Accuracies of genomic predictions for disease resistance of striped catfish to *Edwardsiella ictaluri* using artificial intelligence algorithms

Nguyen Thanh Vu[a,b,c,*], Tran Huu Phuc[c], Kim Thi Phuong Oanh[d], Nguyen Van Sang[c], Trinh Thi Trang[a,b,e], Nguyen Hong Nguyen[a,b,*]

[a]Genecology Research Center, [b]School of Science, Technology and Engineering, University of the Sunshine Coast, 90 Sippy Downs drive, Sippy Downs, QLD 4556, Australia; [c]Research Institute for Aquaculture No.2, 116 Nguyen Dinh Chieu, District 1, Ho Chi Minh city, Vietnam; [d]Institute of Genome Research, Vietnam Academy of Science and Technology, Cau Giay, Hanoi, Vietnam; [e]Vietnam National University of Agriculture, Gia Lam, Vietnam.

*Correspondence to N.T. Vu vunt.ria2@mard.gov.vn or ThanhVu.Nguyen@research.usc.edu.au and N.H. Nguyen nnguyen@usc.edu.au

Abstract

Assessments of genomic prediction accuracies using machine and deep learning methods are currently not available or very limited in aquaculture species. The principal aim of this study was to examine the predictive performance of these new methods for disease resistance to *Edwardsiella ictaluri* in a population of striped catfish *Pangasianodon hypophthalmus* and to make comparisons with four common methods, i.e., pedigree-based best linear unbiased prediction (PBLUP), genomic-based best linear unbiased prediction (GBLUP), single-step GBLUP (ssGBLUP) and a non-linear Bayesian approach (notably BayesR). Our analyses using machine learning (i.e., KAML) and deep learning (i.e., DeepGP) together with the four common methods (PBLUP, GBLUP, ssGBLUP and BayesR) were conducted for two main disease

resistance traits (i.e., survival status coded as 0 and 1 and survival time, i.e., days that the animals were still alive after the challenge test) in a pedigree consisting of 560 individual animals (490 offspring and 70 parents) genotyped for 14,154 Single Nucleotide Polymorphism (SNPs). The results showed that KAML outperformed GBLUP and ssGBLUP, with the increases in the prediction accuracies for both traits by 5.1 – 47.7%. However, the prediction accuracies obtained from KAML were comparable to those estimated using BayesR. Imputation of missing genotypes using AlphaFamImpute increased the prediction accuracies by 0.2 – 33.2% in all the methods used. On the other hand, there were no significant increases in the prediction accuracies for both survival status and survival time when multivariate models were used in comparison to univariate analyses. Interestingly, the genomic prediction accuracies based on only highly significant SNPs ($P < 0.00001$) were not largely different from those obtained from the whole set of 14,154 SNPs. In all our analyses, the accuracies of genomic prediction were somewhat higher for survival time than survival status (0/1 data). It is concluded that there are prospects for the application of genomic selection to increase disease resistance to *Edwardsiella ictaluri* in striped catfish breeding programs.

1. Introduction

Genomic selection has been increasingly practiced in genetic improvement programs for farmed animals and plants, using a range of different statistical methods from genomic-based best linear unbiased prediction (GBLUP) and its extension known as single-step GBLUP (ssGBLUP) to non-linear Bayesian approaches (BayesA, BayesB, BayesC, BayesC-π and notably BayesR) (Lourenco et al. 2020; Moser et al. 2015; VanRaden 2008; Daetwyler et al. 2013; Al-Tobasei et al. 2021; Vallejo et al. 2020). Recently, there has been a growing interest in using Artificial Intelligent (AI) algorithms (i.e., Machine Learning or Deep Learning) to choose optimal genome-wide models without prior consumptions to determine genomic prediction accuracies for quantitative complex traits, especially for disease resistance (tolerance or

resilience). Examples using simulated and real animal and plant data showed that the prediction accuracies (*r*) using machine learning were 10% greater than linear (i.e. GBLUP) and 1.3% greater than non-linear methods (i.e. BayesR), for instance, r = 0.732 – 0.758 for binary and continuous traits using GBLUP vs 0.801 – 0.832 (Yin et al. 2020). To date, however, no (or very limited) studies have used AI methods to estimate genomic breeding values for aquaculture species.

Almost all studies in aquatic animal species have employed GBLUP, ssGBLUP or Bayesian methods to examine genomic prediction accuracies for two main groups of traits, i.e., body weight and disease resistance. A synthesis of the published information shows that the genomic prediction accuracy ranged from 0.38 to 0.89 for growth-related traits (Houston et al. 2020). On the other hand, the predictions of genomic breeding values for disease resistance were varied with studies, pathogens, populations and methods used, with the estimates ranging from 0 (Silva et al. 2019) to 0.8 or greater (Barría et al. 2018). For both growth and disease traits, the majority of the studies demonstrated that either linear (GBLUP, ssGBLUP) or non-linear Bayesian methods increased the prediction accuracies of animal breeding values relative to Pedigreed-based BLUP (PBLUP) by 22 - 24% (Houston et al. 2020). In addition to growth and disease traits, three recent studies performed genomic predictions for meat quality in banana shrimp (Nguyen et al. 2020) or Portuguese Oyster (*Crassostrea angulata*) (Vu et al. 2021) as well as behavior traits (i.e., cannibalism) in Asian seabass (Nguyen and Khang 2021).

Despite the economic importance of the striped catfish in aquaculture, e.g., valuing 2.36 billion USD that accounts for 1% GDP of Vietnam, there has been a paucity of knowledge in using genomic information in breeding programs for high growth (Vu et al. 2019b) or increased disease resistance, especially Bacillary Necrosis of Pangasius (Vu et al. 2019a). The Bacillary Necrosis of Pangasius (BNP) due to *Edwardsiella ictaluri* has been the main contributor to 56 – 92% mortality during larval and fingerling rearing in this species (Vũ et al. 2017). Our earlier studies showed that there is heritable genetic component for BNP resistance, but the heritability for this trait was low, around 0.10 across statistical methods used (Vu et al.

70 2019a; Pham et al. 2021; Dinh Pham et al. 2021). While these results suggests there are prospects for
71 genetic improvement of BNP resistance using conventional selective breeding, the pathogen test has
72 posed substantial challenges in terms of biosecurity issues, times, costs and environmental impacts (Pham
73 et al. 2020; Nguyen 2014). Due to these limitations, genomic selection has emerged as an alternative
74 option to increase resistance of striped catfish to BNP, one of the most severe diseases that has caused
75 significant economic loss for the sector world-wide.

76 Therefore, the principal of this study was to assess genomic prediction accuracies of machine and deep
77 learning methods for two main disease resistance traits (i.e., survival status and survival time) and to make
78 comparisons with four common methods (PBLUP, GBLUP, ssGBLUP and BayesR). Additionally, we
79 explored if imputation of missing genotypes, multiple traits analyses and significant genome-wide markers
80 can improve the prediction accuracies for the disease resistant traits. Our results open new opportunities
81 for genome-based selection to increase the animal resistance to *E. ictaluri* in striped catfish as well as
82 other aquaculture species infected by this highly infectious pathogen.

83 **2. Materials and Methods**

84 **2.1. Fish and challenge test**

85 The animal samples used in this study originated from a selective breeding program for improved disease
86 resistance of striped catfish to *Edwardsiella ictaluri* (Vu et al. 2019a). In 2020, the first generation was
87 produced based on a nested mating design with a ratio of one male to two females. A total of 166 families
88 (32 full- and 134 half-sib families) were successfully produced following the breeding protocol as detailed
89 in Van Sang et al. (2012) and Vu et al. (2019b). Fry of each family were kept in separate fiber glass tanks
90 up to 3 weeks before they were transferred to stock in net hapa installed in earthen pond. When the
91 fingerlings reached an average body weight of 15 – 20 g, a random sample of 100 fish per family was
92 individually identified using PIT (Passive Integrated Transponder) tag. After tagging, a half size of each

93 family was sent to ponds for performance testing and another half was used in pathogen challenge tests
94 for *E. ictaluri* resistance.

95 The challenge test involved a total of 5328 individuals from 166 families (averaging 32 individuals per
96 family). The experimental fish were initially acclimatized in cement tanks for about two weeks. Then the
97 same number of fish from each family was randomly allocated to six cement tanks (10 m^3) for the
98 challenge test using cohabitant method (Vu et al. 2019a). The cohabitant fish (16.7 ± 6.1 g) were firstly
99 inoculated with the bacteria *E. ictaluri* pathogen (10^6 CFU/0.2 ml per fish). Two days after the injection,
100 they were released into the cement tanks to rear with the experimental fish with a ratio of 1 to 3 (or
101 roughly 30% cohabitant fish in each tank). The experiment was conducted over a period of 23 days. During
102 this period, the feeding rate was reduced to 1.5% of the total biomass in tank. Mortality was highest in
103 day 5 and dead fish were sampled for laboratory PCR test to verify that their death symptoms (white spots
104 in spleen, liver and kidney) were due to *E. ictaluri* pathogen. At the conclusion of the experiment, all alive
105 fish were biosecure-buried, following the regulations of the national veterinary authority (Department of
106 Animal Health, Vietnam).

107 **2.2. Phenotype data**

108 During the challenge test, death fish were collected every 3 hours and their clinical symptoms were also
109 recorded. The data were used to calculate two measures of *E. ictaluri* resistance, i.e., survival status and
110 survival time. Survival status was expressed as a binary trait in which dead fish were designated as zero
111 (0) and alive animals at the end of the challenge test were assigned a number 1. Furthermore, survival
112 time was defined as the continuous trait from the start of test until the animal death. Both survival status
113 and survival time were analyzed using linear mixed model to estimate breeding values (EBVs) for all
114 individuals and families in the pedigree. Based on the EBVs ranking, 20 highest resistance families and 20
115 lowest resistance families were chosen from 166 families. Only families with the number of offspring

116 greater than 9 were selected to ensure that the EBVs were estimated with a high level of reliability. Next,
117 we randomly collected fin tissue samples of 12 – 15 fish per family per disease resistance group for
118 genome sequencing using Diversity Arrays Technology (DArTseq™).

119 **2.3. Genotype data**

120 A total of 564 DNA samples (from 564 individual fish) were sent to a commercial service provider in
121 Canberra, Australia for genotyping by sequencing using DArTseq™ technology. The DArTseq™ was based
122 on the genome complexity reduction method in combination with high throughput next generation
123 sequencing using Illumina platform. The sequencing protocol including choice of restricted enzymes and
124 library preparation was optimized for striped catfish (Pangasianodon DarTseq 1.0) using 96 independent
125 DNA samples of striped catfish in our earlier study (Vu et al. 2020). Briefly, the PstI-SphI method (Kilian
126 et al. 2012) was used and the compatible adaptors were designed to include Illumina flow-cell attachment
127 sequence, sequencing primer sequence and capturing variant length of barcode regions (Elshire et al.
128 2011). Then, only mixed fragments (PstI-SphI) were amplified in 30 rounds of PCR, followed by sequencing
129 on Illumina Hiseq2500 (77 cycles per single read). Next, sequences generated in each lane of 96-well
130 microliter were processed by proprietary DArT analysis protocol. The genotype calling were described in
131 detail in our previous studies (Nguyen et al. 2020; Nguyen et al. 2018). The average variant call-rate was
132 99% and the sample calling rate was 92%. Given with our strict quality control, four out of 564 samples
133 were discarded, and 560 samples (490 offspring and 70 parents) remained for downstream analyses.

134 **2.4. Genotype imputation**

135 The missing genotypes were about 10.0% in this study. They were imputed using AlphaFamImpute
136 (Whalen et al. 2020). The imputation was based on offspring-parent (i.e., pedigree) information.
137 Specifically, this method used parental genotypes to impute the missing values of offspring genotypes and

138 then, the offspring genotypes were used to fulfill the missing values of their parents (Whalen et al. 2020).
139 The pedigree used in our analysis was traced back to the base population.

140 **2.5. Estimation of heritability using genotype data**

141 Variances components of the two studied traits (survival status and survival time) were estimated using
142 linear mixed model under the Best Linear Unbiased Prediction (BLUP) framework in AIREMLf90 sub-
143 program of the BLUPF90 family package (Misztal et al. 2002). AIREMLf90 uses Average Information-
144 Restricted Maximum Likelihood method that can provide less computational resources and higher
145 accuracy of variance estimates (Masuda et al. 2014). The models included the fixed effects of spawning
146 batch (4 levels, 1 – 2 weeks interval between successive spawning batches) and pathogen challenge test
147 tanks (2 levels) and the random effect of additive genetics of the individual fish in the pedigree. Age from
148 birth to tagging (124 – 167 days) was also fitted as a linear covariate. The SNP heritability (h^2) was
149 calculated as a ratio of the additive genetic variance (σ_a^2) to total phenotypic variance (σ_p^2), where
150 σ_p^2=σ_a^2+σ_e^2 (σ_e^2 = environmental variance).

151 **2.6. Statistical methods**

152 Accuracies of genomic prediction for the disease resistance traits (survival status and survival time) were
153 estimated using BLUP-family methods (PBLUP, GBLUP, ssBLUP), BayesR, machine learning (i.e., KAML) and
154 deep learning (i.e., DeepGP).

155 **2.6.1. PBLUP, GBLUP and ssGBLUP (using BLUPF90 program)**

156 The mixed model used in PBLUP, GBLUP and ssGBLUP is written in a matrix notation as follows:

157 $$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e} \qquad \text{(equation 1)}$$

158 Where ***y*** is the vector of phenotypic values (survival status or survival time), ***b*** is the vector of fixed effects
159 (i.e., spawning batches and experimental tanks and a linear covariate of age from birth to tagging), ***a*** is

160 the vector of the random term (i.e., the additive genetic effect of individual fish in the pedigree). ***X*** and ***Z***
161 are the incidence matrices related to the fixed and random effects. The letter ***e*** refers to residual variance
162 or error of the estimates corresponded to each phenotypic value.

163 The main difference among BLUP-family methods (PBLUP, GBLUP and ssGBLUP) relates to relationship
164 matrices (**A**, **G** or **H**) used to solve the mixed model equation 1 above.

165 **PBLUP:** In PBLUP, ***a*** is the additive genetic variance of each individual fish with its corresponding matrix **Z**
166 following a normal distribution ~ N(0, $\mathbf{A}\sigma_a^2$), with **A** is the numerator relationship matrix calculated from
167 the pedigree records and σ_a^2 is the additive genetic variance.

168 **GBLUP**: In GBLUP, ***a*** is the random additive genetic effect underlying polygenic effects of SNPs with a
169 normal distribution ~ N(0, $\mathbf{G}\sigma_g^2$) where genomic relationship matrix **G** was generated from 14,154 SNPs
170 and polygenic genetic variance σ_g^2. The GBLUP method assumes that each SNP contributes little to
171 phenotypic variance of the traits studied.

172 **ssGBLUP**: In ssGBLUP, **A** and **G** matrices are blended to produce realized matrix **H** (Aguilar et al. 2010).
173 The inverse of H matrix is expressed as below:

174 $$\boldsymbol{H^{-1}} = \boldsymbol{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{-1} \end{bmatrix} \qquad \text{(equation 2)}$$

175 Where $\mathbf{A}^{-1}$ and $\mathbf{G}^{-1}$ are the inverse of matrices **A** and **G** as described above. $\mathbf{A}^{-1}_{22}$ is the inverse of matrix of
176 genotyped individuals after combining of pedigree and genomic information.

177 The three BLUP methods (PBLUP, GBLUP and ssGBLUP) were implemented in BLUPF90 package (Misztal
178 et al. 2002). For survival status, we used generalized (threshold) mixed model in THRGIBBF1f90 (Tsuruta
179 and Misztal 2006) known as threshold Gibb Sampling method, whereas survival time was analyzed using
180 linear mixed models in AIREMLf90 (Masuda et al. 2014). In the Gibbs sampling, we used 200,000 iterations
181 with 20,000 iterations as burn-in for univariate analysis. The convergence of the parameters estimates

182 was checked using POSTGIBBSf90 program. In each sub-program, we performed 3 methods: traditional
183 pedigree BLUP (PBLUP) which utilizes only individual relationship records (pedigree information only)
184 (Henderson 1985), Genomic BLUP (GBLUP) which uptakes only genomic information (Meuwissen et al.
185 2001), and single step GBLUP (ssGBLUP) (Lourenco et al. 2020) which blended pedigree and genomic
186 information into a realized H matrix (Misztal et al. 2009; Lourenco et al. 2020).

187 **2.6.2. Non-linear Bayesian method (BayesR)**

188 BayesR assumes that SNP effects β_i follow a four-component normal mixture and the effect of SNP *i* is
189 assumed to be distributed as:

190 $$\beta_i \sim \pi_1(\beta_i = 0) + \pi_2 N\left(0.0001\sigma_g^2\right) + \pi_3 N\left(0, 0.001\sigma_g^2\right) + \pi_4 N\left(0, 0.01\sigma_g^2\right) \quad \text{(equation 3)}$$

191 Where σ_g^2 as defined above, σ_g^2 represents the total additive genetic variance (i.e., the cumulative
192 variance of all SNP effects) and $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ the mixing proportions such that $\sum_{i=1}^{4} \pi_i = 1$. The
193 mixing proportions π are assumed to follow a Dirichlet prior, $\pi \sim Dirichlet\ (\alpha + \beta)$, with α representing
194 a vector of pseudo counts and β the cardinality of each component. In this work, we used a flat Dirchlet
195 distribution, with $\alpha = (1,1,1,1)$ for the prior. As suggested by Moser et al. (2015), σ_g^2 is assumed to be a
196 random variable following an Inverse—χ2 distribution.

197 The analysis of BayesR consisted of three mains steps. First, numeric coding genotypes of 560 animals
198 (including 490 offspring and 70 parents) obtained from DArT sequencing were converted to ATGC format
199 (i.e., biallelic variant). Second, the trimmed sequences of each variant were blasted to nucleotide non-
200 redundant striped catfish database using Blast2GO (Conesa et al. 2005), named GENO_Phyp_1.0 (Kim et
201 al. 2018). This analysis matched 13,766 SNPs to 30 autosomal chromosomes. The remaining SNPs (388
202 SNPs) were randomly assigned to 30 chromosomes (12 – 13 SNPs/chromosome). Third, PLINK v1.9 (Purcell
203 et al. 2007) that utilizes genotype and mapping files was used to generate a full dataset for BayesR

204 (https://github.com/syntheke/BayesR). The default running cycles of BayesR included 50,000 steps with
205 20,000 steps discarded as burn-in. We used predicted phenotypes (computed in AIREMLf90 using
206 equation 1), as this package does not allow fitting covariates in the model.

207 **2.6.3. Machine learning and deep learning methods**

208 **KAML (Kinship Adjusted Multiple Loci)**

209 KAML is a machine learning-based method that incorporates cross-validation, multiple regression, grid
210 search and bisection algorithm to improve genomic prediction accuracy for complex quantitative traits
211 (Yin et al. 2020). KAML extends linear mixed model (equation 1) by incorporating quantitative trait
212 nucleotides - QTNs (e.g., traits controlled by genes with large and moderate effects) as covariates and a
213 SNP-weighted trait-specific kinship matrix as the (co)-variance assumption corresponding to the random
214 effect. The selections of QTNs and SNP weights are optimized by machine learning. In a matrix notation,
215 equation 1 becomes:

216 $\boldsymbol{Y = Xb + Qq + Za^* + e}$ with $\boldsymbol{a}^* \sim N(0, K_w\sigma_g^2)$, $\boldsymbol{e} \sim N(0, I\sigma_e^2)$ (equation 4)

217 Where σ_g^2 and σ_e^2 are as described above; $\boldsymbol{Q}$ is covariates matrix of QTNs detected throughout multiple
218 regression analyses; $\boldsymbol{q}$ is a vector of random effect of each QTN corresponding to $\boldsymbol{Q}$ and K_w is SNP-weighted
219 Kinship that is optimized by grid search and bisection procedure. The implementation of KAML included
220 two main steps, i.e., define $\boldsymbol{Q}$ and K_w and then predicting genomic breeding values of the validation set. In
221 the training step, KAML optimized parameters for determining QTNs and SNPs weights which used all
222 testing subset data provided throughout cross-validation. Next, in the prediction step, the optimized
223 parameters were used to predict genomic breeding values of each individual in validation set. A detailed
224 description of the step-wise procedures and the underlying algorithms of KAML is given in Yin et al. (2020).
225 In our analyses, we used PLINK (Purcell et al. 2007) to convert the genotype data into the right format to
226 be analyzed in KAML (https://github.com/YinLiLin/KAML) in R environment (R Core Team 2015). KAML

227 utilized adjusted phenotypes (i.e., phenotypic y-hats) for testing sets and masked (i.e., 'NA' values)
228 phenotypes of validation sets. To define the best parameters (i.e., pseudo QTNs, the optimal SNPs
229 weights, the optimal Log value and the optimized kinship matrix) for mixed models of each testing set, we
230 used five-fold cross-validation with 100 replications.

231 **Deep learning – Multilayer Perceptron (DL-MLP)**

232 For deep learning analysis, we used DeepGP package (https://github.com/lauzingaretti/deepGP/). A
233 detailed description of the theoretical framework is given in Zingaretti et al. (2020) and Pérez-Enciso and
234 Zingaretti (2019). In our analyses, we used two different algorithms of deep learning: Multilayer
235 perceptron and convolutional neural network.

236 DL-MLP is fully connected networks consisting of an input layer, one or several hidden layers, and an
237 output layer. In the context of genomic prediction, input layer receives SNPs data (i.e., 14,154 SNPs). Each
238 of SNP information in the input layer is transferred its value to the first neuron of the first hidden layer by
239 the Z function:

240 $Z^{(k)} = b_{k-1} + W^{(k-1)} f^{(k-1)}(x)$ (equation 5)

241 Where W is weighted SNP information, $f(x)$ is non-linear function and b is bias (i.e., constant). In the case
242 of genomic prediction, the input layer is SNPs genotype (i.e., input variables = number of SNPs) of
243 individual fully connected to the first hidden layers. As a result, the value in the first neuron of the first
244 hidden layer is the sum effect of Z function over 14,154 SNPs ($\sum_{i=1}^{14154} Z$). Likewise, information in the first
245 hidden layer is the input for transferring its value to the second hidden layers and finally to the output
246 layers. The significance step in obtaining best DL-MLP model is to define MLP structure network (i.e,
247 number of hidden layer and neuron of each layer) and hyperparameters which related to equation 5.
248 These parameters and their values in our study include learning rate (--lr 0.025 0.01), type of drop out (--
249 dr 0 0.01), type of regularization (--reg 0.0001), activation function (--act linear tanh), number of hidden

250 layer (--h 1 5), optimizer function (--optimizer Adam) or number of epochs (--epochs 50). These scripts
251 provided in DeepGP (https://github.com/lauzingaretti/DeepGP) were used to obtain hyperparameters in
252 our DL-MLP analyses.

253 **Deep learning – Convolutional Neural Network (DL-CNN)**

254 Convolutional neural network comprises of 3 types of layers: (1) convolutional layers, (2) fully connected
255 layers and (3) input layers. In the context of genomic prediction, first convolutional layer receives data of
256 14154 SNPs and write it to the first layer using kernel (i.e., a definite number of SNPs is grouped or
257 combined). Then, the pooling step is responsible to reduce spatial size of SNPs data of the convolutional
258 layer into the second convolutional layer. There are two types of pooling such as max pooling and average
259 pooling corresponding to maximum values and mean values of SNPs portion covered by the kernel. The
260 process was repeated to make update value of the last convolutional layer. At the last convolutional layer,
261 data were flattened (i.e., convert data of pooled feature of the last convolutional layer into single column
262 feeding to fully connected layers). The fully connected layer works similarly the MLP. The
263 hyperparameters used in our study are size of kernel (--ks 3), learning rate (--lr 0.025 0.01), dropout in the
264 first layer (-dr1 0 0.01), dropout in the hidden layer (--dr2 0 0.1 0.2), number of pooling size (--ps 1),
265 number of convolutional layers (--ncov 1 3), number of strides (--ns 1), number of convolutional
266 operations (--nfilters 8). These hyperparameters are fine-tuned under DeepGP with grid search provided
267 in DeepGP package (Zingaretti et al. 2020). As DeepGP cannot handle missing values, we used imputed
268 genotype for all analyses.

269 **2.6.4. Analysis of SNP subsets, highly significant SNPs and multi-trait mixed model**

270 **SNP subsets**

271 In addition to the analysis of the whole genotype data (14,154 SNPs), we also analyzed two different SNP
272 subsets. The first genotype subset (8,481 SNPs) was obtained through the default quality control (i.e.,

273 excluding genotypes with call rate <90% and minor allele frequency >0.05) in the ssGBLUP of blup90
274 package. The second subset of 834 SNPs remained after strict filtering applied (i.e., excluding genotypes
275 with call rate <90%, minor allele frequency >0.05 and markers that show significant deviations from Hardy-
276 Weinberg Equilibrium).

277 **ssGWAS to select significant SNPs for prediction**

278 Single step genome-wide association analysis (ssGWAS) was accomplished in BLUPF90 suite using full
279 phenotype data for disease resistance traits. We used preGSf90 and postGSf90 together with BLUPF90
280 program to compute SNP effect and its P-values. A panel of 402 SNPs were selected based on the P-value
281 < 0.0001. This genotype subset was used to test the predictive performance of five genomic prediction
282 methods used in our study.

283 **Multi-trait mixed model**

284 Multi-trait analyses are only available in BLUPF90. In this study, two studied traits were jointly analyzed
285 using PBLUP, GBLUP and ssGBLUP models in THRGIBBS1F90 (Tsuruta and Misztal 2006). This involves using
286 Gibbs Sampling with MCMC updates including 1,000,000 iterations with 100,000 burn-in steps.

287 **2.7. Five-fold cross validation**

288 The predictive performances of all five methods (PBLUP, GBLUP, ssGBLUP, BayesR, KAML and DeepGP)
289 used in this study were evaluated using repeated 5-fold cross validation. The procedure involved a random
290 partitioning of the full data into 5 subsets (each subset has 98 individuals including representatives of all
291 the families). The phenotypes of one subset were masked and GEBV/EBV (Genomic/Estimated Breeding
292 Value) of this set were predicted using the information from the other four subsets. This means one
293 phenotype was used once in the validation process of one repetition. Each analysis was repeated 5 times,
294 giving 25 validation sets in total and the correlation coefficient was averaged over 25 values. To compute

295 evaluation metric in each validation set, the GEBV/EBVs of the predicted phenotypes were correlated to
296 the actual phenotype. We chose Pearson correlation coefficient (Benesty et al. 2009) for continuous trait
297 (survival time) and Mathew correlation coefficient, MCC (Matthews 1975) for binary trait (survival status).
298 MCC is proved to be less bias amongst common metrics and giving more meaningful interpretation for
299 binary classification evaluation (Chicco and Jurman 2020). Specifically, we used *PresenceAbsence* R
300 package (Freeman and Moisen 2008) to produce 4x4 confusion matrix which assigned GEBV/EBV values
301 against the actual phenotypes for computing MCC (Supplementary file S1). To obtain the consistency and
302 comparable results among tested software/packages, we used the same subsets of testing/validation data
303 in BLUPF90 family program, BayesR, ML-KAML, DL-MLP and DL-CNN.

304 The genomic prediction accuracy for survival status and survival time was calculated as $Accuracy = \frac{r_{y,\hat{y}}}{\sqrt{h^2}}$;
305 where $r_{y,\hat{y}}$ is the correlation between the predicted breeding values $(\hat{y})$ and actual phenotypes(y).

306 **Data availability**

307 Data are available via Figshare portal provided by the journal.

308 **3. Results**

309 **3.1. Data, variance component and heritability**

310 The average survival rate and survival time after the pathogen challenge test were 32.2% and 6.83 days,
311 respectively (Table 1). When exposed to the *Edwardsiella ictaluri* pathogen, mortality occurred after 2
312 days/or 32.5 hours and the highest death rate was observed after 5 days/or 97.3 hours (Figure 1). There
313 were significant differences (P < 0.0001, Figure 1) in both survival rate (60.00% vs. 4.49%) and survival
314 time (8.66 vs. 5.00 days) between 20 highest and 20 lowest resistant families (Table 1). The variations in
315 the disease resistant traits among families were also significant (Supplementary Figures S1 & S2).

316 The heritabilities estimated for survival status and survival time using traditional PBLUP (0.65 – 0.71) were
317 greater than those obtained from GBLUP (0.55 – 0.66) and ssGBLUP (0.45 – 0.46) but they were lower
318 than those estimated by BayesR (0.61 – 0.86) (Table 2). Across statistical models (methods) used, the
319 estimates of heritability for survival status were slightly higher than those obtained for survival time (Table
320 2).

321 **3.2. Accuracy of genomic prediction using original genotype (un-imputed) data**

322 Figure 2 presents the accuracies of genomic breeding values for survival status and survival time using
323 Machine learning (ML-KAML) and Deep Learning (DL-MLP and DL-CNN) together with four other methods
324 (PBLUP, GBLUP, ssGBLUP and BayesR). When the original genotype data (i.e., un-imputed data) were used,
325 the prediction accuracies for both survival status and survival time obtained from KAML was higher than
326 those calculated from GBLUP and ssGBLUP, ranging from 8.3 to 9.5% and 48.8 to 64.5%, respectively (also
327 see Supplementary Table S1). Additionally, KAML outperformed BayesR but the difference in the
328 prediction accuracy was small (0.574 vs 0.557 for survival status and 0.715 vs. 0.699 for survival time). For
329 both traits, ssGBLUP was slightly better GBLUP. Interestingly, both GBLUP and ssGBLUP did not improve
330 the prediction accuracies for both survival status and survival time as compared with conventional PBLUP
331 method.

332 **3.3. Accuracy of genomic prediction using imputed genotype data**

333 Imputation of missing genotypes increased the accuracies of genomic prediction for both survival status
334 and survival time across the five methods used (Figure 3). The improvements for survival status were from
335 8.9% (GBLUP) to 14.3% (BayesR). For survival time, the increase in the prediction accuracy when the
336 imputed genotypes were used as compared with original data was largest for GBLUP (33.2%), followed by
337 BayesR (9.9%) and KAML (5.0%). Deep learning (DL) using multilayer perceptron (DL-MLP) and

338 convolutional neural network (DL-CNN) had lower prediction accuracies for survival time (0.596 – 0.602)
339 than other methods but the reverse trend was observed for survival status (Supplementary Table S1).

340 **3.4. Univariate vs. multi-variate analysis**

341 Our multivariate analyses used PBLUP, GBLUP and ssGBLUP, and the prediction accuracies for both
342 survival status and survival time are shown in Table 3. There was only an increase in the prediction
343 accuracy when PBLUP was used to estimate genomic breeding values (GEBV) for survival status (0.627 vs
344 0.597) while other multivariate analyses slightly decreased the accuracy of GEBV relative to univariate
345 model (3.4 to 15.0%).

346 **3.5. Accuracy using different subsets of genotype**

347 Table 4 presents the accuracies of GEBV for both survival status and survival time using different subsets
348 of genotype data (or different level of quality control in BLUPF90 program). Either when 8,481k SNPs (after
349 a standard quality control) or 834 SNPs (deep filtering) was used, the prediction accuracies were similar
350 to those estimated using whole genotype dataset (14,154 SNPs). In both traits, the predictive abilities with
351 machine learning (KAML) and/or deep learning methods (DL-MLP and DL-CNN) were slightly better than
352 those of other methods, 0.544 – 0.774 vs 0.513 – 0.646 respectively.

353 **3.6. Genomic prediction in combination with GWAS**

354 The prediction accuracies for the disease resistance traits that incorporated only significant SNPs
355 identified from genome-wide association analysis (GWAS) are given in Table 5. When un-imputed
356 genotypes were used, the prediction accuracies were generally greater to those of the full set of 14,154
357 SNPs. One exception was that GBLUP increased the accuracy by 54.3% for survival time. Furthermore,
358 there were also no differences in the prediction accuracies for both traits between the subset of 402
359 significant SNPs and the full dataset, when imputed genotypes were analyzed. In all analyses irrespective

360 of using un-imputed or imputed genotypes, ML-KAML and BayesR outperformed other methods by 11.5
361 – 29.1% for survival time (PBLUP, GBLUP and ssGBLUP).

362 **4. Discussion**

363 Development of disease resistance lines is essential for the aquaculture sector due to the combined
364 effects of aquaculture intensifications and environmental changes. Our study demonstrated that genomic
365 selection can be practiced improving resistance to *Edwardsiella ictaluri* disease in striped catfish
366 populations, using a range of different statistical methods especially machine and deep learning, and
367 BayesR. Salient findings from our study are discussed as follows.

368 First, the genomic prediction accuracies for the disease resistance traits (survival status and survival time)
369 were substantially improved when machine (i.e., KAML) and deep learning (DeepGP) methods were used
370 relative to BLUP-family methods (GBLUP, ssGBLUP). They also outperformed traditional pedigree-based
371 method (PBLUP) in this study. Hence, genomic selection to improve BNP disease is expected to achieve
372 greater genetic gain than the conventional selective breeding approach using pedigree and phenotype
373 information. The genomic selection can also shorten the breeding cycle of striped catfish from 3-4 to 1.5-
374 2 years as expected to reduce at least a half of generation interval in dairy cattle (Hayes et al. 2009). Our
375 results are consistent with those reported in farmed animal (Baker et al. 2020), livestock (Li et al. 2018;
376 González-Recio and Forni 2011) and plants (Montesinos-López et al. 2019; Montesinos-López et al. 2018;
377 Zingaretti et al. 2020). Despite the superiority of the machine learning methods (i.e., KAML) to PBLUP,
378 GBLUP and ssGBLUP, they had similar predictive power to BayesR for both traits in our study. Reports in
379 farmed animals concluded that BayesR is the best amongst Bayesian and GBLUP methods (Erbe et al.
380 2012) but no published information is available in aquaculture species to compare with our study. To date,
381 only two studies, one in shrimp (Palaiokostas 2021) that used extreme gradient boost method (one of the
382 machine learning approaches) and another in seabream that used support vector machines and linear

383 bagging classification (Bargelloni et al. 2021) and reported that these methods improved the prediction
384 accuracy for survival traits by 1-4% and 20 – 70% relative to GBLUP and Bayesian methods. Therefore,
385 machine or deep learning should be used for genomic evaluation of disease resistance traits, at least in
386 this population of striped catfish. The predictive powers of these methods using AI algorithms are
387 expected to be greater when big data (e.g., hundreds of thousands or million animals sequenced) are
388 analyzed for this population in the future.

389 Second, imputation of missing genotypes (using offspring and parent information available in the pedigree
390 of striped catfish) increased the predictive ability of all the five methods used, especially GBLUP by 9.9 to
391 33.2% relative to when un-imputed data were used. Also note that the imputation of missing genotypes
392 in our study differed from others that used software to compute missing genotypes without considering
393 parental relationship. With the use of offspring-parent information, the imputation likely increased
394 accuracy of G and H matrices used in GBLUP and single-step methods (i.e., ssGBLUP). Published
395 information across species also showed that imputation from low to high density SNP panels or from
396 commercial SNP arrays to whole genome sequence increased the genomic prediction accuracies for
397 resistance to parasite (i.e. sea lice count) and body traits in Atlantic salmon (Tsai et al. 2017), and
398 agronomic traits in maize (Crossa et al. 2013). However, some other studies also showed that imputed
399 genotype did not improve prediction accuracies in livestock animal breeding (Heidaritabar et al. 2016;
400 Chen et al. 2014; Van Binsbergen et al. 2015). Imputation has several benefits, firstly it can be made after
401 sequencing with no extra cost. Second, we can reduce sequence coverage to genotype a larger number
402 of individuals and eventually imputation can increase selection intensity (and selection accuracy), and this
403 can lead to increases in genetic progress made in selected populations. Thus, imputation of low-density
404 sequence can support our future incorporation of genomic selection for BNP disease in this striped catfish
405 population.

406 Third, we found that multivariate analysis improved the predictive ability for the two studied traits only
407 when PBLUP was used. However, both GBLUP and ssGBLUP did not increase the prediction accuracies for
408 both survival time and survival status. These results are expected due to the high and positive genetic
409 correlations (r_g) between the two traits (r_g = 0.81). A bivariate analysis of harvest body weight and fillet
410 yield in Nile tilapia also showed no improvement in the prediction accuracies for these traits (Joshi et al.
411 2020). Multi-trait analyses generally improve accuracy of genetic parameter estimates and animal
412 breeding values for traits whose genetic architectures are different (e.g., low vs high heritable) or
413 genetically antagonistic associated (i.e., unfavorably correlated). Examples of multivariate genomic
414 prediction models involving multi-trait in cassava and wheat plants gave better prediction accuracy over
415 univariate models (Okeke et al. 2017; Sun et al. 2017). To utilize the benefits of multi-trait analysis, we
416 continue collecting other traits of economic importance in striped catfish, namely disease tolerance,
417 disease resilience as well as immunological parameters to increase overall resistant capacity of the
418 animals in this catfish population.

419 Fourth, we also suggest using the full set of SNPs to obtain a reasonable level of prediction accuracy for
420 genomic selection to improve disease resistance traits in striped catfish. It has been well documented that
421 the prediction accuracy increased with number of markers and sequencing depth or genome coverage.
422 Due to characteristics of RAD-sequencing methods, there are still limited number of markers in our study
423 as compared with those reported in farmed animals or plants (Robledo et al. 2018; Davey and Blaxter
424 2010). It is necessary to examine prediction ability of whole genome data when sequencing costs are
425 reduced to an affordable level to sequence thousands (or hundreds of thousands) of animals in a near
426 future. However, there are studies in farmed animals reporting that whole genome sequence data
427 improved the predictive powers for complex traits by only 0.4 – 5.0% (Al Kalaldeh et al. 2019; VanRaden
428 et al. 2017).

429 Finally, the utilization of highly significant SNPs from ssGWAS can improve the prediction accuracies for
430 the two traits studied, for instance with un-imputed genotype data. On the one hand, these SNPs can be
431 used to develop a small SNP panel to genotype a large number of animals in both the training and
432 validation populations. On the other hand, the costs of genotyping animals for 402 SNPs may not be a lot
433 cheaper than DArT sequencing as used in this study. The use of significant SNPs can be beneficial when
434 their biological functions (pathways) are known. In species where reference genome or high quality
435 genome assembly is available, the incorporation of SNPs with biological functions (i.e. multi-omics)
436 improved the prediction accuracy, for instance, about 5 - 19% for fertility traits in dairy cattle (Abdollahi-
437 Arpanahi et al. 2017; Nani et al. 2019), or 27.4 – 60.7% for traits in inbred lines of Drosophila (Ye et al.
438 2020). Unfortunately, a good genome assembly is currently not available for striped catfish (Kim et al.
439 2018), this area thus deserves future studies to enable the efficient utilization of genomic information in
440 the genetic selection program for this economically important species in the aquaculture sector.

441 **5. Concluding remarks**

442 The genomic prediction accuracies for disease resistance traits of striped catfish were moderate to high,
443 suggesting possibilities for the application of genomic selection to improve resilience to BNP (Bacillary
444 Necrosis of Pangasius) disease caused by *Edwardsiella ictaluri* in this population. Machine learning
445 methods outperformed GBLUP-family in all our analyses, regardless of the original or imputed genotype,
446 full or subset data or only significant SNPs used. However, the prediction accuracies using machine
447 learning for both survival status and survival time were almost similar to those obtained from BayesR.
448 There were no clear advantages of deep learning to conventional methods (PBLUP, GBLUP and ssGBLUP).
449 Therefore, either machine learning or BayesR could be used for genomic evaluation of the disease
450 resistance traits in this striped catfish population. Furthermore, breeding to improve resistance to *E.*
451 *ictaluri* can use survival status or survival time as alternative selection criterion as there was no significant
452 difference in the prediction accuracies across the five different methods used in our study.

453 **References**
454

455 Abdollahi-Arpanahi, R., G. Morota, and F. Peñagaricano, 2017 Predicting bull fertility using genomic data
456 and biological information. *Journal of Dairy Science* 100 (12):9656-9666.
457 Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta *et al.*, 2010 Hot topic: A unified approach to
458 utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein
459 final score. *Journal of Dairy Science* 93 (2):743-752.
460 Al-Tobasei, R., A. Ali, A.L. Garcia, D. Lourenco, T. Leeds *et al.*, 2021 Genomic predictions for fillet yield
461 and firmness in rainbow trout using reduced-density SNP panels. *BMC genomics* 22 (1):1-11.
462 Al Kalaldeh, M., J. Gibson, N. Duijvesteijn, H.D. Daetwyler, I. MacLeod *et al.*, 2019 Using imputed whole-
463 genome sequence data to improve the accuracy of genomic prediction for parasite resistance in
464 Australian sheep. *Genetics Selection Evolution* 51 (1):1-13.
465 Baker, L.A., M. Momen, K. Chan, N. Bollig, F.B. Lopes *et al.*, 2020 Bayesian and Machine Learning Models
466 for Genomic Prediction of Anterior Cruciate Ligament Rupture in the Canine Model. *G3: Genes,*
467 *genomes, genetics* 10 (8):2619-2628.
468 Bargelloni, L., O. Tassiello, M. Babbucci, S. Ferraresso, R. Franch *et al.*, 2021 Data imputation and
469 machine learning improve association analysis and genomic prediction for resistance to fish
470 photobacteriosis in the gilthead sea bream. *Aquaculture Reports* 20:100661.
471 Barría, A., K.A. Christensen, G.M. Yoshida, K. Correa, A. Jedlicki *et al.*, 2018 Genomic predictions and
472 genome-wide association study of resistance against Piscirickettsia salmonis in coho salmon
473 (Oncorhynchus kisutch) using ddRAD sequencing. *G3: Genes, genomes, genetics* 8 (4):1183-
474 1194.
475 Benesty, J., J. Chen, Y. Huang, and I. Cohen, 2009 Pearson correlation coefficient, pp. 1-4 in *Noise*
476 *reduction in speech processing*. Springer.
477 Chen, L., C. Li, M. Sargolzaei, and F. Schenkel, 2014 Impact of genotype imputation on the performance
478 of GBLUP and Bayesian methods for genomic prediction. *PloS one* 9 (7):e101544.
479 Chicco, D., and G. Jurman, 2020 The advantages of the Matthews correlation coefficient (MCC) over F1
480 score and accuracy in binary classification evaluation. *BMC genomics* 21 (1):1-13.
481 Conesa, A., S. Götz, J.M. García-Gómez, J. Terol, M. Talón *et al.*, 2005 Blast2GO: a universal tool for
482 annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21
483 (18):3674-3676.
484 Crossa, J., Y. Beyene, S. Kassa, P. Pérez, J.M. Hickey *et al.*, 2013 Genomic prediction in maize breeding
485 populations with genotyping-by-sequencing. *G3: Genes, genomes, genetics* 3 (11):1903-1926.
486 Daetwyler, H.D., M.P. Calus, R. Pong-Wong, G. de Los Campos, and J.M. Hickey, 2013 Genomic
487 prediction in animals and plants: simulation of data, validation, reporting, and benchmarking.
488 *Genetics* 193 (2):347-365.
489 Davey, J.W., and M.L.J.B.i.f.g. Blaxter, 2010 RADSeq: next-generation population genetics. 9 (5-6):416-
490 423.
491 Dinh Pham, K., J. Ødegård, S. Van Nguyen, H. Magnus Gjøen, and G. Klemetsdal, 2021 Genetic analysis of
492 resistance in Mekong striped catfish (Pangasianodon hypophthalmus) to bacillary necrosis
493 caused by Edwardsiella ictaluri. *Journal of fish diseases* 44 (2):201-210.
494 Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto *et al.*, 2011 A robust, simple genotyping-by-
495 sequencing (GBS) approach for high diversity species. *PloS one* 6 (5):e19379.
496 Erbe, M., B. Hayes, L. Matukumalli, S. Goswami, P. Bowman *et al.*, 2012 Improving accuracy of genomic
497 predictions within and between dairy cattle breeds with imputed high-density single nucleotide
498 polymorphism panels. *Journal of Dairy Science* 95 (7):4114-4129.

Freeman, E.A., and G. Moisen, 2008 PresenceAbsence: An R package for presence absence analysis. *Journal of Statistical Software. 23 (11): 31 p.*

González-Recio, O., and S. Forni, 2011 Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genetics Selection Evolution* 43 (1):1-12.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard, 2009 Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* 92 (2):433-443.

Heidaritabar, M., M.P. Calus, H.J. Megens, A. Vereijken, M.A. Groenen *et al.*, 2016 Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. *Journal of Animal Breeding and Genetics* 133 (3):167-179.

Henderson, C., 1985 Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. *Journal of Animal Science* 60 (1):111-117.

Houston, R.D., T.P. Bean, D.J. Macqueen, M.K. Gundappa, Y.H. Jin *et al.*, 2020 Harnessing genomics to fast-track genetic improvement in aquaculture. *Nature Reviews Genetics* 21 (7):389-409.

Joshi, R., A. Skaarud, M. de Vera, A.T. Alvarez, and J. Ødegård, 2020 Genomic prediction for commercial traits using univariate and multivariate approaches in Nile tilapia (Oreochromis niloticus). *Aquaculture* 516:734641.

Kilian, A., P. Wenzl, E. Huttner, J. Carling, L. Xia *et al.*, 2012 Diversity arrays technology: a generic genome profiling technology on open platforms, pp. 67-89 in *Data production and analysis in population genomics*. Springer.

Kim, O.T.P., P.T. Nguyen, E. Shoguchi, K. Hisata, T.T.B. Vo *et al.*, 2018 A draft genome of the striped catfish, Pangasianodon hypophthalmus, for comparative analysis of genes relevant to development and a resource for aquaculture improvement. *BMC genomics* 19 (1):733.

Li, B., N. Zhang, Y.-G. Wang, A.W. George, A. Reverter *et al.*, 2018 Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Frontiers in genetics* 9:237.

Lourenco, D., A. Legarra, S. Tsuruta, Y. Masuda, I. Aguilar *et al.*, 2020 Single-step genomic evaluations from theory to practice: using SNP chips and sequence data in BLUPF90. *Genes* 11 (7):790.

Masuda, Y., I. Aguilar, S. Tsuruta, and I. Misztal, 2014 Acceleration of Computations in AI REML for Single-step GBLUP Models in *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production*.

Matthews, B.W., 1975 Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405 (2):442-451.

Meuwissen, T.H., B.J. Hayes, and M.E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157 (4):1819-1829.

Misztal, I., A. Legarra, and I. Aguilar, 2009 Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science* 92 (9):4648-4655.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet *et al.*, 2002 BLUPF90 and related programs (BGF90), pp. 743-744 in *Proceedings of the 7th world congress on genetics applied to livestock production*.

Montesinos-López, O.A., J. Martín-Vallejo, J. Crossa, D. Gianola, C.M. Hernández-Suárez *et al.*, 2019 New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. *G3: Genes, genomes, genetics* 9 (5):1545-1556.

Montesinos-López, O.A., A. Montesinos-López, J. Crossa, D. Gianola, C.M. Hernández-Suárez *et al.*, 2018 Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. *G3: Genes, genomes, genetics* 8 (12):3829-3840.

Moser, G., S.H. Lee, B.J. Hayes, M.E. Goddard, N.R. Wray *et al.*, 2015 Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet* 11 (4):e1004969.

Nani, J.P., F.M. Rezende, and F. Peñagaricano, 2019 Predicting male fertility in dairy cattle using markers with large effect and functional annotation data. *BMC genomics* 20 (1):258.

Nguyen, N.H., and P.V. Khang, 2021 Genome-Wide Marker Analysis for Traits of Economic Importance in Asian Seabass Lates calcarifer. *Journal of Marine Science and Engineering* 9 (3):282.

Nguyen, N.H., C. Phuthaworn, and W. Knibb, 2020 Genomic prediction for disease resistance to Hepatopancreatic parvovirus and growth, carcass and quality traits in Banana shrimp Fenneropenaeus merguiensis. *Genomics* 112 (2):2021-2027.

Nguyen, N.H., P. Rastas, H. Premachandra, and W. Knibb, 2018 First high-density linkage map and single nucleotide polymorphisms significantly associated with traits of economic importance in Yellowtail Kingfish Seriola lalandi. *Frontiers in genetics* 9:127.

Nguyen, N.P., 2014 Enviromental factors affecting the pathogenesis of Edwardsiella ictaluri in striped catfish Pangasianodon hypophthalmus (Sauvage).

Okeke, U.G., D. Akdemir, I. Rabbi, P. Kulakow, and J.-L. Jannink, 2017 Accuracies of univariate and multivariate genomic prediction models in African cassava. *Genetics Selection Evolution* 49 (1):1-10.

Palaiokostas, C., 2021 Predicting for disease resistance in aquaculture species using machine learning models. *Aquaculture Reports* 20:100660.

Pérez-Enciso, M., and L.M. Zingaretti, 2019 A guide on deep learning for complex trait genomic prediction. *Genes* 10 (7):553.

Pham, K.D., S.V. Nguyen, J. Ødegård, H.M. Gjøen, and G. Klemetsdal, 2020 Case study development of a challenge test against Edwardsiella ictaluri in Mekong striped catfish (Pangasianodon hypophthalmus), for use in breeding: Estimates of the genetic correlation between susceptibilities in replicated tanks. *Journal of fish diseases*.

Pham, K.D., J. Ødegård, S.V. Nguyen, H.M. Gjøen, and G. Klemetsdal, 2021 Genetic correlations between challenge tested susceptibility to bacillary necrosis, caused by Edwardsiella ictaluri, and growth performance tested survival and harvest body weight in Mekong striped catfish (Pangasianodon hypophthalmus). *Journal of fish diseases* 44 (2):191-199.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 81 (3):559-575.

R Core Team, 2015 R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org.

Robledo, D., C. Palaiokostas, L. Bargelloni, P. Martínez, and R. Houston, 2018 Applications of genotyping by sequencing in aquaculture breeding and genetics. *Reviews in aquaculture* 10 (3):670-682.

Silva, R.M., J.P. Evenhuis, R.L. Vallejo, G. Gao, K.E. Martin *et al.*, 2019 Whole-genome mapping of quantitative trait loci and accuracy of genomic predictions for resistance to columnaris disease in two rainbow trout breeding populations. *Genetics Selection Evolution* 51 (1):42.

Sun, J., J.E. Rutkoski, J.A. Poland, J. Crossa, J.L. Jannink *et al.*, 2017 Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. *The plant genome* 10 (2):plantgenome2016.2011.0111.

Tsai, H.-Y., O. Matika, S.M. Edwards, R. Antolín–Sánchez, A. Hamilton *et al.*, 2017 Genotype imputation to improve the cost-efficiency of genomic selection in farmed Atlantic salmon. *G3: Genes, genomes, genetics* 7 (4):1377-1383.

Tsuruta, S., and I. Misztal, 2006 THRGIBBS1F90 for estimation of variance components with threshold and linear models. *Threshold* 3 (4).

Vallejo, R.L., B.O. Fragomeni, H. Cheng, G. Gao, R.L. Long *et al.*, 2020 Assessing Accuracy of Genomic Predictions for Resistance to Infectious Hematopoietic Necrosis Virus With Progeny Testing of Selection Candidates in a Commercial Rainbow Trout Breeding Population. *Frontiers in Veterinary Science* 7.

Van Binsbergen, R., M.P. Calus, M.C. Bink, F.A. van Eeuwijk, C. Schrooten *et al.*, 2015 Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution* 47 (1):1-13.

Van Sang, N., G. Klemetsdal, J. Ødegård, and H.M. Gjøen, 2012 Genetic parameters of economically important traits recorded at a given age in striped catfish (Pangasianodon hypophthalmus). *Aquaculture* 344:82-89.

VanRaden, P.M., 2008 Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91 (11):4414-4423.

VanRaden, P.M., M.E. Tooker, J.R. O'connell, J.B. Cole, and D.M. Bickhart, 2017 Selecting sequence variants to improve genomic predictions for dairy cattle. *Genetics Selection Evolution* 49 (1):1-12.

Vu, N.T., T.T.T. Ha, V.T.B. Thuy, V.T. Trang, and N.H. Nguyen, 2020 Population Genomic Analyses of Wild and Farmed Striped Catfish Pangasianodon Hypophthalmus in the Lower Mekong River. *Journal of Marine Science and Engineering* 8 (6):471.

Vu, N.T., N.V. Sang, T.Q. Trong, N.H. Duy, N.T. Dang *et al.*, 2019a Breeding for improved resistance to Edwardsiella ictaluri in striped catfish (Pangasianodon hypophthalmus): Quantitative genetic parameters. *Journal of fish diseases* 42 (10):1409-1417.

Vũ, N.T., T.Q. Trọng, L.T. Đỉnh, H.T.N. Nga, and N.V. Hiệp, 2017 Draft report: Survey results from Project "Investigation for develop national standard regulation: freshwater fish - striped catfish broodtsock and fingerlings - quality requirements". *Draft report to Fishery Department by Research Institute for Aquacultute No 2*:25.

Vu, N.T., N. Van Sang, T.H. Phuc, N.T. Vuong, and N.H. Nguyen, 2019b Genetic evaluation of a 15-year selection program for high growth in striped catfish Pangasianodon hypophthalmus. *Aquaculture*.

Vu, S.V., C. Gondro, N.T. Nguyen, A.R. Gilmour, R. Tearle *et al.*, 2021 Prediction Accuracies of Genomic Selection for Nine Commercially Important Traits in the Portuguese Oyster (Crassostrea angulata) Using DArT-Seq Technology. *Genes* 12 (2):210.

Whalen, A., G. Gorjanc, and J.M. Hickey, 2020 AlphaFamImpute: high-accuracy imputation in full-sib families from genotype-by-sequencing data. *Bioinformatics* 36 (15):4369-4371.

Ye, S., J. Li, and Z. Zhang, 2020 Multi-omics-data-assisted genomic feature markers preselection improves the accuracy of genomic prediction. *Journal of animal science and biotechnology* 11 (1):1-12.

Yin, L., H. Zhang, X. Zhou, X. Yuan, S. Zhao *et al.*, 2020 KAML: improving genomic prediction accuracy of complex traits using machine learning determined parameters. *Genome Biology* 21 (1):1-22.

Zingaretti, L.M., S.A. Gezan, L.F.V. Ferrão, L.F. Osorio, A. Monfort *et al.*, 2020 Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Frontiers in plant science* 11:25.

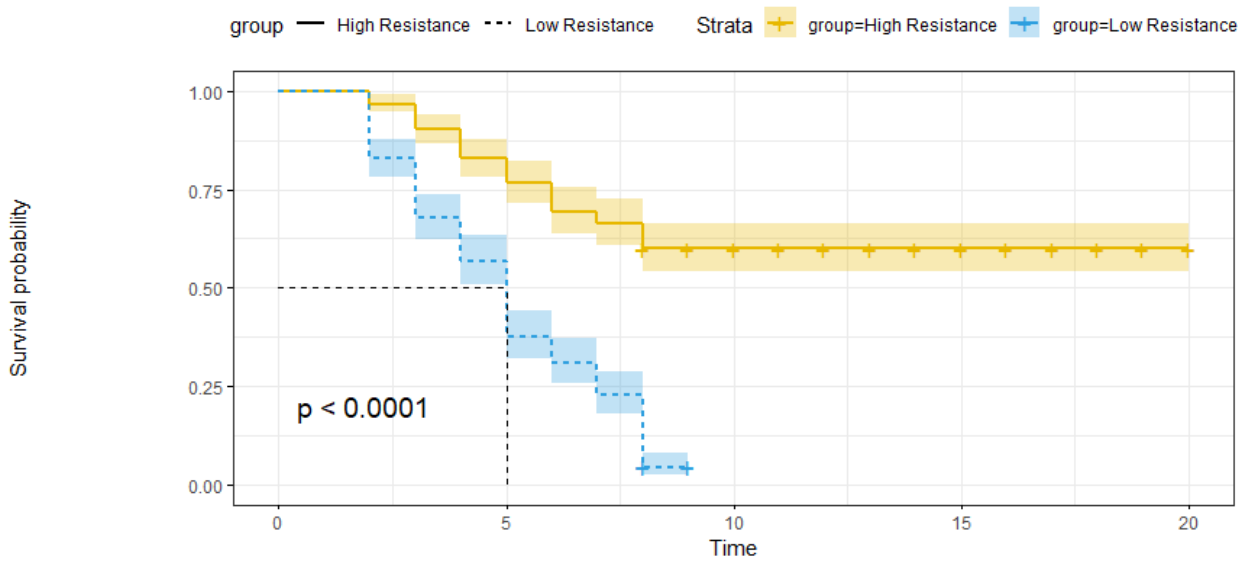

Figure 1. Survival trend in day post challenge of high and low resistance groups in genotyped individuals of striped catfish.

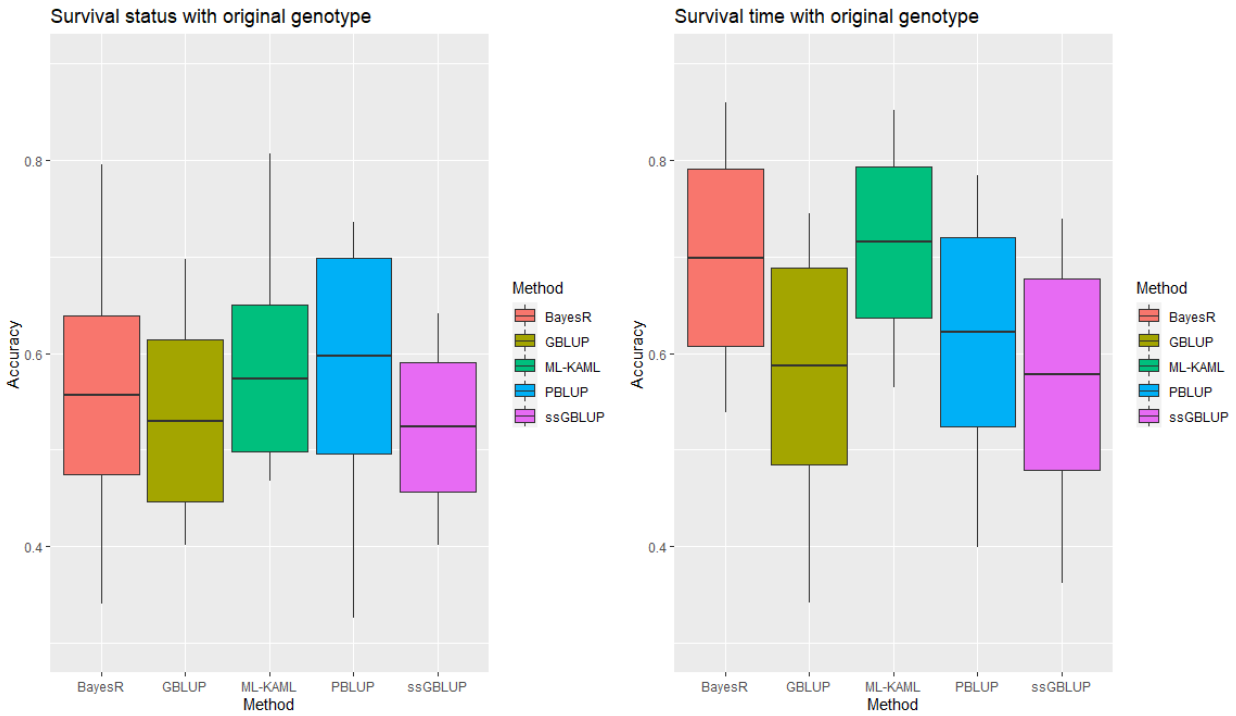

Figure 2. Accuracy of survival status and survival time traits using original genotype. Middle line of the box is mean accuracy; top and bottom lines of the box is accuracy ± one standard deviation. End points of vertical line represent min and max values. Note that PBLUP uses phenotype and pedigree information only.

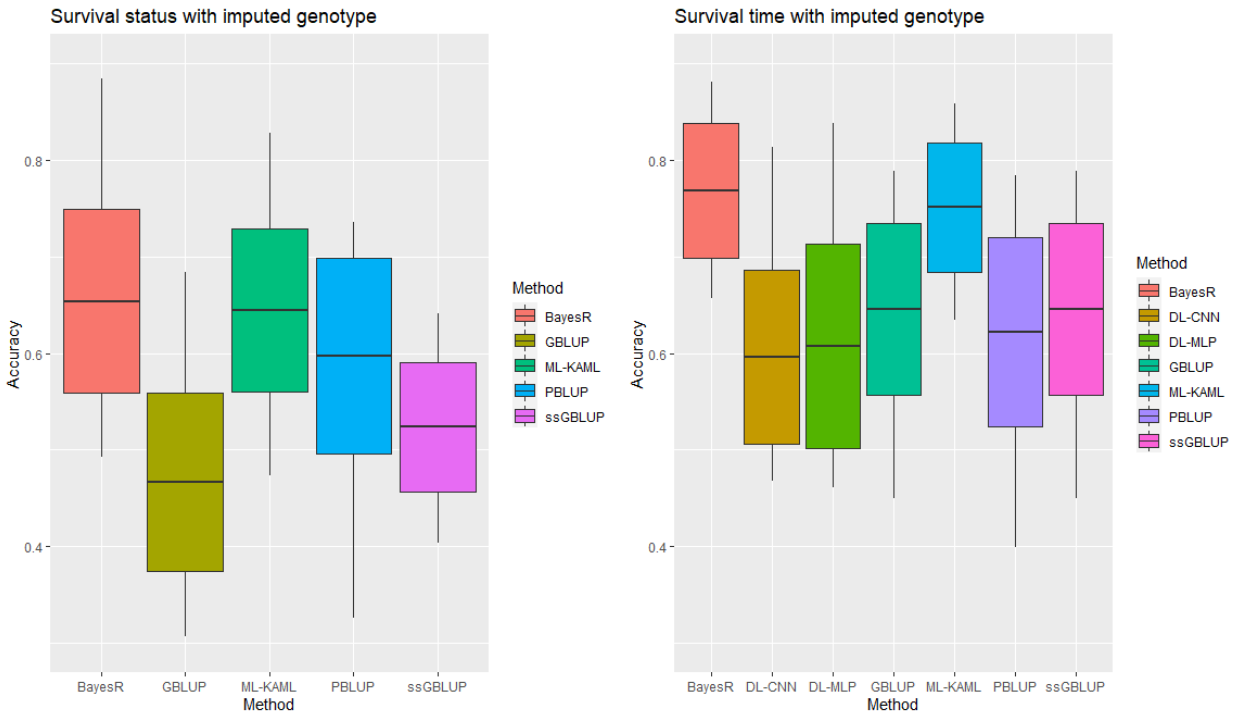

Figure 3. Accuracy of survival status and survival time traits using imputed genotype. Middle line of the box is mean accuracy; top and bottom lines of the box is accuracy ± one standard deviation. End points of vertical line represent min and max values. Note that PBLUP uses phenotype and pedigree information only.

Table 1. Number of animals (n) and phenotypes in the disease resistant and susceptible groups used for genotyping.

Groups	20 Lowest survival family		20 Highest survival families		All 40 families	
Trait	Individuals	Mean ± SD	Ind/family	Mean ± SD	n	Mean ± SD
Survival status (%)	245	4.49 ± 20.75	245	60.00 ± 49.09	490	32.22 ± 46.79
Survival time (day)	245	5.00 ± 2.20	245	8.66 ± 4.33	490	6.83 ± 3.89

Table 2. Variance components and heritability for survival status and survival time using AIREMLf90 and THRGIBBS1f90 sub-program in BLUPF90 and BayesR

Trait	Models	Additive genetic variance	Phenotypic variance	heritability
Survival status	PBLUP	0.147	0.207	0.71
	GBLUP	0.128	0.194	0.66
	ssGBLUP	0.085	0.190	0.45
	BayesR	0.016	0.027	0.61
Survival time	PBLUP	9.579	14,736	0.65
	GBLUP	7.395	13,450	0.55
	ssGBLUP	6.408	13.893	0.46
	BayesR	10,470	12,175	0.86

Note that estimate of genetic variance was not provided by ML-KAML or DeepGP Including DL-MLP and DL-CNN

Table 3. Genomic prediction accuracy for survival status and survival time using uni- and multivariate models

Trait	Method	Univariate		Multivariate		Increasement (%)
		Accuracy	Standard deviation	Accuracy	Standard deviation	
Survival status	PBLUP	0.597	0.101	0.627	0.095	+ 5.0
	GBLUP	0.577	0.095	0.542	0.087	- 6.1
	ssGBLUP	0.602	0.089	0.556	0.090	- 7.6
Survival time	PBLUP	0.622	0.098	0.601	0.083	- 3.4
	GBLUP	0.645	0.089	0.548	0.094	- 15.0
	ssGBLUP	0.646	0.089	0.558	0.094	- 13.6

Note: multivariate analysis is not available in BayesR, KAML and DeepGP.

Table 4. Accuracy of genomic prediction using different subsets of imputed genotypes

Trait	Method	14,154 SNPs		8,481 SNPs		834 SNPs	
Survival status		Accuracy	Standard deviation	Accuracy	Standard deviation	Accuracy	Standard deviation
	GBLUP	0.577	0.090	0.603	0.136	0.513	0.106
	ssGBLUP	0.600	0.082	0.611	0.141	0.530	0.107
	BayesR	0.654	0.095	0.646	0.107	0.606	0.126
	ML-KAML	0.645	0.085	0.650	0.101	0.592	0.094
	DL-MLP	0.739	0.087	0.628	0.117	0.622	0.088
	DL-CNN	1.059	0.038	0.711	0.132	0.643	0.080
Survival time	GBLUP	0.645	0.072	0.649	0.088	0.555	0.104
	ssGBLUP	0.646	0.089	0.649	0.088	0.560	0.104
	BayesR	0.768	0.069	0.757	0.076	0.747	0.081
	ML-KAML	0.751	0.067	0.774	0.073	0.722	0.086
	DL-MLP	0.607	0.087	0.642	0.121	0.612	0.101
	DL-CNN	0.596	0.090	0.544	0.094	0.599	0.100

Whole genotype data = 14,154 SNPs. Subset 1 (8,481 SNPs) resulted from standard QC (i.e., filtering minor allele frequency (>0.05), call-rate for SNP and animal (>0.90) and removal of monomorphic SNPs). Subset 2 (834 SNPs) from deep QC (as default QC but also remove markers that showed significant deviation from Hardy Weinberg Equilibrium and offspring-parent conflicting genotypes)

Table 5. Accuracy of genomic prediction for survival status and survival using highly significant SNPs (P < 0.0001)

Trait	Method	Original (un-imputed) genotype			Imputed genotype		
Survival status		14K SNPs	402 SNPs	%	14K SNPs	402 SNPs	%
	PBLUP	0.597	-	-	-	-	-
	GBLUP	0.530	0.533 ± 0.083	+ 0.0	0.577	0.521 ± 0.080	- 9.7
	ssGBLUP	0.545	0.541 ± 0.092	- 0.0	0.600	0.530 ± 0.083	- 11.7
	BayesR	0.572	0.578 ± 0.125	+ 1.0	0.654	0.654 ± 0.095	+ 0.0
	ML-KAML	0.628	0.584 ± 0.112	- 7.0	0.645	0.652 ± 0.088	+ 1.1
	DeepGP-MLP*	-	-	-	0.739	0.615 ± 0.111	- 16.7
	DeepGP-CNN*	-	-	-	1.059	0.747 ± 0.077	- 29.5
Survival time	PBLUP	0.622	-	-	-	-	-
	GBLUP	0.484	0.747 ± 0.102	+ 54.3	0.645	0.595 ± 0.091	- 7.8
	ssGBLUP	0.645	0.744 ± 0.105	+ 13.7	0.646	0.601 ± 0.091	- 7.0
	BayesR	0.699	0.833 ± 0.103	+ 19.2	0.768	0.768 ± 0.069	+ 0.0
	ML-KAML	0.715	0.855 ± 0.106	+ 19.6	0.751	0.761 ± 0.089	+ 1.3
	DeepGP-MLP	n.e.	-	-	0.607	0.620 ± 0.152	+ 3.1
	DeepGP-CNN	n.e.	-	-	0.596	0.665 ± 0.106	+ 11.6

* Pearson correlation were used instead of Mathew's coefficient.