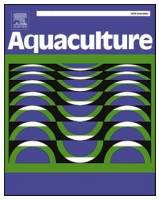
Aquaculture 585 (2024) 740677

Contents lists available at ScienceDirect

Aquaculture

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



Single-step genomic prediction for body weight and maturity age in Finnish rainbow trout (*Oncorhynchus mykiss*)

Andrei A. Kudinov [a,*], Antti Nousiainen [b], Heikki Koskinen [b], Antti Kause [a]

[a] Natural Resources Institute Finland (Luke), Tietotie 4, 31600 Jokioinen, Finland
[b] Natural Resources Institute Finland (Luke), FI-70210 Kuopio, Finland

ARTICLE INFO

Keywords:
Genomic prediction
SNP
Aquaculture
Validation
ssGBLUP

ABSTRACT

The use of genomic information has been proven to be a highly effective in predicting genomic breeding values (GEBV) across various species, including aquatic organisms. In the Finnish national rainbow trout breeding programme, the integration of genomic selection holds particular significance for the traits recorded on sibling fish reared in the main commercial sea production environment, given the selection occurs among the breeding candidates reared in the freshwater nucleus. In the programme, family tanks allow to maintain a pedigree for a large number of fish, and genotyping of a portion of the fish accompanied with a single-step genomic evaluation (ssGBLUP) would maintain high selection intensity and simultaneously make use of possibilities of genomic selection. In this study we used three different statistical approaches to quantify the selection accuracy of ssGBLUP evaluation of body weight and maturity age, relative to the evaluation based on the traditional sire-dam-offspring pedigree (PBLUP). The data included 600,409 fish in the pedigree among which 214,410 and 4573 were phenotyped for the reported traits and genotyped, respectively. Firstly, a phenotypic cross validation study showed that ssGBLUP had a slightly better prediction power for body weight and maturity age recorded at the sea, with an average 2.7% relative increase in accuracy compared to PBLUP. Secondly, a linear regression (LR) of GEBVs computed using either full or reduced dataset demonstrated that the ssGBLUP model had a consistently lower bias and dispersion compared to the PBLUP model, underscoring its efficacy in dealing with complex datasets like ours. When considering the reliability of [G]EBV predictions, the use of ssGBLUP model resulted in a significant improvement. There is, on average, a notable 50% relative increase in the reliability of predictions for the sea-recorded traits. Thirdly, the enhancement in reliability was further evidenced by the individual assessment of [G]EBVs computed using the reverse reliability methodology. Notably, genotyped individuals experienced an average increase of 0.27 units in reliability, while ungenotyped individuals experienced a corresponding increase of 0.03 units. The results show that the ssGBLUP method had higher prediction accuracy for both sea and freshwater traits compared to PBLUP. The developed ssGBLUP model will be instrumental in Finland's rainbow trout breeding, facilitating precise and efficient selection of new candidates.

1. Introduction

Genomic prediction is a method to predict genomic estimated breeding values (GEBV) of breeding candidates using genomic marker information, and a set of phenotyped and genotyped individuals known as a reference population (Meuwissen et al., 2001). Genomic selection in aquaculture species has been shown to be especially useful for hard-to-record-traits, such as disease resistance and product quality traits that are typically recorded from sibs of the breeding candidates (Houston et al., 2020). Aquaculture breeding programmes are shifting from pedigree-based schemes to genomic evaluations in which genotyping with thousands of DNA markers is needed. Genotyping is costly, and not all fish in a breeding programme can be always genotyped. One solution would be to reduce the number of fish in a breeding programme but this may slow down genetic gain, e.g. via reduction in selection intensity. An alternative is selective genotyping of the most interesting breeding candidates, for instance, based on the pre-existing information on their sibs' genotypes and performance for the hard-to-record traits. Consequently, fish breeding programmes may benefit from a method where GEBVs are predicted based on both a traditional sire-dam pedigree and

* Corresponding author.
E-mail address: andrei.kudinov@luke.fi (A.A. Kudinov).

https://doi.org/10.1016/j.aquaculture.2024.740677
Received 15 September 2023; Received in revised form 8 January 2024; Accepted 15 February 2024
Available online 18 February 2024

genotypes obtained from a portion of fish.

Single-step genomic evaluation (ssGBLUP) is a method to combine phenotypes, pedigree and genomic information from a reference population and candidates in a single statistical model to estimate GEBVs (Legarra et al., 2009; Aguilar et al., 2010; Christensen and Lund, 2010). Genotyping itself typically leads to more accurate breeding values of the genotyped individuals, but additionally, the combination of pedigree (**A**) and genomic (**G**) relationship matrix is especially valuable because it may also improve prediction in ungenotyped sibs, reduce bias and predict GEBVs instead of sum of marker effects (aka. Direct Genomic Value, Christensen and Lund, 2010; Kachman et al., 2013; Mäntysaari et al., 2020).

Validation of genomic prediction is an important instrument to understand how appropriate the developed prediction model is for routine genomic evaluation. It is also widely used tool to compare genomic prediction models. In a nutshell, validation stands for regression of corrected phenotype or GEBV calculated accounting phenotypic data on predicted GEBV. In a validation study, predicted GEBVs rely on pedigree and/or genomic information. To estimate predicted GEBVs, phenotypic data is truncated to simulate the absence of data in the candidates.

In aquaculture species, the most common validation method has been phenotypic validation, in which phenotype, or phenotype corrected for fixed effects, is used as response variable. A similar statistical approach is used in livestock breeding (Jairath et al., 1998; Strandén and Mäntysaari, 2010) but it was realized that alternative methods are needed for complex structured populations, models with multiple random factors, phenotypes that are solutions of statistical models (e.g. maternal effects, regression parameters) and for single-step models used on routine basis in evaluations (Interbull, 2023). For such cases Legarra and Reverter (2018) suggested a more general validation approach - linear regression (LR) of the GEBVs computed using the full data on the GEBVs from the truncated data.

In the Finnish national breeding programme for rainbow trout, family tanks are used at the initial phase of growth which allows to maintain a sire-dam-offspring pedigree for large number of fish, and breeding value evaluation has been based on this pedigree (PBLUP) (Kause et al., 2005, 2022). Growth, maturity age, body shape, viscera percentage, survival, skeletal deformations, and eye cataract are selected in two environments, at sea (main commercial production environment) and at the freshwater nucleus. Fraslin et al., 2022a, 2022bshowed that genomic selection for disease resistance against *Flavobacterium columnare* infection is possible in this population, when sibs of the breeding candidates are genotyped and tested outside the nucleus for survival under a natural outbreak. Genotyping also a portion of the breeding candidates accompanied with a single-step genomic evaluation would maintain high selection intensity for all the currently recorded traits and simultaneously make use of possibilities of genomic selection, e.g. for disease resistance.

The aim of this study was 1) to implement ssGBLUP model in the Finnish rainbow trout breeding programme; 2) to quantify the prediction power of the developed ssGBLUP model using different validation methods, and 3) to compute exact selection accuracies for both genotyped and non-genotyped individuals. We focus on growth and maturity traits that so far have not been the most urgent focus because the phenotypic recording of these traits is extensive both at sea and freshwater (Kause et al., 2003, 2005, 2022) but breeding of these traits will likely benefit from genomic information. Our work is an example of implementation of genomic prediction in a real rainbow trout breeding programme.

2. Material and methods

2.1. Data

2.1.1. Phenotypic data

The data was obtained from the Finnish national breeding programme maintained by Luke at the Enonkoski freshwater nucleus and multiple sea stations (Kause et al., 2005, 2022). Hatched fish were kept in a full-sib family tanks until individually tagged - allowing pedigree recording. At tagging each family was split and placed to freshwater and sea testing stations. The freshwater nucleus is a flow-through farm with tanks and raceways, and the water comes from a nearby lake. The sea stations, one or two each year, are fully commercial farms with open net pens, located along the coast of Finland and Åland islands at the Baltic Sea.

Pedigree included 600,409 fish and 6234 families, born between 1992 and 2019. Two populations were present in the pedigree: population I consisted of generation born in 1989, 1992, 1995, 1998, 2001, 2004, 2007, 2010, 2013, population II has two subpopulations: II_a consists of fish born 1990, 1993, 1996, 1999, 2002, 2005, 2008, 2011, and 2014 from which both 2018 and 2019 were generated, and II_b consists of 1997, 2000, 2003, 2006, 2009, 2012, and 2015. Population II_b was established using fish of II_a from year 1993. There were no fish in year classes of 2016 and 2017. The base population was created with fish assumed to be unrelated, born in 1989 and 1990 (Kause et al., 2005). Unknown parent groups formed separately by year and sex were included at the beginning of the pedigree to minimize incompatibility issue in single-step genomic prediction (Misztal et al., 2013).

Recorded traits were four body weight traits recorded at the ages of 1, 2, and 3 years at freshwater ($Weight_1$, $Weight_2$, $Weight_3$) and at the sea at the age of 2 years ($Sea\ weight_2$), and three binary maturation traits (0 = late maturity age, 1 = early maturity age) recorded for the males and females at freshwater ($Maturity_{male}$, $Maturity_{female}$) and for the males at sea ($Sea\ maturity_{male}$; Kause et al., 2005). Males are recorded to mature at the age of 2 and 3 years, and females at the age of 3 and 4 years. At freshwater live fish, and at sea gutted fish, are scored for maturity status by inspection by ultrasound and visual observation, respectively. Female maturity trait is not available from sea because at sea the fish are recorded at age 2 years. Number of records for all traits are shown in Table 1.

The population has been selected based on pedigree-based EBVs estimated using MiX99 software (Strandén and Lidauer, 1999; Kause et al., 2022), and the rate of inbreeding has been controlled by the optimal genetic contribution method and by avoiding mating of relatives (Kause et al., 2005).

2.1.2. Genomic data

Genotypes were available from 4573 fish born 2014, 2018, and 2019. In year class 2014, all the fish were genotyped. In year class 2018 and at sea, the genotyped fish were randomly sampled regards to the phenotypic value. In year class 2019, all early maturing males were genotyped, otherwise the sampling for genotyping was random. Number of genotyped fish with associated phenotypic data is presented in Table 1. DNA samples from fin clips were collected and genotyped using 57 K Axiom Trout Array (https://www.thermofisher.com/order/catalog/product/550571). The proportion of genotyped fish for the year 2018 and 2019 was 16% and 28%, respectively. Quality control was performed in Plink 1.9 software (Purcell et al., 2007) with following filtering criteria: SNPs mapping to a single position in the genome (Fraslin et al., 2022a), average call rate for passing SNPs $\geq$0.90, average call rate for passing samples $\geq$0.70, Hardy-Weinberg equilibrium exact test p-value <1E-50, and minor allele frequency < 0.01. After quality control 40,374 markers remained for imputation with AlphaImpute software (Hickey et al., 2012). Imputation of missing SNPs in the genotyped individuals was performed using family-based imputation approach. For the genotyped fish without genotyped family members missing, alleles were imputed by the most common allele in all genotyped fish.

2.2. Mixed model equation

Multitrait pedigree BLUP (PBLUP) and single-step genomic BLUP

Table 1
Number of records and heritability of the traits.

Trait	Number of records in the full data	Number of records in the 1-year truncated data	Number genotyped fish	Heritability (h^2)	Genetic standard deviation
$Weight_2$	96,544	95,038	2237	0.33	113
$Weight_3$	101,509	98,770	2497	0.34	230
Sea $weight_2$	85,927	83,750	1532	0.33	165
$Maturity_{female}$	55,404	53,640	1130	0.27	0.26
$Maturity_{male}$	41,202	40,014	1064	0.28	0.16
Sea $maturity_{male}$	32,157	31,270	673	0.42	0.20

(ssGBLUP) were used to predict EBVs and GEBVs, correspondingly. The mixed model equations used were:

$$\begin{bmatrix} y_{Weight_1} \\ y_{Weight_2} \\ y_{Weight_3} \\ y_{Sea\ weight_2} \\ y_{Maturity_{male}} \\ y_{Maturity_{female}} \\ y_{Sea\ maturity_{male}} \end{bmatrix} = \begin{bmatrix} by_{Weight_1} \\ \\ \\ \\ by_{Maturity_{male}} \\ by_{Maturity_{female}} \end{bmatrix} + \begin{bmatrix} \\ \\ \\ \\ byts_{Sea\ maturity_{male}} \end{bmatrix} + \begin{bmatrix} \\ bytsms_{Weight_2} \\ bytsms_{Weight_3} \\ bytsms_{Sea\ weight_2} \end{bmatrix} + \begin{bmatrix} \\ bytsc_{Weight_2} \\ bytsc_{Weight_3} \end{bmatrix} + \begin{bmatrix} \\ bytsd_{Weight_2} \\ bytsd_{Weight_3} \\ bytsd_{Sea\ weight_2} \end{bmatrix} + \begin{bmatrix} \\ bytst_{Weight_2} \\ bytst_{Weight_3} \\ \\ bytst_{Maturity_{male}} \\ bytst_{Maturity_{female}} \end{bmatrix} + [byft] + [a] + [e.]$$

where y_i are vectors of observations on traits;

fixed effects are: *by* is birth year, *byts* = birth year and testing station, *bytsms* = birth year, testing station, maturity, and sex, *bytsc* = birth year, testing station, and cataract disease, *bytsd* = birth year, testing station, and spinal deformation, *bytst* = birth year, testing station, and tank;

random effects are: *byft* = birth year and family tank, *a* = additive genetic effect, and *e* = residual.

Maturity, and spinal deformation were coded as 0 in case late maturity and no disease observed, and 1 in opposite cases. Cataract due to *Diplostomum* spp. eye flukes was visually scored as 0 = healthy eyes, 1 = one opaque eye, and 2 = both eyes opaque. Males were coded as 1 and females as 2. A missing fixed effect record was coded as 9 (Kause et al., 2022). In two newest year classes, the fish at the nucleus were reared in two tanks into which the fish were randomly allocated to. Number of effects levels shown in Appendix1a. Breeding value prediction was performed with MiX99 software (Strandén and Lidauer, 1999). $Weight_1$ is recorded on all fish and included in the model only to account for selection bias and any missing observations in traits recorded later in life (Martinez et al., 2006; Janhunen et al., 2014). The genetic correlations of the traits are presented in the Appendix 1. The variance components used in routine breeding evaluations were estimated using phenotypic data of seven year classes, 2001–2007, with the sample size of 200,737 for $Weight_1$, 36,937 for $Weight_2$, 30,744 for $Weight_3$, 33,481 for Sea $weight_2$, 13,197 for $Maturity_{male}$, 19,098 for $Maturity_{female}$, and 10,075 for Sea $maturity_{male}$. The statistical model described in Section 2.2. was used.

2.3. Single-step genomic BLUP

In ssGBLUP full pedigree relationship matrix $\mathbf{A}^{-1}$ is replaced by a joint relationship matrix $\mathbf{H}^{-1}$ (Aguilar et al., 2010; Christensen and Lund, 2010) computed as:

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{pmatrix}$$

where, $\mathbf{A_{22}}$ is a part of **A** for the genotyped animals only and **G** is the genomic relationship matrix. The **G** matrix was computed using hginv v1.0 software (Strandén and Mäntysaari, 2018) as:

$$\mathbf{G} = s_t\,(1-w)\,\mathbf{G}_{05} + w\mathbf{A}_{22}$$

where w is the residual polygenic proportion equal to 0.05, $\mathbf{G_{05}} = 2\left(\frac{\mathbf{M}_{101}\mathbf{M}'_{101}}{m}\right)$ with $\mathbf{M}_{101}$ as an n by m marker matrix with the genotypes coded by {−1,0,1}, m is the number of SNP markers, n is the number of genotyped animals, i.e. assuming allele frequencies = 0.5, and s_t is a scaling factor computed as $\left(\frac{trace(\mathbf{A}_{22})}{trace(\mathbf{G_{05}})}\right)$. The scaling factor was used to make the average of the diagonals of the **G** matrix equal to the average of the diagonal of the $\mathbf{A}_{22}$ matrix. The unknown parent groups were included into $\mathbf{H}^{-1}$ matrix as shown by Misztal et al. (2013) and Matilainen et al. (2018). Inbreeding was computed using RelaX2 software (Strandén and Vuori, 2006) and accounted during $\mathbf{A}^{-1}$ construction in MiX99.

2.4. Model validation

Four validation approaches were tested: *CrossV_Y**, *CrossV_BV*, *ForwardP_Y**, and *ForwardP_BV* (Table 2). The data of year classes 2018 and 2019 were modified to perform different validation tests. The approaches were different on the fraction of phenotypic information used to estimate reduced [G]EBV ([G]EBVr) and information used as a response variable. Reduced data was obtained either by repeated five-fold cross validation across the two year classes (*CrossV*) or by the deletion of the whole last year class of the phenotypic data (*ForwardP*; forward prediction). Phenotype adjusted for fixed effects and random family tank effect (*Y**) or [G]EBV from the full data ([G]EBVf) were used as the response variable in a regression model. In the reduced dataset, groups of genotyped fish with and without own phenotypic records are termed as training and test sets, respectively. Every trait had specific

Table 2
Explanation of methods used for model validation.

Method	TBV[a]	Fish truncated from the data
*CrossV_Y**	Y*[b]	5-fold cross-validation + full sibs
CrossV_BV	[G]EBVf[c]	5-fold cross-validation + full sibs
*ForwardP_ Y**	[G]EBVf	Forward prediction (all fish born 2019)
ForwardP_BV	[G]EBVf	Forward prediction (all fish born 2019)

[a] TBV = True breeding value.
[b] Y* = Phenotype adjusted for the fixed effects and random tank effect.
[c] [G]EBVf = [G]EBV predicted using full phenotypic data.

Table 3
Ratio of the fish in test and training set by trait in the 5-fold cross validation (*CrossV*) and forward prediction validation (*ForwardP*).

Trait	Validation method	
	CrossV [a]	*ForwardP* [a]
$Weight_2$	447 / 1790	804 / 1433
$Weight_3$	391 / 2106	1071 / 1426
Sea $weight_2$	306 / 1226	1148 / 384
$Maturity_{female}$	226 / 904	665 / 465
$Maturity_{male}$	212 / 852	583 / 481
Sea $maturity_{male}$	134 / 539	495 / 178

[a] Test/training fish.

testing group due to different number of records available (Table 3).

2.4.1. Five-fold sib cross-validation (CrossV)

Genotyped and phenotyped fish born 2018 and 2019 were randomly split into five folds. Every fold was used as a testing set while the rest four acted as training set. Phenotypic records of all traits were masked not only for the fish in a fold, but also for their genotyped and ungenotyped full sibs outside the fold, however only the fish in the fold were considered as the test set. Avoidance of full-sib data was done to prevent overtraining of the model. Data reduction was always followed by the prediction of [G]EBVr and regression of Y^* or [G]EBV (on y-axis) on [G]EBVr (on x-axis). Overall procedure was repeated 20 times. The presented results are the mean of 100 runs (i.e., 5 × 20).

2.4.2. Forward prediction (ForwardP)

Similar to the practice in livestock (Mäntysaari et al., 2010) all phenotypes of all traits from the latest year class 2019 were removed from the data regardless of whether the fish was genotyped or not. This strategy imitates situation when prediction is performed for a future generation without own phenotypic records. All genotyped fish with masked records for a particular trait was considered as a test set. Number of records in the training set is presented in Table 1.

2.4.3. Y* as function of [G]EBVr

For the fish in the test group, regression of Y* on [G]EBVr was performed using formula: $Y^* = b_0 + b_1[G]EBVr$, where intercept b_0 and slope b_1 are measures of bias and dispersion. Prediction accuracy (*acc*) was estimated as the Pearson correlation between Y* and [G]EBVr divided by the square root of heritability ($\frac{cor(Y^*,[G]EBVr)}{\sqrt{h^2}}$). These are parameters that reflect population- and model- specific levels of accuracy (Legarra and Reverter, 2018).

2.4.4. EBV or GEBV as function of [G]EBVr

For the fish in the test group, linear regression of [G]EBVf on [G]EBVr was performed using formula: $[G]EBV_f - \overline{[G]EBV_r} = b_0 + b_1([G]EBV_r - \overline{[G]EBV_r})$ (Legarra and Reverter, 2018), where b_0 is a mean difference between GEBVf and GEBVr, b_1 is a dispersion of the model, and R^2 of the model is correlation squared or the predictive ability of the model. These are again parameters that reflect population- and model- specific levels of accuracy (Legarra and Reverter, 2018).

2.5. Reliability approximation

Individual reliabilities were calculated for EBVs (PBLUP) and GEBVs (ssGBLUP). Reliability of EBVs were calculated using Tier and Meyer (2004) approach, while for GEBVs the multistep reverse reliability approximation approach described by Ben Zaabza et al. (2022) was used. Computations were done using the ApaX99 (Strandén and Lidauer, 1999) and MiX99 software. The multistep method was based on a separate calculation of reliabilities for genotyped and non-genotyped fish in the following steps: 1) reliabilities were estimated using PBLUP for all fish using Tier and Meyer (2004) approach; 2) effective record contributions (ERCs) were calculated using reverse reliability approach for genotyped fish; 3) reliabilities were estimated using GBLUP for genotyped fish; 4) ERCs were calculated for all the fish; 5) ERCs were corrected for double counting of information in the genotyped fish; 6) final reliabilities were calculated using weighting for ERC for all the fish. Benefit of the method is the avoidance of double counting of information in genotyped fish and the increased quality of prediction in non-genotyped fish due to genomic information from sibs.

3. Results

3.1. Validation study

The results of validation study for PBLUP and ssGBLUP models are presented in Table 4. In general, ssGBLUP had better prediction power than PBLUP. Accuracies obtained using *CrossV_Y** method were slightly higher for ssGBLUP in all the traits except $Weight_2$ and $Weight_3$. For $Weight_2$ accuracies were equal in PBLUP and ssGBLUP, and for $Weight_3$ 0.05 units lower in ssGBLUP. The bias (b_0) was lower in ssGBLUP model

Table 4
Validation bias (b_0), dispersion (b_1), accuracy (*acc*), and correlation squared (R^2) in pedigree-based (PBLUP) and single-step genomic (ssGBLUP) models obtained using different validation approaches.

Model and trait	*CrossV_Y** [a]			*CrossV_BV* [b]			*ForwardP_Y** [c]			*ForwardP_BV* [d]		
PBLUP	b_0	b_1	*acc*	b_0	b_1	R^2	b_0	b_1	*acc*	b_0	b_1	R^2
$Weight_2$	27	1.04	0.63	10	1.02	0.65	70	0.85	0.51	13	0.83	0.39
$Weight_3$	50	1.01	0.71	6	0.94	0.65	97	0.87	0.65	12	0.87	0.56
Sea $weight_2$	40	0.95	0.51	6	0.91	0.46	86	0.77	0.41	10	0.82	0.33
$Maturity_{female}$	0.05	0.91	0.53	0.01	1.02	0.65	0.04	1.00	0.56	0.01	0.93	0.48
$Maturity_{male}$	−0.04	0.58	0.46	−0.003	0.83	0.66	−0.04	0.24	0.56	0.01	0.61	0.69
Sea $maturity_{male}$	0.02	0.96	0.39	0.01	0.99	0.41	0.01	0.89	0.38	0.02	0.85	0.32
ssGBLUP												
$Weight_2$	10	1.04	0.63	9	1.02	0.74	89	0.84	0.47	11	0.88	0.48
$Weight_3$	69	0.96	0.66	−0.7	0.94	0.75	170	0.81	0.57	7	0.88	0.63
Sea $weight_2$	16	1.01	0.53	4	0.96	0.64	63	0.90	0.45	7	0.91	0.47
$Maturity_{female}$	0.07	1.06	0.60	0.001	1.07	0.79	0.05	1.12	0.60	−0.01	1.00	0.64
$Maturity_{male}$	−0.08	0.66	0.51	−0.004	0.92	0.84	−0.14	0.27	0.60	0.01	0.75	0.74
Sea $maturity_{male}$	0.01	0.97	0.42	0.001	0.99	0.66	−0.04	0.76	0.35	0.01	0.83	0.51

[a] ***CrossV_Y**** = 5-fold cross-validation with phenotype adjusted for the fixed effects and random tank effect used as TBV.
[b] ***CrossV_BV*** = 5-fold cross-validation with [G]EBV predicted using full phenotypic data used as TBV.
[c] ***ForwardP_Y**** = Forward prediction with phenotype adjusted for the fixed effects and random tank effect used as TBV.
[d] ***ForwardP_BV*** = Forward prediction with [G]EBV predicted using full phenotypic data used as TBV.

for all the traits except Weight$_3$. In both PBLUP and ssGBLUP models the dispersion (b_1) in different traits was close to 1.0 except lower Maturity$_{male}$ trait.

CrossV_BV method showed higher R^2 and lower bias for ssGBLUP model in all the traits, suggesting better predictive ability of the model (Table 4). When compared to PBLUP model, lower dispersion was observed in ssGBLUP model for Weight$_3$, Maturity$_{female}$, and Maturity$_{male}$ traits, but similar dispersion in the other traits.

*ForwardP_Y** method showed 0.04 units higher validation accuracy for ssGBLUP, compared to PBLUP, for Sea weight$_2$, Maturity$_{female}$, and Maturity$_{male}$ traits (Table 4). For Weight$_2$, Weight$_3$, and Sea maturity$_{male}$, accuracy was on 0.04, 0.08, and 0.03 higher in PBLUP model. Underdispersion was observed in both PBLUP and ssGBLUP models for Maturity$_{male}$, similar to what was observed in *CrossV _Y** method. In contrast, Maturity$_{female}$ trait showed noticeable overdispersion in ssGBLUP. In *ForwardP_BV,* ssGBLUP showed better prediction abilities in all the traits (Table 4).

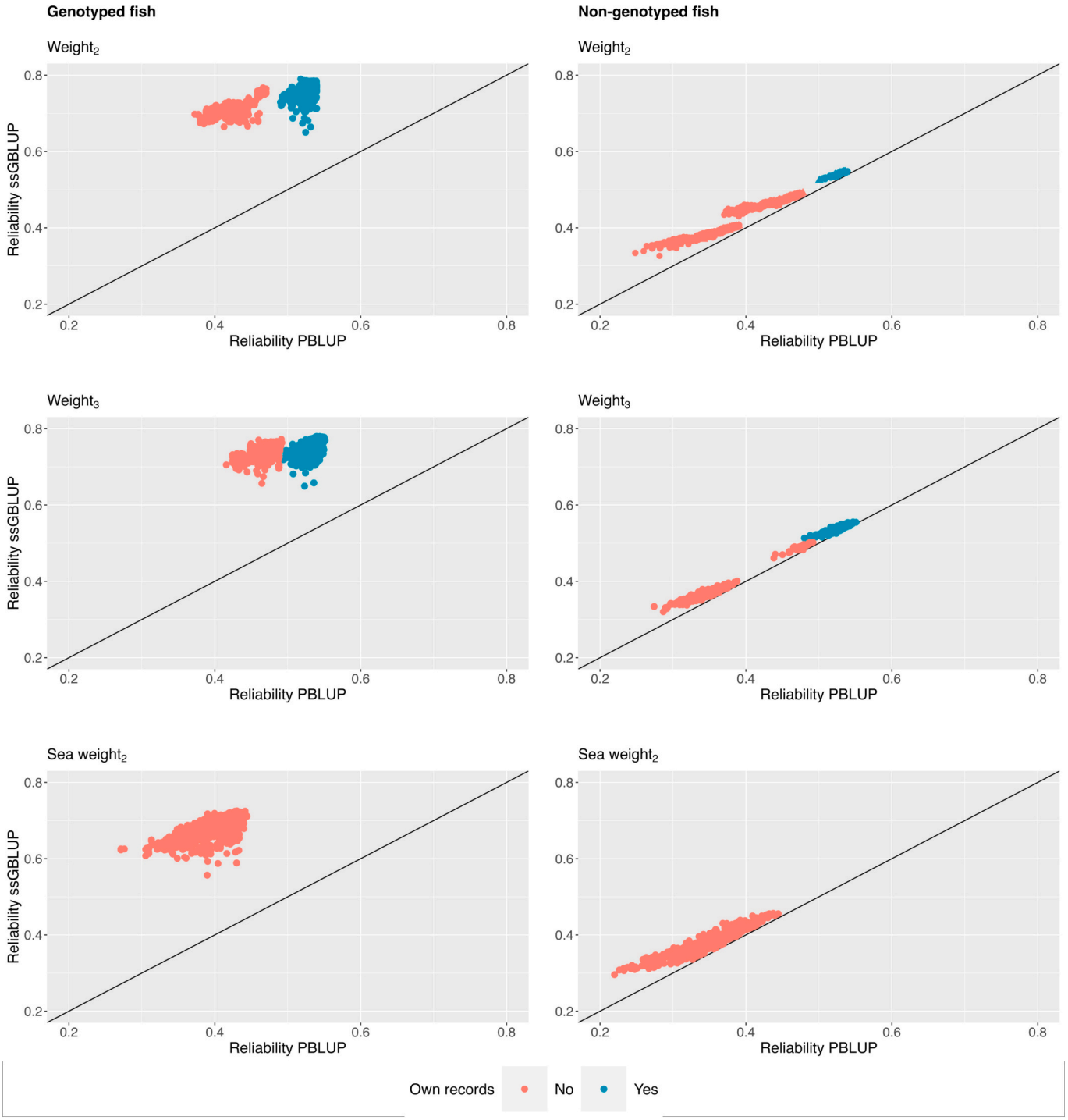


Fig. 1. Comparison of reliability obtained using pedigree-based (PBLUP) and single-step genomic (ssGBLUP) models for body weight traits of genotyped and non-genotyped animals at nucleus.

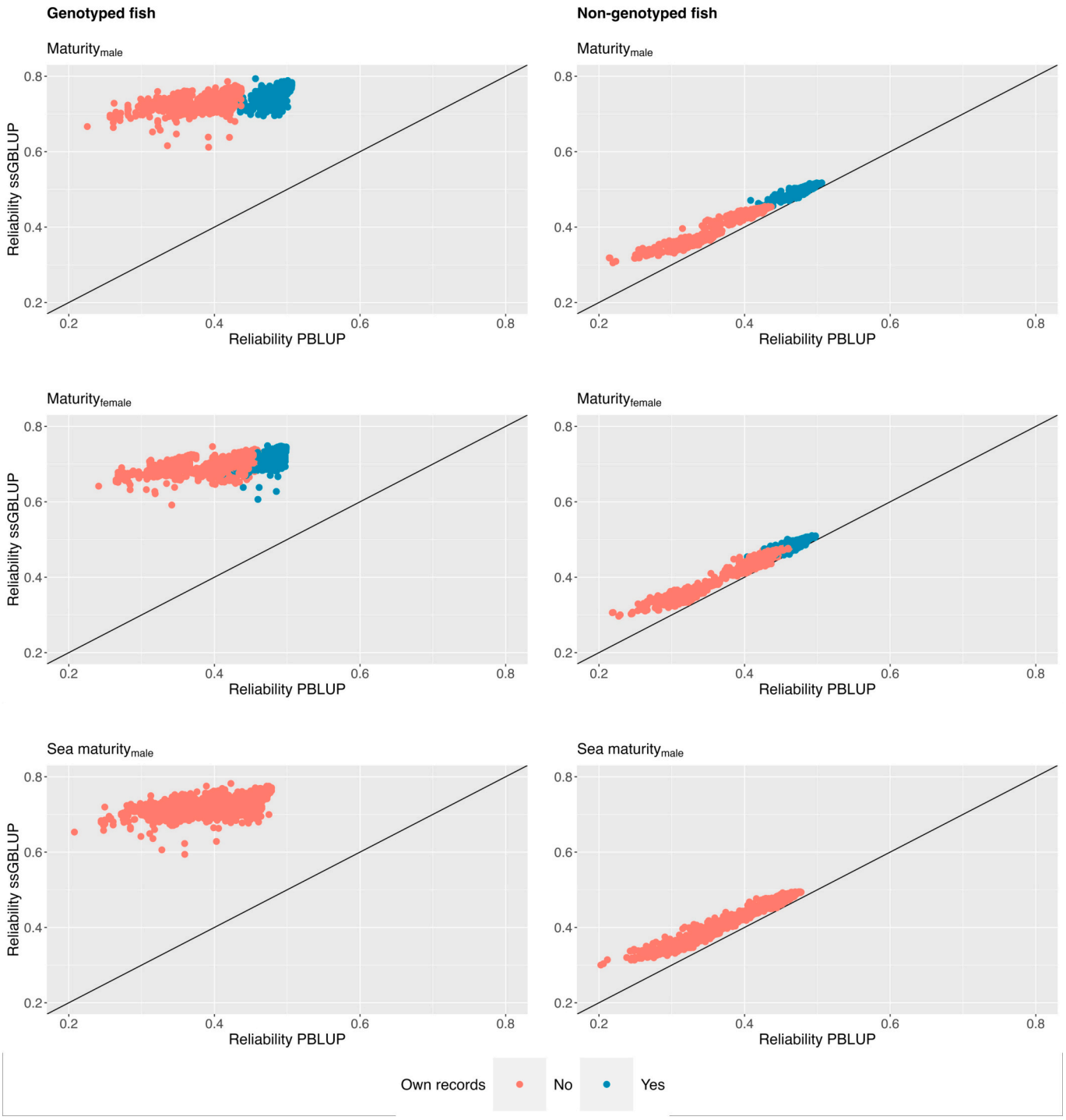


Fig. 2. Comparison of reliability obtained using pedigree-based (PBLUP) and single-step genomic (ssGBLUP) models for maturity traits of genotyped and non-genotyped animals at nucleus.

3.2. Result from reverse reliability estimation

The comparison of reliability of individual fish obtained through PBLUP and ssGBLUP for the body weight and maturity traits are presented in Figs. 1 and 2. Average reliability of prediction in genotyped individuals at the nucleus increased by 0.23, 0.23, 0.31, and 0.27 units for $Weight_2$, $Weight_3$, $Maturity_{male}$, and $Maturity_{female}$ traits, correspondingly. Average reliability of prediction for the sea recorded traits in the freshwater reared fish increased from 0.40 to 0.68 and from 0.40 to 0.73 in Sea $weight_2$ and Sea $maturity_{male}$ traits.

In Figs. 1 and 2, the fish were also divided into subgroups based on the presence or absence of own phenotypic data. For ungenotyped individuals at the nucleus, reliability of prediction was improved on average by only 0.03 units. Segregation of animals in ungenotyped part was explained by presence of genotyped sibs with and without own records.

4. Discussion

4.1. Single-step genomic BLUP

In the current study, single-step genomic evaluation (ssGBLUP) provided in general a higher population prediction accuracy and less biased results than the traditional pedigree-based PBLUP model. This implies that the marker information had a positive influence on the predictive ability of breeding values of body weight and maturity age in rainbow trout. When compared to pedigree based BLUP, single-step genomic evaluation showed moderate, little or no improvement in the accuracies in the phenotypic validation with *CrossV _Y** and *ForwardP_Y**, while the accuracies based on LR approach and reversed reliabilities showed much higher increases when single-step evaluation was applied. Single-step has been shown to be a powerful tool for GEBV prediction in agriculture species (Christensen et al., 2012; Legarra et al., 2014; Mäntysaari et al., 2020) and similar evidence is currently accumulation in aquaculture species (present study; Garcia et al., 2018, 2023).

Improvement was especially considered for sea recorded traits where phenotypes are not available for the breeding candidates in the nucleus. In freshwater fish, the average individual reliability of GEBV was 0.28 and 0.33 units higher than for EBV of Sea maturity$_{male}$ and Sea weight$_2$ traits. In ssGBLUP, breeding candidates reared in the nucleus benefit considerably from their own and their sea-reared sib's genotype information. Moreover, it should be noted that the accuracy in which the freshwater traits aid in the prediction of sea water traits, and vice versa, is impacted by the degree of genotype-by-environment interaction (GxE). In our data, the genetic correlation between sea and freshwater environment is 0.66 for body weight and 0.70 for male maturity (Appendix 1b). This moderate correlation implies significant GxE but simultaneously increases the accuracy of GEBVs of sea water traits estimated for the breeding candidates in the nucleus. On individual level, the prediction accuracy computed using reverse reliability approach was improved by at least 0.23 units for the freshwater traits. This is an important observation because this occurs even though these traits are well phenotyped on the freshwater breeding candidates at the nucleus. Prediction power was slightly improved also for non-genotyped fish.

In the Finnish breeding programme, full-sib families are held separated in family tanks until the fish are big enough for tagging. The maintenance of sire-dam-offspring pedigree for large number of fish is hence easy without any genotyping, but simultaneously genotyping and genomic selection can be integrated into this scheme. For instance, all families are routinely tested outside the freshwater nucleus for slaughter traits (Kause et al., 2007) and performance at commercial sea farms (Kause et al., 2003; current study), and can be recorded for disease resistance (Fraslin et al., 2022). These traits are likely to benefit the most from genomic evaluation (Houston et al., 2020; García-Ballesteros et al., 2022). The current study showed that genotyping of the proportion of fish in the programme increases the selection accuracy even for body weight and maturity status, i.e. traits that are already extensively phenotyped on both breeding candidates and their sibs reared outside the nucleus. Consequently, single-step genomic evaluation fits to such a breeding scheme well, and more added value is expected when applied to traits that cannot be recorded from breeding candidates.

Construction of **G** matrix can be performed using allele frequencies derived from the data (observed), estimated for a base population, or fixed to 0.5. In our study, **G** matrix was constructed with the assumption that average allele frequency was equal to 0.5. The use of base and observed allele frequencies was not suitable for the current population as the main batch of genotyped fish in our study represent only genetics of the recent years and are progeny of the same parental year class. Similarly, Garcia et al. (2023) did not use observed allele frequencies in the genomic evaluation of rainbow trout due to the high relatedness of the genotyped fish. The authors used base population allele frequencies but reported that as a potential source of bias in the validation studies.

4.2. Validation study

Reliable and convenient validation approach is important component of accessing prediction power of a breeding value evaluation. Development of validation methods for ssGBLUP has not received much attention in commercial aquaculture breeding programmes. We compared four validation tests used to assess prediction power of PBLUP and ssGBLUP models. The first method was modification of the most used in aquaculture (Garcia et al., 2018; Al-Tobasei et al., 2021; Song et al., 2022; Fraslin et al., 2022), the 5-fold cross-validation with corrected phenotype used as TBV (*CrossV_Y**). The modification was on masking phenotypes not only in the test animals but also in their full sibs. This helped to avoid overtraining of the model through the common random family tank effect and the relationship matrix. The closer the relationship between the fish in the training and test groups, the higher the accuracy in a validation study (Fraslin et al., 2022). We masked the sibs of the individuals in the reference group to ensure that the change in the accuracy is especially due to the genomic information.

The second approach was done by replacing the corrected phenotype in *CrossV_Y** by [G]EBV computed from the full data (*CrossV_BV*). Linear regression of GEBVf on GEBVr was performed like presented by Legarra and Reverter (2018). In general, ssGBLUP model showed better prediction ability then PBLUP model. However, surprisingly, *CrossV_Y** prediction results for Weight$_3$ trait were slightly better when using PBLUP. This may be explained by the low number of genotypes and records available for the trait.

As an alternative to *CrossV*, two forward prediction approaches were tested (*ForwardP_Y** and *ForwardP_BV*). Phenotypic data was masked from all fish born in 2019. It is important to note that in this approach, the year class 2018 (training set) and 2019 (test set) were created from the same parental year class, yet they do not share any individual sires or dams. In forward prediction, the fish in the test set were not directly influenced by full and half sib information of the training set as the whole year class was deleted. This should reduce, yet not totally remove, model overtraining. The observed accuracies in *ForwardP_BV* were in the range of 0.39 to 0.69 in PBLUP and 0.47 to 0.74 in ssGBLUP with average increase of 0.12. A similar approach was used by Fraslin et al. (2022) but in their study the year classes of Atlantic salmon were not closely related and hence the accuracies were close to zero. In our data, deletion of both year classes 2018 and 2019 was impossible as it would make reference population uninformative.

For freshwater body weight traits, ~0.10 and 0.08 units improvement in predictive ability of ssGBLUP was observed for the *CrossV_BV* and *ForwardP_BV* models, respectively. At the same time nearly no improvement was detected in *CrossV_Y** and *ForwardP_Y** methods. A similar pattern was observed for body weight in rainbow trout in the study by Garcia et al. (2023). Studies on residual carcass weight in channel catfish reported 0.07 units absolute prediction improvement for residual carcass weight using CrossV validation and Y* (Garcia et al., 2018).

It is a challenge to choose a proper validation model especially when genomic evaluation is being launched in a breeding programme. Both under- and overprediction of GEBVs are vitally important to know as wrong selection decisions will cause economical losses. Despite genomic prediction has been implemented already over a decade, the development of validation tests is still a fundamental topic in both large and small populations of livestock (Interbull, 2023). Two main concerns are what parameter should be used as response variable in validation regression model, and which group of animals should be used as the test set. Phenotypic validation (Tsai et al., 2015) that is common in aquaculture studies may not be an optimal tool for complex multigeneration data sets because pre-corrected phenotypes may bias validation due to small contemporary groups (Legarra and Reverter, 2018) and due to unclear behavior of binary traits. Hence results of bias and dispersion are

difficult to interpret. Use of the LR method is more justified for a small and complicated populations because the comparison is based on [G]EBV solutions from both full and reduced models. The method can be applied for binary traits as regression performed on two [G]EBV solutions (Leite et al., 2021). The comparison of [G]EBV from two sources is an elegant approach with understandable bias and dispersion values. Current data set should not suffer from selection bias as two-step genomic prediction not in the routine use and genotyping was performed mostly randomly. The LR method does not require 'true' breeding values to be known. However, it is good to remember that LR method can show high accuracy not because prediction power of GEBVr has increased but because both GEBVf and GEBVr were initially biased and alike. For example, in case where erroneous model taking too little advantage of own phenotype (due to a low heritability), both GEBVf and GEBVr are in the similar distance from true value.

Originally LR validation approach was presented as statistics describing the change in genomic predictions from old to newly developed evaluations. In this sense, the forward prediction (*ForwardP*) type of validation is a proper method to mimic consecutive evaluations. It can be expected that in *CrossV_BV* scenario, the reduced model will be well trained due to existing family links, and thus better prediction will be observed. However, we were not able to detect a large difference between the *CrossV_BV* and *ForwardP_BV* scenarios. This may be explained by the close relation between the training and the testing set, as generations 2018 and 2019 originate from the same parental generation born in 2014. Presumably there is no strait answer which method is the best for the model validation and both can be equally used to understand properties of the model.

5. Conclusions

Genomic prediction has been effectively integrated into the Finnish rainbow trout breeding programme. Based on our results, the ssGBLUP model is expected to have superior prediction accuracy compared to PBLUP. The validation of the ssGBLUP model through linear regression of GEBVs, computed from both the full and reduced data, is an appealing approach for complex data sets like ours.

Authors contributions

The experiment was planned by AK and AAK. The experiment performed by AK. Breeding programme maintenance, fish sampling, and collection of the data were performed by AN and HK. Manuscript was drafted by AAK and critically examined and revised by AK, AN, and HK. All authors have read and approved the final manuscript.

CRediT authorship contribution statement

Andrei A. Kudinov: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. **Heikki Koskinen:** Writing – review & editing, Methodology, Investigation, Data curation, Conceptualization. **Antti Kause:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Antti Kause reports financial support was provided by Horizon Europe.

Data availability

Data is not available as it belongs to Finnish National Breeding Programme.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 818367—'AquaIMPACT—Genomic and Nutritional Innovations for Genetically Superior Farmed Fish to Improve Efficiency in European Aquaculture', and from the Statutory Services of Natural Resources Institute Finland. The staff at Luke's Enonkoski fishfarm are thanked for extensive help with fish management. Professor Ismo Strandén (Luke) thanked for scientific advises.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aquaculture.2024.740677.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J., 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93, 743–752. https://doi.org/10.3168/jds.2009-2730.

Al-Tobasei, R., Ali, A., Garcia, A.L.S., Laurenco, D.A.L., Leeds, T., Salem, M., 2021. Genomic predictions for fillet yield and firmness in rainbow trout using reduced-density SNP panels. BMC Genomics 22, 92. https://doi.org/10.1186/s12864-021-07404-9.

Ben Zaabza, H., Taskinen, M., Mäntysaari, E.A., Pitkänen, T., Aamand, G.P., Strandén, I., 2022. Breeding value reliabilities for multiple-trait single-step genomic best linear unbiased predictor. J. Dairy Sci. 105, 6. https://doi.org/10.3168/jds.2021-21016.

Christensen, O.F., Lund, M.S., 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42, 2. https://doi.org/10.1186/1297-9686-42-2.

Christensen, O.F., Madsen, P., Nielsen, B., Ostersen, T., Su, G., 2012. Single-step methods for genomic evaluation in pigs. Animal. 6, 1565–1571. https://doi.org/10.1017/S1751731112000742.

Fraslin, C., Koskinen, H., Nousiainen, A., Houston, R.D., Kause, A., 2022a. Genome-wide association and genomic prediction of resistance to Flavobacterium columnare in a farmed rainbow trout population. Aquaculture. 557, 738332 https://doi.org/10.1016/j.aquaculture.2022.738332.

Fraslin, C., Yáñez, J.M., Robledo, D., Houston, R.D., 2022b. The impact of genetic relationship between training and validation populations on genomic prediction accuracy in Atlantic salmon. Aquac. Rep. 23, 101033 https://doi.org/10.1101/2021.09.14.460263.

Garcia, A.L.S., Bosworth, B., Waldbieser, G., Misztal, I., Tsuruta, S., Lourenco, D.A.L., 2018. Development of genomic predictions for harvest and carcass weight in channel catfish. Genet. Sel. Evol. 50, 66. https://doi.org/10.1186/s12711-018-0435-5.

Garcia, A.L.S., Tsuruta, S., Gao, G., Palti, Y., Lourenco, D.A.L., Leeds, T., 2023. Genomic selection models substantially improve the accuracy of genetic merit predictions for fillet yield and body weight in rainbow trout using a multi-trait model and multi-generation progeny testing. Genet. Sel. Evol. 55, 11. https://doi.org/10.1186/s12711-023-00782-6.

García-Ballesteros, S., Fernández, J., Kause, A., Villanueva, B., 2022. Predicted genetic gain for carcass yield in rainbow trout from indirect and genomic selection. Aquaculture 554, 738119. https://doi.org/10.1016/j.aquaculture.2022.738119.

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

Houston, R.D., Bean, T.P., Macqueen, D.J., Gundappa, M.K., Jin, Y.H., Jenkins, T.L., Selly, S.L.C., Martin, S.A.M., Stevens, J.R., Santos, E.M., Davie, A., Robledo, D., 2020. Harnessing genomics to fast-track genetic improvement in aquaculture. Nat. Rev. Genet. 21, 389–409. https://doi.org/10.1038/s41576-020-0227-y.

Interbull, 2023. https://interbull.org/static/web/Session_II.pdf.

Jairath, L., Dekkers, J.C., Schaeffer, L.R., Liu, Z., Burnside, E.B., Kolstad, B., 1998. Genetic evaluation for herd life in Canada. J. Dairy Sci. 81, 550–562. https://doi.org/10.3168/jds.S0022-0302(98)75607-3.

Janhunen, M., Kause, A., Vehviläinen, H., Nousiainen, A., Koskinen, H., 2014. Correcting within-family pre-selection in genetic evaluation of growth – a simulation study on rainbow trout. Aquaculture. 434, 220–226. https://doi.org/10.1016/j.aquaculture.2014.08.020.

Kachman, S.D., Spangler, M.L., Bennett, G.L., Hanford, K.J., Kuehn, L.A., Snelling, W.M., Thallman, R.M., Saatchi, M., Garrick, D.J., Schnabel, R.D., Taylor, J.F., Pollak, E.J., 2013. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. Genet. Sel. Evol. 45, 1297–9686. https://doi.org/10.1186/1297-9686-45-30.

Kause, A., Ritola, O., Paananen, T., Mäntysaari, E.A., Eskelinen, U., 2003. Selection against early maturity in farmed rainbow trout: the quantitative genetics of sexual dimorphism and genotype-by-environment interactions. Aquaculture. 228, 53–68. https://doi.org/10.1016/S0044-8486(03)00244-8.

Kause, A., Ritola, O., Paananen, T., Wahlroos, H., Mäntysaari, E.A., 2005. Genetic trends in growth, sexual maturity and skeletal deformations, and rate of inbreeding in a

breeding programme for rainbow trout. Aquaculture. 247, 177–187. https://doi.org/10.1016/j.aquaculture.2005.02.023.
Kause, A., Paananen, T., Ritola, O., Koskinen, H., 2007. Direct and indirect selection of visceral lipid weight, fillet weight and fillet percent in a rainbow trout breeding programme. J. Anim. Sci. 85, 3218–3227. https://doi.org/10.2527/jas.2007-0332.
Kause, A., Nousiainen, A., Koskinen, H., 2022. Improvement in feed efficiency and reduction in nutrient loading from rainbow trout farms: the role of selective breeding. J. Anim. Sci. 100, 8. https://doi.org/10.1093/jas/skac214.
Legarra, A., Reverter, A., 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Genet. Sel. Evol. 50, 53. https://doi.org/10.1186/s12711-018-0426-6.
Legarra, A., Aguilar, I., Misztal, I., 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92, 4656–4663. https://doi.org/10.3168/jds.2009-2061.
Legarra, A., Misztal, I., Aguilar, I., 2014. Single step methods with a view towards poultry breeding. In: 10. World Congress of Genetics Applied to Livestock Production (WCGALP), Aug 2014, Vancouver, Canada.
Leite, N.G., Knol, E.F., Garcia, A.L.S., Lopes, M.S., Zak, L., Tsuruta, S., Silva, F.F.E., Lourenco, D., 2021. Investigating pig survival in different production phases using genomic models. J. Anim. Sci. 99, 8. https://doi.org/10.1093/jas/skab217.
Mäntysaari, E.A., Liu, Z., VanRaden, P.M., 2010. Interbull validation test for genomic evaluations. Interbull Bull. 41, 17–22.
Mäntysaari, E.A., Koivula, M., Strandén, I., 2020. Symposium review: single-step genomic evaluations in dairy cattle. J. Dairy Sci. 103, 5314–5326. https://doi.org/10.3168/jds.2019-17754.
Martinez, V., Kause, A., Mäntysaari, E.A., Mäki-Tanila, A., 2006. The use of alternative breeding schemes to enhance genetic improvement in rainbow trout: II. Two-stage selection. Aquaculture. 254, 195–202. https://doi.org/10.1016/j.aquaculture.2005.11.011.
Meuwissen, T.H., 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.
Misztal, I., Vitezica, Z.G., Legarra, A., Aguilar, I., Swan, A.A., 2013. Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed. Genet. 130, 253–258. https://doi.org/10.1111/jbg.12025.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. https://doi.org/10.1086/519795.
Song, H., Dong, T., Hu, M., Yan, X., Xu, S., Hu, H., 2022. First single-step genomic prediction and genome-wide association for body weight in Russian sturgeon (*Acipenser gueldenstaedtii*). Aquaculture. 561 https://doi.org/10.1016/j.aquaculture.2022.738713.
Strandén, I., Lidauer, M., 1999. Solving large mixed models using preconditioned conjugate gradient iteration. J. Dairy Sci. 82, 2779–2787. https://doi.org/10.3168/jds.S0022-0302(99)75535-9.
Strandén, I., Mäntysaari, E.A., 2010. A recipe for multiple trait deregression. Interbull Bull 42, 21–24.
Strandén, I., Mäntysaari, E.A., 2018. HGinv Program. Natural Resources Institute Finland (LUKE).
Strandén, I., Vuori, K., 2006. RelaX2: pedigree analysis program. In: Proceedings of the 8th World Congress on Genetics Applied to Livestock Production: 13-18 August 2006; Belo Horizonte, MG, Brazil, pp. 27–30.
Tier, B., Meyer, K., 2004. Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. J. Anim. Breed. Genet. 121, 77–89. https://doi.org/10.1111/j.1439-0388.2003.00444.x.
Tsai, H.Y., Hamilton, A., Tinch, A.E., Guy, D.R., Gharbi, K., Steer, M.J., Matika, O., Bishop, S.C., Houston, R.D., 2015. Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. BMC Genomics 16, 969. https://doi.org/10.1186/s12864-015-2117-9.