Aquaculture Reports 47 (2026) 103464

Contents lists available at ScienceDirect

Aquaculture Reports

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

Genomic prediction of survival traits for scuticociliatosis resistance in a vaccinated olive flounder cohort: Comprehensive evaluation and optimization of statistical and machine learning models

Yasara Kavindi Kodagoda[a,b,1], Gaeun Kim[a,b,1], D.S. Liyanage[a,b], W.K.M. Omeka[a,b], Cheonguk Park[a,b], Jeongeun Kim[a,b], H.A.C.R. Hanchapola[a,b], M.A.H. Dilshan[a,b], D.C.G. Rodrigo[a,b], G.A.N.P. Ganepola[a,b], David B. Jones[c], Cecile Massault[c], Dean R. Jerry[c,d], Jihun Lee[a,b,*], Jehee Lee[a,b,*]

[a] *Department of Marine Life Sciences & Center for Genomic Selection in Korean Aquaculture, Jeju National University, Jeju 63243, Republic of Korea*
[b] *Marine Life Research Institute, Jeju National University, Jeju 63333, Republic of Korea*
[c] *Centre for Sustainable Tropical Fisheries and Aquaculture, College of Science and Engineering, James Cook University, Townsville, QLD 4811, Australia*
[d] *Tropical Futures Institute, James Cook University, Singapore*

ARTICLE INFO

Keywords:
Aquaculture
Genomic prediction
Genomic selection
Olive flounder
Vaccination
Disease resistance
Scuticociliatosis resistance
M. avidus

ABSTRACT

Genomic prediction utilizes genome-wide markers to facilitate genomic selection in aquaculture, enabling more precise and accelerated genetic gains compared to traditional pedigree-based methods. Training population size and marker density influence genomic prediction accuracy, and both contribute to the overall cost of selection programs. Scuticociliatosis, caused by the parasite *Miamiensis avidus*, poses a major threat to olive flounder aquaculture and can be managed through selective breeding. We evaluated the effects of marker density and training population size on prediction accuracy using 10 genomic prediction models, including pedigree-based best linear unbiased prediction (BLUP [PBLUP]), genomic BLUP (GBLUP), Bayesian methods (BayesA, BayesB, BayesC, Bayesian Lasso, Bayesian Ridge Regression), regularized regression (Ridge Regression and Elastic Net [EN]), and random forest (RF). The analysis was conducted using 474 genotyped individuals comprising 60 paternal half-sib families. Model performance was evaluated using 5-fold cross-validation repeated 10 times. GBLUP and Bayesian methods consistently outperformed PBLUP, EN, and RF across all survival traits. GBLUP achieved a significantly higher prediction accuracy (0.64–0.72), which is more than twice that of PBLUP (0.06–0.31). The highest predictive ability (0.558 ± 0.006) was achieved with the top 1000 GWAS-ranked markers for the GBLUP model, highlighting the importance of informed marker selection. Contrastingly, random marker selection exhibited no clear gains in predictive performance. Predictive ability was improved by increasing training population size, with no significant differences between 3- and 5-fold cross-validation at larger sample sizes ($p < 0.001$). These findings underscore the importance of selecting appropriate genomic prediction models tailored to the genetic architecture of survival traits, and the benefit of GWAS-informed marker selection to maximize genomic prediction ability in olive flounder aquaculture breeding programs.

1. Introduction

Genomic selection is revolutionizing aquaculture breeding by using genome-wide markers to predict the genetic merit of individuals, enabling the early and accurate identification of superior broodstock and accelerating genetic gains for economically important traits such as growth, disease resistance, and environmental adaptability (Yáñez et al., 2014). Contrary to traditional pedigree-based methods, genomic selection uses genome-wide markers to calculate genomic estimated breeding values (gEBVs), enabling higher accuracy in candidate selection and greater genetic gain, with reported improvements of up to approximately 10 % for body-weight traits (Zenger et al., 2019). This approach

* Correspondence to: Marine Molecular Genetics Lab, Jeju National University, 102 Jejudaehakno, Jeju 63243, Republic of Korea.
E-mail addresses: jihun137@jejunu.ac.kr (J. Lee), jehee@jejunu.ac.kr (J. Lee).

[1] These authors contributed equally to this work

https://doi.org/10.1016/j.aqrep.2026.103464
Received 8 October 2025; Received in revised form 5 February 2026; Accepted 13 February 2026
Available online 20 February 2026

is particularly advantageous in aquaculture species with large family sizes, such as shrimp (*Litopenaeus vannamei*) and Atlantic salmon (*Salmo salar*), because genomic selection enables large-scale communal breeding while mitigating realized inbreeding risk by up to 81 % when combined with optimized mating and contribution strategies (Vandeputte and Haffray, 2014; Verbyla et al., 2022). Pedigree-based best linear unbiased prediction (BLUP) [PBLUP] uses expected relationships from the pedigree and can capture both between- and within-family genetic differences; however, when many full-sibs lack their own phenotypes (as in lethal disease challenges), its ability to distinguish Mendelian sampling differences among them is limited. In contrast, genomic prediction uses dense genome-wide marker data to build a genomic relationship matrix (GRM) that more accurately reflects realized relationships, enabling better capture of within-family genetic variance and more accurate estimation of breeding values (Zenger et al., 2019). Genomic prediction can be implemented either by replacing the pedigree relationship matrix with a marker-based GRM (GBLUP) or by estimating single nucleotide polymorphism (SNP) effects in a reference population and applying them to selection candidates (SNP-BLUP or ridge regression [RR]-BLUP), both yielding gEBVs that enable earlier and more accurate selection (Hosoya et al., 2021; Somsiam et al., 2024). By integrating low-cost genotyping and advanced statistical models, genomic selection optimizes traits critical for sustainability, including disease resistance and environmental adaptability (Allal and Nguyen, 2022). Its commercial implementation, for example, in Tasmanian salmon breeding, demonstrated marked gains in productivity and genetic diversity, indicating that genomic selection is critical for resilient aquaculture systems considering climate change (Verbyla et al., 2022).

Genomic prediction models enable more accurate estimates of breeding values by capturing the contribution of genome-wide markers to complex traits, offering particular advantages in populations where direct evaluation of disease resistance is difficult or expensive (Song et al., 2023). Despite these advances, the optimal application of these models remains challenging, as several factors such as marker density, training population size, and algorithm choice substantially affect prediction accuracy and the design of cost-effective breeding programs (Zhang et al., 2019). Higher SNP densities and larger training populations increase the accuracy of genomic prediction, while model and trait-specific factors (e.g., heritability, genotyping error, algorithm choice) are also crucial (Massault et al., 2025). Comprehensive evaluations confirm that the model performance depends on the underlying genetic architecture of the traits (Song and Hu, 2022). Prior research has shown that advanced models such as GBLUP and Bayesian methods frequently outperform pedigree-based BLUP, particularly when GWAS-informed SNPs are used to predict complex traits in aquaculture. For example, Dong et al., (2016) found increased prediction accuracy for growth traits in large yellow croaker when SNPs with the largest effects were used. Yoshida and Yáñez, (2022) reported enhanced prediction for growth under thermal stress in rainbow trout using preselected GWAS variants. Kriaridou et al., (2020) showed that increasing the training population size raised genomic prediction accuracy but gains diminished and plateaued when the training set reached approximately 1200–1600 individuals, indicating an inflection point for cost-effective cohort design.

In aquaculture, genomic prediction for disease resistance traits, especially under vaccinated conditions that simulate production environments, remains relatively underdeveloped (Jiang et al., 2025). Scuticociliatosis, due to *Miamiensis avidus*, is recognized as a major cause of economic loss in olive flounder (*Paralichthys olivaceus*) production, where selective breeding for genetic resistance is increasingly viewed as a sustainable complement to vaccination strategies (Kodagoda et al., 2025). The high prevalence of *M. avidus* and the routine adoption of vaccination as a management strategy underscore the need to evaluate genomic prediction performance directly in vaccinated cohorts in order to optimize selected broodstock for the environments and biosecurity regimes encountered in real-world production systems. Evidence from salmonids shows that resistance expressed with and without vaccination is only partially correlated. This is demonstrated by the relatively low genetic correlations for resistance to bacterial diseases such as furunculosis when comparing vaccinated and unvaccinated fish. These findings indicate that these responses represent partly distinct traits and that selection in unvaccinated challenge tests may not maximize realized gains in vaccinated farms (Drangsholt et al., 2012, 2011). Vaccination can modify both phenotypic variability and the genetic architecture of resistance, potentially altering the genetic markers and models most effective for selection. Host genetic variation is a major factor influencing vaccine efficacy and disease resistance in aquaculture species, with differences in vaccine response among families reflecting underlying genetic diversity. Some families display pronounced survival benefits following immunization, whereas others derive limited protection (Figueroa et al., 2020). Building on this insight, selective breeding programs in olive flounder offer the potential to harness host genetic variation to improve vaccine responsiveness and strengthen disease resistance across farmed populations.

While genomic selection has been recently implemented in national olive flounder breeding programs in South Korea, its application has focused mainly on growth, thermal tolerance, and viral resistance traits (Liyanage et al., 2025; Omeka et al., 2024; Udayantha et al., 2025). Studies on genomic selection targeting parasitic resistance remain limited, despite scuticociliatosis being the most prevalent parasitic disease affecting olive flounder aquaculture on Jeju Island, South Korea. While multivalent vaccines against *M. avidus* have helped to reduce mortality, their efficacy remains constrained by antigenic diversity and host genetic variation in immune responsiveness (Shivam et al., 2021; Sohn et al., 2023). This underscores the need for complementary genetic strategies to accelerate genetic gain in resistance, as our previous GWAS on a vaccinated population revealed low-to-moderate heritability, a polygenic basis, and significant marker–trait associations for scuticociliatosis resistance (Kodagoda et al., 2025).

This study, being the first of its kind, aimed to comprehensively evaluate genomic prediction for scuticociliatosis resistance in a vaccinated olive flounder breeding population by: comparing alternative prediction models and trait definitions, and assessing the impact of marker density and training population size on the accuracy of survival-based genomic predictions to inform cost-efficient selection programs. We compared 10 models, including conventional approaches such as PBLUP and GBLUP, with Bayesian, regularized regression, and machine learning methods, and explored the use of GWAS-informed markers to develop computationally efficient genomic selection strategies. We also assessed correlated survival-related traits, binary survival, days post-challenge to death (DPC Date), and hours post-challenge to death (DPC Time), as potential proxies for predicting scuticociliatosis resistance. Overall, our work provides practical insights into model selection and data requirements for effective genomic selection and integration into commercial olive flounder breeding programs targeting improved disease resistance.

1.1. Methodology

1.1.1. Base population and husbandry

The base population (dF0) was produced by crossbreeding parent candidates from a broodstock of olive flounder individuals maintained under commercial breeding conditions at a breeding center on Jeju Island, South Korea, as outlined in our previous study (Kodagoda et al., 2025). Briefly, a controlled crossbreeding design was implemented using 16 mating groups, each consisting of 12 dams and 5 sires. GRMs were constructed from SNP genotypes using the gaston R package (Perdry and Dandine-Roulland, 2015) following the method proposed by VanRaden (2008). Mate allocation was optimized using these GRM-based genomic relatedness coefficients, with maximum thresholds of 0.15 between sexes and 0.25 within sexes. Fertilization was achieved via strip spawning by pooling gametes from all individuals within each group and incubating them in tanks. All broodstock underwent high-density SNP genotyping using a custom-designed 70 K Axiom myDesign SNP array (Thermo Fisher Scientific, MA, USA) and were individually tagged with radio-frequency identification microchips (Trovan, London, UK) to enable precise parentage assignment and GRM estimation for progeny analysis (Kodagoda et al., 2025).

Larvae of mixed families were reared under standardized aquaculture conditions until 90 days post-hatching. Subsequently, the fingerlings were raised until they were approximately 10 cm long. The water temperature was maintained at 18–22 ℃ using filtered underground seawater. Rigorous biosecurity protocols, including routine health checks and water quality monitoring, ensured disease-free conditions throughout the rearing period.

1.2. Experimental design and dataset

The phenotypic and SNP data were obtained from the Flounder Genomic Selection Project (2022), as described in our previous work (Kodagoda et al., 2025). Data from 474 fish, phenotyped and genotyped, were used to develop and validate genomic prediction models. Phenotypic data were collected following vaccination with a formalin-killed *M. avidus* vaccine and subsequent challenge with live *M. avidus* (JJB1403) strain via intraperitoneal injection at 4×10^5 cells per fish. Additionally, 100 unvaccinated fish were included as controls and challenged together with vaccinated fish following the same procedure. The challenge experiment was conducted for up to 13 days, with fish mortality recorded every 2 h. Monitoring was terminated when survival curves plateaued with no mortalities for 24–48 consecutive hours and surviving fish showed no clinical signs of active infection. Unvaccinated controls ended on day 7 (77 % cumulative mortality), while vaccinated fish reached criteria on day 13 (52 % cumulative mortality by day 12) to capture complete survival distribution and vaccine protection kinetics.

Fish that survived until the end of the experiment were classified as survivors. Three survival traits were evaluated: Binary Survival (1 for survivors, 0 for mortalities), days post-challenge to death (DPC Date, e.g., a value of 2 for fish surviving 2 days post-challenge), and hours post-challenge to death (DPC Time, e.g., a value of 54 for fish surviving 2 days and 6 h post-challenge) (Kodagoda et al., 2025). These three survival traits provide a comprehensive measure of disease resistance as the ultimate breeding objective, capturing both complete resistance (Binary Survival) and disease progression timing (DPC Date and DPC Time) while integrating outcomes from all immune pathways. Fin clips were collected from all the vaccinated and challenged fish for subsequent genomic DNA extraction. Morphological records were collected from all fish following the disease challenge experiment, including measurements from fish upon mortality and from survivors at the end of the challenge period. Morphological records included body weight (g), measured using a digital balance; standard length (cm), measured from the tip of the snout to the caudal peduncle using digital calipers; and maximum body width (cm), measured at the widest point of the body using digital calipers. Clinical signs, including ascites, ulceration, fin loss, and hemorrhage, were recorded to confirm disease-related mortality. Tank number was also recorded to account for environmental effects. All experimental procedures were conducted considering the institutional animal care guidelines of the Jeju National University (Approval Number 2024–0054).

1.3. Genomic DNA extraction, genotyping, quality control, and genotype matrix construction

Genomic DNA was extracted from the fin clips (∼ 50 mg) collected from 581 fish (474 progeny and 107 broodstock) using the DNeasy Blood & Tissue Kit (Qiagen, Munich, Germany), following the manufacturer's instructions. DNA purity and concentration were determined using a Thermo Scientific Multiskan Sky Microplate Spectrophotometer (Thermo Fisher Scientific) and samples were diluted to 50 ng/μL with nuclease-free water (Thermo Fisher Scientific).

All 581 individuals were genotyped using the custom Affymetrix Axiom myDesign 70 K SNP array at BluGen Aquaculture Genomics, South Korea (see Section 2.3) (Kodagoda et al., 2025; Liyanage et al., 2022). Raw genotype data were processed using the Axiom Analysis Suite 5.3, with standard quality filters applied to exclude poorly clustered variants using established criteria: dish quality control (DQC $\geq$0.80), sample call rate ($\geq$0.95), and plate-level passing rates ($\geq$0.95). Only high-confidence polymorphic SNPs classified in the Poly-HighResolution category were retained after assessing clustering performance and allele distributions.

Further quality control (QC) was performed using PLINK v1.9 (https://www.cog-genomics.org/plink/) to exclude the SNPs with a minor allele frequency (MAF) < 0.05, genotyping rate < 0.10, or Hardy–Weinberg equilibrium deviations ($p \leq 0.01$). After QC filtering, a final dataset comprising 52,046 high-quality SNPs and 581 samples (474 progeny and 107 broodstock) was retained, yielding a call rate of 0.99. The genomic SNP density across the genome (24 chromosomes) was visualized using the CMPlot R package (Yin et al., 2021), with green indicating conserved regions of low SNP density and red highlighting regions of high SNP density (Supplementary Figure 1).

A standardized genotype matrix Z (individuals × markers) was constructed from the QC-filtered SNPs to prepare the dataset for genomic prediction analyses. SNP genotypes were encoded additively as -1 (homozygous for the reference allele), 0 (heterozygous), and 1 (homozygous for the alternative allele), representing allele dosage suitable for downstream statistical modeling. This matrix was used to construct GRMs, fit Bayesian regression models, and implement penalized and machine learning approaches (S. Wang et al., 2025).

A GRM, quantifying the proportion of the genome shared between individuals, was computed from the standardized genotype matrix (Z), which was imported using the bed.matrix() function from the gaston package in R (https://cran.r-project.org/package=gaston). The GRM was calculated using the GRM() function in the same package, following the method proposed by VanRaden (2008) (VanRaden, 2008). The inverse of the GRM was subsequently obtained using the ginv() function from the MASS package (https://cran.r-project.org/package=MASS). The GRM was used to visualize population structure through a heatmap and to construct a phylogenetic tree (VanRaden, 2008).

1.4. Parentage assignment and relationship Matrix construction

Parentage assignments were determined using QC-filtered genotyping data from 1000 SNPs across parents and progenies. A custom R script implemented the opposite homozygote count method (Hayes, 2011), followed by validation using Cervus software v3.0.7 (http://www.fieldgenetic.com/pages/aboutCervus_Overview.jsp) to resolve ambiguous assignments.

The corresponding parentage assignments were used in PBLUP for pedigree reconstruction. The additive relationship matrix (A) was computed using the A.matrix() function from the AGHmatrix R package (https://cran.r-project.org/package=AGHmatrix), following the Henderson method, with support for flexible ploidy and inverse matrix estimation (Kalinowski et al., 2007).

1.5. Estimation of gEBV and model optimization

For the estimation of single-trait gEBVs, nine prediction methods (statistical and machine learning methods) were evaluated for survival traits in this study. They included PBLUP, GBLUP, and five Bayesian methods: BayesA (BA), BayesB (BB), BayesC (BC), Bayesian Lasso (BL), and Bayesian Ridge Regression (BRR), two regression approaches: ridge regression (RR) and elastic net (EN), and the machine-learning method: random forest (RF). Model performance was assessed as the correlation coefficient between predicted and observed values, using 5-fold cross-validation, based on varying numbers of high-quality SNP data.

1.5.1. GBLUP and PBLUP

gEBVs for each survival trait were obtained using single-trait PBLUP and GBLUP models, which were implemented through the mixed.solve() function from the rrBLUP package in R (https://cran.r-project.org/package=rrBLUP) (Endelman, 2011), applying the mixed linear animal model formulated as:

$$y = \mu + Zu + e \tag{1}$$

where y is the vector of phenotypic observations, μ is the overall mean, Z is the incidence matrix relating observations to random genetic effects, u is the vector of individual breeding values, and e is the vector of random residuals. The breeding values u were assumed to follow u ∼ N (0, Kσ_a^2), where σ_a^2 represents the additive genetic variance and K is the relationship matrix defining the covariance structure among individuals. To assess the influence of different relationship matrices, two versions of K were tested: (i) A matrix constructed from traditional parentage records for PBLUP, and (ii) the GRM derived from SNP marker data for GBLUP (see Sections 2.4 and 2.5). Residuals were assumed to follow e ∼ N (0, Iσ_e^2), where σ_e^2 is the residual variance, and I is the identity matrix. Narrow-sense univariate heritability (h^2) was estimated for each trait using both PBLUP and GBLUP models following the formula:

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

where h^2 is the narrow sense heritability, σ_a^2, σ_p^2, and σ_e^2 are the additive genetic, phenotypic, and residual variances, respectively.

Additionally, genetic correlations were estimated for all pairwise combinations of six traits (Binary Survival, DPC Date, DPC Time, Weight, Length, and Width) using bivariate restricted maximum likelihood (REML) models implemented in GCTA (v1. 93.2), with the GRM constructed from genome-wide SNP data. The bivariate model was expressed using the following formula:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{3}$$

where y_1 and y_2 are vectors of phenotypes for the two traits, μ_1 and μ_2 are the intercepts or fixed effects, Z_1 and Z_2 are incidence matrices relating observations to random genetic effects u_1 and u_2, and e_1 and e_2 are vectors of residuals. The genetic correlation (r_g) was calculated as the genetic covariance (σ_{g12}) between two traits divided by the square root of the product of their genetic variances (σ_{g1}^2 and σ_{g2}^2):

$$r_g = \frac{\sigma_{g_{12}}}{\sqrt{\sigma_{g_1}^2 \sigma_{g_2}^2}} \tag{4}$$

1.5.2. Bayesian model

Bayesian generalized linear regression models (BA, BB, BC, BL, and BRR) were implemented using the BGLR package (https://cran.r-project.org/package=BGLR) (Pérez and De Los Campos, 2014). Unlike GBLUP, standard BRR assumes a homogeneous Gaussian prior for all marker effects. The Bayesian framework models phenotypic variation as:

$$y = \mu + \Sigma \text{ (from } j = 1 \text{ to } n) \; Z_j \, g_j + e \tag{5}$$

where Z_j is the genotype vector for SNP j, g_j is its marker effect size, and e ∼ N (0,Iσ_e^2) is the residual error. Each Bayesian model employed a distinct prior on the marker effects, g_j.

BA assumed that all SNPs contributed to phenotypic variation, modeling marker effects via a scaled t-distribution; $g_j \sim t_\upsilon$ (0,S), where υ (degrees of freedom) and S (scale parameter) governed the genetic variance (σ^2g) and tail thickness. BB and BC presumed sparse genetic architectures, assigning non-zero effects via mixture priors (a point mass at zero combined with Gaussian slabs). BL applied a Laplace prior for marker-wise shrinkage: $g_j \sim$ Laplace (0, λ), λ controlling the sparsity level. BRR assumed a constant variance structure $\left(g_i \sim N\left(0, \sigma_g^2\right)\right)$, inducing stronger shrinkage for SNPs with extreme minor allele frequencies. While contrasting with GBLUP's locus-specific variance assumptions, BRR's Gaussian prior aligns with the infinitesimal model underlying GBLUP and RR (Thavamanikumar et al., 2015).

1.5.3. Penalized regression and machine learning models

RR was employed using a penalized least-squares approach with an L2-norm penalty to estimate regression coefficients using the rrBLUP package in R. The GBLUP, RR, and BRR models were considered theoretically equivalent under specific parametrizations, although differences were observed in hyperparameter tuning and software implementations. EN was employed to combine L1- and L2-norm penalties, where α = 1 corresponds to lasso (sparse selection) and α = 0 to RR (shrinkage), using the glmnet package in R (https://cran.r-project.org/package=glmnet). As a supervised machine learning approach, RF was used with decision trees to model additive genetic effects, formulated as Y = F(x) + μ + Ψ, where μ captures polygenic random effects and Ψ represents residual noise (https://cran.r-project.org/package=randomForest). F(x) represents additive genetic contributions through ensemble tree-based predictions (Zhang et al., 2024).

1.5.4. Marker configuration and density optimization

To evaluate the effect of marker density on predictive performance, all prediction models were evaluated using SNP subsets of varying sizes (500, 1 K, 5 K, 10 K, 20 K, 30 K, 40 K, and 50 K markers). For each density, the standardized genotype matrix Z (encoded as –1, 0, and 1) was reconstructed using only the selected SNP subset, enabling direct comparison across marker selection scenarios. Two strategies were employed to generate these SNP subsets. First, a random subsampling approach (random selection) simulated reduced-density SNP panels by randomly selecting markers from the full set while maintaining representative genome-wide coverage.

Second, a trait-informed selection strategy (GWAS-ranked) ranked SNPs by ascending *p*-values from three separate GWAS analyses performed for each survival trait (Binary Survival, DPC Date, and DPC Time) to identify markers associated with scuticociliatosis resistance, prioritizing markers with the strongest statistical associations to each target trait (Habier et al., 2009; Wang et al., 2018). This resulted in three trait-specific marker sets for genomic prediction evaluation.

1.5.5. Training population size and evaluation of genomic predictive ability

To assess the predictive performance of genomic prediction models, repeated cross-validation strategies were implemented with systematic variation in training population size. The full dataset of 474 individuals was randomly partitioned into training (reference) and testing (validation) sets using 3- and 5-fold cross-validation, each repeated 10 times with different random seeds to minimize partitioning bias. Within each replicate, phenotypes for validation individuals were masked, and predictions were generated using the respective models (Meuwissen et al., 2001).

Predictive ability (r) was calculated as the Pearson correlation coefficient between predicted gEBVs and observed phenotypes (y) in the validation sets (r = cor(gEBV, y)). Predictive accuracy was calculated by scaling predictive ability by the square root of heritability: $r/\sqrt{h^2}$ (or cor (gEBV, y)/$\sqrt{h^2}$) to estimate the correlation between predicted breeding values and true breeding values. By varying the number of folds, we systematically evaluated model performance across different training population sizes.

1.5.6. Statistical analysis

The nonparametric Friedman test was employed to evaluate statistical significance in predictive performance among genomic prediction methods, evaluated over repeated cross-validation on the same dataset, with Kendall's coefficient of concordance (W) calculated to assess the consistency of method rankings across the cross-validation folds (Schrauf et al., 2021). Pairwise differences between methods and their rankings were further examined using the post-hoc Nemenyi test.

2. Results

2.1. Descriptive statistics, parentage assignment, and genomic relatedness

The challenge experiment was conducted for 12 days, during which survival analysis revealed a stepwise decline over time, with a median survival of 12 days, indicating that 50 % of the subjects remained alive after 12 days (Figure 1The vaccinated group exhibited a higher survival probability over time and a longer median survival compared with those of the unvaccinated control (3 days) (Fig. 1). From the vaccinated cohort, 474 phenotyped and genotyped fish were selected to develop and validate genomic prediction models. Their descriptive statistics are summarized in Table 1.

Parentage assignment of the study population (474) from dF0 progeny identified 60 paternal half-sib families of approximately 10 fish per family (Fig. 2A). Based on visual inspection of the GRM heatmap, five subpopulations were identified as distinct blocks along the diagonal, indicating clusters of more closely related individuals within families. Genomic relationships averaged near zero and ranged from −0.27 to 1.28 (Fig. 2B), with positive values indicating above-average relatedness and negative values indicating below-average relatedness within the population.

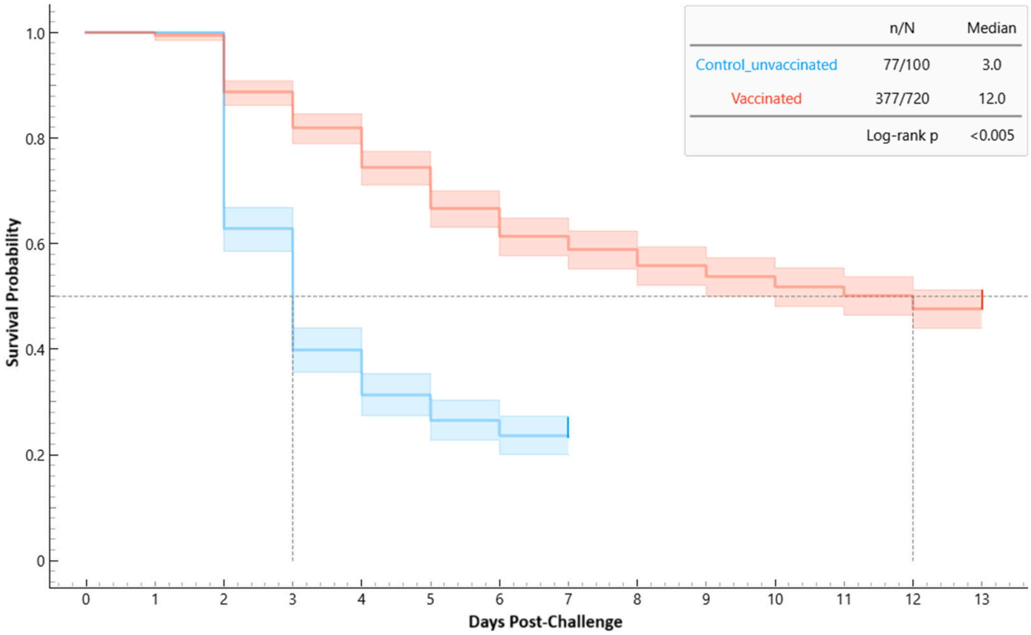
Heritability estimates were obtained from univariate GBLUP models fitted separately for each trait. Morphological traits exhibited moderate heritabilities (Length: 0.39 ± 0.07, Width: 0.34 ± 0.07, and Weight: 0.37 ± 0.07), while survival traits exhibited lower heritabilities (binary survival: 0.08 ± 0.05, DPC Date: 0.10 ± 0.05, and DPC Time: 0.11 ± 0.06). Genetic correlation among traits was estimated using bivariate models (Fig. 2C). Strong positive genetic correlations were observed among morphological traits, and between survival and time-based traits (Survival–DPC Date, $r_g = 0.84$; Survival–DPC Time, $r_g = 0.84$, and DPC Date–DPC Time, $r_g = 0.75$). Moderate positive correlations were observed for Width and Survival ($r_g = 0.57$), with weaker or non-significant correlations between other trait combinations (Fig. 2C). Heritability estimates from bivariate models with Width were highly consistent with univariate estimates for all survival traits (Binary Survival: 0.10 ± 0.05, DPC Date: 0.11 ± 0.05, DPC Time: 0.11 ± 0.05), with differences within the range of standard errors. The moderate positive genetic correlation between width and survival suggests that selection for increased body width could simultaneously improve disease resistance, enabling joint genetic improvement of morphological and survival traits through multi-trait genomic selection.

Table 1
Descriptive statistics for the recorded binary survival, days post-challenge to death (DPC Date), and hours post-challenge to death (DPC Time).

Trait	No of samples	Min	Max	Mean	SD
Binary Survival	474	0	1	0.47	0.50
DPC Date	474	1	13	8.88	4.42
DPC Time	474	0.79	12.13	197.08	102.40

2.2. Prediction model evaluation for gEBV estimation

The predictive performance of 10 genomic prediction methods for gEBV estimation was evaluated based on the data from 474 genotyped individuals and 52,046 quality-controlled SNPs, using 5-fold cross-validation repeated 10 times (Fig. 3). Statistically significant differences in predictive abilities were observed across methods for the three survival traits: Binary Survival, DPC Date, and DPC Time ($p < 0.001$). Mean predictive ability values for three traits are presented in Supplementary Table 1. For all traits, Bayesian regression methods and GBLUP consistently outperformed traditional pedigree-based and machine learning approaches. Specifically, for Binary Survival, BRR was the highest performing method with a predictive ability value of 0.204 ± 0.03, followed by BL (0.203 ± 0.02) and BA (0.195 ± 0.01) (Fig. 3). Contrastingly, PBLUP, EN, and RF exhibited significantly lower performance than Bayesian and GBLUP methods ($p < 0.001$). For DPC Date, BB achieved the highest predictive ability (0.230 ± 0.03), followed by BL (0.226 ± 0.04) and GBLUP (0.226 ± 0.03), whereas EN and PBLUP exhibited the lowest precision ($p < 0.001$). For DPC Time, BA yielded the highest predictive ability (0.240 ± 0.02), followed by BRR (0.238



	n/N	Median
Control_unvaccinated	77/100	3.0
Vaccinated	377/720	12.0
	Log-rank p	<0.005

Fig. 1. Kaplan–Meier survival probabilities overtime following challenge experiment (days post-challenge) for vaccinated versus unvaccinated control groups. The log-rank test was performed to assess differences in survival, with a $p < 0.005$ indicating a statistical difference between groups.

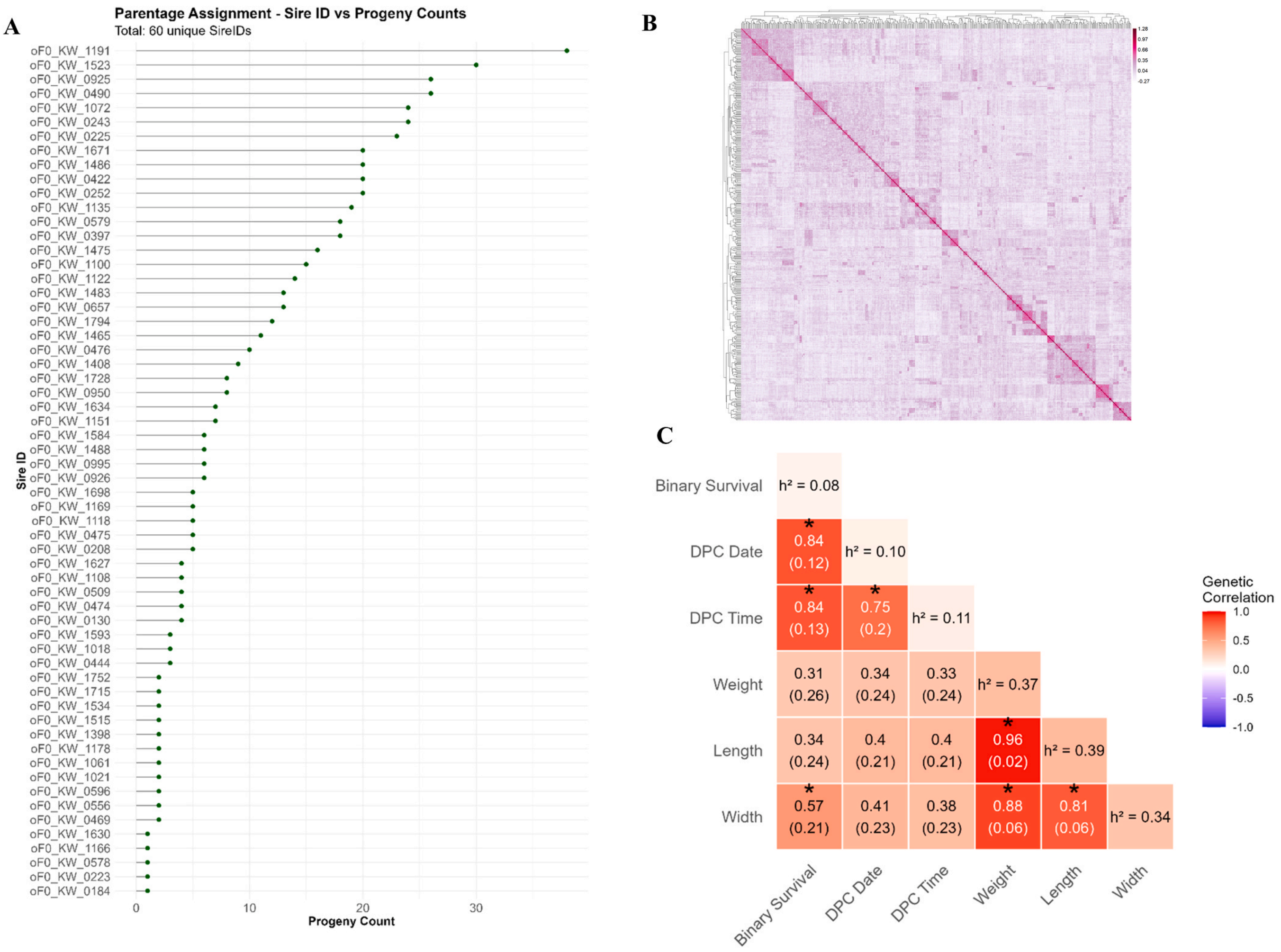


Fig. 2. Overview of genetic structure and trait correlations related to scuticociliatosis resistance in olive flounder. (A) Summary of parentage assignments with paternal identities (Sire ID). (B) Heatmap of the genomic relatedness of fish based on genomic relationship matrix. (C) Heatmap of genetic correlations among six scuticociliatosis resistance-related traits, estimated using bivariate restricted maximum likelihood (REML) in genome-wide complex trait analysis (GCTA). Diagonal values indicate heritability (h^2) with standard errors. Color intensity reflects correlation strength; asterisks indicate statistical significance ($p < 0.05$).

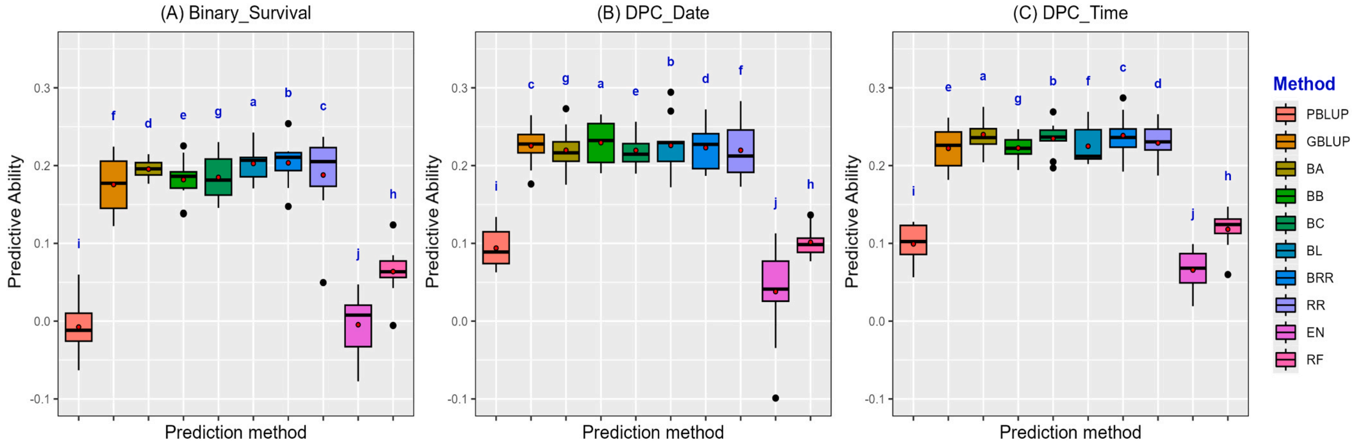


Fig. 3. Comparative predictive performance of 10 genomic prediction methods (PBLUP, GBLUP, BA, BB, BC, BL, BRR, RR, EN, RF) for three survival traits: **(A)** Binary Survival, **(B)** DPC Date, and **(C)** DPC Time. Predictive ability was estimated using 5-fold cross-validation, based on data from 474 genotyped individuals and 52,046 quality-controlled SNPs. The boxes illustrate second and third quartiles, and the whiskers represent interquartile ranges. Red dots represent the mean predictive ability. For each trait, Significant differences among prediction methods were assessed using the Friedman test ($p < 0.001$), followed by Nemenyi post-hoc comparisons. Methods sharing the same letter are not significantly different ($p < 0.05$). Letters are assigned alphabetically with "a" being the best performing group.

± 0.03) and BC (0.235 ± 0.02), whereas EN, RF, and PBLUP exhibited significantly lower values ($p < 0.001$).

Table 2 presents predictive ability and prediction accuracy alongside the estimated heritabilities for the three survival traits, comparing PBLUP and GBLUP methods. Overall, predictive abilities ranged from 0.01 to 0.10 for PBLUP and 0.18–0.23 for GBLUP, while prediction accuracies ranged from 0.06 to 0.31 and 0.64–0.72, respectively (Table 2). Notably, GBLUP exhibited significantly higher prediction accuracies than PBLUP ($p < 0.05$) for all three survival traits: Binary Survival (0.64 ± 0.03), DPC Date (0.72 ± 0.05), and DPC Time (0.68 ± 0.05).

2.3. Effect of marker configuration and density optimization

The effect of different marker densities and marker selection strategies (random selection or GWAS-ranked) on predictive ability was evaluated for three survival traits using the GBLUP method with 5-fold cross validation (Fig. 4). GWAS-based marker selection consistently outperformed random marker selection across all three survival traits using GBLUP. Predictive ability increased with the number of top GWAS-ranked markers, peaking at 1000 markers (Binary Survival: 0.534 ± 0.03; DPC Date: 0.553 ± 0.03; DPC Time: 0.558 ± 0.06) before declining with additional markers (Fig. 4). In contrast, random marker selection showed no clear improvement with increasing marker numbers and consistently performed significantly lower ($p < 0.001$) (Fig. 4). The highest predictive abilities were achieved with top 1000 GWAS-ranked markers using the GBLUP method: 0.534 ± 0.03, 0.553 ± 0.03, and 0.558 ± 0.06 for the three traits: Binary Survival, DPC Date, and DPC Time, respectively (Fig. 4A–C).

Furthermore, the predictive abilities with top 1000 GWAS-ranked markers were compared across nine genomic prediction methods. Based on their mean rank scores, the overall performance ranking for the three survival traits was: GBLUP, BA, BRR, RR, BC, BB, BL, EN, and RF (Fig. 5). The mean predictive ability values across the nine genomic prediction methods, using the top 1000 GWAS-ranked markers, are presented in Table 3. Binary survival trait exhibited comparatively lower performance than the DPC Date and DPC Time traits across all prediction methods (Fig. 5). The data on the 16 top-ranked GWAS-identified markers for three survival traits are summarized in Supplementary Table 2.

2.4. Effect of population size on predictive ability

The effect of training population size on the predictive performance of the GBLUP model with the top 1000 GWAS markers was evaluated under 3- and 5-fold cross-validation schemes (Fig. 6). Training population size significantly affected GBLUP predictive ability in a trait-dependent manner (Friedman test, $p < 0.001$) (Fig. 6). For DPC Time, populations ≥ 200 individuals achieved optimal performance (0.50–0.55 predictive ability) with no significant improvement beyond 200 individuals for both cross-validation schemes. In contrast, DPC Date showed continued improvement up to 300 individuals with 0.53–0.58 predictive ability. Binary Survival exhibited the highest variability and required populations ≥ 350 individuals for stable predictions, with moderate predictive ability (0.50–0.55). Small training populations ($n = 50$–100) resulted in significantly reduced accuracy across all traits, with 5-fold CV providing more robust predictions than 3-fold CV at these sample sizes. These findings indicate trait-specific minimum population requirements ranging from 300 to 350 individuals for effective genomic prediction in scuticociliatosis resistance.

Table 2
Predictive ability, prediction accuracy, and estimated heritability (h^2) for three survival traits from Pedigree-based best linear unbiased prediction (PBLUP), Genomic-based BLUP (GBLUP) models, using 52,046 quality-controlled SNPs and 5-fold cross-validation.

Traits	Estimated Heritability (h^2)		Predictive Ability [cor(gEBV,y)]		Prediction Accuracy [cor(gEBV,y)/$\sqrt{h2}$]	
	PBLUP	GBLUP	PBLUP	GBLUP	PBLUP	GBLUP
Binary Survival	0.033 ± 0.016	0.079 ± 0.050	0.01 ± 0.03	0.18 ± 0.04*	0.06 ± 0.02	0.64 ± 0.03*
DPC Date	0.103 ± 0.026	0.103 ± 0.055	0.09 ± 0.02	0.23 ± 0.03*	0.28 ± 0.04	0.72 ± 0.05*
DPC_Time	0.104 ± 0.027	0.106 ± 0.056	0.10 ± 0.03	0.22 ± 0.03*	0.31 ± 0.04	0.68 ± 0.05*

Notes: For each trait, statistically significant differences between the two prediction methods were evaluated using the Wilcoxon signed-rank test. Star marks (*) in the table indicate significant differences at $p < 0.05$.

3. Discussion

Genomic prediction models are crucial for accurately estimating breeding values of complex polygenic traits with moderate heritability, such as disease resistance in aquaculture (Zenger et al., 2019). These models leverage dense SNP information to explain the genetic variance of complex traits more precisely than pedigree-based methods, accounting for major and minor effect loci (Griot et al., 2021). This leads to higher prediction accuracy and enables the selection of candidates with precise breeding values, accelerating genetic gain even for traits that are difficult or expensive to measure directly, such as survival traits after disease challenge (Houston et al., 2020). While pedigree records provide expected relationships based on Mendelian inheritance, genomic data reveal the actual segregation of chromosomal segments, accounting for Mendelian sampling variation and detecting cryptic relatedness or admixture that may not be evident from pedigree alone (Hayes, 2011; VanRaden, 2008). This improved characterization of realized genetic relationships enhances prediction accuracy, particularly in aquaculture populations where pedigree records may be incomplete or population substructure exists due to breeding schemes or wild introgression (Mao et al., 2023).

Phenotyping for disease resistance often requires expensive and labor-intensive challenge tests or field evaluations, complicating the establishment of sufficient, well-phenotyped training populations essential for reliable genomic selection (Griot et al., 2021; Ødegård et al., 2011). Moreover, the genetic architecture of resistance traits is complex and polygenic, involving many loci with small to large effects. While this complexity poses challenges for marker-assisted selection that relies on identifying specific causal variants, genomic selection can effectively capture these distributed genetic effects through genome-wide marker information (Meuwissen et al., 2001).

The present study aimed to optimize genomic prediction for scuticociliatosis survival traits, despite their low heritability (0.08–0.11), by combining a 70 K high-density SNP array with GWAS-informed putative causative markers and evaluating multiple genomic and machine-learning models to improve prediction accuracy and reliability. In highly polygenic traits, model choice, sample size, and marker density play crucial roles, enabling accuracy to exceed expectations based on heritability estimates (Calus et al., 2008). Although low heritability constrains baseline prediction accuracy, genomic prediction can still be improved by incorporating causative variants (if known) or markers in strong linkage disequilibrium with the traits of interest (de los Campos et al., 2013), utilizing flexible statistical models such as Bayesian approaches that better accommodate complex genetic architectures, and expanding or refining the training population to include closely related individuals (Guarini et al., 2018). Dense SNP coverage enables genomic models to capture small-effect QTL distributed genome-wide, which collectively explain substantial genetic variance even when individual effects are modest (Gebreyesus et al., 2021).

Overall, integrating genomic information with optimized modeling strategies can substantially improve the prediction of low-heritability traits, as consistently shown across species. Despite low–moderate heritability (~0.14) for Vibriosis resistance in Pacific oysters, genome-wide SNP data in a Bayesian genomic selection model increased progeny survival rate by 18.42 % and survival time by 12.73 % compared to

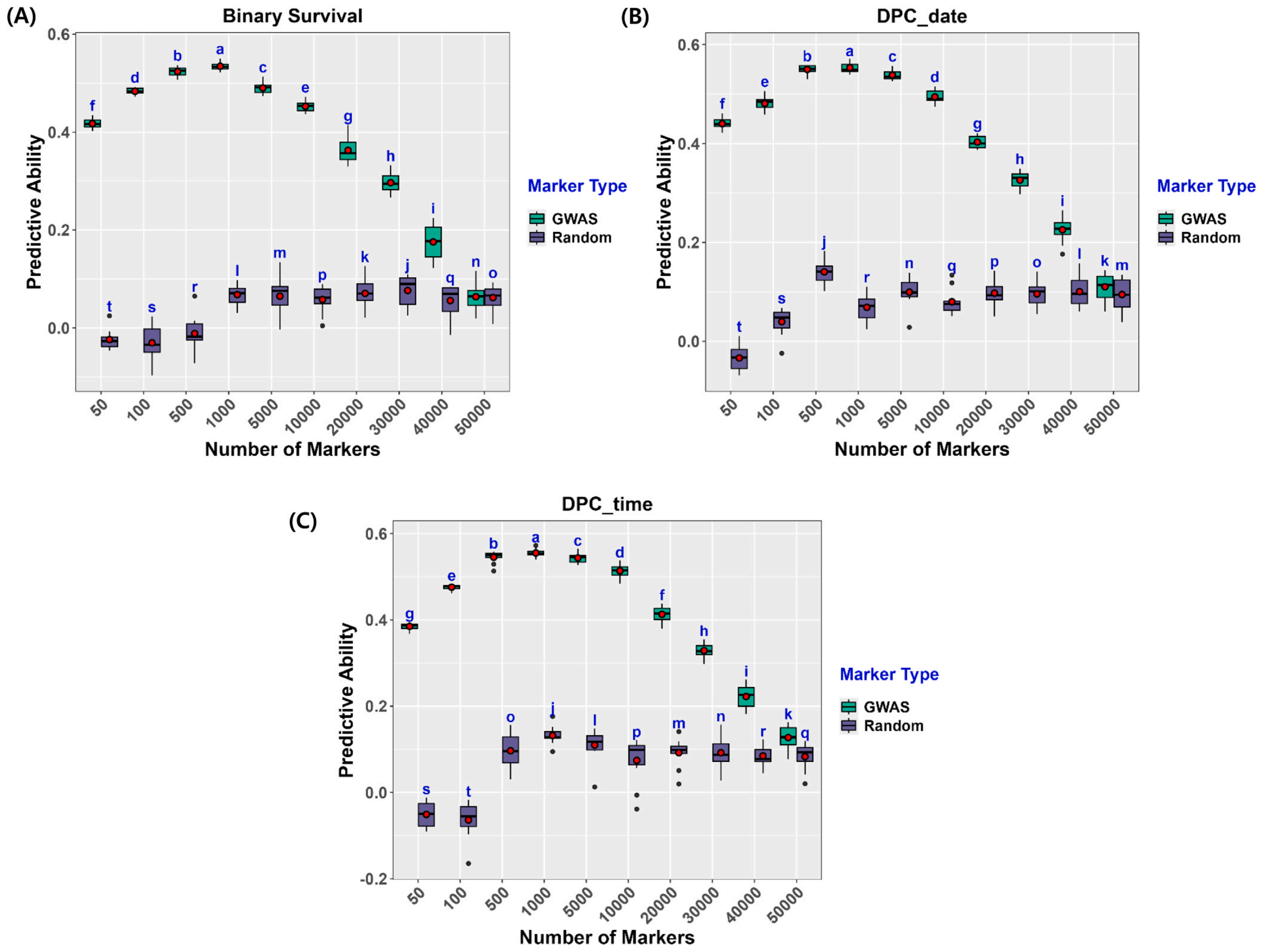


Fig. 4. Comparative predictive performance of the GBLUP model across different marker configurations (50, 100, 500, 1000, 5000, 10,000, 20,000, 30,000, 40,000, and 50,000 SNPs) selected either at random (light green) or based on GWAS significance (purple) for three traits: **(A)** Binary Survival, **(B)** DPC Date, and **(C)** DPC Time. Predictive ability was estimated using 5-fold cross-validation in 474 genotyped individuals. Boxplots show the second and third quartiles, whiskers represent interquartile ranges, and red dots indicate mean predictive ability. Significant differences among marker sizes were detected using the Friedman test ($p < 0.001$), followed by Nemenyi post-hoc comparisons. Marker sizes sharing the same letter are not significantly different ($p < 0.05$). Letters are assigned alphabetically with “a” being the best performing group.

controls (Yang et al., 2024). In striped catfish, genomic and AI models using ~6.5 K SNPs substantially outperformed pedigree-based evaluation for predicting low-heritability *Edwardsiella ictaluri* resistance, with prediction accuracy further improved by incorporating large-effect SNPs (Vu et al., 2021).

Aquaculture breeding programs increasingly aim to enhance disease resistance to reduce dependence on vaccines over the long term. However, current production systems continue to rely heavily on vaccination, and host genetic variation in vaccine-mediated protection implies that selection must be evaluated under conditions representative of commercial practice (Figueroa et al., 2020). In many marine finfish species, including salmonids, vaccines remain fundamental due to high pathogen pressure, and selection based solely on unvaccinated challenge tests may yield suboptimal gains under farm conditions where fish are typically vaccinated (Nguyen, 2024).

While the long-term goal is to reduce dependence on vaccination by increasing innate resistance, current production systems rely on vaccines, and host genetic variation in vaccine-mediated protection means that selection decisions must be based on performance in vaccinated populations to ensure realized gains under farm conditions (Figueroa et al., 2020). In many marine finfish systems including salmonid, vaccines remain essential due to high pathogen pressure and selection based only on unvaccinated challenge tests can result in suboptimal gains under field conditions where fish are vaccinated (Nguyen, 2024). Given the current high prevalence of *M. avidus* in Korean olive flounder production and the widespread use of vaccination as standard management, evaluating genomic prediction directly in vaccinated cohorts ensures that selected broodstock are optimized for the conditions that they will experience in commercial production (Song et al., 2025). Host genetic variation also explains heterogeneity in vaccine protection; some families or genotypes respond much better to vaccination than others, reinforcing the need to evaluate resistance in vaccinated rather than assuming uniform vaccine effects (Safonova et al., 2022). In Atlantic salmon, where low genetic correlations (0.32 ± 0.13) were observed between resistance to furunculosis in vaccinated and unvaccinated fish (Drangsholt et al., 2012, 2011), and vaccine protective capacity against *Piscirickettsia salmonisis* exhibited significant variation (Figueroa et al., 2020), indicating that resistance with and without vaccination is partly different traits; thus selection on unvaccinated fish may not maximize response in vaccinated farms.

Genomic selection and vaccination should therefore be viewed as complementary strategies; genomic selection can increase baseline resistance and reduce outbreak risk and mortality, potentially allowing vaccine use to be reduced in the long term, but in the short–medium

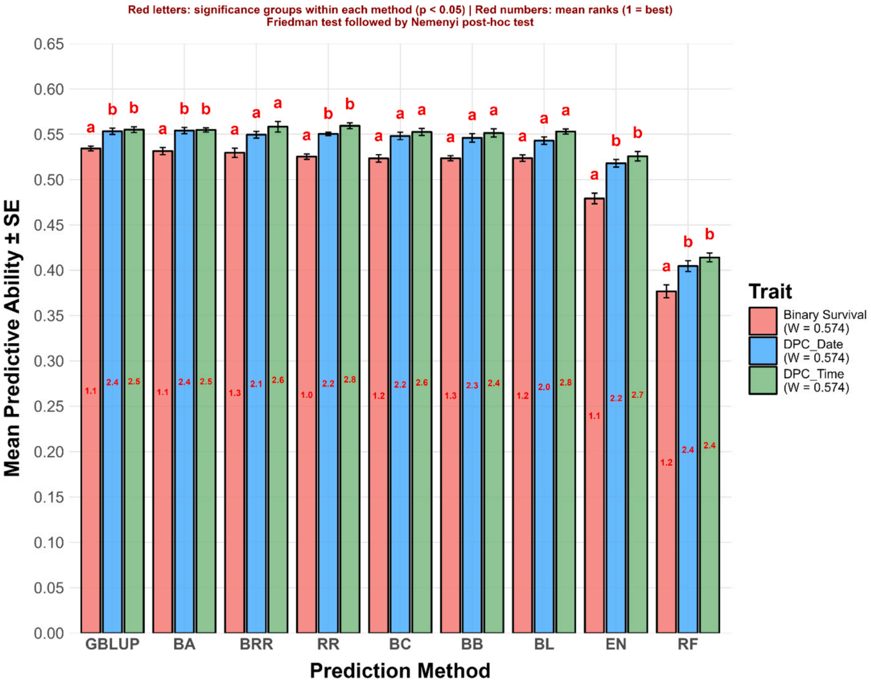


Fig. 5. Comparative predictive performance of genomic prediction methods (GBLUP, BA, BB, BC, BL, BRR, RR, EN, RF) for 3 survival traits: Binary Survival, DPC Date, and DPC Time, using the top 1000 GWAS-ranked SNPs. Predictive ability was estimated via 5-fold cross-validation in 474 genotyped individuals. Significant differences among traits within each prediction method were assessed using the Friedman test ($p < 0.001$), followed by Nemenyi post-hoc comparisons. Red letters above bars indicate significance groups for each method, where traits sharing the same letter are not significantly different ($p < 0.05$). Red numbers denote the mean ranks (1 = best) assigned by the Friedman test. Error bars represent standard errors. Kendall's W values in the legend reflect concordance effect sizes within traits, indicating the consistency of method rankings.

Table 3
Mean predictive ability values of nine genomic prediction methods using top 1000 GWAS-ranked SNPs and 5-fold cross validation.

Prediction Method	Binary Survival	DPC Date	DPC Time
GBLUP	0.534 ± 0.03^{a}	0.553 ± 0.03^{a}	0.558 ± 0.03^{a}
BA	0.532 ± 0.04^{a}	0.554 ± 0.03^{a}	0.555 ± 0.02^{a}
BB	0.524 ± 0.03^{b}	0.546 ± 0.05^{b}	0.551 ± 0.02^{a}
BC	0.523 ± 0.04^{b}	0.548 ± 0.04^{b}	0.553 ± 0.02^{a}
BL	0.524 ± 0.04^{b}	0.543 ± 0.04^{b}	0.553 ± 0.06^{a}
BRR	0.530 ± 0.05^{b}	0.550 ± 0.04^{b}	0.555 ± 0.03^{a}
RR	0.525 ± 0.03^{b}	0.550 ± 0.02^{a}	0.559 ± 0.03^{a}
EN	0.479 ± 0.06^{b}	0.518 ± 0.04^{b}	0.526 ± 0.03^{b}
RF	0.377 ± 0.07^{b}	0.405 ± 0.06^{b}	0.414 ± 0.02^{b}

Notes: For each trait, statistically significant differences among the methods were detected using the Friedman test ($p < 0.001$) with Post-hoc Nemenyi tests grouped methods based on significant differences. Methods sharing the same superscript letter are not significantly different ($p < 0.05$), while different letters indicate significant differences in predictive performance. Superscript letters are assigned alphabetically, with "a" representing the best performing group. Abbreviations: Genome-wide association studies (GWAS), Single nucleotide polymorphisms (SNPs), Pedigree-based best linear unbiased prediction (PBLUP), Genomic-based BLUP (GBLUP), BayesA (BA), BayesB (BB), BayesC (BC), Bayesian Lasso (BL), Bayesian Ridge Regression (BRR), Ridge regression (RR), elastic net (EN), random forest (RF).

term, breeding must be evaluated under current biosecurity regimes. Studies in terrestrial livestock, where genomic prediction of vaccine-mediated immune responses and antibody traits (for example, Newcastle disease vaccination in chickens and PRRSV vaccination in pigs), have shown that these traits are heritable and predictable, support the idea that survival or resistance measured under vaccination is an appropriate breeding goal, and position the present study as an analogous effort focused on survival under vaccination as the relevant expression of resistance in olive flounder aquaculture (Hako Touko et al., 2021; Hickmann et al., 2021).

In our study, predictive abilities were compared across different models to identify the best-fit approach for survival traits. The evaluated models included PBLUP (traditional), GBLUP (comprehensive), Bayesian methods (flexible), regularized regressions, such as EN and RR (penalized), and RF (nonparametric). GBLUP and Bayesian methods consistently outperformed PBLUP, EN, and RF across all survival traits analyzed, corroborating previous reports showing that genome-wide marker information (GBLUP) and adaptive shrinkage with variable effect selection (Bayesian methods) enhance prediction accuracy for survival traits compared with pedigree- or machine learning-based approaches (Haque et al., 2025; Liyanage et al., 2025; Vu et al., 2022).

Despite low heritability, GBLUP remains effective for survival traits by exploiting dense marker information, within-family segregation, and model flexibility in ways that pedigree-based selection cannot, resulting in realized prediction accuracy that often exceeds expectations based on heritability alone (Joshi et al., 2021). Dense genome-wide SNP panels capture Mendelian sampling variation within and across families, which is largely invisible to PBLUP, thereby increasing the proportion of genetic variance that can be exploited even when heritability is low (Gebreyesus et al., 2021). Although heritability estimates for DPC Time were similar between methods, GBLUP exhibited significantly higher prediction accuracy (~129 % increase over PBLUP), reflecting the ability of GBLUP to capture realized genetic relationships from actual allelic variation (G-matrix) rather than expected relationships based on pedigree records (A-matrix), thereby accounting for Mendelian sampling and linkage disequilibrium with causal loci. This advantage is particularly pronounced in populations with shallow pedigrees, such as ours spanning a single generation, where parents are treated as unrelated in PBLUP despite sharing common ancestry from the domesticated founder population (Jerry et al., 2022). By estimating realized genetic relationships from genome-wide SNPs, GBLUP more accurately captures true relatedness and improves genetic evaluation accuracy. Similar improvements have been reported in aquaculture: GBLUP increased prediction accuracy by 8–13 % for viral nervous necrosis resistance in

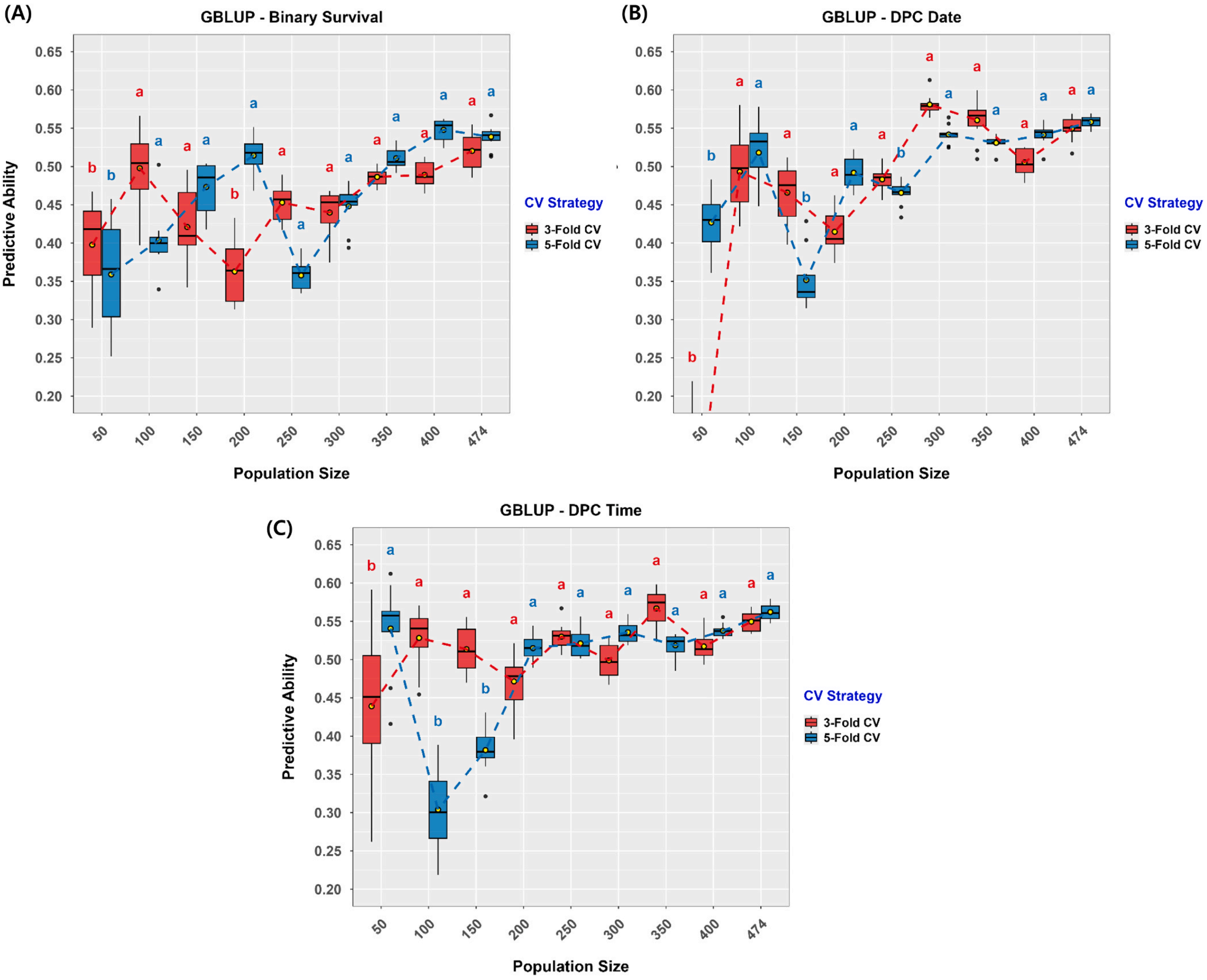


Fig. 6. Comparative predictive performance of the GBLUP model using the top 1000 GWAS-ranked SNPs across various population sizes (50, 100, 150, 200, 250, 300, 350, 400, and 474) evaluated by 3-fold (red) and 5-fold (blue) cross-validation for three traits: **(A)** Binary Survival, **(B)** DPC Date, and **(C)** DPC Time. Boxplots depict the interquartile range (second and third quartiles), whiskers show the data range excluding outliers, and yellow dots indicate mean predictive ability. Significant differences among population sizes were detected using the Friedman test ($p < 0.001$), followed by Nemenyi post-hoc comparisons. Red and blue letters indicate statistically distinct groups for 3- and 5-fold cross-validation, respectively. Population sizes sharing the same letter within each CV strategy are not significantly different ($p < 0.01$).

European sea bass compared to pedigree-based selection (Palaiokostas et al., 2018). Similar studies in gilthead seabream demonstrated 27–53 % accuracy improvements for pasteurellosis resistance using genomic selection methods, including GBLUP (Palaiokostas et al., 2016). GBLUP assumes normally distributed additive effects with equal variance across all markers and uses a GRM to capture realized relationships beyond pedigree records. In contrast, Bayesian methods model marker effects with flexible prior distributions, effectively capturing traits influenced by loci with large or variable effects and complex genetic architectures. (Abdollahi-Arpanahi et al., 2022).

According to our previous study, resistance to scuticociliatosis in olive flounder is a polygenic trait, characterized by multiple small-effect loci distributed across several chromosomes, supporting the strengths of GBLUP and Bayesian methods in capturing complex genetic architectures through modeling numerous small effects (Kodagoda et al., 2025). Conversely, machine learning methods such as RF typically effective in datasets characterized by strong non-linear interactions or large effect sizes but not effective in aggregating small, additive genetic signals across thousands of SNP markers, which are common in survival trait prediction (Peng et al., 2025). In scenarios where trait architecture is predominantly additive, the advantage of these machine learning models diminishes, resulting in reduced predictive accuracy relative to linear approaches such as GBLUP or Bayesian regressions (Saadat et al., 2024). Moreover, the RF method is more prone to overfitting, particularly in situations where the number of markers greatly exceeds the number of genotyped individuals (Montesinos-López et al., 2021). EN exhibited reduced predictive accuracy for traits with a predominantly polygenic architecture involving many small additive effects, as its variable selection and shrinkage may inadequately aggregate these subtle signals compared to methods such as GBLUP that explicitly model genome-wide relationships, consistent with previous findings that revealed that parametric models outperform penalized regression in such contexts (Meuwissen et al., 2001; Montesinos-López et al., 2025).

The underperformance of EN and RF relative to GBLUP/Bayesian methods can be attributed to mismatches between these methods' assumptions and the polygenic architecture of scuticociliatosis resistance traits. EN's sparsity-inducing L1/L2 penalties, designed to perform variable selection by shrinking many coefficients to zero, are poorly suited to highly polygenic traits where genetic variance is distributed across thousands of small-effect SNPs rather than concentrated in a few moderate-to-large effect loci (Wang et al., 2019). In the context of our dataset (52,046 SNPs, N = 474), EN likely discarded many small but truly causal variants, resulting in over-shrinkage and loss of additive genetic signal that GBLUP and Bayesian approaches retain by distributing shrinkage across all markers. Similarly, RF struggled in this "large p, small n" (many more SNPs than individuals) setting where tens of thousands of correlated SNPs and limited training samples lead to overfitting and high variance; most decision trees were built on largely uninformative predictors and could not efficiently exploit the weak, diffuse additive signal as effectively as linear whole-genome regression (Costa et al., 2022). While EN and RF can outperform linear genomic models when traits exhibit sparser genetic architectures with major QTLs, strong non-additive effects (dominance, epistasis, G×E interactions), or when training populations exceed 1000 individuals, these conditions were not met for our survival traits. Collectively, these findings demonstrate that linear genomic models are better suited for predicting scuticociliatosis resistance in olive flounder, where polygenic architecture with predominantly additive effects, and limited training population size relative to high SNP density.

Unlike single immune markers like antibody response that capture only specific pathways and often correlate poorly with actual survival outcomes, survival traits integrate all defense mechanisms and directly reflect the ultimate breeding objective of disease resistance under practical production conditions (Fjalestad et al., 1993). Consistent with previous reports, a higher predictive ability was observed for continuous and time-to-event traits (DPC Date and DPC Time) than that for Binary Survival in this study. Continuous traits typically exhibit higher heritability and additive genetic variance, as they capture richer phenotypic variation, enabling models such as GBLUP or Bayesian methods to better resolve polygenic signals (Toghiani et al., 2017; Vallejo et al., 2017; Vu et al., 2022). For example, previous studies on aquaculture species have demonstrated that survival time achieved higher prediction accuracy than that of binary survival status due to reduced phenotypic truncation and improved modeling of genetic architecture (Ren et al., 2022). Binary survival compresses complex biological variation into two categories (0/1 outcomes), reducing heritable variance and introducing challenges such as class imbalance and over-shrinkage of small effects (Dai et al., 2021; Schubach et al., 2017). Although survival models like Cox regression can handle censored data, they are seldom integrated into genomic prediction, limiting their utility for binary traits (Irlmeier et al., 2022). Improving predictions for binary survival may require threshold models or resampling strategies, as shown in previous studies (Nguyen and Vu, 2022).

Despite the advances in low-cost genotyping platforms and genotype imputation strategies, optimizing training population design and effectively utilizing dense SNP data is crucial for overcoming current challenges and maximizing genetic gain in disease resistance breeding programs (Huang et al., 2025; Nguyen, 2024). For example, Lu et al., (2020) demonstrated that preselecting SNPs through GWAS and integrating them into single-step GBLUP and Bayesian models significantly improved genomic prediction accuracy for disease resistance in flounder. Our results demonstrate that predictive ability improved with marker selection based on GWAS significance, achieving the highest performance with the top 1000 GWAS-ranked markers. Beyond this threshold, performance declined due to the inclusion of markers with small or null effects that introduce noise rather than signal when using standard GBLUP with uniform SNP weighting. Weighted GBLUP (wGBLUP) that assign differential weights to SNPs based on their estimated effects could potentially mitigate this issue by down-weighting less informative markers while retaining the benefits of higher marker density (Wang et al., 2012; Zhang et al., 2010). Future studies that explore these weighted genomic prediction approaches may extend the advantage of larger marker panels beyond the 1,000-SNP threshold and further improve prediction accuracy for disease resistance. Contrastingly, randomly selected markers exhibited no clear increasing trend, highlighting the importance of informed marker selection. GWAS-prioritized SNPs incrementally improved the correlation between observed and predicted phenotypes, achieving superior predictive performance with fewer markers than using the full marker set or solely top GWAS hits. This approach effectively balances model simplicity and accuracy, highlighting the value of integrating GWAS-derived marker prioritization into prediction pipelines (Jeong et al., 2020). Informed selection, particularly based on GWAS, enhances genomic prediction by focusing on phenotype-relevant variants and reducing data dimensionality.

The practical implications of GWAS-informed marker selection depend on breeding program objectives and economic constraints. For trait-specific selection programs focused on a single breeding goal (e.g., disease resistance alone), custom low-density chips containing 1000–5000 trait-relevant SNPs could provide cost-effective genotyping while maintaining high prediction accuracy. However, for multi-trait selection programs, which are typical in aquaculture breeding where growth, disease resistance, and quality traits are simultaneously improved, a “combined chip” approach is more advantageous. Such chips would integrate markers associated with multiple traits of economic importance, identified through GWAS or linkage analysis across different phenotypes (Boichard et al., 2012). Our results across three scuticociliatosis resistance traits (binary survival, DPC Date, DPC Time) support this combined approach, while some QTL regions were trait-specific, others showed pleiotropic effects across survival-related phenotypes, suggesting shared genetic architecture that could be captured by a unified marker panel.

Marker selection can improve prediction accuracy for low-heritability traits by focusing the GRM on SNPs that track the latent genetic structure shared across trait. For low-heritability traits, genetic correlations with other traits and SNP–trait associations are often estimated imprecisely because environmental noise dominates. Selecting a subset of markers that best capture the common variation across traits reduces noise in the marker set and yields a GRM that better reflects the underlying genetic covariances, which increases realized prediction accuracy for the focal low-heritability trait (Klápště et al., 2020). Selecting markers based on GWAS significance to optimize predictive ability can be effective for moderate to lowly heritable traits, as prioritizing informative markers strengthens the genomic relationship signal. For traits with complex genetic architecture, weaker marker-trait associations, and higher environmental noise, gains in genomic prediction accuracy can be limited. Previous studies have demonstrated that considering the specific genetic architecture, such as incorporating major effect loci or weighting markers by their functional relevance, can improve prediction outcomes (Lin et al., 2022; Morgante et al., 2018). Morgante et al. (2018) reported that prediction accuracy improved when genomic models are informed by the genetic architecture deduced from mapping the top variants with major effects and interactions in the training the data, particularly when mapping resolution is sufficient (Morgante et al., 2018). However, when the genetic control is highly polygenic and non-additive, or environmental variance is substantial, prediction accuracy typically decreases despite model improvements (Wang et al., 2025; Yin et al., 2020).

A sufficiently large and well-phenotyped training population is crucial for reliable genomic prediction, but sample size represents a critical cost factor in disease resistance breeding. Challenge tests necessitate animal sacrifice, significant infrastructure, and labor, making large-scale phenotyping expensive (Houston et al., 2020). Our analysis revealed that predictive ability plateaus at approximately 300 individuals, beyond which, marginal accuracy gains diminish substantially relative to additional costs. Consistent with our findings, previous studies have shown that accuracy gains diminish once the training population adequately captures genetic variation and relationships. (Akdemir and Isidro-Sánchez, 2019; Edwards et al., 2019). Therefore, in this study, a training population of ~300 represents a cost-effective breeding design, allowing resources beyond this threshold to be redirected toward improving phenotyping precision, evaluating additional families, or shortening generation intervals, and investments likely to yield greater genetic gain per unit cost than further expanding training population size (Fernández-González et al., 2023; Werner et al., 2020).

4. Conclusion

We optimized a genomic prediction framework to implement genomic selection for complex low to moderately heritable survival traits of scuticociliatosis resistance in vaccinated olive flounder. Genomic prediction (GBLUP and Bayesian models) significantly outperformed pedigree-based prediction, with survival time traits achieving higher ability than that of binary survival. GWAS-based marker prioritization substantially improved genomic prediction ability by enriching models with phenotype-relevant SNPs, with predictive gains increasing up to an optimal set of ~1000 top-ranked markers. Predictive ability increased with training population size but plateaued at approximately 300 individuals, representing the use of a sufficient training data set for reliable evaluation. Collectively, the results highlight the critical need for optimizing phenotyping strategies, marker configuration, and training population design to advance selective breeding programs aimed at complex disease resistance traits in aquaculture species.

CRediT authorship contribution statement

H.A.C.R. Hanchapola: Methodology, Investigation. **M.A.H Dilshan:** Methodology, Investigation. **Cheonguk Park:** Methodology, Investigation. **Jeongeun Kim:** Methodology, Investigation. **Jihun Lee:** Writing – review & editing, Supervision, Investigation. **Jehee Lee:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **W.K.M. Omeka:** Writing – review & editing, Investigation, Formal analysis. **Dean R. Jerry:** Writing – review & editing, Validation, Supervision, Software. **Gaeun Kim:** Writing – review & editing, Methodology, Investigation. **D.S. Liyanage:** Writing – review & editing, Software, Investigation, Formal analysis. **David B. Jones:** Writing – review & editing, Validation, Supervision, Software. **Cecile Massault:** Validation, Supervision, Software. **Yasara Kavindi Kodagoda:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **D.C.G. Rodrigo:** Methodology, Investigation. **G.A.N.P. Ganepola:** Methodology, Investigation.

Declaration of Competing Interest

Authors declare no conflicts of interest.

Acknowledgments

This research was supported by the Korea Institute of Marine Science and Technology Promotion (KIMST), funded by the Ministry of Oceans and Fisheries (RS-2022-KS221670).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.aqrep.2026.103464.

Data availability

Data will be made available on request.

References

Abdollahi-Arpanahi, R., Lourenco, D., Misztal, I., 2022. A comprehensive study on size and definition of the core group in the proven and young algorithm for single-step GBLUP. Genet. Sel. Evol. 54, 34. https://doi.org/10.1186/s12711-022-00726-6.

Akdemir, D., Isidro-Sánchez, J., 2019. Design of training populations for selective phenotyping in genomic prediction. Sci. Rep. 9, 1446. https://doi.org/10.1038/s41598-018-38081-6.

Allal, F., Nguyen, N.H., 2022. Genomic selection in aquaculture species. Genom. Predict. Complex Traits Hum. N. Y. NY 469–491. https://doi.org/10.1007/978-1-0716-2205-6_17.

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., Gietzen, K.J., Hayes, B.J., Lawley, C.T., Sonstegard, T.S., van Tassell, C.P., VanRaden, P.M., Viaud-Martinez, K.A., Wiggans, G.R., 2012. Design of a bovine low-density snp array optimized for imputation. PLoS One 7, e34130. https://doi.org/10.1371/journal.pone.0034130.

Calus, M.P.L., Meuwissen, T.H.E., de Roos, A.P.W., Veerkamp, R.F., 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178, 553–561. https://doi.org/10.1534/genetics.107.080838.

Costa, W.G., da, Celeri, M., de, O., Barbosa, I., de, P., Silva, G.N., Azevedo, C.F., Borem, A., Nascimento, M., Cruz, C.D., 2022. Genomic prediction through machine learning and neural networks for traits with epistasis. Comput. Struct. Biotechnol. J. 20, 5490–5499. https://doi.org/10.1016/j.csbj.2022.09.029.

Dai, X., Fu, G., Zhao, S., Zeng, Y., 2021. Statistical learning methods applicable to genome-wide association studies on unbalanced case-control disease data. Genes. https://doi.org/10.3390/genes12050736.

Dong, L., Xiao, S., Chen, J., Wan, L., Wang, Z., 2016. Genomic selection using extreme phenotypes and pre-selection of SNPs in large yellow croaker (Larimichthys crocea). Mar. Biotechnol. 18, 575–583. https://doi.org/10.1007/s10126-016-9718-4.

Drangsholt, T.M.K., Gjerde, B., Ødegård, J., Finne-Fridell, F., Evensen, Ø., Bentsen, H.B., 2011. Quantitative genetics of disease resistance in vaccinated and unvaccinated Atlantic salmon (*Salmo salar* L.). Hered. (Edinb.) 107, 471–477. https://doi.org/10.1038/hdy.2011.34.

Drangsholt, T.M.K., Gjerde, B., Ødegård, J., Finne-Fridell, F., Evensen, Ø., Bentsen, H.B., 2012. Genetic correlations between disease resistance, vaccine-induced side effects and harvest body weight in Atlantic salmon (*Salmo salar*). Aquaculture 324325 312–314. https://doi.org/10.1016/j.aquaculture.2011.11.007.

Edwards, S.M., Buntjer, J.B., Jackson, R., Bentley, A.R., Lage, J., Byrne, E., Burt, C., Jack, P., Berry, S., Flatman, E., Poupard, B., Smith, S., Hayes, C., Gaynor, R.C., Gorjanc, G., Howell, P., Ober, E., Mackay, I.J., Hickey, J.M., 2019. The effects of training population design on genomic prediction accuracy in wheat. Theor. Appl. Genet 132, 1943. https://doi.org/10.1007/s00122-019-03327-y.

Endelman, J.B., 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4, 250–255. https://doi.org/10.3835/PLANTGENOME2011.08.0024.

Fernández-González, J., Akdemir, D., Isidro y Sánchez, J., 2023. A comparison of methods for training population optimization in genomic selection. Theor. Appl. Genet 136, 30. https://doi.org/10.1007/s00122-023-04265-6.

Figueroa, C., Veloso, P., Espin, L., Dixon, B., Torrealba, D., Elalfy, I.S., Afonso, J.M., Soto, C., Conejeros, P., Gallardo, J.A., 2020. Host genetic variation explains reduced protection of commercial vaccines against Piscirickettsia salmonis in Atlantic salmon. Sci. Rep. 10, 18252. https://doi.org/10.1038/s41598-020-70847-9.

Fjalestad, K.T., Gjedrem, T., Gjerde, B., 1993. Genetic improvement of disease resistance in fish: An overview. In: Genetics in Aquaculture. Elsevier, pp. 65–74. https://doi.org/10.1016/B978-0-444-81527-9.50011-7.

Gebreyesus, G., Lund, M.S., Sahana, G., Su, G., 2021. Reliabilities of genomic prediction for young stock survival traits using 54K SNP chip augmented with additional single-nucleotide polymorphisms selected from imputed whole-genome sequencing data. Front. Genet 12, 667300. https://doi.org/10.3389/fgene.2021.667300.

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., 2021. 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.

Guarini, A.R., Lourenco, D.A.L., Brito, L.F., Sargolzaei, M., Baes, C.F., Miglior, F., Misztal, I., Schenkel, F.S., 2018. Comparison of genomic predictions for lowly heritable traits using multi-step and single-step genomic best linear unbiased predictor in Holstein cattle. J. Dairy Sci. 101, 8076–8086. https://doi.org/10.3168/jds.2017-14193.

Habier, D., Fernando, R.L., Dekkers, J.C.M., 2009. Genomic selection using low-density marker panels. Genetics 182, 343–353. https://doi.org/10.1534/genetics.108.100289.

Hako Touko, B.A., Kong Mbiydzenyuy, A.T., Tumasang, T.T., Awah-Ndukum, J., 2021. Heritability estimate for antibody response to vaccination and survival to a newcastle disease infection of native chicken in a low-input production system. Front. Genet 12, 666947. https://doi.org/10.3389/fgene.2021.666947.

Haque, M.A., Jang, E.-B., Lee, H.-D., Shin, D.-H., Jang, J.-H., Kim, J.-J., 2025. Performance of weighted genomic BLUP and Bayesian methods for Hanwoo carcass traits. Trop. Anim. Health Prod. 57, 38. https://doi.org/10.1007/s11250-025-04293-y.

Hayes, B.J., 2011. Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. J. Dairy Sci. 94, 2114–2117. https://doi.org/10.3168/jds.2010-3896.

Hickmann, F.M.W., Braccini Neto, J., Kramer, L.M., Huang, Y., Gray, K.A., Dekkers, J.C.M., Sanglard, L.P., Serão, N.V.L., 2021. Host genetics of response to porcine reproductive and respiratory syndrome in sows: antibody response as an indicator trait for improved reproductive performance. Front. Genet 12. https://doi.org/10.3389/fgene.2021.707873.

Hosoya, S., Yoshikawa, S., Sato, M., Kikuchi, K., 2021. Genomic prediction for testes weight of the tiger pufferfish, *Takifugu rubripes*, using medium to low density SNPs. Sci. Rep. 11, 1–10. https://doi.org/10.1038/s41598-021-99829-1.

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. https://doi.org/10.1038/s41576-020-0227-y.

Huang, Y., Li, Z., Li, M., Zhang, X., Shi, Q., Xu, Z., 2025. Fish genomics and its application in disease-resistance breeding. Rev. Aquac. https://doi.org/10.1111/raq.12973.

Irlmeier, R., Hughey, J.J., Bastarache, L., Denny, J.C., Chen, Q., 2022. Cox regression is robust to inaccurate EHR-extracted event time: an application to EHR-based GWAS. Bioinformatics 38, 2297–2306. https://doi.org/10.1093/bioinformatics/btac086.

Jeong, S., Kim, J.Y., Kim, N., 2020. GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. Sci. Rep. 10, 1–12. 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.

Jiang, Z., Bai, Y., Li, N., Chen, L., Zhao, J., Li, Y., Zhou, T., Xu, P., 2025. Genomic Approaches to Enhance Disease Resistance in Large Yellow Croaker: Recent Progress and Future Perspectives. Rev. Aquac. 17, e70084. https://doi.org/10.1111/RAQ.70084;REQUESTEDJOURNAL:JOURNAL:17535131;WGROUP:STRING:PUBLICATION.

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

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

Klápště, J., Dungey, H.S., Telfer, E.J., Suontama, M., Graham, N.J., Li, Y., McKinley, R., 2020. Marker selection in multivariate genomic prediction improves accuracy of low heritability traits. Front. Genet 11, 499094. https://doi.org/10.3389/fgene.2020.499094.

Kodagoda, Y.K., Kim, G., Liyanage, D.S., Omeka, W.K.M., Park, C., Kim, J., Lee, J.H., Hanchapola, H.A.C.R., Dilshan, M.A.H., Rodrigo, D.C.G., Jones, D.B., Massault, C., Jerry, D.R., Lee, J., 2025. Genome-wide association mapping of scuticociliatosis resistance in a vaccinated population of olive flounder (*Paralichthys olivaceus*). Fish. Shellfish Immunol. 162, 110339. https://doi.org/10.1016/j.fsi.2025.110339.

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/FULL.

Lin, Q., Teng, J., Cai, X., Li, J., Zhang, Z., 2022. Utilization strategies of two environment phenotypes in genomic prediction. Genes 13, 722. https://doi.org/10.3390/genes13050722.

Liyanage, D.S., Lee, S., Yang, H., Lim, C., Omeka, W.K.M., Sandamalika, W.M.M.G., Udayantha, H.M.V., Kim, G., Ganeshalingam, S., Jeong, T., Oh, S.R., Won, S.H., Koh, H.B., Kim, M.K., Jones, D.B., Massault, C., Jerry, D.R., Lee, J., 2022. Genome-wide association study of VHSV-resistance trait in *Paralichthys olivaceus*. Fish. Shellfish Immunol. 124, 391–400. https://doi.org/10.1016/j.fsi.2022.04.021.

Liyanage, D.S., Lee, S., Yang, H., Lim, C., Omeka, W.K.M., Sandamalika, W.M.M.G., Udayantha, H.M.V., Kim, G., Hanchapola, H.A.C.R., Ganeshalingam, S., Jeong, T., Oh, S.R., Won, S.H., Koh, H.B., Kim, M.K., Jones, D.B., Massault, C., Jerry, D.R., Lee, J., 2025. Genomic prediction of survival traits in the response of olive flounder (*Paralichthys olivaceus*) to viral hemorrhagic septicemia virus: comparing machine learning models and traditional approaches. Aquaculture 595, 741685. https://doi.org/10.1016/j.aquaculture.2024.741685.

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., Calus, M.P.L., 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193, 327–345. https://doi.org/10.1534/genetics.112.143313.

Lu, S., Liu, Y., Yu, X., Li, Y., Yang, Y., Wei, M., Zhou, Q., Wang, J., Zhang, Y., Zheng, W., Chen, S., 2020. Prediction of genomic breeding values based on pre-selected SNPs using ssGBLUP, WssGBLUP and BayesB for Edwardsiellosis resistance in Japanese flounder. Genet. Sel. Evol. 52, 49. https://doi.org/10.1186/s12711-020-00566-2.

Mao, J., Tian, Y., Liu, Q., Li, D., Ge, X., Wang, X., Hao, Z., 2023. Revealing genetic diversity, population structure, and selection signatures of the Pacific oyster in Dalian by whole-genome resequencing. Front. Ecol. Evol. 11, 1337980. https://doi.org/10.3389/fevo.2023.1337980.

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.

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

Montesinos-López, O.A., Montesinos-López, A., Mosqueda-Gonzalez, B.A., Montesinos-López, J.C., Crossa, J., Ramirez, N.L., Singh, P., Valladares-Anguiano, F.A., 2021. A zero altered Poisson random forest model for genomic-enabled prediction. G3 Genes|Genomes|Genet. 11. https://doi.org/10.1093/g3journal/jkaa057.

Montesinos-López, A., Montesinos-López, O.A., Ramos-Pulido, S., Mosqueda-González, B. A., Guerrero-Arroyo, E.A., Crossa, J., Ortiz, R., 2025. Artificial intelligence meets genomic selection: comparing deep learning and GBLUP across diverse plant datasets. Front. Genet 16, 1568705. https://doi.org/10.3389/FGENE.2025.1568705/BIBTEX.

Morgante, F., Huang, W., Maltecca, C., Mackay, T.F.C., 2018. Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals. Hered. (Edinb.) 120, 500–514. https://doi.org/10.1038/s41437-017-0043-0.

Nguyen, N.H., 2024. Genetics and genomics of infectious diseases in key aquaculture species. Biology. https://doi.org/10.3390/biology13010029.

Nguyen, N.H., Vu, N.T., 2022. Threshold models using Gibbs sampling and machine learning genomic predictions for skin fluke disease recorded under field environment in yellowtail kingfish *Seriola lalandi*. Aquaculture 547, 737513. https://doi.org/10.1016/j.aquaculture.2021.737513.

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

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.

Palaiokostas, C., Cariou, S., Bestin, A., Bruant, J.-S., Haffray, P., Morin, T., Cabon, J., Allal, F., Vandeputte, M., Houston, R.D., 2018. Genome-wide association and genomic prediction of resistance to viral nervous necrosis in European sea bass (*Dicentrarchus labrax*) using RAD sequencing. Genet. Sel. Evol. 50, 30. https://doi.org/10.1186/s12711-018-0401-2.

Palaiokostas, C., Ferraresso, S., Franch, R., Houston, R.D., Bargelloni, L., 2016. Genomic Prediction of Resistance to Pasteurellosis in Gilthead Sea Bream (*Sparus aurata*) Using 2b-RAD Sequencing. G3 Genes Genomes Genet. 6, 3693–3700. https://doi.org/10.1534/g3.116.035220.

Peng, J., Lei, X., Liu, T., Xiong, Yi, Wu, J., Xiong, Yanli, You, M., Zhao, J., Zhang, J., Ma, X., 2025. Integration of machine learning and genome-wide association study to explore the genomic prediction accuracy of agronomic trait in oats (*Avena sativa* L.). Plant Genome 18. https://doi.org/10.1002/tpg2.20549.

Perdry, H., Dandine-Roulland, C., 2015. gaston: genetic data handling (QC, GRM, LD, PCA) &amp. Linear Mixed Models CRAN Contrib. Packag. https://doi.org/10.32614/CRAN.package.gaston.

Pérez, P., De Los Campos, G., 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198, 483–495. https://doi.org/10.1534/genetics.114.164442.

Ren, S., Mather, P.B., Tang, B., Hurwood, D.A., 2022. Insight into selective breeding for robustness based on field survival records: New genetic evaluation of survival traits in pacific white shrimp (*Penaeus vannamei*) breeding line. Front. Genet 13, 1–14. https://doi.org/10.3389/fgene.2022.1018568.

Saadat, H.B., Torshizi, R.V., Manafiazar, G., Masoudi, A.A., Ehsani, A., Shahinfar, S., 2024. Comparing machine learning algorithms and linear model for detecting significant SNPs for genomic evaluation of growth traits in F2 chickens. J. Agric. Sci. Technol. 26, 1261–1274. https://doi.org/10.22034/JAST.26.6.1261.

Safonova, Y., Shin, S.B., Kramer, L., Reecy, J., Watson, C.T., Smith, T.P.L., Pevzner, P.A., 2022. Variations in antibody repertoires correlate with vaccine responses. Genome Res 32, 791–804. https://doi.org/10.1101/gr.276027.121.

Schrauf, M.F., de los Campos, G., Munilla, S., 2021. Comparing genomic prediction models by means of cross validation. Front. Plant Sci. 12, 734512. https://doi.org/10.3389/fpls.2021.734512.

Schubach, M., Re, M., Robinson, P.N., Valentini, G., 2017. Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants. Sci. Rep. 7, 1–12. https://doi.org/10.1038/s41598-017-03011-5.

Shivam, S., El-Matbouli, M., Kumar, G., 2021. Development of fish parasite vaccines in the OMICs Era: progress and opportunities. Vaccines 9, 179. https://doi.org/10.3390/vaccines9020179.

Sohn, H., Kwon, H., Lee, S., Wan, Q., Lee, J., 2023. Development of a trivalent vaccine for prevention of co-infection by Miamiensis avidus and Tenacibaculum maritimum in farmed olive flounder. Fish. Aquat. Sci. 26, 605–616. https://doi.org/10.47853/FAS.2023.e52.

Somsiam, P., Sukhavachana, S., Pattarapanyavong, N., Tunkijjanukij, S., Phuthaworn, C., Poompuang, S., 2024. Genomic predictions for daily gain and fillet weight using correlated size and body area measurements in Asian seabass (*Lates calarifer*, Bloch 1790). Aquaculture 591, 741133. https://doi.org/10.1016/j.aquaculture.2024.741133.

Song, H., Dong, T., Yan, X., Wang, W., Tian, Z., Sun, A., Dong, Y., Zhu, H., Hu, H., 2023. Genomic selection and its research progress in aquaculture breeding. Rev. Aquac. 15, 274–291. https://doi.org/10.1111/RAQ.12716;CTYPE:STRING:JOURNAL.

Song, H., Dong, T., Yan, X., Wang, W., Zhang, Q., Hu, H., 2025. Advancing aquaculture breeding through genomic selection: models, tools, and challenges. Water Biol. Secur, 100494. https://doi.org/10.1016/J.WATBS.2025.100494.

Song, H., Hu, H., 2022. Strategies to improve the accuracy and reduce costs of genomic prediction in aquaculture species. Evol. Appl. 15, 578–590. https://doi.org/10.1111/eva.13262.

Thavamanikumar, S., Dolferus, R., Thumma, B.R., 2015. Comparison of genomic selection models to predict flowering time and spike grain number in two hexaploid wheat doubled haploid populations. G3 Genes Genomes Genet 5, 1991–1998. https://doi.org/10.1534/g3.115.019745.

Toghiani, S., Hay, E., Sumreddee, P., Geary, T.W., Rekaya, R., Roberts, A.J., 2017. Genomic prediction of continuous and binary fertility traits of females in a composite beef cattle breed. J. Anim. Sci. 95, 4787–4795. https://doi.org/10.2527/jas2017.1944.

Udayantha, H.M.V., Kim, J., Kim, G., Lee, Jihun, Lee, S., Park, C.-U., Jones, D.B., Massault, C., Jerry, D.R., Liyanage, D.S., Lee, Jehee, 2025. Genome-wide association and genomic prediction of thermal tolerance in olive flounders (Paralichthys olivaceus): a validation study. Aquac. Rep. 45, 103205. https://doi.org/10.1016/j.aqrep.2025.103205.

Vallejo, R.L., Leeds, T.D., Gao, G., Parsons, J.E., Martin, K.E., Evenhuis, J.P., Fragomeni, B.O., Wiens, G.D., Palti, Y., 2017. Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree-based model in rainbow trout aquaculture. Genet. Sel. Evol. 49, 1–13. https://doi.org/10.1186/s12711-017-0293-6.

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, 118682. https://doi.org/10.3389/fgene.2014.00432.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423. https://doi.org/10.3168/jds.2007-0980.

Verbyla, K.L., Kube, P.D., Evans, B.S., 2022. Commercial implementation of genomic selection in Tasmanian Atlantic salmon: scheme evolution and validation. Evol. Appl. 15, 631–644. https://doi.org/10.1111/eva.13304.

Vu, N.T., Phuc, T.H., Oanh, K.T.P., Sang, N., Van, Trang, T.T., Nguyen, N.H., 2021. Accuracies of genomic predictions for disease resistance of striped catfish to Edwardsiella ictaluri using artificial intelligence algorithms. bioRxiv 2021.05.10.443499. https://doi.org/10.1101/2021.05.10.443499.

Vu, N.T., Phuc, T.H., Phuong Oanh, K.T., van Sang, N., Trang, T.T., Nguyen, N.H., 2022. Accuracies of genomic predictions for disease resistance of striped catfish to *Edwardsiella ictaluri* using artificial intelligence algorithms. G3 Genes Genomes Genet. 12. https://doi.org/10.1093/G3JOURNAL/JKAB361.

Wang, P., Meng, F., Del Azodi, C.B., Segura Abá, K.E., Casler, M.D., Shiu, S.-H., 2025. Optimizing genomic prediction for complex traits via investigating multiple factors in switchgrass. Plant Physiol. 198, 188. https://doi.org/10.1093/plphys/kiaf188.

Wang, X., Miao, J., Chang, T., Xia, J., An, B., Li, Y., Xu, L., Zhang, L., Gao, X., Li, J., Gao, H., 2019. Evaluation of GBLUP, BayesB and elastic net for genomic prediction in Chinese Simmental beef cattle. PLoS One 14, e0210442. https://doi.org/10.1371/journal.pone.0210442.

Wang, H., Misztal, I., Aguilar, I., Legarra, A., Muir, W.M., 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94, 73–83. https://doi.org/10.1017/S0016672312000274.

Wang, S., Wei, Y., Liu, D., Zhang, X., Wang, Q., Pan, Y., Ma, P., 2025. Impact of different genomic relationship matrix construction methods on the accuracy of genomic prediction in different species. Front. Genet 16, 1576248. https://doi.org/10.3389/fgene.2025.1576248.

Wang, X., Xu, Y., Hu, Z., Xu, C., 2018. Genomic selection methods for crop improvement: Current status and prospects. Crop J. https://doi.org/10.1016/j.cj.2018.03.001.

Werner, C.R., Gaynor, R.C., Gorjanc, G., Hickey, J.M., Kox, T., Abbadi, A., Leckband, G., Snowdon, R.J., Stahl, A., 2020. How population structure impacts genomic selection accuracy in cross-validation: implications for practical breeding. Front. Plant Sci. 11, 592977. https://doi.org/10.3389/FPLS.2020.592977/BIBTEX.

Yáñez, J.M., Houston, R.D., Newman, S., 2014. Genetics and genomics of disease resistance in salmonid species. Front. Genet 5, 1–13. https://doi.org/10.3389/fgene.2014.00415.

Yang, B., Zhi, C., Li, P., Xu, C., Li, Q., Liu, S., 2024. Genomic selection accelerates genetic improvement of resistance to Vibriosis in the Pacific oyster, Crassostrea gigas. Aquaculture 584, 740679. https://doi.org/10.1016/j.aquaculture.2024.740679.

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., Liu, X., 2021. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genom. Proteom. Bioinforma. 19, 619–628. https://doi.org/10.1016/j.gpb.2020.10.007.

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, 1–22. https://doi.org/10.1186/s13059-020-02052-w.

Yoshida, G.M., Yáñez, J.M., 2022. Increased accuracy of genomic predictions for growth under chronic thermal stress in rainbow trout by prioritizing variants from GWAS using imputed sequence data. Evol. Appl. 15, 537–552. https://doi.org/10.1111/EVA.13240;WGROUP:STRING:PUBLICATION.

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., Liu, J., Ding, X., Bijma, P., de Koning, D.J., Zhang, Q., 2010. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS One 5, 1–8. https://doi.org/10.1371/journal.pone.0012648.

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, 441967. https://doi.org/10.3389/fgene.2019.00189.

Zhang, C., Zhang, Y., Liu, C., Wang, L., Dong, Y., Sun, D., Wen, H., Zhang, K., Qi, X., Li, Y., 2024. Genome-wide association study and genomic prediction for growth traits in spotted sea bass (*Lateolabrax maculatus*) using insertion and deletion markers. Anim. Res. One Heal 2, 400–416. https://doi.org/10.1002/ARO2.87.