.au (B. Liang).

https://doi.org/10.1016/j.aquaculture.2025.743314
Received 4 June 2025; Received in revised form 6 October 2025; Accepted 16 October 2025
Available online 17 October 2025

1. Introduction

Traditional selective breeding in aquaculture is typically pedigree-based, with families maintained separately until tagging or verified through DNA parentage (e.g., microsatellite or low-density SNP panels) (Gjedrem, 2005). Breeding decisions use phenotypes and genetic relationships, often via pedigree-based BLUP to estimate breeding values, assuming average relatedness from Mendelian expectations (e.g., 50 % for full siblings, 25 % for half-siblings). This ignores variation from Mendelian sampling and recombination, limiting precision. Progress is slow since selection occurs after maturity, and pedigree maintenance is labor-intensive and error-prone, especially in large or hard-to-tag populations. Genomic selection (GS) improves efficiency and accuracy by using high-density SNP markers to capture linkage disequilibrium with causal variants (Meuwissen et al., 2001), enabling combined effects of most QTLs to be estimated and genetic variance among individuals, including siblings, to be more accurately reflected. GS has proven effective for traits such as growth in barramundi (*Lates calcarifer*) (Jerry et al., 2022) and olive flounder (*Paralichthys olivaceus*) (Omeka et al., 2024), growth and sea lice resistance in Atlantic salmon (*Salmo salar*) (Tsai et al., 2016; Tsai et al., 2015), bacterial cold-water disease resistance in rainbow trout (*Oncorhynchus mykiss*) (Vallejo et al., 2017), and growth in yellowtail kingfish (*Seriola lalandi*) (Nguyen et al., 2018). GS is now implemented in a number of high-value finfish, crustaceans and molluscs, accelerating genetic improvement industry-wide (Abdelrahman et al., 2017; Bangera et al., 2025; D'Agaro et al., 2021; Houston et al., 2020; Kudinov et al., 2024; Zenger et al., 2019).

Statistical models are central to GS, enabling breeding values to be estimated by capturing the relationship between markers and traits. Parametric models include ridge regression, GBLUP, kinship-adjusted multiple-loci mixed model (KAML), LASSO, and Bayesian regressions, while non-parametric options include random forest (RF), support vector machines (SVM), neural networks (NN), and other machine learning approaches (Abdollahi-Arpanahi et al., 2020; Wang et al., 2022). GBLUP, the simplest genomic LMM, estimates EBVs from a genomic relationship matrix (VanRaden, 2008) assuming all SNPs have equal variance and normally distributed small effects (Dong et al., 2016; Habier et al., 2007), which suits most polygenic traits but limits accuracy for oligogenic traits. KAML integrates GWAS-derived pseudo-QTNs and trait-specific kinship matrices to accommodate diverse genetic architectures (Yin et al., 2020). LASSO assumes additive marker effects but performs variable selection by shrinking small-effect SNPs to zero. Bayesian models allow heterogeneous variances and non-normal distributions of SNP effects; BayesA assigns each SNP a unique variance, BayesB and BayesCπ set some to zero, and BayesR groups SNPs into large, moderate, small, or negligible effect categories (Moser et al., 2015), performing well for traits influenced by major genes, such as viral nervous necrosis (VNN) resistance in European sea bass (*Dicentrarchus labrax*) (Griot et al., 2021a), or skewed-effect traits in humans (Chatterjee et al., 2013). Non-parametric models make no distributional assumptions, capturing non-linear relationships via decision trees (RF), kernel functions (SVM), or deep learning (NN). In general, parametric models excel for additive architectures, while non-parametric models perform better for traits with epistatic or dominance effects (Abdollahi-Arpanahi et al., 2020; Howard et al., 2014).

SNP marker density and training population size strongly influence GS prediction accuracy, cost efficiency, and breeding program outcomes. Higher SNP density and larger training sets generally improve accuracy (Griot et al., 2021b) by better capturing LD patterns, but substantially increase genotyping costs. High-density panels or whole-genome sequencing offer precise, stable LD detection across generations, enhancing prediction reliability, yet are often impractical for large-scale aquaculture due to expense. Low-density panels are more cost-effective but risk reduced accuracy, particularly for highly polygenic traits or in populations with low LD persistence. For instance, Heath et al. (2010) suggested that low-density SNP panels with low LD between markers and QTLs could result in inaccurate predictions of genetic values for human height. Conversely, in rainbow trout, prediction accuracy was good with only 500 SNPs, attributed to the species' long-range LD (Vallejo et al., 2018). Moreover, prediction accuracy may respond differently to changes in SNP density and population size across statistical models (Wang et al., 2022). Thus, evaluating various SNP marker densities and training population sizes across various models is crucial to ensure GS remains both effective and economically viable, considering the trade-off between genotyping costs and the benefits of improved prediction accuracy.

Feature selection in genomic prediction involves identifying informative markers, such as SNPs, that capture most of the genetic variation for a trait while excluding redundant or irrelevant ones. This can improve prediction accuracy, reduce genotyping costs, and lower computational demands (Scutari et al., 2013). However, selected marker subsets often perform best only under the specific trait, population, or environment used for training, limiting broader applicability (Massault et al., 2025). Selecting SNP subsets based on GWAS *p*-values is a widely used feature selection approach in genomic prediction (Al-Mamun et al., 2025; Luo et al., 2021). Common strategies include selecting SNPs with p-values below predefined thresholds (e.g., 1×10^{-6} or 5×10^{-8}), choosing the top 100 to 10000 GWAS-ranked SNPs, or incorporating SNP–SNP interactions identified from GWAS results into GS.

Malabar red snapper (*Lutjanus malabaricus*) is an emerging aquaculture species in Southeast Asia, yet no formal selective breeding program currently exists for this species. Our previous studies have demonstrated significant potential for improving harvest traits such as growth, body shape and nutritional value through selective breeding initiatives (Liang et al., 2025; Purushothaman et al., 2025). Building on this foundation, the present study aimed to explore and optimize GS approaches to advance genetic improvement programs for Malabar red snapper. In this study, a total of 2547 Malabar red snappers cultured across three rearing sites were genotyped using a custom Axiom™ myDesign™ 70K Red Snapper SNP array (ThermoFisher Scientific™). Harvest traits including body weight (BW), total length (TL), and body depth (BD) were measured, alongside derived traits such as Fulton's condition factor (K) and body shape index (BSI). Three representative models were selected to evaluate GS performance. GBLUP was included as one of the most widely used GS models, serving as a standard benchmark. BayesR was chosen as a representative Bayesian approach, known to deliver similar or higher accuracy than other Bayesian alphabet models (Meher et al., 2025; Xu et al., 2021). Pure machine learning models were excluded because they typically excel at traits influenced by epistatic or dominance effects, whereas the focal trait in this study, growth, is primarily governed by additive genetic effects in many fish species. Instead, KAML was selected to evaluate a LMM framework augmented with machine learning, to test whether it could outperform traditional LMM and to explore its potential for more complex traits such as K and BSI. GS was performed across six training population sizes (80, 400, 800, 1200, 1600, and 2038 samples) and six SNP marker densities (500, 1000, 5000, 10000, 30,000, and 56,378). Both random and GWAS-ranked SNP selection strategies were applied to assess their impact on prediction accuracies. The prediction accuracies of different combinations of models, SNP densities, and training population sizes were analysed and compared to identify the optimal GS approach for each trait. This study provides critical insights into the design of future genetic improvement programs for Malabar red snapper, with the overall goal of enhancing aquaculture productivity and marketability in Singapore and the region.

2. Materials and methods

2.1. Fish rearing, phenotyping, and genotyping

This study used the same dataset as Liang et al. (2025), consisting of genotype and phenotype data from 2547 Malabar red snapper (aged 1.5

years) for five traits: body weight (BW), total length (TL), body depth (BD), Fulton's condition factor (K), and body shape index (BSI). Briefly, 28,000 fingerlings from three hatcheries in Malaysia and Singapore were stocked in January 2022 at two commercial fish farms in Singapore (Farm 1: open net cages; Farm 2: flow-through seawater tanks). An additional 3000 fingerlings from Malaysia were stocked at the Marine Aquaculture Centre (MAC), Singapore, in July 2022. Fish were reared to harvest size over approximately 1.5 years according to each farm's protocols. Seawater parameters were similar across all sites, with temperatures of 29–31 °C, pH 7.9–8.2, salinity 27–30 ppt, and dissolved oxygen levels of 6–8 mg/L at Farm 1 and above 5 ppm at Farm 2 and MAC. Fish were fed commercial pellets containing 43–44 % crude protein two to three times daily. A total of 956 fish from Farm 1, 955 from Farm 2 (sampled in May 2023), and 668 from MAC (sampled in November 2023) were collected for phenotyping and genotyping prior to harvest. BW and TL were measured using a Biomark® phenotyping station, while BD was assessed from standardized photographs using ImageJ (Schneider et al., 2012). Spinal deformities were recorded as a binary trait. Fulton's condition factor (K) was calculated as $K = (10^5 \times BW(g)) / TL(mm)^3$ (Fulton, 1904), and body shape index (BSI) was derived using $BSI = (10 \times BD(mm)) / TL(mm)$ (Domingos et al., 2021). For genotyping, fin clips from all 2579 fish were submitted to Neogen (New Zealand) for DNA extraction and analysis using a custom Axiom™ myDesign™ 70 K Red Snapper SNP array (ThermoFisher Scientific™). Genotyping was conducted on the GeneTitan™ MC Instrument following the Axiom™ Propel XPRES 384HT Workflow. A total of 2547 samples (98.8 %) passed quality control thresholds (DQC $\geq$ 0.82, call rate $\geq$ 0.97), including 946 from Farm 1, 948 from Farm 2, and 653 from MAC. SNP filtering using Axiom Analysis Suite 5.3 and PLINK 2.0 retained 56,378 high-quality SNPs (79.5 % of the initial 70,874) for downstream analyses. In summary, the full dataset comprised phenotype data for five traits and genotype data for 56,378 SNPs from 2547 samples. Both rearing site and deformity status were included as fixed effects in all statistical models to account for environmental differences across the three sites, thereby enabling more accurate and unbiased comparisons of traits and genomic predictions across the combined dataset. To examine the population structure of fish reared at the three sites, principal component analysis (PCA) was conducted using PLINK 2.0 (www.cog-genomics.org/plink/2.0/), and the results were visualized in R with *ggplot2* (Wickham, 2016).

2.2. Genome-wide association study

In this study, within-population GS refers to scenarios where the training and test sets are derived from the same dataset, whereas across-population GS involves the use of distinct datasets for training and testing. To evaluate the effectiveness of GWAS-ranked SNP subsets in both contexts, two separate GWAS analyses were performed for the body weight trait under the following scenarios: **Scenario I** represented within-population GS. Here, 80 % of individuals from the full dataset were randomly assigned to the training set, and the remaining 20 % formed the test set. GWAS was conducted using the entire dataset to rank SNPs based on association with body weight. Genomic estimated breeding values (GEBVs) were then predicted for the test set, with phenotypes masked during prediction. GS accuracy was assessed by correlating predicted GEBVs with observed phenotypic values. **Scenario II** encompassed both within- and across-population GS. In this case, 80 % of individuals from Farm 1 and Farm 2 were used as the training set, while two independent test sets were defined: (1) the remaining 20 % of individuals from Farm 1 and Farm 2 (Test Set 1, representing within-population prediction), and (2) the entire MAC population (Test Set 2, representing across-population prediction). GWAS was performed exclusively on individuals from Farm 1 and Farm 2 to rank SNPs, ensuring that SNP selection for Test Set 2 originated from a genetically distinct population, thus simulating an across-population prediction framework.

Both GWAS analyses employed the GCTA-LOCO (leave-one-chromosome-out) method to identify significant QTLs and rank SNPs based on *p*-value for subsequent subset selection (Yang et al., 2014). The GWAS was conducted using a univariate LMM, where phenotypes were modelled as:

$$y = a + bx + g^{-} + e \quad (1)$$

In this model, y represents the vector of phenotypes (body weight), a is the mean term, and b represents the additive effect (fixed effects) of the candidate SNP being tested for association. The SNP genotype indicator variable, x, is coded as 0, 1 or 2. The term g^{-} represents the cumulative effect of all SNPs, excluding those located on the chromosome containing the candidate SNP, with the $var(g^{-})$ re-estimated each time a chromosome is omitted during the GRM calculation. Lastly, e represents residual error. The leave-one-chromosome-out approach was applied to minimize confounding effects by recalculating the GRM for each chromosome, excluding SNPs on the chromosome being analysed. SNP effects for each chromosome were independently estimated using phenotypic and genotypic data, with the GRM derived from the remaining chromosomes. SNP associations with body weight were tested using a likelihood ratio test, and *p*-values were corrected for multiple testing using the Bonferroni method. A genome-wide significance threshold was determined using the Bonferroni correction, which adjusts the significance level (0.05) by dividing it by the total number of quality-control filtered SNPs (~56K), to identify statistically significant associations (Kaler and Purcell, 2019). The distribution of association *p*-values was visualized using Manhattan plots to identify significant SNPs and assess trends across the genome. Significant SNPs were further analysed by aligning the extended DNA sequence of each significant SNP (10000 bp to both flanking regions) against the NCBI core nucleotide database using BLASTN 2.16.1+ (Zhang et al., 2000) to identify potential candidate genes.

2.3. Creating various SNP marker densities and training population size subsets

To evaluate the effects of SNP marker density and training population size on GEBV prediction accuracy, six marker densities (500, 1000, 5000, 10000, 30,000, and 56,378 SNPs) and six training population sizes (80, 400, 800, 1200, 1600 and 2038 individuals) were tested. The highest density and largest training population corresponded to the full dataset, while the other values were selected based on a preliminary GBLUP analysis of body weight. In this analysis, a broad range of arbitrarily chosen values was screened to identify representative points that captured the overall trend in prediction accuracy and the plateau at which further increases offered little improvement. The chosen SNP densities also reflect the range of commercial array options: from low- and medium-density arrays, which are more affordable, to the more expensive high-density array used in this study, thereby covering a realistic spectrum of costs and marker resolutions. For SNP marker density evaluations, the full sample dataset ($n = 2547$) was used for all SNP subsets, while for population size evaluations, the full SNP dataset ($n = 56,378$) was used for all population subsets. Two series of SNP subsets were generated based on different selection criteria: random SNP subsets and GWAS-ranked SNP subsets. In the random selection approach, SNPs were sampled from the full dataset without replacement to match the designated densities, ensuring even coverage across the genome and enabling direct comparison of prediction performance across densities by minimizing other sources of variation. To account for stochastic variation in SNP effect, the random sampling process was repeated five times for each marker density. In the GWAS-ranked approach, SNPs were ordered according to their significance levels (*p*-values) from the GWAS of body weight trait using the full dataset, and subsets were constructed by sequentially selecting the most significant markers to match the designated densities. Populations of varying sizes were created by randomly subsetting the full dataset, with a single

random subset generated for each designated size. All SNP and training populations subsets were evaluated with three models to predict GEBVs: GBLUP, BayesR, and KAML. For the analyses comparing within- and across-population prediction, Random 5000 SNP Replicate 1 subset was used, along with GWAS-ranked 5000 SNPs selected based on the GWAS of body weight traits using only Farm 1 and Farm 2 data.

2.4. Statistical GS models

The SNP effects and GEBVs were calculated using SNP genotype data with three different GS models: GBLUP, BayesR, and KAML. these environmental differences were accounted for in the analysis by including farm as a fixed effect in the statistical models. This approach helps control for environmental variation across farms, ensuring more accurate and unbiased comparison of growth traits and genomic predictions across the combined dataset.

2.5. GBLUP

GBLUP was performed using average information restricted maximum likelihood (AIREMLF90) algorithm in BLUPF90 (Misztal et al., 2002) fitting the single-trait animal model as described in Eq. (2):

$$y = X\beta + Zg + e \text{ with } g \sim N\left(0, G\sigma_g^2\right) \text{ and } e \sim N\left(0, I\sigma_e^2\right) \quad (2)$$

where y is a vector of observed phenotypic values (i.e., five traits), X is a fixed effects design matrix (including rearing site with three levels and deformity status with two levels), and β is a vector of fixed effect coefficients. Z is a random effects design matrix, g represents the additive genetic effect of each individual, whereby G is the genomic relationship matrix generated by VanRaden method (VanRaden, 2008) and e is a vector of residuals. GBLUP model captures genetic variance solely through the random effect Zg and assumes all SNPs have equal contributions to phenotypic variance.

2.6. BayesR

Phenotypes are modelled to relate to markers using a univariate mixed model:

$$y = 1_n\mu + W\alpha + X\beta + e \text{ with } e \sim N\left(0, I\sigma_e^2\right) \quad (3)$$

where y is a vector of observed phenotypic values, 1_n is a vector of ones of length n, and μ represents the overall mean. The matrix W contains covariates associated with fixed effects, with their corresponding effects represented by α. The matrix X is an $n \times p$ matrix containing genotypes coded as 0, 1, or 2 copies of the reference allele, and β is the vector of SNP effects. The residual term e is an n-dimensional vector distributed as $e \sim N\left(0, I\sigma_e^2\right)$, where I is an $n \times n$ identity matrix, and σ_e^2 represents the residual variance. BayesR assumes that the SNP effects β follow a mixture of normal distributions with varying variances, representing SNPs with large, moderate, small, or no effects on the trait, as in Eq. (4):

$$p\left(\beta_j|\pi, \sigma_g^2\right) = \pi_1 \times N\left(0, 0 \times \sigma_g^2\right) + \pi_2 \times N\left(0, 10^{-4} \times \sigma_g^2\right) + \pi_3 \times N\left(0, 10^{-3} \times \sigma_g^2\right) + \pi_4 \times N\left(0, 10^{-2} \times \sigma_g^2\right) \quad (4)$$

where σ_g^2 represents the additive genetic variance explained by SNPs, with sparsity incorporated by setting the effect and variance of the first mixture component to zero. π represents the proportion of different SNP effects and constrained to be positive and sum to unity. The unknown parameters (μ, π, β, σ_g^2, σ_e^2) were estimated using a Markov Chain Monte Carlo (MCMC) approach, employing a Gibbs sampling scheme to derive values from each parameter's conditional posterior distribution. In this study, PLINK2 (Chang et al., 2015) was used to convert genotype data into the suitable format for analysis with BayesR software (https://github.com/syntheke/bayesR). All analyses were conducted with a chain length of 50,000, using the first 20,000 samples as burn-in. Posterior parameter estimates were obtained from 30,000 samples by selecting every 10th sample (thinning) after the burn-in phase to reduce autocorrelation and storage requirements, ensuring that the remaining samples are more independent and representative of the true posterior distribution. BayesR incorporates prior probabilities for each SNP effect category, facilitating Bayesian inference to estimate posterior distributions of the model parameters.

2.7. KAML

KAML extends Eq. (2) by incorporating pseudo QTNs (Q) as covariates and optimizes a SNP-weighted kinship matrix (K_w) for the random effect to better account for the remaining genetic variance, as described in Eq. (5):

$$y = Xb + Qq + Z\mu + e \text{ with } \mu \sim N\left(0, K_w\sigma_g^2\right) \text{ and } e \sim N\left(0, I\sigma_e^2\right) \quad (5)$$

where q represents the vector of fixed effects associated with the covariate matrix Q. Machine learning techniques including cross-validation, multiple regression, grid search, and bisection algorithms are used to optimize parameters for pseudo QTNs and K_w. In this study, PLINK2 (Chang et al., 2015) was used to convert genotype data into the suitable format for analysis with KAML package (https://github.com/YinLiLin/KAML) in the R environment (R Core Team 2024). KAML.Impute function was used to impute missing values in the genotype, and GWAS was performed using a Mixed Linear Model (MLM) with 5-fold cross-validation and 5 replicates to determine optimal parameters for pseudo QTNs and SNP weights within the KAML package.

2.8. Evaluation of GS accuracy

Prediction accuracies for each trait were assessed using a repeated five-fold cross-validation (CV) procedure with 10 independent repetitions, yielding a total of 50 CV runs per trait to account for stochastic variation. In each repetition, individuals with both phenotypic and genotypic data were randomly partitioned into five equal folds; in each fold, one subset (20 %) served as the validation set with phenotypes masked, and the remaining four subsets (80 %) formed the training set for GEBV prediction. To evaluate the impact of within- and across-population predictions on accuracy, the same five-fold CV procedure with 10 independent repetitions was applied using the GBLUP model with 5000 randomly selected SNPs. In each repetition, 80 % of the combined Farm 1 and Farm 2 populations were used as the training set to predict the remaining 20 % of Farm 1 and Farm 2 populations (representing within population prediction), as well as 100 % of the MAC population (representing across population prediction). Phenotypes for all validation individuals were masked prior to model fitting. For across-population prediction using GWAS-ranked 5000 SNP subset, a preliminary analysis using all Farm 1 and Farm 2 individuals as the training set yielded zero accuracy for predicting the MAC population. Thus, no further cross-validation was performed for this subset.

The heritability $\left(h^2\right)$ was estimated using the GBLUP model applied

to the full dataset and is defined as:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \quad (6)$$

To determine the prediction accuracy of GEBVs, Pearson correlation coefficient between the predicted GEBVs and the deviation of observed phenotypes from the population mean was calculated. Notably, this is mathematically equivalent to the correlation between the predicted GEBVs and the observed phenotypes. The result was then divided by the square root of heritability ($\sqrt{h^2}$) as shown in Eq. (7):

$$Accuracy = \frac{cor(y, GEBV)}{\sqrt{h^2}} \quad (7)$$

A common h^2 value, derived from GBLUP using the full dataset for each trait, was uniformly applied across all models, training population sizes, and SNP marker densities for the corresponding traits. The values for h^2 were 0.29 for BW, 0.30 for TL, 0.39 for BD, 0.21 for K, and 0.33 for BSI, as reported in Liang et al. (2025). This approach ensured consistency and provided a standardized benchmark for evaluating GEBV prediction accuracy across traits and models. Negative prediction accuracies were set to zero, as accuracy cannot be below zero. Accuracy results were averaged across cross-validation folds and replicates, then compared across different marker densities, population sizes, selection strategies, and models. To quantify uncertainty, 95 % confidence intervals for prediction accuracies were calculated using Fisher's z-transformation, where correlation coefficients were first transformed to z-scores, standard deviations were derived, and intervals were subsequently back-transformed to the correlation scale.

2.9. Cost–accuracy trade-off analysis

To support practical implementation in breeding programs, prediction accuracy was plotted against per-sample genotyping cost for each SNP density. Genotyping costs for reference SNP panel sizes were obtained from the literature, with reported prices of €7.5 per sample for 300 SNPs, €15 per sample for 3000 SNPs, and €20 per sample for 57,000 SNPs (Fraslin et al., 2023). These reference values were converted to US dollars and subsequently used to estimate the costs of intermediate SNP densities (600, 1200, 6000, 12,000, 37,000, and 70,000), which approximately correspond to the 500, 1000, 5000, 10000, 30,000, and 56,378 high-quality SNPs retained after quality control in this study. Log-linear interpolation between the nearest reference points was applied to derive per-sample costs for each density. This approach was selected because in array-based genotyping, small SNP panels are relatively expensive per SNP, while adding more SNPs increases the total per-sample cost at a progressively smaller rate. Log transformation of the cost data provides a smooth, approximately linear relationship suitable for interpolation, reflecting the diminishing incremental cost of additional SNPs. The interpolated costs were then combined with observed prediction accuracies to generate the cost–accuracy trade-off curve.

To describe this relationship quantitatively, a logarithmic regression model was fitted with prediction accuracy as the response variable and the natural logarithm of per-sample genotyping cost as the predictor. The fitted model follows the equation:

$$Prediction\ accuracy = \beta_0 + \beta_1 \bullet \ln(Cost\_USD) \quad (8)$$

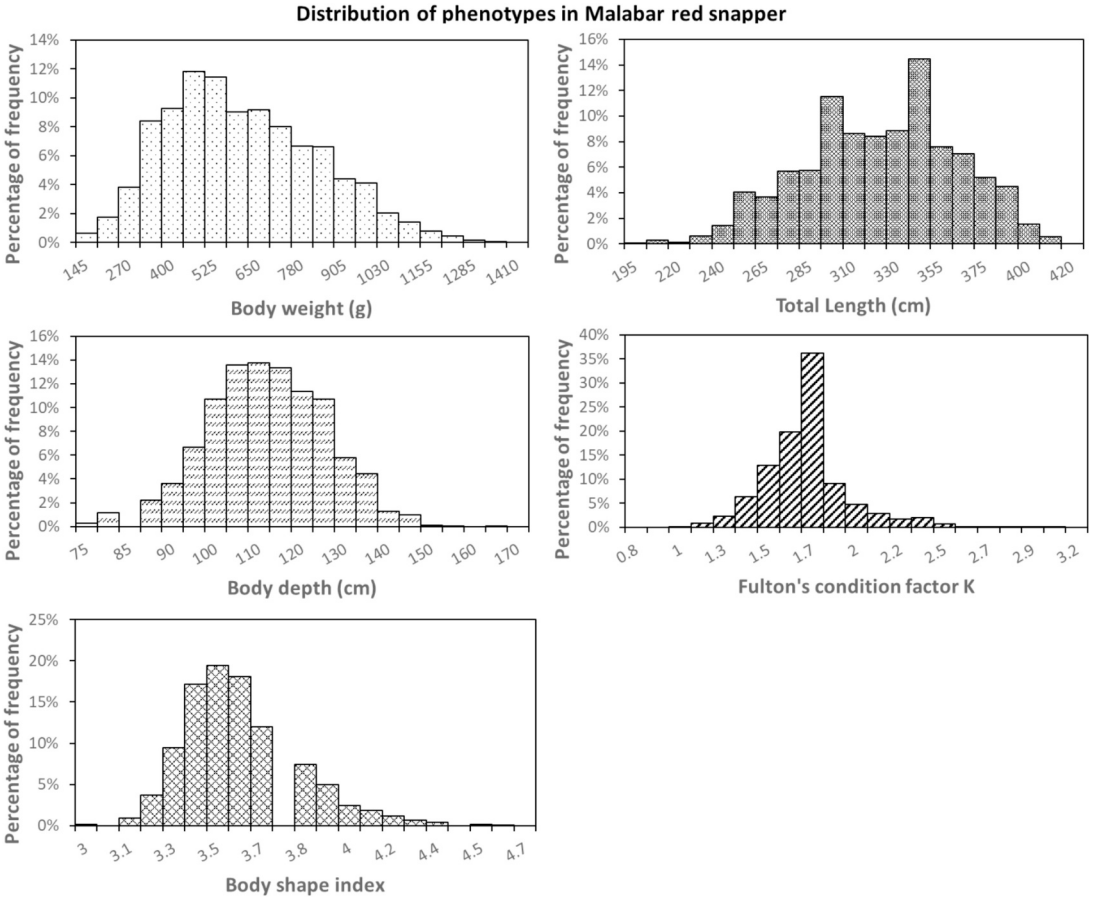


Fig. 1. Distribution of phenotypes in Malabar red snapper. Individuals were grouped into bins with fixed intervals for each trait, and the frequency of each bin was normalized to percentage rather than counts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where β_0 is the intercept and β_1 is the slope. The model was visualized as a blue dashed line in the cost–accuracy figure, with the R^2 value indicating the proportion of variance in prediction accuracy explained by genotyping cost.

3. Results

3.1. Descriptive statistics of morphometric traits

The distribution of harvest traits in Malabar red snapper, including BW, TL, BD, K, and BSI is illustrated in Fig. 1. The histogram for BW displayed a unimodal distribution with a slight right skew, where most individuals weighing between 412 g and 858 g. TL exhibited a slightly left-skewed unimodal pattern, with the majority of values ranging from 288 mm to 366 mm. BD shows a relatively symmetrical distribution, with measurements concentrated between 101 mm and 130 mm. K demonstrated a moderately peaked distribution, with most values lying between 1.5 and 2.0, indicating consistent body condition across the population. BSI displayed a narrow, symmetric distribution, with values centred around 3.6.

The PCA plot of red snapper from the three rearing sites is presented in **Supplementary Fig. 2**. Fish from Farm 1 and Farm 2 showed complete overlap, reflecting their shared origin and random distribution between the two farms. In contrast, only a small fraction of MAC fish overlapped with Farm 1 and Farm 2, indicating genetic divergence.

3.2. Significant SNP associations identified via GWAS

Manhattan plots illustrating the GWAS results for BW, TL, BD, K and BSI are shown in **Supplementary Fig. 1**. Two SNPs (SNP ID: AX-686371523 and AX-686370822) reached the suggestive significance threshold and potentially associated with body weight in Malabar red snapper. Both SNPs were on Scaffold 8 with position of 17,153,291 bp and 23,015,586 bp respectively (unpublished data). BLAST analysis of the DNA sequences within a 10000 bp region surrounding the significant SNPs revealed that SNP AX-686371523 aligned with the *rho GTPase-activating protein 42* gene in a couple of fish species such as striped seabass *Morone saxatilis* and European seabass *Dicentrarchus labrax*, while SNP AX-686370822 corresponded to the microsatellite sequence *MiSaTPW112* in largemouth bass *Micropterus salmoides*.

3.3. GS accuracies across different models, training population sizes and SNP marker densities

Overall, genomic prediction accuracies for growth-related traits (i.e., BW, TL, and BD) ranged from 0.40 to 0.62 using full dataset across the three GS models. BayesR outperformed other models for ratio traits (i.e., K and BSI), achieving accuracies of 0.76 for K and 0.74 for BSI with the full dataset. The effect of training population size on prediction accuracy is illustrated in **Supplementary Table 1** and Fig. 2. To minimize potential bias in the GS model due to low minor allele frequency (MAF) and monomorphic SNPs, we ensured that the MAF in the smallest training population (80 samples) ranged from 0.07 to 0.5, a range considered acceptable for genomic prediction. While this small training population

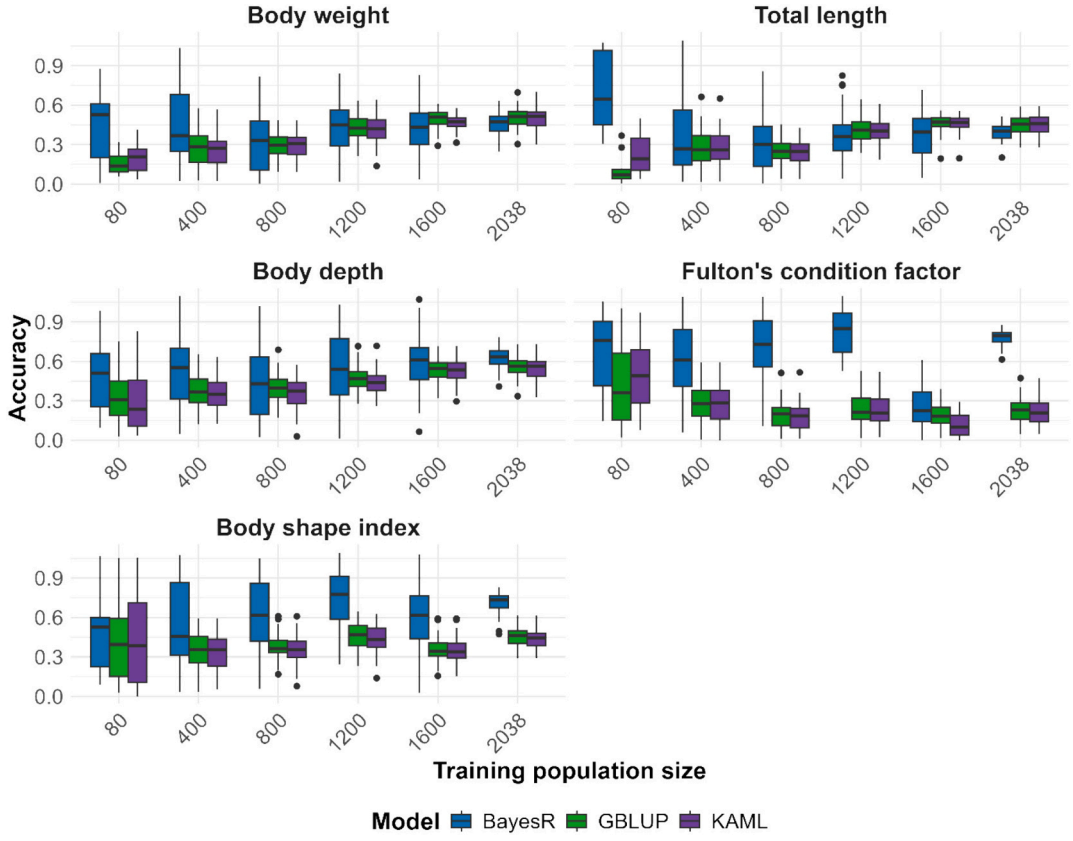


Fig. 2. Accuracy of estimated breeding values using BayesR, GBLUP, and KAML models for body weight, total length, body depth, Fulton's condition factor, and body shape index across different training population sizes (i.e., 80, 400, 800, 1200, 1600, and 2038 samples) in Malabar red snapper. Each boxplot represents the average accuracy across 50 replicates. The boxes indicate the interquartile range (IQR) from the 25th to the 75th percentile, the line inside the box represents the median (50th percentile), and the whiskers extend to the smallest and largest data points within 1.5 × IQR from the lower and upper quartiles. Dots outside the whiskers represent outliers, which are values that fall beyond 1.5 × IQR from the quartiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

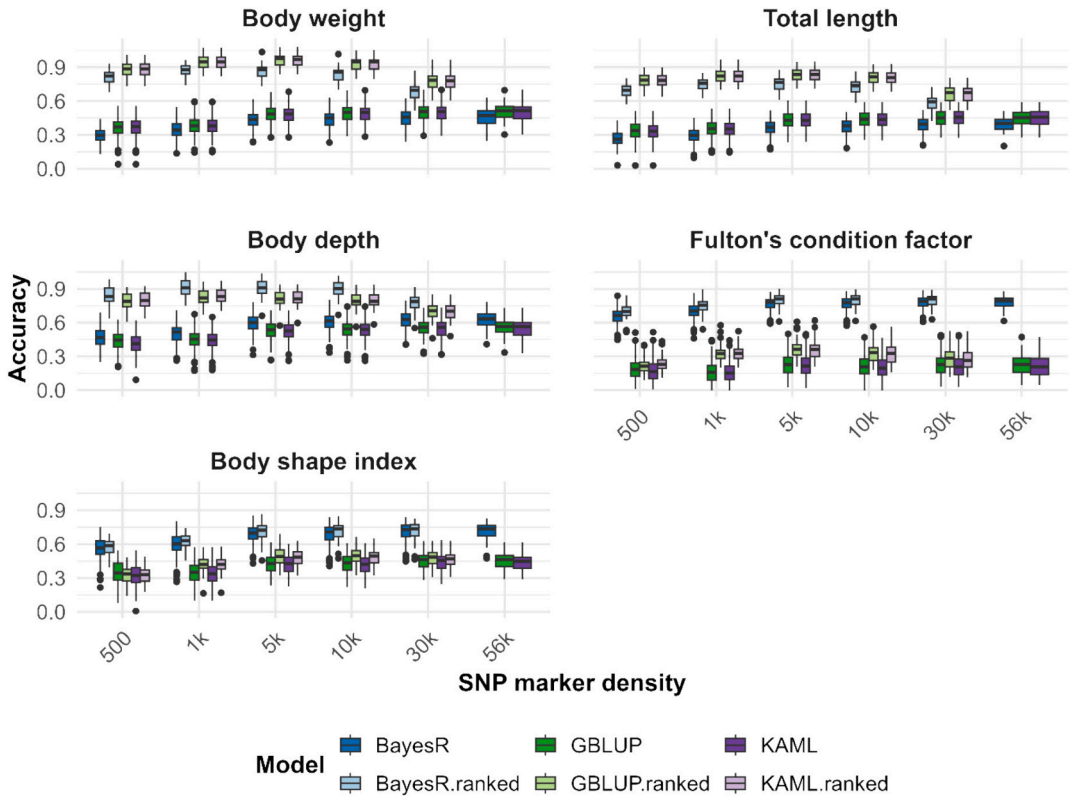


Fig. 3. Accuracy of estimated breeding values using BayesR, GBLUP, and KAML models for body weight, total length, body depth, Fulton's condition factor, and body shape index as a function of different SNP marker densities (i.e., 500, 1000, 5000, 10000, 30,000, and 56,378 SNP markers) in Malabar red snapper. SNP subsets were either randomly selected with five repeats (blue, green, and purple) or ranked based on GWAS *p*-values for body weight (light blue, light green, and light purple). Each boxplot represents the average accuracy across replicates (250 for random subsets and 50 for GWAS-ranked subsets). Boxes indicate the interquartile range (IQR) from the 25th to the 75th percentile, the line inside the box represents the median (50th percentile), and whiskers extend to the smallest and largest data points within 1.5 × IQR from the lower and upper quartiles. Dots outside the whiskers represent outliers, which are values that fall beyond 1.5 × IQR from the quartiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

size resulted in poor predictions, accuracy improved consistently with increasing training population sizes across all models, reaching a plateau at 1600 samples. For example, GEBV prediction for body weight increased from approximately 0.22–0.25 with 400 samples to 0.45–0.50 with 2038 samples across all models. Standard deviations of accuracy also decreased from 0.17 to 0.41 with 400 samples to 0.07–0.08 with 2038 samples, indicating greater stability in predictions. The effect of SNP marker density on prediction accuracy is illustrated in **Supplementary Table 2** and Fig. 3. For random SNP subsets, generally the accuracy increased with denser SNP sets, plateauing beyond 5000 SNPs. For instance, GBLUP and KAML achieved an accuracy of 0.5 for body weight using 56,378 SNPs, compared to 0.36–0.37 with 500 SNPs. The analysis also compared random SNP subsets with GWAS-ranked subsets for within-population GS. For growth-related traits, GWAS-ranked subsets consistently outperformed random subsets, improving prediction accuracy by more than twice for lower SNP densities such as 500 and 1000 SNPs. However, for ratio traits, the accuracy of GWAS-ranked subsets was just a bit higher than random subsets. Overall, the results indicate that prediction accuracy is influenced by both training population size and SNP marker density, with larger datasets and denser SNP sets yielding higher accuracy. Finally, within-population genomic prediction yielded higher accuracy than across-population prediction for all traits studied when using GBLUP with the random 5000 SNPs subset (Fig. 4 and **Supplementary Table 3**). Although the GWAS-ranked 5000 SNP subset yielded high prediction accuracy in within-population genomic selection, it failed to produce meaningful predictions in the across-population context, resulting in zero prediction accuracy. These values are not presented in Fig. 4 due to their null outcome. This finding underscores the limited generalisability of GWAS-ranked SNP panels when applied to genetically distant populations.

Correlation of prediction accuracies among BayesR, GBLUP, and KAML GS models across five harvest traits is shown in Fig. 5. GBLUP and KAML showed consistently high correlations across all traits, with coefficients exceeding 0.95, indicating strong agreement in their ability to predict EBVs. In contrast, BayesR exhibited variable correlations with GBLUP and KAML depending on the trait. For BW and BD, BayesR demonstrated high correlations with both models, exceeding 0.93, suggesting similar predictive performance for these growth traits. However, for TL and BSI, the correlations were moderate, ranging from 0.54 to 0.74, indicating some divergence in predictions relative to GBLUP and KAML. Notably, no significant correlation was observed between BayesR and the other two models for K, reflecting differences in how BayesR captures genetic variation for this ratio-based trait. These findings highlight that while GBLUP and KAML perform similarly across traits, BayesR may better account for specific genetic architectures, particularly for ratio traits like K and BSI.

3.4. *Cost–accuracy trade-off analysis*

Results are shown in Fig. 6. While the prediction accuracy increased with SNP density and plateaued beyond 5000 SNPs, the cost–accuracy curve indicates that moderate-density panels (5000–10000 SNPs) achieve a practical balance between genotyping cost (US$ 18.8–20.1 per sample) and predictive performance (0.48–0.49), with minimal additional gain from very high-density panels.

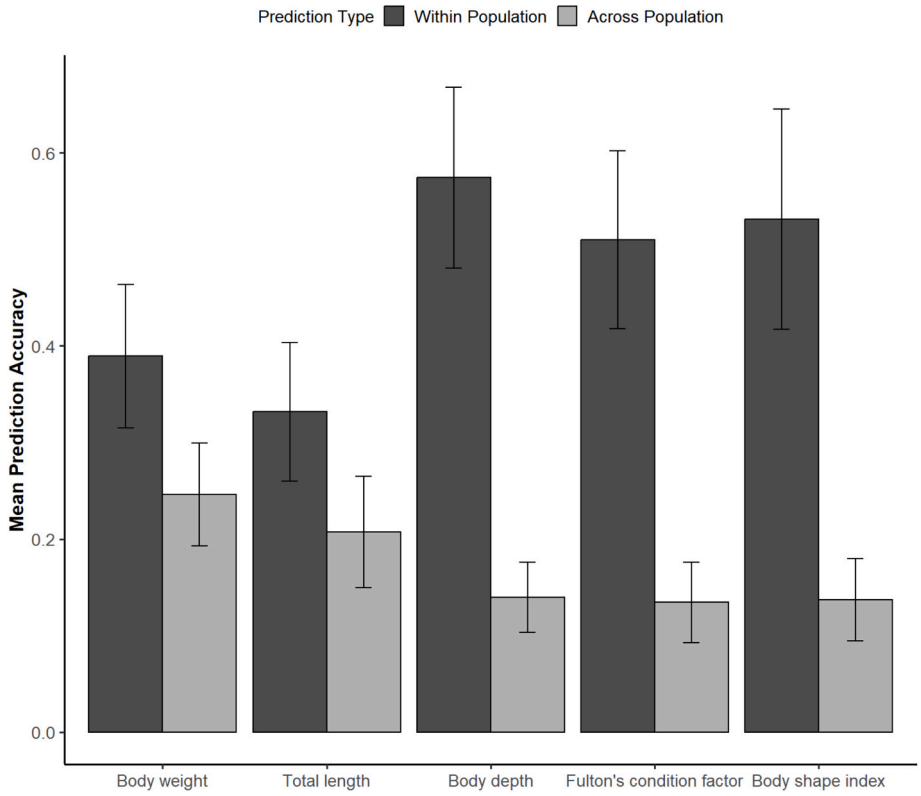


Fig. 4. Comparison of within- and across-population genomic prediction accuracies for body weight, total length, body depth, Fulton's condition factor (K), and body shape index, using random SNP subsets. Mean prediction accuracies are shown based on 50 replicates, with error bars representing standard deviations. Prediction accuracies obtained using GWAS-ranked SNP subsets were zero in the across-population scenario and are therefore not displayed.

4. Discussion

Our previous study demonstrated that the h^2 of growth traits in Malabar red snapper (*L. malabaricus*) ranged from 0.29 ± 0.03 to 0.39 ± 0.05, and the h^2 of body shape traits was 0.21 ± 0.03, indicating that both can be improved through selective breeding programs (Liang et al., 2025). GS offers a powerful tool for enhancing breeding programs in aquaculture. Optimizing GS strategies is crucial for its effective implementation, particularly in emerging species such as L. *malabaricus* in Southeast Asia. Although reducing SNP density and training population size lowers genotyping costs, it often comes at the expense of accuracy, requiring a careful balance between cost efficiency and genetic gain. This study evaluated the efficiency of GS methodologies for improving key harvest traits (i.e., BW, TL, BD, K, BSI) by comparing multiple models, SNP marker densities, and training population sizes. Our findings identified 5000 SNPs as the optimal marker density and 1200 individuals as the optimal training population size, with GBLUP being the most accurate and efficient GS model for improving growth traits in L. *malabaricus*. These results provide a practical framework for enhancing genetic improvement in Malabar red snapper breeding programs while minimizing costs.

A larger training population generally enhances both prediction accuracy and stability (i.e., lower standard deviations across replicates). Our results showed that the full dataset consistently outperformed other subsets, achieving the highest prediction accuracy and stability across all models and traits. However, the relationship between accuracy and training population size varied uniquely across different models and traits. For BW trait, the accuracy of GBLUP plateaued with the 1600 training population size, while the accuracy of BayesR and KAML further improved when training population increased from 1600 to 2038 individuals. For TL and BD trait, the accuracy of GBLUP and KAML plateaued with 1600 training population size, while the accuracy of BayesR further improved when training population size increased to 2038. In both cases, BayesR required a larger training population to reach the best performance. According to the literature, prediction accuracy in aquaculture breeding programs typically plateau when the training population size exceeds 4000 individuals (Dagnachew and Meuwissen, 2019; Lillehammer et al., 2013). While this threshold could not be tested in this study due to limited sample size, the findings highlight the importance of continuously collecting genotype and phenotype data as the selective breeding program progresses. Regular updates to performance records and an expanding dataset are critical for improving prediction accuracy over time. For Malabar red snapper, an initial training population of 1200 individuals is expected to provide a reasonable level of prediction accuracy with GBLUP and 56K SNPs for a selective breeding program targeting growth traits. Interestingly, the prediction accuracy for K and BSI decreased when training population increased from 1200 to 1600 for all models. This is likely due to the complex genetic architecture and non-linear ratio nature of K and BSI traits, as well as possible interactions between loci, non-additive effects, and environmental influences, which make K and BSI more challenging to model. As a result, increasing the training population size may not necessarily increase the prediction accuracy, because larger training population sizes might capture more noise or genetically distant samples. Nevertheless, the full dataset of 2038 samples still yielded the best accuracy for K and BSI for all models considering both accuracy and standard deviation.

Careful optimization of SNP densities can balance cost with prediction accuracy. In the present study, prediction accuracy showed minimal improvement beyond a density of 5000 SNPs across all traits and models for random SNP subsets, indicating that 5000 SNP markers were sufficient for retaining high prediction accuracy. Using a low-density SNP panel, such as a 6K SNP array that retains approximately 5000 high-quality SNP markers after quality control, can reduce genotyping costs

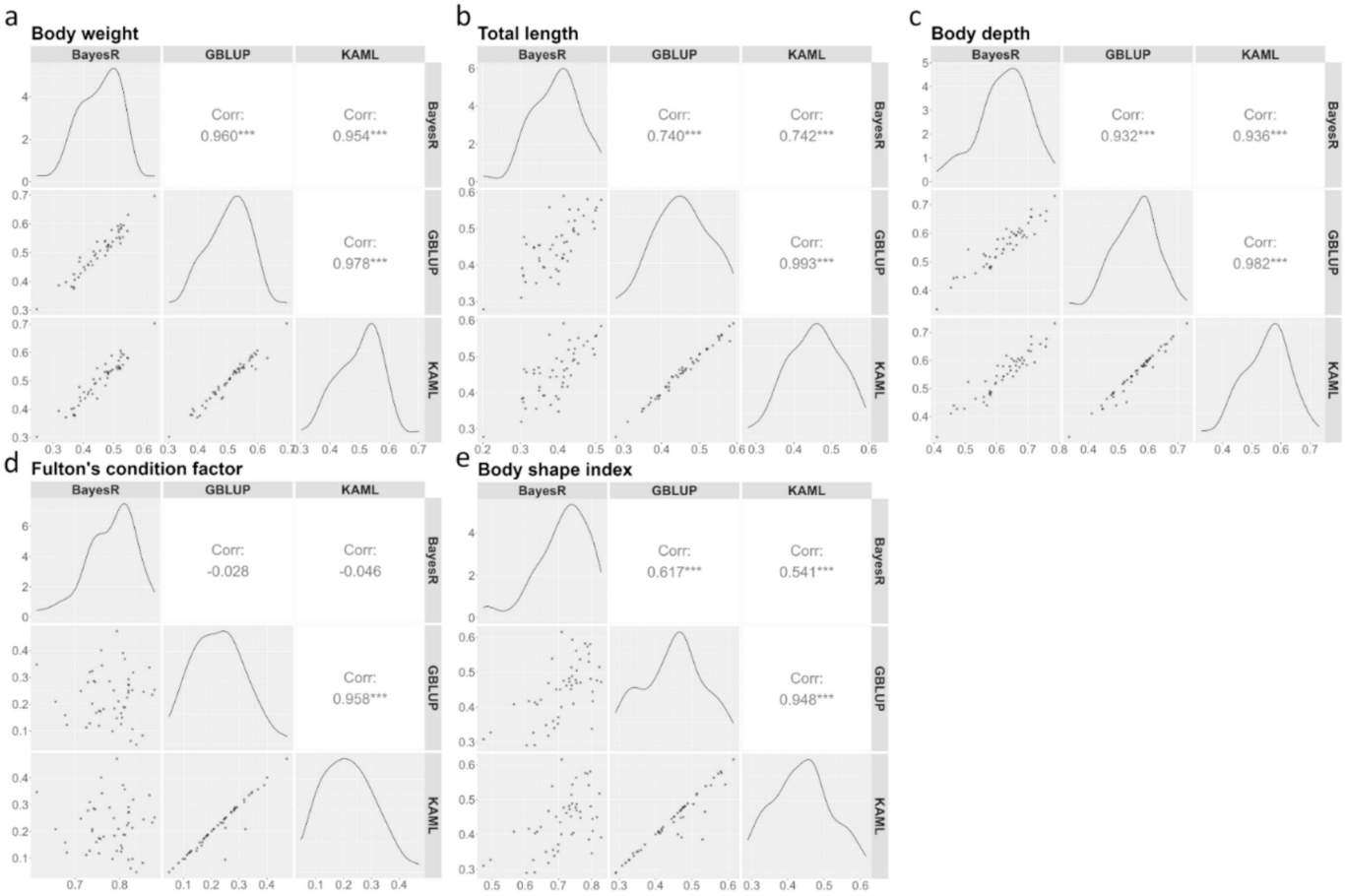


Fig. 5. Pairwise scatterplot matrix for prediction accuracy among BayesR, GBLUP, and KAML genomic selection models across five harvest traits: body weight, total length, body depth, Fulton's condition factor K, and body shape index. Diagonal panels display histograms of prediction accuracy distributions for each model. Lower off-diagonal panels show pairwise scatterplots between models, while upper off-diagonal panels display the corresponding Pearson correlation coefficients. Statistical significance of the correlations is indicated as follows: no asterisk (not significant), * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$).

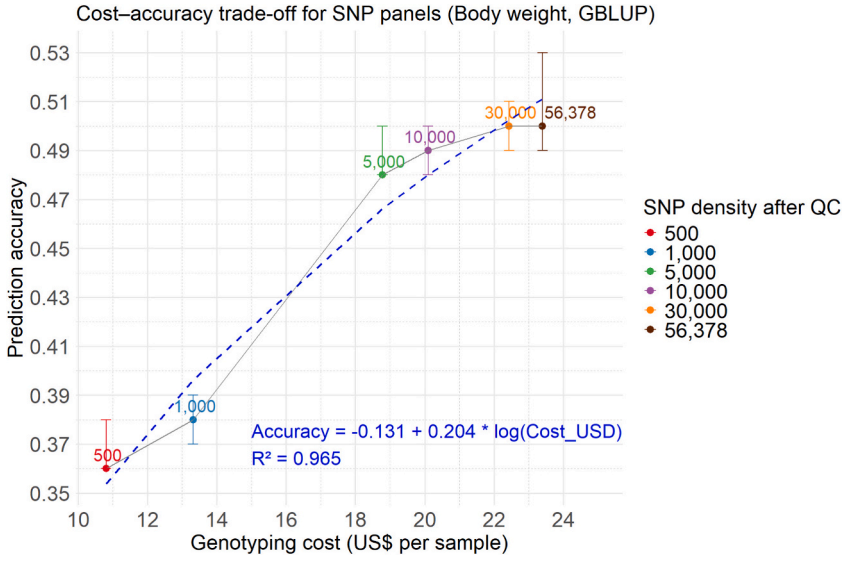


Fig. 6. Cost–accuracy trade-off for genomic prediction of body weight in Malabar red snapper using GBLUP across different SNP panel densities. Prediction accuracy (mean ± 95 % confidence interval) is plotted against per-sample genotyping cost in USD. Points are colour-coded and labelled according to SNP density after QC. The fitted log model (blue dashed line) illustrates the relationship between genotyping cost and prediction accuracy, with R^2 representing the proportion of variance in accuracy explained by the model. The model equation and R^2 value are shown on the plot. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

by approximately 20 % compared to the current 70K SNP panel (Fig. 6), while resulting in only a minor reduction in accuracy of 4 %. This trade-off makes it a cost-effective option for commercial breeding programs. Our findings are consistent with previous studies in other aquaculture species. In European sea bass (*Dicentrarchus labrax*) and gilthead sea bream (*Sparus aurata*), SNP densities ranging from 3000 to 6000 were reported to be optimal for genomic prediction of resistance to VNN and vibriosis (Griot et al., 2021b). Similarly, a SNP density of 4000 was found to be sufficient for accurate genomic relationship estimation in black tiger shrimp *Penaeus monodon* (Guppy et al., 2020), while genomic prediction accuracy for body weight in barramundi plateaued at densities between 5000 and 10000 SNPs (Massault et al., 2025). In Atlantic salmon, a density of 5000 SNPs was previously reported as optimal for predicting host resistance to sea lice (Tsai et al., 2016). In contrast, Kriaridou et al. (2020) reported that SNP densities of 1000 to 2000 were able to achieve similar accuracies to those obtained with higher SNP densities (7000 to 10000 SNPs) across several species and traits such as amoebic gill disease in Atlantic salmon, growth traits in Common carp *Cyprinus carpio*, pasteurellosis in Sea bream *Sparus aurata*, and ostreid herpesvirus resistance in Pacific oyster *Crassostrea gigas*. The variation in optimal SNP densities may stem from differences in trait genetic architecture, population structure, and LD patterns (e.g., the relevance of SNPs to the traits) across traits, species, and populations. This complexity makes it challenging to define a universally optimal SNP subset for GS. Furthermore, GWAS-ranked SNP subsets consistently outperformed corresponding random SNP subsets to predict growth-related traits such as BW, TL, and BD for within-population GS, while showing comparable performance to random SNP subsets for other traits like K and BSI. This suggests that although GWAS-based SNP selection can enhance prediction accuracy, its effectiveness is highly trait-dependant and may be less effective for traits which are not highly genetically correlated to the trait which GWAS was conducted on. This finding is consistent with a previous study on barramundi, where GWAS-ranked SNPs based on body weight enhanced genomic prediction for body weight but not for other traits such as K and BSI (Massault et al., 2025). In addition to GWAS-based SNP selection, there are other approaches which can enhance GS accuracy, ranging from simple filtering techniques, such as removing uninformative or redundant SNPs based on variance thresholds or LD pruning, to more advanced algorithms like Elastic Net, BayesA, LASSO, and gradient boosting machine (GBM). Several specialized tools and algorithms have also been developed for optimisation of SNP selection, including GMStools (Jeong et al., 2020) and machine learning-based approaches such as incremental feature selection using random forests (Heinrich et al., 2023). These methods and tools offer valuable opportunities for further optimizing genomic prediction in Malabar red snapper.

A practical question when designing a low-density SNP panel for cost-effective applications is whether random or GWAS-ranked SNP subsets should be used. In this study, across-population prediction analyses showed that while the random 5000 SNP subsets provided reasonable accuracy for predicting GEBVs in the MAC dataset using Farm 1 and Farm 2 as the training set, GWAS-ranked 5000 SNP subsets resulted in zero prediction accuracy for the MAC dataset under the same training conditions. This is likely because most GWAS signals represent SNPs, which are markers in high linkage disequilibrium with, but not identical to, the underlying causal variants. The PCA plot further highlighted genetic differentiation between MAC population and the combined Farm 1 and Farm 2 populations. Differences in allele frequencies and LD patterns between populations can weaken the ability of SNPs prioritised in one population to capture the underlying causal variants in another. In addition, GWAS-based selection in a population can capture environmental associations that do not generalize across populations or production environments. Even when causal variants are shared, genotype-by-environment (G x E) interactions can modify effect sizes between rearing sites, further reducing cross-population predictive power. Consequently, SNPs selected via GWAS for body weight in Farm 1 and Farm 2 are unlikely to be informative for MAC fish. This explains why the top 5,00 GWAS-ranked SNPs achieved high within-population accuracy 0.81–0.97 for growth traits with GBLUP) but failed to deliver predictive accuracy across genetically distinct populations. Therefore, when designing low-density SNP chips, random 5000 SNPs offer broader compatibility with genetically distant or unknown samples, whereas GWAS-ranked SNPs are better suited for within-population predictions as part of a targeted feature selection approach. Several strategies may help overcome the limited transferability of prediction accuracy across populations. Expanding the training population to include individuals from multiple populations can capture a broader spectrum of genetic variation, thereby producing more robust prediction models. This approach has been shown to improve prediction accuracy in plants when combining data from three distinct populations (Sverrisdóttir et al., 2018) and in pigs, where joint genomic prediction across populations of the same breed with certain genetic links further enhanced accuracy (Zhou et al., 2019). However, differences in phenotypic distributions, LD patterns, heritability, and genetic variance across populations may constrain the effectiveness of a combined reference population. An alternative solution is the development of multi-population GS models that explicitly account for population structure by treating the same trait in different populations as distinct but genetically correlated traits. In this framework, multi-population prediction is formulated as a multi-trait problem, leveraging between-population genetic correlations to improve predictive performance. Evidence suggests that such approaches outperform conventional multi-population models that simply pool populations together (Teng et al., 2024). Another complementary strategy involves constructing reference panels across populations and optimizing the relative proportion of each to prevent large populations from dominating prediction outcomes, thereby improving cross-population prediction accuracy (Zhou et al., 2025). Furthermore, various strategies can optimize the use of low-density SNP markers in genetic breeding programs. These include applying GS within families (Lillehammer et al., 2013), imputing low-density SNPs to high-density (Tsai et al., 2017), using SNP subsets identified by machine learning (Li et al., 2018), increasing MAF thresholds, and selecting evenly spaced markers (Robledo et al., 2018). These approaches could serve as potential avenues for further refining GS in Malabar red snapper. Furthermore, our results demonstrated that within-population prediction accuracies were consistently higher than those of across-population predictions for all traits when using 5000 randomly selected SNPs. For example, the accuracy of predicting body weight in the MAC population using data from Farm 1 and Farm 2 was only 64 % of the accuracy obtained when predicting body weight within the combined Farm 1 and Farm 2 populations. This reduction in predictive performance is likely due to the lower genetic relatedness between the MAC population and those from Farm 1 and Farm 2, as well as the presence of moderate G × E interactions between these rearing sites (ranging from 0.45 to 0.60), as reported in our previous study (Liang et al., 2025). These findings are consistent with previous studies showing lower accuracies for across-population predictions of host resistance to sea lice in Atlantic salmon (Tsai et al., 2016). To maintain high accuracy, across-population predictions would require a higher SNP density to better capture genetic variation (Kriaridou et al., 2020). Additionally, our results showed that the difference in prediction accuracy between within-population and across-population GS was more pronounced for traits with lower heritability (**Supplementary Table 3**), consistent with findings from a previous study in silico (de Roos et al., 2009).

Comparing genomic selection models is crucial for determining the most effective approach for predicting key aquaculture traits. In this study, both KAML and GBLUP outperformed BayesR in predicting BW and TL across all SNP marker densities and training population sizes. While KAML incorporated GWAS and machine learning for a more comprehensive SNP selection approach, its accuracy remained comparable to GBLUP for the traits analysed. This similarity is likely due to both models utilizing LMM, with KAML's additional selection methods

providing no significant advantage for traits with a predominantly polygenic architecture, as indicated by GWAS results. In fact, most traits targeted for selection in aquaculture follow a polygenic inheritance model, including growth rate (Houston et al., 2009), salinity tolerance (Norman et al., 2012), and spawning time (Colihueque et al., 2010); while a limited number of major QTLs have been identified and applied in marker-assisted selection (MAS) programs, such as those associated with sexual maturation in Atlantic salmon (Ayllon et al., 2015; Barson et al., 2015) and resistance to infectious pancreatic necrosis (IPN) in Atlantic salmon (Houston et al., 2012). Given that KAML requires substantially greater computational resources, GBLUP emerges as the more efficient choice for predicting BW and TL breeding values in Malabar red snapper.

In contrast, although BayesR required more computational resources than GBLUP, it outperformed both GBLUP and KAML in predicting EBVs for BD and demonstrated significantly stronger performance in predicting EBVs for K and BSI. Body shape traits, including K and BSI, are generally influenced by a combination of many loci with small effects and a few major QTLs of moderate-to-large effect (Reid and Peichel, 2010), and involve dominant and epistatic interactions (Husemann et al., 2017). For example, in cichlid fishes, 34 major QTLs were found to explain 3.2–8.6 % of phenotypic variation in body shape, alongside numerous small-effect loci (DeLorenzo et al., 2023). This genetic architecture aligns well with the BayesR framework, which models SNP effects as a mixture of four distributions representing no effect, small, medium, and large effects, thereby enabling it to capture a spectrum of variant influences from negligible to substantial. In contrast, GBLUP assumes that all SNPs contribute equally small additive effects, which can limit its ability to capture large-effect variants. The advantage of BayesR in such scenarios has also been reported in plant and livestock species, particularly for traits influenced by rare or large-effect variants (Zhang et al., 2019). This likely explains BayesR's superior performance for K and BSI in our study, where the underlying genetic architecture appears to fit this mixed-effect model.

Furthermore, previous studies have shown that low heritability generally leads to reduced prediction accuracy (D'Agaro et al., 2021; Nirea et al., 2012; Vela-Avitúa et al., 2015). In this study, the prediction accuracy of different traits across all GS models followed a similar trend. However, an exception was observed with BayesR, which achieved a high accuracy of 0.78 for K despite its lower heritability ($h^2_K = 0.21$) when compared to other traits such as body weight ($h^2_{BW} = 0.29$) (**Supplementary Table 1** and **Supplementary Table 2**). This likely reflects BayesR's ability to extract meaningful predictive information from genomic regions with moderate-to-large effects, and its potential usefulness for traits with lower heritabilities.

GWAS is a widely used method for identifying SNPs and genes associated with economically important traits in aquaculture. Targeted genotyping of known causative SNPs or assigning greater weight to these variants in GS models could enhance prediction accuracy by focusing on genomic regions with confirmed functional relevance. In this study, two significant SNPs associated with body weight were identified: one annotated to the *Rho GTPase-activating protein 42* (*ARHGAP42*) gene and the other to the microsatellite sequence *MiSaTPW112*. While *MiSaTPW112* serves as a microsatellite marker, *ARHGAP42* encodes a Rho GTPase-activating protein, which plays a key role in cytoskeletal organization and is involved in essential cellular functions such as growth, differentiation, neuronal development, and synaptic activity (Moon and Zheng, 2003). The function of Rho GTPases is highly context-dependent, varying by cell type and environmental conditions (Pedersen and Brakebusch, 2012). These findings highlight potential genetic regions linked to growth traits and provide valuable insights for further studies on validation of their functions in relation to growth traits in L. *malabaricus*. In addition, the overall GWAS results suggested that all traits examined in this study followed a polygenic inheritance pattern, with no major QTL detected in these populations.

5. Conclusion

This study aimed to optimize genomic selection strategies for harvest traits in Malabar red snapper (*Lutjanus malabaricus*) by balancing prediction accuracy with genotyping cost. The findings demonstrate that using 5000 SNPs and a training population of 1200 individuals with the GBLUP model offers a cost-effective and accurate approach for predicting growth-related traits. These values should be interpreted as practical guidelines rather than universal optimal thresholds, as the optimal values are dataset-specific and can vary with factors such as trait genetic architecture, linkage disequilibrium, population structure, and relatedness between training and validation sets. Although GWAS-informed SNP selection achieved higher accuracy for within-population predictions, its utility was trait-dependent—performing well only for the specific trait used in the GWAS and its genetically correlated traits. In contrast, randomly selected SNP subsets provided more consistent performance across all traits and demonstrated greater robustness in across-population prediction scenarios.

CRediT authorship contribution statement

Bing Liang: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Dean R. Jerry:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Vu Nguyen:** Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation. **Purushothaman Kathiresan:** Writing – review & editing, Investigation, Formal analysis. **David B. Jones:** Writing – review & editing, Investigation, Formal analysis. **Xueyan Shen:** Writing – review & editing, Methodology, Investigation, Funding acquisition, Conceptualization. **Joyce Koh:** Writing – review & editing, Investigation. **Celestine Terence:** Writing – review & editing, Investigation. **Maria G. Nayfa:** Writing – review & editing, Investigation. **Maura Carrai:** Writing – review & editing, Investigation. **Rachel Jia Wen Ho:** Writing – review & editing, Investigation. **Hazim Mohamed:** Writing – review & editing, Investigation. **Saraphina Dianne Rwei Qing Tneo:** Writing – review & editing, Investigation. **Grace Loo:** Writing – review & editing, Methodology, Investigation, Funding acquisition, Conceptualization. **Shubha Vij:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization. **Jose A. Domingos:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Ethics statement

The study was carried out at James Cook University in Singapore, following the approval from the Institute Animal Care and Use Committee (IACUC) under the reference number 2021-A010.

Funding

This work was funded by the Singapore Food Story (SFS) R&D Programme in 'Sustainable Urban Food Production' (NRF-000190-00; Proposal ID: SFSRNDSUFP1-0097).

Declaration of competing interest

The authors declare no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aquaculture.2025.743314.

Data availability

Data will be made available on request.

References

Abdelrahman, H., ElHady, M., Alcivar-Warren, A., Allen, S., Al-Tobasei, R., Bao, L., Beck, B., Blackburn, H., Bosworth, B., Buchanan, J., Chappell, J., Daniels, W., Dong, S., Dunham, R., Durland, E., Elaswad, A., Gomez-Chiarri, M., Gosh, K., Guo, X., Zhou, T., 2017. Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research. BMC Genomics 18 (1), 191. https://doi.org/10.1186/s12864-017-3557-1.

Abdollahi-Arpanahi, R., Gianola, D., Peñagaricano, F., 2020. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet. Sel. Evol. 52 (1), 12. https://doi.org/10.1186/s12711-020-00531-z.

Al-Mamun, H.A., Danilevicz, M.F., Marsh, J.I., Gondro, C., Edwards, D., 2025. Exploring genomic feature selection: a comparative analysis of GWAS and machine learning algorithms in a large-scale soybean dataset. Plant Genome 18 (1), e20503. https://doi.org/10.1002/tpg2.20503.

Ayllon, F., Kjærner-Semb, E., Furmanek, T., Wennevik, V., Solberg, M.F., Dahle, G., Taranger, G.L., Glover, K.A., Almén, M.S., Rubin, C.J., Edvardsen, R.B., Wargelius, A., Houston, R., 2015. The vgll3 locus controls age at maturity in wild and domesticated Atlantic Salmon (*Salmo salar* L.) males. PLoS Genet. 11 (11), e1005628. https://doi.org/10.1371/journal.pgen.1005628.

Bangera, R., Thorland, I., Hillestad, B., Johansen, H., Refstie, T., 2025. Accuracy of genomic predictions for yield traits, flesh colour and fillet fat in Atlantic salmon (*Salmo salar*).

Barson, N.J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G.H., Fiske, P., Jacq, C., Jensen, A.J., Johnston, S.E., Karlsson, S., Kent, M., Moen, T., Niemelä, E., Nome, T., Næsje, T.F., Orell, P., Romakkaniemi, A., Sægrov, H., Urdal, K., Primmer, C.R., 2015. Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. Nature (London) 528 (7582), 405–408. https://doi.org/10.1038/nature16062.

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). https://doi.org/10.1186/s13742-015-0047-8.

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S.J., Park, J.-H., 2013. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat. Genet. 45 (4), 400–405. https://doi.org/10.1038/ng.2579.

Colihueque, N., Cárdenas, R., Ramírez, L., Estay, F., Araneda, C., 2010. Analysis of the association between spawning time QTL markers and the biannual spawning behavior in rainbow trout (*Oncorhynchus mykiss*). Genet. Mol. Biol. 33 (3), 578–582. https://doi.org/10.1590/S1415-47572010000300032.

D'Agaro, E., Favaro, A., Matiussi, S., Gibertoni, P.P., Esposito, S., 2021. Genomic selection in salmonids: new discoveries and future perspectives. Aquac. Int. 29 (5), 2259–2289. https://doi.org/10.1007/s10499-021-00747-w.

Dagnachew, B., Meuwissen, T., 2019. Accuracy of within-family multi-trait genomic selection models in a sib-based aquaculture breeding scheme. Aquaculture 505, 27–33. https://doi.org/10.1016/j.aquaculture.2019.02.036.

de Roos, A.P., Hayes, B.J., Goddard, M.E., 2009. Reliability of genomic predictions across multiple populations. Genetics 183 (4), 1545–1553. https://doi.org/10.1534/genetics.109.104935.

DeLorenzo, L., Mathews, D., Brandon, A.A., Joglekar, M., Carmona Baez, A., Moore, E.C., Ciccotto, P.J., Roberts, N.B., Roberts, R.B., Powder, K.E., 2023. Genetic basis of ecologically relevant body shape variation among four genera of cichlid fishes. Mol. Ecol. 32 (14), 3975–3988. https://doi.org/10.1111/mec.16977.

Domingos, J.A., Goldsbury, J.A., Bastos Gomes, G., Smith, B.G., Tomlinson, C., Bade, T., Sander, C., Forrester, J., Jerry, D.R., 2021. Genotype by environment interactions of harvest growth traits for barramundi (*Lates calcarifer*) commercially farmed in marine vs. freshwater conditions. Aquaculture 532, 735989. https://doi.org/10.1016/j.aquaculture.2020.735989.

Dong, L., Xiao, S., Wang, Q., Wang, Z., 2016. Comparative analysis of the GBLUP, emBayesB, and GWAS algorithms to predict genetic values in large yellow croaker (*Larimichthys crocea*). BMC Genomics 17 (1), 460. https://doi.org/10.1186/s12864-016-2756-5.

Fraslin, C., Robledo, D., Kause, A., Houston, R.D., 2023. Potential of low-density genotype imputation for cost-efficient genomic selection for resistance to *Flavobacterium columnare* in rainbow trout (*Oncorhynchus mykiss*). Genet. Sel. Evol. 55 (1), 59. https://doi.org/10.1186/s12711-023-00832-z.

Fulton, T.W., 1904. The rate of growth of fishes. Twenty-second Ann. Rep. 0, 141–241. https://cir.nii.ac.jp/crid/1570291225855291904.

Gjedrem, T., 2005. Selection and Breeding Programs in Aquaculture, 1. Aufl. ed. Springer-Verlag. https://doi.org/10.1007/1-4020-3342-7.

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., 2021a. 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.

Griot, R., Allal, F., Phocas, F., Brard-Fudulea, S., Morvezen, R., Haffray, P., François, Y., Morin, T., Bestin, A., Bruant, J.-S., Cariou, S., Peyrou, B., Brunier, J., Vandeputte, M., 2021b. Optimization of genomic selection to improve disease resistance in two marine fishes, the European Sea bass (*Dicentrarchus labrax*) and the Gilthead Sea bream (*Sparus aurata*). Front. Genet. 12, 665920. https://doi.org/10.3389/fgene.2021.665920.

Guppy, J.L., Jones, D.B., Kjeldsen, S.R., Le Port, A., Khatkar, M.S., Wade, N.M., Sellars, M.J., Steinig, E.J., Raadsma, H.W., Jerry, D.R., 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. https://doi.org/10.1186/s12864-020-06960-w.

Habier, D., Fernando, R.L., Dekkers, J.C.M., 2007. Impact of genetic relationship information on genome-assisted breeding values. Genetics (Austin) 177 (4), 2389. https://doi.org/10.1534/genetics.107.081190.

Heath, A.C., Benyamin, B., Nyholt, D.R., Goddard, M.E., Gordon, S., Martin, N.G., Henders, A.K., Madden, P.A., Visscher, P.M., McEvoy, B.P., Yang, J., Montgomery, G. W., 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42 (7), 565–569. https://doi.org/10.1038/ng.608.

Heinrich, F., Lange, T.M., Kircher, M., Ramzan, F., Schmitt, A.O., Gültas, M., 2023. Exploring the potential of incremental feature selection to improve genomic prediction accuracy. Genet. Sel. Evol. 55 (1), 78. https://doi.org/10.1186/s12711-023-00853-8.

Houston, R.D., Bishop, S.C., Hamilton, A., Guy, D.R., Tinch, A.E., Taggart, J.B., Derayat, A., McAndrew, B.J., Haley, C.S., 2009. Detection of QTL affecting harvest traits in a commercial Atlantic salmon population. Anim. Genet. 40 (5), 753–755. https://doi.org/10.1111/j.1365-2052.2009.01883.x.

Houston, R.D., Davey, J.W., Bishop, S.C., Lowe, N.R., Mota-Velasco, J.C., Hamilton, A., Guy, D.R., Tinch, A.E., Thomson, M.L., Blaxter, M.L., Gharbi, K., Bron, J.E., Taggart, J.B., 2012. Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. BMC Genomics 13 (1), 244. https://doi.org/10.1186/1471-2164-13-244.

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. 21 (7), 389–409. https://doi.org/10.1038/s41576-020-0227-y.

Howard, R., Carriquiry, A.L., Beavis, W.D., 2014. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 Genes Genomes Genetics 4 (6), 1027–1046. https://doi.org/10.1534/g3.114.010298.

Husemann, M., Tobler, M., McCauley, C., Ding, B., Danley, P.D., 2017. Body shape differences in a pair of closely related Malawi cichlids and their hybrids: effects of genetic variation, phenotypic plasticity, and transgressive segregation. Ecol. Evol. 7 (12), 4336–4346. https://doi.org/10.1002/ece3.2823.

Jeong, S., Kim, J.-Y., Kim, N., 2020. GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. Sci. Rep. 10 (1), 19653. https://doi.org/10.1038/s41598-020-76759-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, 737761. https://doi.org/10.1016/j.aquaculture.2021.737761.

Kaler, A.S., Purcell, L.C., 2019. Estimation of a significance threshold for genome-wide association studies. BMC Genomics 20 (1), 618. https://doi.org/10.1186/s12864-019-5992-7.

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.

Kudinov, A.A., Nousiainen, A., Koskinen, H., Kause, A., 2024. Single-step genomic prediction for body weight and maturity age in Finnish rainbow trout (*Oncorhynchus mykiss*). Aquaculture 585, 740677. https://doi.org/10.1016/j.aquaculture.2024.740677.

Li, B., Zhang, N., Wang, Y.-G., George, A.W., Reverter, A., Li, Y., 2018. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front. Genet. 9, 237. https://doi.org/10.3389/fgene.2018.00237.

Liang, B., Jerry, D.R., Nguyen, V., Shen, X., Koh, J., Terence, C., Nayfa, M.G., Carrai, M., Purushothaman, K., Wen, R.H.J., Mohamed, H., Qing, S.D.T.R., Loo, G., Vij, S., Domingos, J.A., 2025. Genetic parameters and genotype by environment interaction for harvest traits of Malabar red snapper (*Lutjanus malabaricus*). Aquaculture. https://doi.org/10.1016/j.aquaculture.2025.742247, 742247.

Lillehammer, M., Meuwissen, T.H.E., Sonesson, A.K., 2013. A low-marker density implementation of genomic selection in aquaculture using within-family genomic breeding values. Genetic. Select. Evol. (Paris) 45 (1), 39. https://doi.org/10.1186/1297-9686-45-39.

Luo, Z., Yu, Y., Xiang, J., Li, F., 2021. Genomic selection using a subset of SNPs identified by genome-wide association analysis for disease resistance traits in aquaculture species. Aquaculture 539, 736620. https://doi.org/10.1016/j.aquaculture.2021.736620.

Massault, C., Jones, D.B., Lillehammer, M., Cate, H., Harrison, P., Strugnell, J.M., Zenger, K.R., Robinson, N.A., Jerry, D.R., 2025. Accuracies of genomic prediction accounting for genotype by environment remain high when using small sets of selected SNPs in barramundi *Lates calcarifer*. Aquaculture 599, 742138. https://doi.org/10.1016/j.aquaculture.2025.742138.

Meher, P.K., Pradhan, U.K., Ray, M., Gupta, A., Parsad, R., Gupta, P.K., 2025. Ensemble of Bayesian alphabets via constraint weight optimization strategy improves genomic prediction accuracy. G3 Genes Genomes Genetics. https://doi.org/10.1093/g3journal/jkaf150.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics (Austin) 157 (4), 1819–1829. https://doi.org/10.1093/genetics/157.4.1819.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., Lee, D., 2002. BLUPF90 and related programs (BGF90). In: Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France. Communication, 28, pp. 743–744.

Moon, S.Y., Zheng, Y., 2003. Rho GTPase-activating proteins in cell regulation. Trends Cell Biol. 13 (1), 13–22. https://doi.org/10.1016/S0962-8924(02)00004-1.

Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R., Visscher, P.M., 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 11 (4), e1004969. https://doi.org/10.1371/journal.pgen.1004969.

Nguyen, N.H., Premachandra, H.K.A., Kilian, A., Knibb, W., 2018. Genomic prediction using DArT-Seq technology for yellowtail kingfish *Seriola lalandi*. BMC Genomics 19 (1), 107. https://doi.org/10.1186/s12864-018-4493-4.

Nirea, K.G., Sonesson, A.K., Woolliams, J.A., Meuwissen, T.H.E., 2012. Strategies for implementing genomic selection in family-based aquaculture breeding schemes: double haploid sib test populations. Genetic. Select. Evol. (Paris) 44 (1), 30. https://doi.org/10.1186/1297-9686-44-30.

Norman, J.D., Robinson, M., Glebe, B., Ferguson, M.M., Danzmann, R.G., 2012. Genomic arrangement of salinity tolerance QTLs in salmonids: a comparative analysis of Atlantic salmon (*Salmo salar*) with Arctic charr (*Salvelinus alpinus*) and rainbow trout (*Oncorhynchus mykiss*). BMC Genomics 13, 420. https://doi.org/10.1186/1471-2164-13-420.

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 optimization for growth traits of olive flounder (*Paralichthys olivaceus*). Aquac. Rep. 36, 102132. https://doi.org/10.1016/j.aqrep.2024.102132.

Pedersen, E., Brakebusch, C., 2012. Rho GTPase function in development: how in vivo models change our view. Exp. Cell Res. 318 (14), 1779–1787. https://doi.org/10.1016/j.yexcr.2012.05.004.

Purushothaman, K., Vu, N.T., Qing, S.D.T.R., Koh, J., Mohamed, M.H.B., Wen, R.H.J., Liang, B., Loo, G., Domingos, J.A., Jerry, D.R., Vij, S., 2025. Genetic evaluation of nutritional traits in Malabar red snapper (*Lutjanus malabaricus*): heritability and genetic correlations of fatty acid composition. Aquaculture 599, 742144. https://doi.org/10.1016/j.aquaculture.2025.742144.

Reid, D.T., Peichel, C.L., 2010. Perspectives on the genetic architecture of divergence in body shape in sticklebacks. Integr. Comp. Biol. 50 (6), 1057–1066. https://doi.org/10.1093/icb/icq030.

Robledo, D., Palaiokostas, C., Bargelloni, L., Martínez, P., Houston, R., 2018. Applications of genotyping by sequencing in aquaculture breeding and genetics. Rev. Aquac. 10 (3), 670–682. https://doi.org/10.1111/raq.12193.

Schneider, C.A., Rasband, W.S., Eliceiri, K.W., 2012. NIH image to ImageJ: 25 years of image analysis. Nat. Methods 9 (7), 671–675. https://doi.org/10.1038/nmeth.2089.

Scutari, M., Mackay, I., Balding, D.J., 2013. Improving the Efficiency of Genomic Selection. Cornell University Library, Ithaca arXiv.org.

Sverrisdóttir, E., Sundmark, E.H.R., Johnsen, H.Ø., Kirk, H.G., Asp, T., Janss, L., Bryan, G., Nielsen, K.L., 2018. The value of expanding the training population to improve genomic selection models in tetraploid potato [original research]. Front. Plant Sci. 9. https://doi.org/10.3389/fpls.2018.01118.

Teng, J., Zhai, T., Zhang, X., Zhao, C., Wang, W., Tang, H., Wang, D., Shang, Y., Ning, C., Zhang, Q., 2024. Improving multi-population genomic prediction accuracy using multi-trait GBLUP models which incorporate global or local genetic correlation information. Brief. Bioinform. 25 (4). https://doi.org/10.1093/bib/bbae276.

Tsai, H.-Y., Hamilton, A., Tinch, A.E., Guy, D.R., Gharbi, K., Stear, M.J., Matika, O., Bishop, S.C., Houston, R.D., 2015. Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. BMC Genomics 16 (1), 969. https://doi.org/10.1186/s12864-015-2117-9.

Tsai, H.-Y., Hamilton, A., Tinch, A.E., Guy, D.R., Bron, J.E., Taggart, J.B., Gharbi, K., Stear, M., Matika, O., Pong-Wong, R., Bishop, S.C., Houston, R.D., 2016. Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations. Genetic. Select. Evol. (Paris) 48 (1), 47. https://doi.org/10.1186/s12711-016-0226-9.

Tsai, H.-Y., Matika, O., Edwards, S.M., 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 (Bethesda, Md.) 7 (4), 1377–1383. https://doi.org/10.1534/g3.117.040717.

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. Genetic. Select. Evol. (Paris) 49 (1), 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. (1986) 135 (4), 263–274. https://doi.org/10.1111/jbg.12335.

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.

Vela-Avitúa, S., Meuwissen, T.H., Luan, T., Ødegård, J., 2015. Accuracy of genomic selection for a sib-evaluated trait using identity-by-state and identity-by-descent relationships. Genetic. Select. Evol. (Paris) 47 (1), 9. https://doi.org/10.1186/s12711-014-0084-2.

Wang, K., Yang, B., Li, Q., Liu, S., 2022. Systematic evaluation of genomic prediction algorithms for genomic prediction and breeding of aquatic animals. Genes 13 (12), 2247. https://www.mdpi.com/2073-4425/13/12/2247.

Wickham, H., 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York. https://ggplot2.tidyverse.org.

Xu, W., Liu, X., Liao, M., Xiao, S., Zheng, M., Yao, T., Chen, Z., Huang, L., Zhang, Z., 2021. FMixFN: a fast big data-oriented genomic selection model based on an iterative conditional expectation algorithm. Front. Genet. 12, 721600. https://doi.org/10.3389/fgene.2021.721600.

Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., Price, A.L., 2014. Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. 46 (2), 100–106. https://doi.org/10.1038/ng.2876.

Yin, L., Zhang, H., Zhou, X., Yuan, X., Zhao, S., Li, X., Liu, X., 2020. KAML: improving genomic prediction accuracy of complex traits using machine learning determined parameters. Genome Biol. 21 (1), 146. https://doi.org/10.1186/s13059-020-02052-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.

Zhang, Z., Schwartz, S., Wagner, L., Miller, W., 2000. A greedy algorithm for aligning DNA sequences. J. Comput. Biol. 7 (1–2), 203–214. https://doi.org/10.1089/10665270050081478.

Zhang, H., Yin, L., Wang, M., Yuan, X., Liu, X., 2019. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Front. Genet. 10, 189. https://doi.org/10.3389/fgene.2019.00189.

Zhou, L., Wang, Y., Yu, J., Yang, H., Kang, H., Zhang, S., Wang, C., Liu, J., 2019. Improving genomic prediction for two Yorkshire populations with a limited size using the single-step method. Anim. Genet. 50 (4), 391–394. https://doi.org/10.1111/age.12806.

Zhou, L., Zhu, L., Ma, F., Gu, M., Na, R., Zhang, W., 2025. Assessing the impact of different mixing strategies on genomic prediction accuracy for beef cattle breeding values in multi-breed genomic prediction. Animals (Basel) 15 (16). https://doi.org/10.3390/ani15162463.