Accuracy of genomic prediction of host resistance to salmon lice in Atlantic salmon (*Salmo salar*) using imputed high-density genotypes

M. H. Kjetså[1], J. Ødegård[2], T. H. E. Meuwissen[1]

[1]Norwegian University of Life Sciences, Faculty of Biosciences, PO Box 5003, 1432 Ås, Norway.

[2]Breeding and Genetics, AquaGen, PO Box 1240, 7462 Trondheim, Norway

*corresponding author: maria.kjetsa@nmbu.no

Abstract

Salmon lice (*Lepeophtheirus salmonis*) is a marine ectoparasite responsible for major losses to the salmon farming industry each year. Salmonids are the primary hosts of the parasite, including the widely farmed species Atlantic salmon (*Salmo salar*) and rainbow trout (*Oncorhynchus mykiss*). Improving resistance towards the parasite in farmed Atlantic salmon could decrease the need for treatments, increase the welfare of the fish, as well as reduce the infection pressure on wild populations. Phenotypic resistance can be recorded in controlled challenge-tests and has been found to be moderately heritable. The aim of the study was to compare three different genomic selection models with respect to within- and across-family prediction accuracy with both moderate and high SNP-chip densities (215K and imputed 750K). The models tested were: Genomic Best Linear Unbiased Prediction

(GBLUP), BayesC and a model combining a polygenic term and a BayesC term (BayesGC). Predictive abilities of the models were compared using five-fold cross-validation.

The trait was found to be highly polygenic. All three models had a similar predictive ability. The BayesGC model had a slight advantage over the GBLUP and BayesC models, however this difference was not significant. For within-family prediction there was no advantage from increasing the SNP density from 215K to 750K genotype density. However, for across-family prediction a slight improvement in predictive ability was observed at the higher density compared to the lower.

Keywords

Genomic prediction, Atlantic salmon, salmon lice, imputation, prediction accuracy

1. Introduction

Genomic Prediction (GP) is being adopted in the fields of plant, animal and aquaculture breeding and human genetics. GP links data on individual phenotypes with genomic data from genome-wide dense marker maps, using a reference population of both genotyped- and phenotyped individuals to predict a population with only genotyped individuals (Meuwissen, Hayes & Goddard, 2001). The accuracy of GP is dependent on the heritability of the trait, the size and quality of the reference population and the genetic relationships between the reference population and the predicted population (Calus & Veerkamp, 2007; Meuwissen, Hayes & Goddard, 2001).

Salmon louse (*Lepeophtheirus salmonis*) is a naturally occurring ectoparasitic copepod that is found on most salmonid species in the *Salmo*, *Onchorhynchus* and *Salvelinus* genera, such as Atlantic salmon (*Salmo salar*), Sea trout (*Salmo trutta*), Pink salmon (*Oncorhynchus gorbuscha)* and Rainbow trout (*Onchorhynchu mykiss*) (Torrissen et al., 2013). The parasite causes large welfare- and economic problems for the Atlantic salmon and rainbow trout farming industries. In 2011, the losses due to the parasite in the Norwegian fish farming industry were estimated to 436 million US dollars (Abolofia et al., 2017), and the losses have increased markedly since then (Overton et al., 2018). The parasite also poses a threat to wild populations, as salmon louse copepods from farmed fish may infect wild salmonids. To reduce impact on wild stocks, treatment of farmed fish is mandatory at low infestation levels in Norway. The treatment costs, rather than damages caused by the parasite itself, are the major problems for the industry. Treatments are performed frequently, have high mortality rates, and cause stress for the fish. In addition, salmon lice are developing resistance to some of the drugs used for treatment (Overton et al., 2018). The effects of salmon lice infestations from fish farms to wild salmon population are hard to quantify but there are definitely sizable negative effects to wild stocks (Torrissen et al., 2013).

Genetic variability in host-resistance to *Lepeophtheirus salmonis* is found in multiple studies (e.g. Gjerde, Ødegård & Thorland, 2011), (H. Y. Tsai et al., 2016) & (Ødegård et al., 2014). The heritability estimates of the trait depend on the recording conditions. In a natural disease outbreak, the heritability estimates range between 0.02±0.02 and 0.14±0.02 (Kolstad et al., 2005). For challenge tests in sea cages the estimates are around 0.14±0.03 (Ødegård et al., 2014), and for challenge tests in land-based tank conditions a heritability of 0.33±0.05 is found (Gjerde et al., 2011). There are also naturally differences in the

susceptibility of different salmonid species, seen especially in the Pacific salmons (*Oncorhynchus spp.)* where the Coho- *(Oncorhynchus kisutch)* and Pink salmon *(Oncorhynchus gorbuscha)* reject the lice more rapidly than the Chinook (*Oncorhynchus tshawytscha*) (Torrissen et al., 2013).

Selective breeding for disease resistance is often dependent on challenge tests performed on siblings for phenotypic data. It can also be performed on disease data collected in the field environment. For challenge tests, the tested individuals are, due to regulative restrictions, excluded as selection candidates when tested fish are not allowed to re-enter the breeding nucleus after being exposed to potential pathogens. Estimates of Breeding Values (EBVs) are predicted for the elite breeding candidates based on the information from their challenge tested full sibs. Because the EBVs are predicted for animals without phenotype data, prediction is mainly based on family information (full- and half-sib). This implies that only the between family component of the EBV can be predicted by traditional Best Linear Unbiased Prediction (BLUP), which reduces both the intensity of selection and the accuracy because there is no information on the within family deviation, which encompasses half of the genetic variation (Gjerde et al., 2011).

When using genomic data and genomic selection, within family deviations can be predicted based on the DNA data (Sonesson and Meuwissen, 2009), and this increases the prediction accuracy as more of the genetic variation can be explained. Ødegård et al. (2014) found that using genomic prediction methods gave a higher reliability than using only pedigree information. However, Sonesson & Meuwissen (2009) found in their simulation study that the accuracy of selection dropped when the challenge test was done only every other

92 generation or only in one generation when using the GBLUP method. This implies that it
93 would be necessary to challenge test every generation to get accurate predictions.
94
95 The accuracy of genomic predictions increases with the number of phenotypes relative to
96 the effective number of genomic segments of the population (Daetwyler et al., 2010).
97 Bayesian variable selection methods (Meuwissen et al., 2001; Verbyla, Bowman, Hayes, &
98 Goddard, 2010) attempt to increase the relative weight of markers being in LD with casual
99 mutation and remove markers that are not linked to causal loci (i.e., not useful for
100 prediction), and thereby reduce the number of marker effects to estimate.
101
102 Bayesian selection approaches such as Bayes (A/B/C/R) have been found to have a higher
103 predictive ability in simulation studies, but differences were smaller in studies using real
104 data (Neves et al., 2012). One of the biggest differences between the Bayesian methods and
105 GBLUP is that GBLUP assumes that genetic variance is evenly distributed over SNPs, whilst
106 the Bayesian methods try to differentiate SNPs with respect to their relative importance. In
107 the current study we investigate the BayesC (Habier et al., 2011), and BayesGC models
108 (Iheshiulor et al., 2017). In BayesGC, a polygenic effect and a Bayesian term are fitted
109 simultaneously, so that we account for both numerous SNPs of small effect, as well as a
110 smaller group of SNPs with a potentially larger effect. In contrast to Iheshiulor et al. (2017),
111 who used an iterative conditional expectation (ICE) algorithm for the BayesGC model, we
112 fitted this model using a Gibbs-sampling approach.
113
114 The aim of this study was to compare three methods of genomic prediction: Genomic Best
115 Linear Unbiased Prediction (GBLUP), using a genomic relationship matrix, two Bayesian

variable selection methods BayesGC (Meuwissen et al., 2020) and BayesC for the trait host resistance to salmon lice in Atlantic salmon, measured as number of lice per fish. Furthermore, prediction accuracies of the GEBVs based on a 215K SNP genotypes and imputed 750K SNP panels were compared using both within-family and across-family prediction scenarios.

2. Methods

The data came from an admixed population of Atlantic salmon (*S. salar*) that were genotyped and challenge tested for susceptibility to *L. Salmonis.* The challenge test was conducted by adding *L. salmonis* in the water of sea-net cages closed off with tarpaulins. After 10-15 days the number of lice were manually counted. The fish were from the 2011 year-class from the AquaGen population as described in (Ødegård et al., 2014). The total number of challenge-tested fish was 2850 from the test conducted in the period July 16-18, 2012. The challenge test is thoroughly described in (Ødegård et al., 2014) and was approved by the Norwegian Animal Research Authority (S-2012/148773).

From the challenge-tested fish, 1385 fish were genotyped and their data was used here. The 1385 phenotyped- and genotyped fish belonged to 99 full-sib families and were offspring from 68 sires and 69 dams. The smallest family consisted of 7 individuals and the largest 21 with a mean size of 14. Lice resistance was recorded as the number of lice counted from each fish (LC). However, this trait was highly skewed and thus the trait was log-transformed and called logLC (Ødegård et al., 2014).

All 1385 fish were genotyped with a 220K Affymetrix genome-wide SNP-chip. The total number of SNPs after quality control was 215610. A group of parents (n = 59) was genotyped with a high-density SNP-chip with 990K SNPs from a custom SNP-chip used by AquaGen. After quality control there was a total 745,998 SNPs remaining.

Our 1385 phenotyped and genotyped fish were imputed to 750K using the FImpute software (Sargolzaei et al., 2014). FImpute is a rule-based, deterministic method for genotype imputation and phasing (Wang et al., 2016). The parental fish had not been challenge-tested, and were only used as reference animals for the imputation and phasing.

Both the original 215K and the 750K imputed genotypes were used to construct two genomic relationship matrices (**G**-matrix; one using 215K and one using 750K), using own software based on VanRaden method 1 (VanRaden, 2008);

$$\boldsymbol{G} = \frac{MM'}{2\sum p_j(1-p_j)},\ M_{ij} = x_{ij} - 2p_j$$

where x_{ij} is the genotype of fish *i* for SNP *j*, with x_{ij}= 0,1 or 2 for the reference homozygote, heterozygote and opposite homozygote, respectively, and p_j is the allele frequency of the alternative allele of SNP *j* for all fish. The **G**-matrices were then used in the genomic predictions described below.

2.1 Calculation of Yield Deviations

LogLC was corrected for fixed effects by calculating Yield Deviations (YD), since the Bayesian variable selection approach models used here could not handle complicated modelling of fixed effects. The model was:

y = **Xb** + **Zu** + **e**

where **y** is a vector of logLC phenotypes, **b** is a vector of fixed effect of overall mean, person counting the lice, the day of count, and a fixed regression on the weight of the fish measured on the day of the count (correcting for the fact that bigger fish may contain more lice due to a larger surface area). **Z** is a design matrix linking individuals to the phenotype. **u** is the random effect of the individual fish (**u**~N(0,**A**σ_a^2) where **A** is the pedigree relationship matrix; **e** is the residual effect, where (**e**~N(0,**I**σ_e^2), where **I** is an identity matrix. This model was analyzed using DMU (Madsen and Jensen, 2013). The DMUAI module was used to estimate the variance components and the DMU4 model to produce individual Yield Deviations (YD) that were used in the further analysis.

2.2 GBLUP

The YD were first analysed by the GBLUP model:

$$\mathbf{YD} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{e}$$

Where **YD** is a vector of the Yield Deviation of LogLC, μ = overall mean, **Z** = design matrix linking individuals to the YD, **u** = vector of random effects of the individual fish (***u***~N(0,**G**σ_u^2), where **G** is the genomic relationship matrix, and **e** = vector of random residuals with variance **e** ~N(0, $\mathbf{I}\sigma_e^2$) and Identity matrix **I**.

2.3 BayesC

The model for BayesC (Habier et al., 2011) was as follows:

$$\mathbf{YD} = \mathbf{1}\mu + \sum_i \mathrm{I_i}\mathbf{X_i}\mathrm{s_i} + \mathbf{e}$$

where YD = Yield Deviation, 1 is a vector of ones, μ is overall mean, $\mathbf{X}_i$ is a vector of genotypes for SNP *i* containing 0 for homozygote individuals, 1 for heterozygotes, and 2 for

185 the alternative homozygote genotype. I_i is an indicator of whether the SNP *i* is in the model
186 in a particular MCMC-cycle or not (0/1). s_i is the SNP effect, where if the SNP *i* is in the
187 model: s_i ~N(0, σ^2_m) and **e** is the residual with variance **e** ~N(0, $\mathbf{I}\sigma^2_e$) where **I** is an identity
188 matrix. The MCMC – chain was run for 20 000 Gibbs-cycles using 4000 burn-in cycles, in two
189 distinct chains. The prior probability of I_i = 1 is π. If the SNP *i* is in the model: s_i ~N(0,
190 σ^2_u/1000). ***e*** ~N(0, $\mathbf{I}\sigma^2_e$) and $\mathbf{I}$ is an identity matrix.
191

192 ## 2.4 BayesGC

193 The BayesGC model fits a polygenic effect and a BayesC term simultaneously. The polygenic
194 effect is fitted using the genomic relationship matrix (**G**) as in the GBLUP model. The BayesC
195 term assumes SNPs to have normally distributed effects with probability (π) or an effect of 0
196 with probability (1-π). The BayesC method is the same as the one used in (Iheshiulor et al.,
197 2017), except that we use a Monte Carlo Markov Chain (MCMC) algorithm for estimation of
198 SNP effects and the polygenic effect whereas they use an iterative conditional expectation
199 (ICE) algorithm to approximate the results from such an MCMC analysis.
200

201 Here we describe how the total genetic variance σ^2_u is partitioned over the fitted SNPs and
202 the polygenic effect. For the Bayes C method;

203 $\sigma^2_m = \frac{Fr * \sigma^2_u}{\overline{HET}}$

204 Where σ^2_m is the genetic variance explained by a single SNP,

205 Fr = the fraction of the total genetic variance explained by a single fitted SNP, i.e. 1/1000
206 because we assume each SNP explain 1/1000th of the genetic variance.

207 $\overline{HET}$ = average heterozygosity = $\frac{2\sum p_i\,(1-p_i)}{N_{loci}}$

For a Bayes C model, this would mean using prior probability of fitting a SNP of:

$$\pi_c = \frac{1000}{N_{loci}}$$

Such that $\sigma_u^2 = \pi_c \cdot N_{loci} \cdot \overline{HET} \cdot \sigma_m^2$

For the BayesGC method we both have a polygenic effect and fitted SNP effects. Again, we also assume that each fitted SNP explains 0.1% of the total genetic variance.

In addition, the total genetic variance σ_u^2 should not be affected by the partitioning of the variance across the SNPs and the polygenic effect. Let q be the fraction of σ_u^2 explained by SNPs, then the variance explained by the polygenic effect is σ_{pol}^2 = (1-q) σ_u^2. Hence,

$$\sigma_u^2 = \sigma_{pol}^2 + q \cdot \pi \cdot loci \cdot \overline{HET} \cdot \sigma_m^2$$

It follows that:

$$\pi_{gc} = q * \pi_c$$

Where π_{gc} is the π value used for the BayesGC model. Four different values of *q* were tested for BayesGC, *q* = 0.05, 0.25, 0.5 and 0.75 corresponding to SNPs explaining 5%, 25%, 50% and 75% of the total genetic variance (denoted BayesGC_05, BayesGC_25, BayesGC_50, BayesGC_75, respectively).

The BayesGC model is thus as follows:

$$\mathbf{YD} = \mathbf{1}\mu + \mathbf{Zu} + \sum_i \mathrm{I}_i \mathbf{X}_i \mathrm{s}_i + \mathbf{e}$$

where **YD** is a vector of the Yield Deviations of LogLC, **1** is a vector of ones, μ is overall mean, **Z** is a design matrix that links individuals to the YD, ***u*** = random polygenic effect with variance V(***u***) = $\mathbf{G}\sigma_{\mathrm{pol}}^2$. $\mathbf{X_i}$ = vector of genotypes for SNP i containing 0 for homozygote individuals, 1 for heterozygots, and 2 for the alternative homozygote genotype. I_i is an indicator of whether SNP *i* is in the model in a MCMC-cycle or not (0/1) and the prior

probability of I_i = 1 is π. s_i is the SNP effect, where if the SNP *i* is in the model: s_i ~N(0, σ^2_m). **e** is the residual with variance **e** ~N(0, **I**σ^2_e) where **I** is an identity matrix. The MCMC – chain was run for 4000 burn-in cycles and a total of 20000 Gibbs-cycles. The EBVs from the two Gibbs-chains had a correlation of >0.9999 and thus the EBVs were assumed to be converged, and the results presented for both BayesC and BayesGC is the average of two Gibbs-chains.

2.5 Cross Validation

We compared the three methods of genomic prediction for their predictive ability obtained from a 5-fold-crossvalidation design. There were two alternative scenarios (see below) and all models and scenarios were analyzed using two different SNP densities (215K and imputed 750K). The cross-validation for each scenario was performed by randomly splitting the data set (with some restrictions depending on the scenario; see below) into five separate subsets. In each "fold" the phenotypes of the corresponding data set were set to missing (masked), while phenotypes of the remaining four subsets were included in the analysis. This way the animals with phenotype included was set as the reference population (training-set) and the animals with missing phenotype were used as a validation population whose phenotypes were predicted (validation-set). Each fish was once included in the validation set over the five folds, i.e. there was no overlap between the validation sets. There were six replications of the five-fold cross-validation. Each five-fold cross-validation produced two Gibbs-chains and thus the results within each replicate is the result of two Gibbs-chains and the results shown is the average of these chains over the six replicates.

We analyzed two different cross-validation scenarios:

Within-family scenario: Evenly distributing the fish within each full-sib group across the five subsets, so all fish have full-sibs in the training data when its own phenotype is masked. *Across-family scenario:* Entire full-sib families are allocated at random to one of the subsets, masking entire families at the same time. Half-siblings may still be present in training and validation sets. The analysis (either BayesC, GBLUP or BayesGC) was then performed for each fold and we extracted the GEBVs from the animals whose records were masked (the records of each individual were masked in one of the 5 folds). The accuracy of prediction was estimated as:

$$r_{pred} = \frac{cor(\text{GEBV}, YD)}{\sqrt{h^2}}$$

Where h^2 is estimated using a pedigree-based model.

2.6 Significance test

To test the models for significant differences in prediction accuracy we used a bootstrapping procedure (Efron, B. Tibishirani, 1994) to test the correlation between GEBV and YD in each model following (Iversen et al., 2019). Two models at a time were compared to find which predicted the YDs best by randomly bootstrap sampling data points triplets (EBVs for each of the two models and the corresponding YD) with replacement. 10,000 bootstrap samples were constructed for each pairwise comparison. We determined which model yielded a higher correlation with the YD for each bootstrap sample. The models were considered significantly different if one of the models had a higher correlation in at least 97.5% of the bootstrap samples (equals a p-value of 5% due to the two-sidedness of the test).

3. Results

The estimates of the variance components of LogLC were $\sigma_e^2 = 0.414$ and $\sigma_u^2 = 0.069$ resulting in a heritability of $h^2 = 0.14$ estimated using the pedigree relationship matrix. For the 215K SNP-chip and the within-family scenario (Table 1) the highest prediction accuracy was 0.675 which was achieved by BayesGC_05 and BayesGC_25. The accuracy of GBLUP and BayesC was 0.671 and 0.672 respectively.

In the 215K SNPchip and across-family scenario (Table2), the highest prediction accuracy was for BayesGC_05 at 0.602 Followed by BayesGC_25 and BayesGC_50 with an accuracy of 0.601. BayesC and GBLUP followed at 0.599 and 0.596 respectively. There were no significant differences between any of the models using 215K genotypes neither within- nor across-family. For the 750K SNPchip and within-family scenario (Table 3). BayesGC_25 had the highest accuracy of 0.673 followed by BayesGC_05 with an accuracy of 0.673. GBLUP and BayesC had an accuracy of 0.669 and 0.670 respectively. The differences between the methods were not significant in the within-family scenario. For the 750K across-family scenario (Table 4), the highest accuracy was obtained from BayesC and BayesGC_75 with an accuracy of 0.611. GBLUP had an accuracy of 0.607 and BayesGC_05 and BayesGC_50 had an accuracy of 0.605, but none of the differences were statistically significant.

Increasing genotype density from 215K to 750K within family (Tables 1 and 3) had no effect on the accuracy of prediction. However, between the 215K and 750K genotype densities for the across family scenarios (Tables 2 and 4), we can see a slightly higher accuracy all of the methods. For GBLUP: 0.596 versus 0.607, for BayesGC_05: 0.602 versus 0.605, for BayesGC_25 0.601 versus 0.610 and for BayesC 0.599 versus 0.611 using genotype densities 215K and 750K respectively. However, there were no significant differences in prediction accuracy between different genotype densities in the across family scenario.

3.1 Regression coefficient

The slopes for the within-family scenarios are 1.1 and for the across-family the slope is 1.2. There were no differences in estimates of the slopes between the methods. A too high slope indicates that the spread of the EBVs is too small. Possibly the estimated genetic variance is too small. The estimated variance is based on a pedigree relationship matrix, while we are using a genomic relationship matrix in our predictions.

3.2 Posterior probabilities

A brief analysis of our posterior probabilities was conducted (Appendix A), and no SNPs with posterior probability higher than 0.02 were detected. Hence, we could not detect any QTLs for the trait, but there was some regions with elevated posterior probabilities, which might indicate that some regions are more associated with the trait than others.

4. Discussion

The accuracy of genomic predictions of host resistance to salmon lice (*Lepeophtheirus salmonis*) was substantial and varied between 0.59-0.68. Within-family predictions yielded higher accuracies than across-family predictions. This was expected as there will be a higher genetic relationship between the test- and training animals in the within-family prediction scenario, and a higher genetic relationship between test- and training set is often connected

to a higher prediction reliability (Wu et al., 2015). Although the across-family scenario does not contain full-sibs in a training set for any animals in the validation set, half-sibs may still be present, and so the relationship between animals in the across-family scenario is lower than for the within-family, but cannot be regarded as very distant. It would be interesting to see if there is a larger difference between the models when the relationship between the animals in a training set and test set is more distant, as the predictions would need to rely more on the LD between markers and not so much the family relationships Unfortunately, the family structure of our data does not allow to test at lower genetic relationships.

Sonesson (2007) studied the decay of prediction accuracy as the relationship between the reference population in a sib-testing scheme decreases over generations. Within a generation, the markers that only explain family effects could be used for the prediction of family means, whereas across generations, the family effects decay and the SNPs that explain the trait variance become more important. Hence, higher SNP density and accounting for single SNP effects in BayesGC is expected to become more important at more distant genetic relationships between training and validation sets.

The main differences between the three models in our study lie in how they model the genetic variance of the SNPs. The GBLUP method explains the variance by assuming all SNPs have an equal variance, and all SNPs are fitted jointly through the G-matrix. The BayesC model assumes that the genetic variance is explained by a relatively small fraction of the SNPs and fits those SNPs explicitly in the model. BayesGC fits all SNPs through the G-matrix, and at the same time fits a few SNPs that explain substantially more genetic variance than the others. The different BayesGC versions differentiate in how the total genetic variance is

348 divided between the G-matrix or the SNP-markers. This is one of the reasons we had hoped
349 to see a bigger difference between the models for the across-family prediction scenario.
350
351 Other studies showed promising results for a BayesGC type of method. Solberg, Sonesson,
352 Woolliams, Odegard, & Meuwissen (2009) fit a polygenic effect using pedigree information
353 and the Bayes B method from Meuwissen, Hayes, & Goddard (2001) to fit SNP effects. They
354 conclude that fitting a polygenic effect has a small impact on the accuracy of genome-wide
355 EBVs in the generation immediately following phenotyping, but as the generations progress,
356 the predictions with a polygenic effect retain a higher accuracy, and that this persistence in
357 accuracy is significant for higher marker densities. Calus & Veerkamp (2007) found an
358 increase in the prediction accuracy when including a polygenic effect when the SNP density
359 and heritability was high. Calus et al. did not predict over generations and generally had a
360 smaller genome size and lower marker densities than Solberg et al., (2009). Hence, it is
361 expected that including a BayesC and polygenic term increases prediction accuracies,
362 especially as the genetic relationships between the training and evaluation animals
363 decrease. However, both these studies are simulation studies. We found from our study
364 with real data, that there was no significant difference between our models in the across-
365 family scenario compared to the within-family scenario at either genotypic densities.
366
367 Ma et al. (2019) found that using a Bayesian model including known QTLs increased the
368 reliability of prediction accuracy regardless of the genetic distance between the reference
369 population and the predicted population. They found that the Bayesian methods had a
370 larger advantage for traits linked to major genes such as milk yield and fat compared to
371 fertility and mastitis that had almost no effect. They also saw that a small reference

372 population (<1000 individuals) could affect the reliability of the prediction. As we have both
373 a relatively small reference population (~1000 individuals) in addition to a highly polygenic
374 trait, this might have had an impact on why the Bayesian methods did not outperform
375 GBLUP.
376
377 Iheshiulor et al. (2017) compared the Bayes GC method with GBLUP and BayesC on real
378 data from cattle. Their BayesGC method used an iterative conditional expectation (ICE)
379 algorithm to fit their BayesC term while we used a Gibbs sampling algorithm. They found
380 that the BayesGC performed marginally better than GBLUP and BayesC for all their traits
381 and for one trait the difference was significant. Iheshiulor et al (2017) finds that BayesC
382 always performs between GBLUP and BayesGC. Our results showed that the BayesC method
383 performed either the same or worse than BayesGC and the same or slightly better than
384 GBLUP. In other words, the BayesC term did not add prediction accuracy compared with the
385 GBLUP model, which may explain why the BayesGC model did not have an advantage over
386 GBLUP. Moreover, the performance of the Bayesian methods may be affected by the
387 assumption that each SNP explains 0.1% of the genetic variance, which limits the number of
388 SNPs fitted. However, fitting more SNPs would make the use of fitting both a polygenic trait
389 and a Bayes C term redundant, as fitting many small SNPs would be practically the same as
390 fitting polygenic effects. On the other hand, fitting fewer and larger SNPs would not agree
391 with the polygenic nature of the trait. We did, however, test different assumptions for the
392 BayesC method, assuming that each SNP explain $\frac{1}{500}$, $\frac{1}{2000}$ and $\frac{1}{10000}$ of genetic variance.
393 None of these assumptions yielded a significantly different accuracy for the BayesC
394 prediction accuracy and thus the results were not included here.
395

Increasing marked densities increased the accuracy slightly for across-family prediction for all methods, but for within family, the accuracy was the same for both marker densities or could even seem slightly lower for the high-density genotype. For highly polygenic traits such as lice resistance, most of the accuracy comes from information on close relatives. Studies have found that these relationships are accurately predicted with marker panels as low as 1000 SNPs across genome (Kriaridou et al., 2020). We had 215K SNPs at our lowest density and so the relationships are expected to be accurately fitted by a 215K marker panel, and thus there is limited effect of increasing the SNP density even more. Still, a small increase in accuracy for across-family predictions may be expected for the higher genotype density, as across-family predictions relies more on LD between markers and causative mutations. However, the benefits of higher density might be reduced due to imputation errors. Our 750K genotypes were imputed, whereas the 215K genotypes were recorded. Our reference population for the imputation was small (59 parents) and did not include all the parents of the animals in our dataset. This means that some of the families were imputed based on parental animals from other families. Close relatives share long haplotypes, which likely results in similar imputation, and possibly similar imputation errors, within the haplotype. Incorrect imputation may thus be more likely to cause bias in across-family than within-family prediction (within-family relationships are still accurately captured by the imputed SNPs). As BayesGC fits a polygenic term in addition to the BayesC term, it could be more robust than BayesC towards these kinds of errors, however differences in accuracy were small and not statistically significant in our study.

4.1 Posterior probabilities

When fitting the BayesC-term we have both a prior and a posterior probability of whether a SNP should be fitted in the model or not. The prior probability is an input parameter, and the posterior probability is determined by the model from the Gibbs-sampling and data. The posterior probability is the probability of how often the SNP was fitted in the model for all the Gibbs samples. If one SNP explains more variance than another it should have a higher posterior probability of inclusion. It is feasible to detect QTLs using the posterior probabilities from Bayes C (van den Berg et al., 2013). However, in order to detect QTLs, the recommendation is to use large datasets and highly heritable traits. For our study, the sample size is limited (n=1385), and the heritability is low to moderate. Tsai et al., (2016) did a GWAS analysis for the trait host resistance to salmon lice (*Lepeophtheirus salmonis*) but did not find any QTL for the trait. However, Rochus et al., (2018) found 2 QTL, on chromosome 1 and 23 respectively using a mixed linear model GWAS, and 70 SNPs using a forward multiple linear regression model that did not correct for population stratification and relatedness, and thus many of the 70 SNPs may be due to population structure. A few small QTL have also been found for sea lice more prevalent in the southern hemisphere (*Caligus rogercresseyi*). Among these, Cáceres et al., (2019) found 7 windows explaining up to 3% of the genetic variance for Atlantic salmon. The regions were associated with immune responses, cytoskeletal factors and cell migrations. Robledo et al., (2019) also found 3 single QTLs that explained approximately 4% of the genetic variance each. 3 QTL regions of 3-5 Mb explaining between 7.8 and 13.4% of the genetic variance of sea lice density for the *C. rogercresseyi* lice. However, it is known that estimates of QTL variances coming from the same data in which they were detected are overestimated by the Beavis effect (Xu, 2003). Hence, some QTL for sea lice resistance were found in the literature, however the genetics

and heritability of lice resistance has also been found to depend on the recording methodology.

5. Concluding remarks

When using Genomic Prediction within-families, a SNP-density of 215K seems to be more than sufficient to achieve a good prediction accuracy. However, if one want to predict across-family one might benefit from a higher density genotype, although, if genotype imputation is required to achieve the higher density, imputation errors might reduce the benefits. Host resistance to salmon lice behaved as a highly polygenic trait in our data with no major QTL regions and there seems to be virtually no benefit in fitting a BayesC term for this trait since the GBLUP, BayesC and BayesGC yielded very similar accuracies.

Acknowledgements

We are grateful to the helpful comments of two anonymous reviewers. Funding from the Norwegian Research Council (project 255297) is gratefully acknowledged. AquaGen AS is acknowledged for providing data and genotype information.

References

Abolofia, J., Asche, F., Wilen, J.E., 2017. The Cost of Lice: Quantifying the Impacts of Parasitic Sea Lice on Farmed Salmon. Mar. Resour. Econ. 32, 329–349. https://doi.org/10.1086/691981

464 Cáceres, P., Barría, A., Christensen, K.A., Bassini, L.N., Correa, K., Lhorente, J.P., Yáñez, J.M.,
465 2019. Genome-scale comparative analysis for host resistance against sea lice between
466 Atlantic salmon and rainbow trout. bioRxiv 624031. https://doi.org/10.1101/624031
467 Calus, M.P.L., Veerkamp, R.F., 2007. Accuracy of breeding values when using and ignoring the
468 polygenic effect in genomic breeding value estimation with a marker density of one SNP
469 per cM. J. Anim. Breed. Genet. 124, 362–368. https://doi.org/10.1111/j.1439-
470 0388.2007.00691.x
471 Daetwyler, H.D., Pong-Wong, R., Villanueva, B., Woolliams, J.A., 2010. The impact of genetic
472 architecture on genome-wide evaluation methods. Genetics 185, 1021–1031.
473 https://doi.org/10.1534/genetics.110.116855
474 Efron, B. Tibishirani, R.J., 1994. An Introduction to the Bootstrap [WWW Document]. Boca
475 Rat. CRC Press LLC. URL
476 https://books.google.no/books?hl=en&lr=&id=gLlpIUxRntoC&oi=fnd&pg=PR14&ots=A9
477 BvU8J7F2&sig=rU1bHQeofAkRYvjRIucY5ei_XkQ&redir_esc=y#v=onepage&q&f=false
478 (accessed 11.19.19).
479 Gjerde, B., Ødegård, J., Thorland, I., 2011. Estimates of genetic variation in the susceptibility
480 of Atlantic salmon (Salmo salar) to the salmon louse Lepeophtheirus salmonis.
481 Aquaculture 314, 66–72. https://doi.org/10.1016/j.aquaculture.2011.01.026
482 Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J., 2011. Extension of the bayesian alphabet
483 for genomic selection. BMC Bioinformatics 12. https://doi.org/10.1186/1471-2105-12-
484 186
485 Iheshiulor, O.O.M., Woolliams, J.A., Svendsen, M., Solberg, T., Meuwissen, T.H.E., 2017.
486 Simultaneous fitting of genomic-BLUP and Bayes-C components in a genomic prediction
487 model. Genet. Sel. Evol. 49, 1–13. https://doi.org/10.1186/s12711-017-0339-9

Iversen, M.W., Nordbø, Ø., Gjerlaug-Enger, E., Grindflek, E., Soares Lopes, M., Meuwissen, T., 2019. Effects of heterozygosity on performance of purebred and crossbred pigs. Genet. Sel. Evol. 51. https://doi.org/10.1186/s12711-019-0450-1

Kolstad, K., Heuch, P.A., Gjerde, B., Gjedrem, T., Salte, R., 2005. Genetic variation in resistance of Atlantic salmon (Salmo salar) to the salmon louse Lepeophtheirus salmonis. Aquaculture 247, 145–151. https://doi.org/10.1016/j.aquaculture.2005.02.009

Kriaridou, C., Tsairidou, S., Houston, R.D., Robledo, D., 2020. Genomic Prediction Using Low Density Marker Panels in Aquaculture: Performance Across Species, Traits, and Genotyping Platforms. Front. Genet. 11, 124. https://doi.org/10.3389/fgene.2020.00124

Ma, P., Lund, M.S., Aamand, G.P., Su, G., 2019. Use of a Bayesian model including QTL markers increases prediction reliability when test animals are distant from the reference population. J. Dairy Sci. 102, 7237–7247. https://doi.org/10.3168/jds.2018-15815

Madsen, P., Jensen, J., 2013. A User's Guide to DMU A Package for Analysing Multivariate Mixed Models.

Meuwissen, T.H., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–29.

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

Neves, H.H.R., Carvalheiro, R., Queiroz, S.A., 2012. A comparison of statistical methods for genomic selection in a mice population. BMC Genet. 13. https://doi.org/10.1186/1471-2156-13-100

Ødegård, J., Moen, T., Santi, N., Korsvoll, S.A., Kjøglum, S., Meuwisse, T.H.E., 2014. Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). Front. Genet. 5, 1–8. https://doi.org/10.3389/fgene.2014.00402

512 Overton, K., Dempster, T., Oppedal, F., Kristiansen, T.S., Gismervik, K., Stien, L.H., 2018.
513 Salmon lice treatments and salmon mortality in Norwegian aquaculture: a review. Rev.
514 Aquac. https://doi.org/10.1111/raq.12299
515 Robledo, D., Gutiérrez, A.P., Barría, A., Lhorente, J.P., Houston, R.D., Yáñez, J.M., 2019.
516 Discovery and Functional Annotation of Quantitative Trait Loci Affecting Resistance to
517 Sea Lice in Atlantic Salmon. Front. Genet. 10. https://doi.org/10.3389/fgene.2019.00056
518 Rochus, C.M., Holborn, M.K., Ang, K.P., Elliott, J.A.K., Glebe, B.D., Leadbeater, S., Tosh, J.J.,
519 Boulding, E.G., 2018. Genome-wide association analysis of salmon lice (*Lepeophtheirus*
520 *salmonis*) resistance in a North American Atlantic salmon population. Aquac. Res. 49,
521 1329–1338. https://doi.org/10.1111/are.13592
522 Sargolzaei, M., Chesnais, J.P., Schenkel, F.S., 2014. A new approach for efficient genotype
523 imputation using information from relatives. BMC Genomics 15, 478.
524 https://doi.org/10.1186/1471-2164-15-478
525 Solberg, T., Sonesson, A., Woolliams, J., Degard, J., Meuwissen, T., 2009. Persistence of
526 accuracy of genome-wide breeding values over generations when including a polygenic
527 effect. Genet. Sel. Evol. 41, 1–8. https://doi.org/10.1186/1297-9686-41-53
528 Sonesson, A.K., 2007. Within-family marker-assisted selection for aquaculture species. Genet.
529 Sel. Evol. 39, 301. https://doi.org/10.1186/1297-9686-39-3-301
530 Sonesson, A.K., Meuwissen, T.H., 2009. Testing strategies for genomic selection in
531 aquaculture breeding programs. Genet. Sel. Evol. 41, 1–9. https://doi.org/10.1186/1297-
532 9686-41-37
533 Torrissen, O., Jones, S., Asche, F., Guttormsen, A., Skilbrei, O.T., Nilsen, F., Horsberg, T.E.,
534 Jackson, D., 2013. Salmon lice - impact on wild salmonids and salmon aquaculture. J. Fish
535 Dis. https://doi.org/10.1111/jfd.12061

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

Tsai, H.Y., Hamilton, A., Tinch, A.E., Guy, D.R., Bron, J.E., Taggart, J.B., Gharbi, K., Stear, M., Matika, O., Pong-Wong, R., Bishop, S.C., Houston, R.D., 2016. Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations. Genet. Sel. Evol. 48, 1–11. https://doi.org/10.1186/s12711-016-0226-9

van den Berg, I., Fritz, S., Boichard, D., 2013. QTL fine mapping with Bayes C(π): a simulation study. Genet. Sel. Evol. 45, 19. https://doi.org/10.1186/1297-9686-45-19

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–23. https://doi.org/10.3168/jds.2007-0980

Verbyla, K.L., Bowman, P.J., Hayes, B.J., Goddard, M.E., 2010. Sensitivity of genomic selection to using different prior distributions. BMC Proc. 4, 2–5. https://doi.org/10.1186/1753-6561-4-s1-s5

Wang, Y., Lin, G., Li, C., Stothard, P., 2016. Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle. Springer Sci. Rev. 4, 79–98. https://doi.org/10.1007/s40362-017-0041-x

Wu, X., Lund, M.S., Sun, D., Zhang, Q., Su, G., 2015. Impact of relationships between test and training animals and among training animals on reliability of genomic prediction. J. Anim. Breed. Genet. 132, 366–375. https://doi.org/10.1111/jbg.12165

Xu, S., 2003. Theoretical Basis of the Beavis Effect. Genetics 165, 2259–2268.

560

561 Table 1. Results from the within-family predictions using 215K genotype density.

	acc	**SE(acc)**	**b**	**π**	$\boldsymbol{\sigma}_{\mathrm{pol}}{}^{2}$	$\boldsymbol{\sigma}_{\mathrm{m}}{}^{2}$	$\mathbf{n}_{\mathrm{mrk}}$
GBLUP	0.671	0.011	1.08	0	0.069	0	0
BayesGC_05	0.675	0.011	1.09	0.0002	0.065	0.00017	50
BayesGC_25	0.675	0.011	1.09	0.0012	0.052	0.00017	250
BayesGC_50	0.674	0.011	1.09	0.0023	0.034	0.00017	500
BayesGC_75	0.673	0.011	1.09	0.0035	0.017	0.00017	750
BayesC	0.672	0.011	1.09	0.0046	0	0.00017	1000

562 **acc** is accuracy of prediction (Pearson correlation between estimated and true breeding value
563 divided by the square root of the heritability).

564 **SE(acc)** is the standard error of the means of the accuracy for each replication.

565 **b** is the regression coefficient. π is the prior probability of a SNP having an effect or not.

566 $\boldsymbol{\sigma}_{\mathrm{pol}}{}^{2}$ is the variance attributed to the polygenic effect.

567 $\boldsymbol{\sigma}_{\mathrm{m}}{}^{2}$ is the variance assumed for a single SNP effect (if fitted in the model).

568 $\mathbf{n}_{\mathrm{mrk}}$ is the estimated number of markers fitted in the model based on the π value multiplied by the
569 total number of markers.

570 Table 2. Results from the across-family predictions using 215K genotype density.

	acc	**SE(acc)**	**b**	**π**	$\boldsymbol{\sigma}_{\mathrm{pol}}{}^{2}$	$\boldsymbol{\sigma}_{\mathrm{m}}{}^{2}$	$\mathbf{n}_{\mathrm{mrk}}$
GBLUP	0.596	0.012	1.18	0	0.069	0	0
BayesGC_05	0.602	0.014	1.23	0.0002	0.065	0.00017	50
BayesGC_25	0.601	0.013	1.19	0.0012	0.052	0.00017	250
BayesGC_50	0.601	0.013	1.19	0.0023	0.034	0.00017	500
BayesGC_75	0.600	0.013	1.19	0.0035	0.017	0.00017	750
BayesC	0.599	0.013	1.19	0.0046	0	0.00017	1000

571 **acc** is accuracy of prediction (Pearson correlation between estimated and true breeding value
572 divided by the square root of the heritability).

573 **SE(acc)** is the standard error of the means of the accuracy for each replication.

574 **b** is the regression coefficient. π is the prior probability of a SNP having an effect or not.

575 $\boldsymbol{\sigma}_{\mathrm{pol}}{}^{2}$ is the variance attributed to the polygenic effect.

576 $\boldsymbol{\sigma}_{\mathrm{m}}{}^{2}$ is the variance assumed for a single SNP effect (if fitted in the model).

577 $\mathbf{n}_{\mathrm{mrk}}$ is the estimated number of markers fitted in the model based on the π value multiplied by the

578 total number of markers.

579

Table 3. Results from the within-family predictions using 750K genotype density.

	acc	**SE(acc)**	**b**	**π**	$\boldsymbol{\sigma}_{\mathrm{pol}}{}^{2}$	$\boldsymbol{\sigma}_{\mathrm{m}}{}^{2}$	$\mathbf{n}_{\mathrm{mrk}}$
GBLUP	0.669	0.010	1.09	0	0.069	0	0
BayesGC_05	0.673	0.011	1.10	0.00007	0.065	0.00027	50
BayesGC_25	0.676	0.012	1.03	0.00034	0.052	0.00027	250
BayesGC_50	0.672	0.010	1.10	0.00067	0.034	0.00027	500
BayesGC_75	0.671	0.011	1.10	0.00101	0.017	0.00027	750
BayesC	0.670	0.011	1.10	0.00134	0	0.00027	1000

acc is accuracy of prediction (Pearson correlation between estimated and true breeding value divided by the square root of the heritability).

SE(acc) is the standard error of the means of the accuracy for each replication.

b is the regression coefficient. π is the prior probability of a SNP having an effect or not.

$\boldsymbol{\sigma}_{\mathrm{pol}}{}^{2}$ is the variance attributed to the polygenic effect.

$\boldsymbol{\sigma}_{\mathrm{m}}{}^{2}$ is the variance assumed for a single SNP effect (if fitted in the model).

$\mathbf{n}_{\mathrm{mrk}}$ is the estimated number of markers fitted in the model based on the π value multiplied by the total number of markers.

590 Table 4. Results from the across-family predictions using 750K genotype density.

	acc	**SE(acc)**	**b**	**π**	$\boldsymbol{\sigma}_{\mathrm{pol}}{}^{2}$	$\boldsymbol{\sigma}_{\mathrm{m}}{}^{2}$	$\mathbf{n}_{\mathrm{mrk}}$
GBLUP	0.607	0.009	1.21	0	0.069	0	0
BayesGC_05	0.605	0.012	1.24	0.00007	0.065	0.00027	50
BayesGC_25	0.610	0.013	1.16	0.00034	0.052	0.00027	250
BayesGC_50	0.605	0.012	1.24	0.00067	0.034	0.00027	500
BayesGC_75	0.611	0.009	1.23	0.00101	0.017	0.00027	750
BayesC	0.611	0.009	1.23	0.00134	0	0.00027	1000

591 **acc** is accuracy of prediction (Pearson correlation between estimated and true breeding value
592 divided by the square root of the heritability).

593 **SE(acc)** is the standard error of the means of the accuracy for each replication.

594 **b** is the regression coefficient. π is the prior probability of a SNP having an effect or not.

595 $\boldsymbol{\sigma}_{\mathrm{pol}}{}^{2}$ is the variance attributed to the polygenic effect.

596 $\boldsymbol{\sigma}_{\mathrm{m}}{}^{2}$ is the variance assumed for a single SNP effect (if fitted in the model).

597 $\mathbf{n}_{\mathrm{mrk}}$ is the estimated number of markers fitted in the model based on the π value multiplied by the

598 total number of markers.

599