Norwegian University of Life Sciences
Faculty of Biosciences
Department of Animal and Aquacultural sciences

Philosophiae Doctor (PhD)
Thesis 2022:33

Genomic prediction using high-density and whole-genome sequence genotypes

Genomisk prediksjon ved bruk av høy tetthets- og hel-genom sekvens genotyper

Maria Valkeneer Kjetså

Genomic prediction using high-density and whole-genome sequence genotypes

Genomisk prediksjon ved bruk av høy tetthets- og hel-genom sekvens genotyper

Philosophiae Doctor (PhD) Thesis

Maria Valkeneer Kjetså

Norwegian University of Life Sciences
Faculty of Biosciences
Department of Animal and Aquacultural sciences

Ås 2022

Thesis number 2022:33
ISSN 1894-6402
ISBN 978-82-575-1905-6

Supervisors and Evaluation Committee

PhD Supervisors

Prof. Theo H. E. Meuwissen
Department of Animal and Aquacultural Sciences
Norwegian University of Life Sciences
P.O. Box 5003, N-1432 Ås
Norway

Assoc. Prof. Jørgen Ødegård
Breeding and Genetics, AquaGen
P.O. Box 1240, N-7462 Trondheim
Norway

PhD Evaluation Committee

Prof. Jörn Bennewitz
Department of Animal Genetics and Breeding
University of Hohenheim
Garbenstr. 17
70599 Stuttgart
Germany

Dr. Aniek Bouwman
Animal Breeding and Genomics
Wageningen University and Research
P.O. Box 338
6700 AH Wageningen
The Netherlands

Dr. Gareth F. Difford
Department of Animal and Aquacultural Sciences
Norwegian University of Life Sciences
P.O. Box 5003, N-1432 Ås
Norway

Acknowledgements

This research was financed by the Norwegian Research Council under project no. 255297 – “From whole-genome sequence data to precision breeding”. AquaGen and Norsvin are also acknowledged for providing the data used in this thesis.

To my supervisors, Theo Meuwissen and Jørgen Ødegård, I would like to express my sincerest gratitude. Thank you for sharing your knowledge, and for your guidance and support through this time. Theo, I appreciate your patience, time, and efforts to support me through this endeavour.

I would also like to thank Eli Grindflek, Arne Gjuvsland and Øyvind Nordbø at Norsvin for your input, guidance, knowledge, and discussions. Your help and support have been invaluable.

I also want to thank all my colleagues and friends at the Animal Breeding and Genetics Group at IHA, thank you for your friendship, encouragement, and coffee-break discussions. You know who you are 😊 Thank you also to my colleagues in NordGen for your support.

To my family and friends: Thank you so much for your support and encouragement. Even though you might not understand exactly what I am doing, I could not have done it without you! And finally, to my beloved husband and partner in life, Peter, thank you for your constant love and support. You help me believe in myself, even when I don’t.

Maria Valkeneer Kjetså
Ås, April 2022

Table of Contents

Papers I-III have individual page numbers.

1 Abbreviations and definitions

BCS – Body Condition Score

BLUP – Best Linear Unbiased Prediction

GBLUP – Genomic Best Linear Unbiased Prediction

GEBV – Genomic Estimated Breeding Values

GS – Genomic Selection

HD – High Density genotypes

LW3W – Litter Weight at 3 Weeks

LD – Linkage Disequilibrium

M3W – Mortality within 3 Weeks

MCMC – Markov Chain Monte Carlo

pCADD – pig Combined Annotation Dependent Depletion

QTL – Quantitative Trait Loci

STB – total number of Still Born piglets

SHL – Shoulder Lesions

SNP – Single Nucleotide Polymorphism

TNB – Total Number of Born piglets

WGS – Whole Genome Sequence

YD – Yield Deviations

2 List of papers

I. Kjetså, M. H., Ødegård, J., and Meuwissen, T. H. E. 2020. **Accuracy of genomic prediction of host resistance to salmon lice in Atlantic salmon (*Salmo salar*) using imputed high-density genotypes**. *Aquaculture, 526, 735415.*

II. Kjetså, M. V., Gjuvsland, A., Nordbø, Ø., Grindflek., E., and Meuwissen, T. H. E. 2022. **Accuracy of genomic prediction of maternal traits in pigs using Bayesian variable selection methods.** *Submitted to Journal of Animal Breeding and Genetics.*

III. Kjetså, M. V., Gjuvsland, A., Grindflek., E., and Meuwissen, T. H. E. 2022. **Whole-genome sequence and pCADD marker-based genomic prediction for maternal traits in two pig lines.** *Manuscript.*

3 Summary

The main objective of this thesis was to investigate genomic prediction methods for high-density and whole-genome sequence genotypes, with emphasis on traits that may have difficulties achieving a high prediction accuracy with pedigree-based predictions, such as disease resistance and maternal traits. A Bayesian variable selection method that combines a polygenic term through a G-matrix and a BayesC term (BayesGC) was compared with Genomic Best Linear Unbiased Prediction (GBLUP), and for Paper I and II, it was also compared to BayesC.

Paper I aimed to investigate genomic prediction accuracy for the trait host resistance to salmon lice in Atlantic salmon (*Salmo salar*). Three genomic prediction methods (GBLUP, BayesC and BayesGC) were compared using 215K and 750K SNP genotypes through both within-family and across-family prediction scenarios. The data consisted of 1385 fish with both phenotype- and genotype, and the prediction accuracy was determined through five-fold cross-validation. The results showed an accuracy of ~0.6 and ~0.61 for across-family prediction with 215K and 750K genotypes and ~0.67 for within-family prediction for both genotypes. BayesGC showed a slightly higher prediction accuracy than GBLUP and BayesC, especially for the across-family predictions, but the differences were insignificant.

Paper II investigated the prediction accuracy of GBLUP, BayesC and BayesGC for six maternal traits in Landrace sows. The data consisted of between 10,000 and 15,000 sows, all genotyped and imputed to a genotype density of 660K SNPs. The effects of different priors for the Bayesian variable selection methods were also investigated. The ~1,000 youngest sows were used as validation animals to validate the prediction accuracy. Results showed a variation in genomic prediction accuracy between 0.31 to 0.61 for the different traits. The accuracy did not vary much between the different methods and priors within traits. BayesGC had a 9.8 and 3% higher accuracy than GBLUP for traits M3W and BCS. However, for the other traits, there were minor differences.

For within-breed prediction marker density and sizes of reference populations are often sufficient. However, when predicting across breeds, one might need a higher density, such as Whole Genome Sequence (WGS), or one could benefit from

functional markers derived from WGS. Paper III investigates prediction accuracy for four maternal traits in two pig populations, a pure-bred Landrace (L) and a Synthetic (S) Yorkshire/Large White line. Prediction accuracy was tested with three different marker data sets: High-Density (HD), Whole Genome Sequence (WGS) and markers derived from WGS based on their pig Combined Annotation Dependent Depletion (pCADD) score. Two genomic prediction methods (GBLUP and BayesGC) were investigated for across- within- and multi-line predictions. For across- and within-line prediction, reference population sizes between 1K and 30K animals were analysed for prediction accuracy. In addition, multi-line reference population consisting of 1K, 3K or 6K animals for each line in different ratios were tested. The results showed that a reference population of 3K-6K animals for within-line prediction was usually sufficient to achieve a high prediction accuracy. However, increasing to 30K animals in the reference population further increased prediction accuracy for two of the traits. A reference population of 30K across-line animals achieved a similar accuracy to 1K within-line animals. For multi-line prediction, the accuracy was most dependent on the number of within-line animals in the reference data. The S-line provided a generally higher prediction accuracy than the L-line. Using pCADD scores to reduce the number of markers from WGS data in combination with the GBLUP method generally reduced prediction accuracies relative to GBLUP_HD analyses. When using BayesGC, prediction accuracies were generally similar when using HD, pCADD, or WGS marker data, suggesting that the Bayesian method selects a suitable set of markers irrespective of the markers provided (HD, pCADD, or WGS).

Overall, these three studies showed that BayesGC seemed to have a slight advantage over GBLUP, especially with large datasets, high-density genotypes, and when relationships between the reference and validation animals were lower. They also showed that the relationship between the animals in the reference and validation population, and the size of the reference population, had a more significant impact on the prediction accuracy than the prediction method.

4 Norsk sammendrag

Hovedmålet med denne oppgaven var å undersøke genomiske prediksjonsmetoder for genotyper med høy markørtetthet og hel-genom sekvens, med vekt på egenskaper som kan ha vanskeligheter med å oppnå høy prediksjonsnøyaktighet med stamtavlebaserte prediksjoner, som sykdomsresistens og maternale egenskaper. En Bayesiansk seleksjonsmetode som kombinerer et polygent element via en G-matrise og et BayesC-element (BayesGC) ble sammenlignet med Genomic Best Linear Unbiased Prediction (GBLUP), og for Paper I og II ble den også sammenlignet med BayesC.

Målet for Artikkel I var å undersøke genomisk prediksjonsnøyaktighet for egenskapen «Vertsresistens mot lakselus» hos atlantisk laks (*Salmo salar*). Tre genomiske prediksjonsmetoder (GBLUP, BayesC og BayesGC) ble sammenlignet ved bruk av 215K og 750K SNP-genotyper gjennom prediksjonsscenarier både innen-familie og på tvers av familie. Dataene besto av 1385 fisk med både fenotype- og genotypeinformasjon, og prediksjonsnøyaktigheten ble bestemt gjennom fem-folds kryssvalidering. Resultatene viste en nøyaktighet på ~0,6 og ~0,61 for tverr-familieprediksjon med 215K og 750K genotyper, og ~0,67 for innen-familieprediksjon for begge genotyper. BayesGC viste en litt høyere prediksjonsnøyaktighet enn GBLUP og BayesC, spesielt for prediksjoner på tvers av familier, men forskjellene var ikke signifikante.

Artikkel II undersøkte prediksjonsnøyaktigheten til GBLUP, BayesC og BayesGC for seks maternale egenskaper hos landrasepurker. Dataene besto av mellom 10 000 og 15 000 purker, alle genotypet med en genotypetetthet på 660K SNP-er. Effektene av forskjellige priorer for de Bayesianske seleksjonsmetodene ble også undersøkt. De ca. 1000 yngste purkene ble brukt som valideringsdyr for å validere prediksjonsnøyaktigheten. Resultatene viste en variasjon i genomisk prediksjonsnøyaktighet mellom 0,31 til 0,61 for de forskjellige egenskapene. Nøyaktigheten varierte ikke mye mellom de forskjellige metodene og priorene innen egenskaper. BayesGC hadde en 9,8 og 3 % høyere nøyaktighet enn GBLUP for egenskapene M3W og BCS. For de andre egenskapene var det imidlertid mindre forskjeller.

Å predikere innenfor rase med tilstrekkelig markørtetthet og størrelse på referansepopulasjonen er én ting. Men når man predikerer på tvers av raser, kan man trenge en høyere markørtetthet, slik som Hel-genom sekvens (HGS), eller man kan dra nytte av funksjonelle markører avledet fra HGS. Artikkel III undersøker prediksjonsnøyaktighet for fire morsegenskaper i to grisepopulasjoner, en renraset Landrase- (L) og en Syntetisk (S) Yorkshire/Stor Hvit-linje. Prediksjonsnøyaktigheten ble testet med tre forskjellige markørdatasett: Høy-Tetthet (HT), Hel-genom sekvens (HGS) og markører avledet fra HGS basert på deres pig Combined Annotation Dependent Depletion (pCADD) score. To genomiske prediksjonsmetoder (GBLUP og BayesGC) ble undersøkt for prediksjoner på tvers av linjer og kombinasjoner av linjer. For på tvers- og innenfor linje ble referansepopulasjonsstørrelser mellom 1K og 30K dyr analysert for prediksjonsnøyaktighet. I tillegg ble kombinert-linje referansepopulasjon bestående av 1K, 3K eller 6K dyr for hver linje i forskjellige forhold testet. Resultatene viste at en referansepopulasjon på 3K-6K dyr for prediksjon innenfor linjen vanligvis var nok til å oppnå en høy prediksjonsnøyaktighet. Økning til 30 000 dyr i referansepopulasjonen økte imidlertid prediksjonsnøyaktigheten for to egenskaper signifikant. En referansepopulasjon på 30 000 dyr på tvers av linjen oppnådde en lignende nøyaktighet som 1 000 dyr innenfor linje. For kombinasjons-linje prediksjonsnøyaktighet var nøyaktigheten mest avhengig av antall dyr innenfor linje i referansedataene. S-linjen ga en generelt høyere prediksjonsnøyaktighet sammenlignet med L-linjen. Bruk av pCADD-score for å redusere antall markører fra HGS-data i kombinasjon med GBLUP-metoden reduserte generelt prediksjonsnøyaktigheten i forhold til GBLUP_HT-analyser. Når du bruker BayesGC, var prediksjonsnøyaktigheten generelt like ved bruk av HT-, pCADD- eller HGS-markørdata, noe som tyder på at den Bayesianske metoden velger et passende sett med markører uavhengig av de angitte markørene (HT, pCADD eller HGS).

Totalt sett viste disse tre studiene at BayesGC så ut til å ha en liten fordel fremfor GBLUP, spesielt med store datasett, genotyper med høy tetthet og når forholdet mellom referanse- og valideringsdyrene var lavere. De viste også at forholdet mellom dyrene i referanse- og valideringspopulasjonen, og størrelsen på referansepopulasjonen, hadde en mer signifikant innvirkning på prediksjonsnøyaktigheten enn prediksjonsmetoden.

5 General Introduction

5.1 Introduction

From the domestication of animals starting over 10,000 years ago, humans have gradually cultivated both animals and plants to suit their needs better. At some point, they must have noticed how offspring tended to look like their parents. However, the way different traits were inherited was a mystery for a very long time. The ancient Greek philosopher Aristotle believed, for example, that the man contributed to the "form" of the offspring, while the woman contributed with the "matter" (Henry, 2006). Even if we did not know how traits were inherited, systematic breeding started in the 18th century, with record-keeping and artificial selection introduced through Sir Robert Bakewell. The first herd-book was established for the thoroughbred horse in 1871 to keep an overview of the animals' relationships with each other (Oldenbroek and van der Waaij 2015).

It was not until Gregor Mendel did his pea plant hybridization experiments that we got a more systemic knowledge of the heredity of traits. He reported his results in 1865 on how traits of the peas were passed down through generations in a systematic way. He died in 1884, but it was not until the early 20th-century that other botanist reported their results and linked them back to his work. Many breakthroughs in genetic research followed in the 20th century. Among them was the determination of DNA as the material of heredity in 1952, and the discovery of the helical structure of DNA in 1953. Furthermore, the Human Genome Project, which established the base-pair sequence of the human genome, started in 1990 and ended in 2003. (Oldenbroek and van der Waaij 2015).

As the knowledge of genetics and heredity developed, theories for utilising the laws of heredity were soon developed for the breeding of our domesticated animals as well. Mendel showed the inheritance for traits affected by a single gene. However, the inheritance of traits affected by many genes, for example, height, was more challenging to dissect. When Fisher published his paper in 1918, showing how many genes could contribute to the variance of the height, where each gene followed Mendelian inheritance laws, he laid the foundation for the field of Quantitative Genetics (Visscher & Goddard, 2019). Lush published the article "animal breeding based on quantitative statistics and genomic information" in 1937. In 1941, a PhD student from Lush, Hazel, published the selection index theory, where several traits could be weighed to produce an index to rate animals against each other to select the best animals for breeding. Henderson developed the Estimated Breeding Values (EBVs) and came up with Best Linear Unbiased Prediction (BLUP) in 1950. However, the computer power at the time was too low to perform the calculations. It was not until the late 1980's that the computer power became significant enough to estimate BLUP breeding values with a complete animal model (Oldenbroek and van der Waaij 2015).

5.2 Genomic Selection

When DNA information first became available, the idea of being able to select animals for breeding based directly on their DNA was intriguing. As DNA information became more widespread, methods were needed to utilise the information for breeding value estimation. At first, most animal breeding research focused on finding the specific markers that would explain each trait, the Quantitative Trait Loci (QTLs), to utilise them in Marker-Assisted Selection. However, the results from these methods explained only a small part of the total genetic variance (de Koning, 2016).

In 2001 the paper "Prediction of total genetic value using genome-wide dense marker maps" by Meuwissen, Hayes, and Goddard (2001) was published. They used a reference population of animals with both known phenotypes and genotypes to estimate the marker effects and use them in the breeding value prediction. Meuwissen et al. also suggested using Single Nucleotide Polymorphisms (SNPs) as the genomic data. As the technology developed, prices of genotyping with SNP markers kept going down, which made the method feasible. Genomic selection was first implemented in dairy cattle breeding, where it could replace the expensive progeny testing scheme while also improving the accuracy of predictions (Schaeffer, 2006). Today, genomic selection is implemented for most animal and plant species and is even used in disease research for humans (de Koning, 2016).

5.3 Accuracy of Genomic Selection

To successfully apply genomic selection, the prediction of the genomic breeding values must be accurate. The accuracy of genomic prediction methods depends on the proportion of genetic variance captured by the markers and the accuracy with which genetic effects captured by the markers can be estimated (Dekkers, 2007). The accuracy of which the effects of markers can be estimated is dependent on the genomic prediction method, the size of the reference population, the heritability of the phenotypes in the reference dataset and the number of independent QTL that affect the trait (Daetwyler, Pong-Wong, Villanueva, & Woolliams, 2010). A critical aspect of genomic prediction accuracy is the effective number of chromosome segments, M_e: if M_e increases, accuracy decreases. When a population is more related, M_e is lower, increasing the accuracy (Daetwyler, Calus, Pong-Wong, De Los Campos, & Hickey, 2013). The accuracy of genomic prediction also increases with the number of phenotypes relative to the population's effective number of genomic segments (Daetwyler et al., 2013). In addition, the additive genetic relationship between the animals in the reference and validation population is essential, where the decay of accuracy with decreasing additive genetic relationship is higher with a small reference population (Habier, Tetens, Seefried, Lichtner, & Thaller, 2010).

To capture genetic variance, the genotype data must contain markers in Linkage Disequilibrium (LD) with QTL. With a higher density genotype, there is a higher chance of markers being in LD with QTL. Across-breed prediction suffers from low across population LD compared to within-breed prediction, i.e., across breeds M_e is larger. Thus, a higher marker density might be needed to predict across breeds compared to within-breed. If the number of QTL is smaller than the effective number of segments, many segments carry no QTL (Daetwyler et al., 2010). Variable selection methods may identify the segments with QTL and concentrate on their prediction. Hence, variable selection methods improve prediction accuracy by estimating the effects of fewer segments, for example, by a priori assuming many segments have no effect. This approach is only effective if there are many segments with no effect. Otherwise, the SNP-BLUP prior assumption that all SNPs have an effect is justified and yields the most accurate predictions.

5.4 Genomic data

5.4.1 Whole Genome Sequence

Whole Genome Sequence data (WGS), in addition to containing Single Nucleotide Polymorphism (SNPs), also includes other causes of variations, such as deletions, duplications, indels, Copy Number variations (CNVs) and other polymorphisms. Thus, WGS data are more likely to include the causative variant or have markers with very close LD with the causative variant. Suppose all the variations of a population could be explained. In that case, the predictions are no longer dependent on LD between SNPs and QTL, leading to increased accuracy of GP (Meuwissen & Goddard, 2010; van Binsbergen et al., 2014). When LD is incomplete, but there is high marker density, WGS could improve GS since it does not need to rely on LD between flanking markers and QTL, thereby increasing the signal in diverged populations, for example across-breed (De Roos, Hayes, & Goddard, 2009; Goddard, 2009; van Binsbergen et al., 2014).

However, some studies have demonstrated that using WGS data did not increase prediction accuracy or increased it only slightly compared to using high-density SNP panel genotypes. For example, van Binsbergen et al. (2015) reported that using imputed WGS data did not increase the accuracy of GP in Holstein-Friesian cattle compared to using HD SNP genotype data. Zhang et al. (2018) also showed that increasing marker density did not increase or only slightly increased the accuracy of GP of feed efficiency component traits in Duroc pigs. Thus, GP with WGS data could be an attractive approach, although to date, the expectation of a higher accuracy has not been realized with real WGS data.

5.4.2 Selection of functional polymorphisms

One option is to select functional markers from the WGS that could later be used in the prediction method. A common approach has been to select significant markers from a genome-wide association study as predictors. However, this approach is likely to select false positive markers and ignores markers below the significance threshold. Nevertheless, Brøndum et al. (2015) showed that the accuracy of GS could be improved by adding several significant QTL that were detected by genome-wide association studies (GWAS) using WGS data. Another way to find functional markers could be through “Combined annotation dependent depletion” (CADD) (Rentzsch, Witten, Cooper, Shendure, & Kircher, 2019). The CADD model was first developed for humans, and it is a method to capture signals of evolutionary selection throughout the genome across many generations and combines this with genomic features, epigenetic data, and other predictors to estimate a deleteriousness score for a given variant. The method was then further developed for pigs, thus named pCADD (Groß et al., 2020). pCADD is a method for prioritizing SNPs in the pig genome with respect to their putative deleteriousness, in correspondence with the biological significance of the genomic region where they are located. The aim for developing pCADD was to help researchers and breeders to evaluate newly observed SNPs, and rank potentially harmful SNPs that are propagated by breeding. But it is also a possible way to find markers that are functional that could be interesting to fit in a model for animal breeding value estimations.

5.4.3 Genotype Imputation

Most animals that are genotyped today are genotyped with a SNP-chip. A lot of breeding companies might have started out with genotyping with lower marker densities and density increased as the technology has developed. Imputation of genotypes has been widely used to upgrade the animals genotyped with a lower density so that they can still be utilised in the breeding evaluations. For all the papers in this dissertation, we utilised imputation in order to increase the number of animals with HD or WGS genotypes. Although most studies on imputation accuracy show quite decent accuracies of imputation on individuals (~.90-.99 for some (van Binsbergen et al., 2014), there are also many issues. For example, the accuracy for each SNP can vary across the genome, giving very low accuracies in certain parts of the genome (Larmer, Sargolzaei, Ventura, & Schenkel, 2011). Many of the same factors that affect prediction accuracy of genomic prediction also affects the accuracy of imputation, such as the size and constitution of the reference animals, and the relationship between the reference animals and the imputed animals. The genotype density that is being imputed also matters, where for instance imputing from 6K directly to HD gives a lower imputation accuracy compared to if you impute first from 6K to 50K and then to HD (van Binsbergen et al., 2014). For smaller breeds, not having closely related animals in the reference group for imputation is also an issue and generally gives them lower prediction accuracy (Ventura et al., 2014).

5.5 Across- and multi- breed prediction

It has been suggested to combine related breeds into one larger reference population for small populations. However, some markers may be in high LD with a QTL in one population and not in the other, especially for populations that diverged many generations ago (Andreescu et al., 2007; de Roos, Hayes, Spelman, & Goddard, 2008; Gautier et al., 2007). The allele substitution effects across populations might also be different, resulting in a difference in genetic variation. A QTL might also segregate in one population and not in the other (Clark, Hickey, Daetwyler, & van der Werf, 2012; De Roos et al., 2009; M. E. Goddard & Hayes, 2009; Habier et al., 2010; Hayes, Bowman, Chamberlain, & Goddard, 2009; Wientjes, Veerkamp, & Calus, 2013). For across-breed prediction, where the reference population is from one population used to predict breeding values for animals in another population, the accuracies reported are zero or close to zero (Erbe et al. 2012; Hozé et al. 2014; L Zhou et al. 2014a; L. Zhou et al. 2014b). Using multi-breed populations with 50 or 777K has only shown minor improvements in prediction accuracy. The use of WGS data could improve this.

5.6 Genomic Prediction methods

Today, 20 years after the introduction of Genomic selection, genotypes with large densities have never been more available. Whole Genome Sequencing (WGS) prices are still decreasing (Wetterstrand KA, 2021), making it probable that WGS data will be more available in the future. This means that there is a need to develop methods to predict genomic breeding values that can optimise the use of high-density genotypes and whole genome sequence data (WGS). For linear methods, such as Genomic Best Linear Unbiased Prediction (GBLUP), all markers have equal weight in the prediction. There is not necessarily a significant increase in accuracy when increasing prediction accuracy from medium- to high-density genotypes and WGS when using GBLUP (van Binsbergen et al., 2014; VanRaden et al., 2012).
Bayesian methods try to differentiate SNPs relative to their importance, giving markers in high LD with causal mutations a higher relative weight (Meuwissen et al., 2001; Verbyla, Bowman, Hayes, & Goddard, 2010). The non-linear Bayesian GS methods are often referred to as the “Bayesian Alphabet” (Gianola, 2013; Gianola, De Los Campos, Hill, Manfredi, & Fernando, 2009). Some of the (many) proposed methods are BayesA and BayesB (Meuwissen et al. 2001), BayesC (Habier, Fernando, Kizilkaya, & Garrick, 2011), BayesR (Erbe et al., 2012) and BayesGC (Meuwissen, Berg, & Goddard, 2021). The main differences between the methods are the type of distributions and the number of distributions for the SNP effects. For example, BayesA uses one t-distribution for SNP-effects, while BayesB has two distributions: one t-distribution with probability p of a SNP having an effect, and one with probability 1-p with 0 effect, i.e., giving many SNPs a null effect. BayesC is similar to BayesB, as both have two distributions, where one has a null effect. However, BayesC uses a normal distribution instead of a t-distribution for SNPs with

effect and assumes a common variance for all these SNPs, while BayesB assumes SNP-specific variances. BayesR uses four normal distributions, where one of them has a null effect. BayesGC fits a polygenic effect through a G-matrix and a BayesC term. Hence, BayesGC fits many SNPs with a small effect through the G-matrix and a group of SNPs selected by the model with more significant effects through the BayesC term.

6 Aim and outline of the thesis

The main aim is to investigate alternative ways to make best use of high-density genotypes and WGS data in genomic prediction under different scenarios. The detailed aims are:

1) To compare alternative methods of genomic prediction for the trait host resistance to salmon lice in Atlantic salmon for prediction accuracies of the GEBVs based on a 215 K SNP genotypes and imputed 750 K SNP panels.

2) To determine the prediction accuracy of maternal traits in Landrace sows using a panel of 660K SNP markers and a9 - 15K reference population and compare the prediction accuracies of alternative methods of genomic prediction.

3) To compare prediction accuracy using a pCADD derived marker panel, a high density (HD) SNP-chip marker panel, and a Whole Genome Sequence (WGS) marker panel, using both a linear prediction method (GBLUP) and a Bayesian variable selection method (BayesGC).

4) To compare the effect of within-, across- and multi-breed genomic predictions at different sizes of reference populations.

7 Brief summary of papers

7.1 Paper I

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

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 (215 K and imputed 750 K). 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.

<u>Main results</u>
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 215 K to 750 K genotype density. However, for across-family prediction a slight improvement in predictive ability was observed at the higher density compared to the lower.

<u>Conclusions</u>
When using Genomic Prediction within-families, a SNP-density of 215 K was sufficient to achieve a good prediction accuracy. However, if one wants 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 was no benefit in fitting a BayesC term for this trait since the GBLUP, BayesC and BayesGC yielded very similar accuracies.

7.2 Paper II

Accuracy of genomic prediction of maternal traits in pigs using Bayesian variable selection methods.

Maternal traits in sows are often more difficult to record and have low heritabilities. Thus, new methods to increase the prediction accuracy such as genomic prediction are of interest. The aim of this study was to compare three methods of genomic prediction: GBLUP, BayesC and BayesGC for genomic prediction of six maternal traits in Landrace sows using a panel of 660K SNPs. The effect of different priors for the Bayesian methods were also investigated. GBLUP does not take the genetic architecture into account as all SNPs are assumed to have equally sized effects and relies heavily on the relationships between the animals for accurate predictions. Bayesian approaches rely on both fitting SNPs that describe relationships between animals in addition to fitting single SNP effects directly. Both the relationship between the animals and single SNP effects are important for accurate predictions.

Main results

The accuracy of genomic prediction on six maternal traits in landrace pigs varied greatly ranging from 0.31 to 0.61. The prediction accuracies did not vary much between the different genomic prediction methods. The two traits Mortality within three weeks (M3W) and Body Condition Score (BCS) could benefit from using a BayesGC approach with a 9.8 and 3.0% increase in accuracy respectively, while the remaining traits only showed minor improvements.

Conclusions

Although GBLUP, BayesC and BayesGC all yielded similar genomic prediction accuracies, the accuracy of BayesGC was always as high as or higher than that of GBLUP. Within the BayesGC method the accuracies could vary depending on the prior distributions. The models were more sensitive to how many markers were fitted in the model through varying the fraction of the total genetic variance explained by a single marker (Fr) compared to the amount of total genetic variance explained by marker effects as a whole (BayesGC_10, BayesGC_50 or BayesGC_90), but overall, most traits were robust against varying the prior distributions.

7.3 Paper III

Whole-genome sequence and pCADD marker-based genomic prediction for maternal traits in two pig lines.

The aim of this study was to investigate prediction accuracy for three different marker data sets (High-Density, pCADD, and Whole Genome Sequence (WGS)) and two different genomic prediction methods (GBLUP and BayesGC) on four maternal traits in pigs. The data consisted of two nucleus pig populations, one pure-bred Landrace (L) and one Synthetic (S) Yorkshire/Large White line. All animals had records on maternal traits and were genotyped, with up to 30K animals in each line. We investigated the necessary size of reference population needed to obtain a sufficient prediction accuracy within- and across-line and the effect of using a multi-line reference population with both a high ratio of within-line and a high ratio of across-line animals in the reference population.

Main Results

A reference population of 3K-6K animals for within-line prediction was sufficient to achieve a high prediction accuracy. However, increasing to 30K animals in the reference population significantly increased prediction accuracy for two traits. A reference population of 30K across-line animals achieved a similar accuracy to 1K within-line animals. For multi-line prediction accuracy, the accuracy was most dependent on the number of within-line animals in the reference data. The S-line provided a generally higher prediction accuracy compared to the L-line. Using pCADD scores to reduce the number of markers from WGS data in combination with the GBLUP method generally reduced prediction accuracies relative to GBLUP_HD analyses. When using BayesGC, prediction accuracies were generally similar when using HD, pCADD, or WGS marker data, suggesting that the Bayesian method selects a suitable set of markers irrespective of the markers provided (HD, pCADD, or WGS).

Conclusions

A large reference population size can help accuracy for both within- and across-line predictions. For multi-line prediction, adding more within-line animals are more important than a larger number of across-line animals. The BayesGC method benefited from a large reference population and was less dependent on the different genotype marker datasets to achieve a high prediction accuracy.

8 General discussion

Genomic prediction is a method that is still under constant development even 20 years after it was proposed. The technological advancement in both computer science, molecular biology and bioinformatics has great potential to propel the development of animal breeding and selection. This thesis explores ways of utilising these advances in genomic selection for traits that traditionally have difficulty obtaining a high prediction accuracy, especially compared to traditional pedigree predictions.

Paper I explored the accuracy of genomic prediction for host resistance to salmon lice in Atlantic salmon with three methods; GBLUP, BayesC and BayesGC, where BayesGC were found to have a slight advantage over GBLUP and a higher density genotype had a slight advantage when predicting across families, but not within families. In **Paper II**, the accuracy of genomic prediction for GBLUP, BayesC and BayesGC was explored for six maternal traits in pigs. The priors used for Bayesian variable selection were also explored, showing that BayesGC had a slight advantage and that the priors used are relatively stable for prediction accuracy, but that it could give a potential increase in accuracy of 9.2% when ~100 markers were fitted with a high variance attributed to each marker**. Paper III** explored the impact of reference population size for within-, across- and multi-breed prediction for up to 30K animals. In addition to looking at the effect of different marker data sets (HD, pCADD and WGS). **Paper III** showed that increasing the reference population substantially affected prediction accuracy. With 30K across-line animals, one could achieve prediction accuracies comparable to 1K within-line animals. There was also a benefit of a multi-line reference population. However, too many animals of a different line in a multi-line reference population could decrease the accuracy, and the value of adding within-line animals to the reference population is much higher than many across-line animals.

This general discussion will address some of the possibilities of utilising large marker densities such as WGS data for animal breeding and other measures to improve prediction accuracy for genomic selection.

8.1 Utilisation of high-density and WGS genotype marker data

As genotyping costs are likely to reduce, we will likely have breeding programs that routinely use whole genome sequencing instead of SNP-chip genotyping. One thing is to consider the cost of WGS genotyping, and another factor is the benefit it could give to a breeding program. WGS is supposed to contain all causative variants and thus be able to explain all variations in a trait and not be dependent on having markers in LD with QTL, as it should contain the causative markers. However, when comparing simulation studies of WGS with studies on actual data, the results have been ambiguous so far.
A few things that might be good to note about the current way of utilising WGS data are that 1) it is often based on genotype imputation, 2) it is often pruned based on Minor Allele Frequencies and LD, and 3) it does not necessarily contain all structural variation and 4) it is more prone to genotyping errors compared to SNP-chip genotyping.

Ad 1) In Paper I, we concluded that part of why increasing the accuracy to a higher density did not increase the prediction accuracy could be due to imputation errors. Many of the same factors that affect genomic prediction accuracy also affect imputation accuracy. The reference population may be too small, or the relationship between the reference animals and the animals to be imputed is low. In that case, imputation accuracy will be reduced (Ventura et al., 2014). Imputing from a low to a high marker density will also be more prone to errors than when the difference in marker densities are smaller (Larmer et al., 2011; van Binsbergen et al., 2014).

Ad 2) One of the arguments for WGS is that it contains all the causative mutations. However, the raw WGS is massive with millions of markers, which is computationally challenging to handle. The markers are usually pruned through, e.g., LD, MAF and other quality control measures, so that the final set of markers is reduced. However, the pruning could potentially remove some of the causative mutations without us realising this.

Ad 3) That not all structural variates are included. WGS used in studies for breeding value estimations has gone through a pipeline of quality checks to make the data more like the other genotypes, meaning that CNVs, indels, deletions, etc., are mostly excluded.

Ad 4) The technology for Whole Genome Sequencing is more fragile than SNP-chip genotyping and more prone to genotyping errors. It also requires a good and stable DNA quality for extractions so that the DNA strands stay long and are not fragmented in the genotyping process (Pérez-Enciso, Rincón, & Legarra, 2015; Taylor et al., 2016)

Does this mean that WGS is unsuitable for animal breeding programs? Absolutely not. Nevertheless, it does mean that there is room for improvement. One thing is that if the cost of WGS goes down, the need for imputation would also be reduced, and the imputation accuracy would be higher with more reference animals. We could also develop new pipelines and software better adapted to WGS to ensure we can get better information from the data and include all the polymorphisms. In addition, it shows that the breeding programs should have a plan for optimising the use of the WGS information and ensure high quality DNA samples when going towards the use of WGS data.

In Paper III, for instance, one could not find a big difference between the WGS and the other two high-density genotype marker data when the BayesGC method was applied. However, WGS seemed to slightly benefit from the use of large reference populations compared to the other marker data, suggesting that we need extensive reference data sets to best use WGS data. One issue with WGS data is the significant number of markers even after pruning with QC. How do we distinguish the markers that affect the traits and those that do not? In paper III, we investigated the use of pCADD scores, a relatively new way of ranking Single Nucleotide Variants (SNVs) based on their putative deleteriousness. In Paper III, we used a pCADD score of about 11, a score selected to have a similar number of SNPs compared to the High-Density genotype (~400K). If the SNPs with a higher pCADD score were more likely to be in biologically active regions, they should be more likely to be close to causative mutations. However, the results in paper III showed hardly any difference in accuracy between pCADD and HD SNPs. Even when paired with a linear model such as GBLUP, the pCADD SNPs had a lower prediction accuracy than the High-Density genotypes. Further research in pCADD scores could use fewer SNPs with a higher pCADD score to find the deleterious variants and then pair these with a set of densely and evenly distributed markers across the genome. The other marker set would account for other potential causative mutations not captured through pCADD but through markers in high LD with QTL.

One option that could have been interesting to explore further with WGS is the Posterior probabilities of a SNP having an effect or not provided by the Bayesian methods. These posterior probabilities showcase which markers were utilised in the model and could have great potential for QTL discoveries (Irene van den Berg, Fritz, & Boichard, 2013).

8.2 Factors affecting the accuracy of genomic prediction

8.2.1 Marker density

Decent prediction accuracies have been found for marker panels as low as 2K (Kriaridou, Tsairidou, Houston, & Robledo, 2019). However, prediction accuracy increases until about 50K marker panels, which are common in, for instance, commercial cattle and pig breeding. The results have not shown as much increase in prediction accuracy when further increasing the marker density to high-density (Su et al., 2012) and WGS marker panels (Van Binsbergen et al., 2015). One of the reasons for this could be that for GBLUP when constructing a genomic relationship matrix, 50K markers are often enough to accurately model the relationship between the animals on a genomic level. Paper I saw a slight increase in prediction accuracy when increasing from 200 to 700K markers only for the across-family reference population. Hence, prediction over more considerable genetic distances requires higher marker densities. Paper III did not find a big difference in accuracy between HD and WGS marker densities for across, multi or within line reference populations.

8.2.2 Relationship between animals

The relationship between the reference population and the validation population is significant for prediction accuracy (Clark et al., 2012; D. Habier, R.L. Fernando, & J.C.M. Dekkers, 2007). **Paper I** shows this effect on the within vs across family reference populations. Having one full-sib in the reference population positively affected the prediction accuracy, exceeding the effects of marker density and prediction method. When the relationship between the reference and validation animals is low, for instance, when predicting across lines, there is evidence that one needs a higher marker density and a more extensive reference population size to get a sufficient prediction accuracy (van den Berg, Meuwissen, MacLeod, & Goddard, 2019). Paper III shows that a reference population of 30K animals from across-line animals was needed to achieve accuracy close to that of a within-line reference population size of 1K. In this case, the value of 1 within-line animal could be equivalent to 30 or more across-line animals in terms of contribution to prediction accuracy, showing that constructing a relevant reference population is essential, not just the number of animals.

Some issues with prediction across lines are that QTLs might segregate differently in the different populations. Either the QTLs do not segregate at all, or the markers have a different LD to the QTL in one line compared to another. Here the genetic constitution of the different populations could also have an effect. For a line with recent high inbreeding, the genomic segments in common between the animals would be large. This means that a smaller number of markers would be required to estimate the genomic segments of the population. For a more admixed population, however, more markers would be necessary, and, in turn, a larger number of animals would be needed to estimate the marker effects. Paper III saw that the two lines, L and S, had a different effect when used as the across-line reference population. The S-line generally had a higher within-line prediction accuracy than L and added more accuracy when used in multi-line predictions. The S-line is a synthetic breed where several lines have been crossed, possibly showing more considerable contrasts between haplotypes. The L-line, however, has been a purebred line for many generations. As our data had a high SNP-density and a high number of animals, the large population size with more chromosomal segments could give more information for prediction across-breed. Nevertheless, the L-line might not have enough variation between haplotypes to be a valuable reference for the more admixed S-line.

8.2.3 Genetic architecture

The genetic architecture can determine the best ways of predicting the trait. Paper I found the trait Host resistance to Lice to be highly polygenic. For this trait, a method like GBLUP, where all markers are set to have an equally small effect, seems sufficient and most of the M_e effective segments defined by Daetwyler et al. contain QTL. Hence, not much is gained from using a Bayesian variable selection method. The accuracy of Bayesian methods has been shown to be affected by the number and size of QTL for the trait, since with an increasing number of QTL more of the M_e effective segments will contain QTL (Clark, Hickey, & Van Der Werf, 2011). Paper II shows that using a different prior for polygenic and SNP effects could improve the prediction accuracy of some traits. These traits could have QTLs that the model detected, and thus the more considerable emphasis on these QTL in the model helped improve the accuracy of prediction.

8.2.4 Genomic prediction methods

This thesis compared GBLUP with BayesGC (**Paper III**) and BayesC (**Paper I** and **II**). GBLUP is one of the most universally accepted GS methods. It is relatively easy to comprehend and use. The difference between GBLUP and regular pedigree BLUP is using a Covariance matrix based on genomic relationships instead of the pedigree relationships. Genomic information makes it possible to distinguish relationships between animals in full-sib families, such as those used in Atlantic Salmon breeding. In **Paper I,** we had both within-family and across-family predictions. In the past, Salmon breeding for disease traits was dependent on challenge testing, and the breeders could not perform challenge tests on animals that were selection candidates. They were dependent on using sibling performance of full-sib families to estimate breeding values and perform the selection. With GS, challenge tests on the selection candidates can still not be performed, but breeding value estimates for the selection candidate have a higher prediction accuracy of up to 100% (Ødegård et al., 2014).

Another standard method of predicting breeding values is ssGBLUP (Christensen & Lund, 2010), which is used when not all animals are genotyped. For ssGBLUP, the covariance matrix is a combination of pedigree and genotypes, referred to as an **H**-matrix (Legarra, Aguilar, & Misztal, 2009). There have been issues with biases of ssGBLUP (Nordbø, Gjuvsland, Eikje, & Meuwissen, 2019). However, for populations where many animals are not genotyped, ssGBLUP can yield genomic breeding value estimates for all animals and not just the genotyped animals. Most of the increase in accuracy from GBLUP and ssGBLUP is from a better estimation of the covariance matrix, i.e., the relationship matrix between the animals, compared to the pedigree relationship matrix, **A**.

Bayesian methods differentiate the markers and try to fit markers with effects and down-weigh markers that do not have an effect. BayesGC (Meuwissen, van den Berg, & Goddard, 2021) fits both a polygenic effect through a covariance matrix (**G**-matrix) and single SNP effects through a BayesC term. This could further increase the prediction accuracy for traits that have QTL with significant effects. Paper II showed that M3W had 9.2% higher accuracy than GBLUP when using BayesGC with few markers fitted together with a polygenic term. Most of the traits did not significantly increase prediction accuracy when using BayesGC compared to GBLUP. However, BayesGC usually performed slightly better than the other methods.

So far, we only tested single trait prediction, but we could expand BayesGC to multi-trait prediction. It could be relatively straightforward if the assumption is that one SNP with a significant effect would have a large effect on all traits (Karaman, Lund, & Su, 2019; Kemper, Bowman, Hayes, Visscher, & Goddard, 2018). However, if one cannot assume this, the modelling would require sampling which combinations of traits are affected by each SNP, which could give many combinations, especially with many traits.

Other possible developments of Bayesian variable selection methods would be to expand the model to include non-genotyped animals. Then the challenge would be how to infer genotypes on non-genotyped animals. There are methods for imputing these genotypes using MCMC methods (Fernando, Dekkers, & Garrick, 2014). It would also be possible to exchange the covariance matrix in BayesGC with an H-matrix instead of using a G-matrix to fit the polygenic trait. One of the most significant drawbacks of Bayesian variable selection methods today is the computational demand of running MCMC chains. Further research into developing even more efficient algorithms and parallel computations could further reduce this and increase computer power efficiency as the computer technology develops.

8.3 Further developments

Animal breeding has always been a field that has been adjacent to other fields of science. From botanist Mendel and statistician Fisher to molecular biology, computer science and bioinformatics that we are relying more and more on today. So far, we have just scratched the surface of the underlying genetics of complex traits. Areas where modelling is under development are, for example, Epigenetics (How the environment can affect the genes) and Epistasis (How the genes interact with each other). The more we learn about DNA and bioinformatics, the more accurate genomic predictions will be. Currently, we are estimating markers as if they have two options: either they affect a trait or not. And then, we try to quantify how much of the trait variance the marker explains. We learn from molecular genetics studies that a marker is not necessarily as simple as that. For example, a gene might be a protein that is part of a process in the body, but we also have genes that regulate other genes, turning them on and off (Watson et al., 2008). Epigenetic effects may imply that an animal has a QTL present in the genome, but the gene the QTL may not be expressed in a particular animal. This, of course, is making QTL detection and mapping quite hard, as results are not consistent.

In the long run, it might not be feasible that all QTLs and markers and their effects are known on a molecular level. It is hard to imagine having a statistical model that could take all main effects and interactions into account. Nevertheless, we still need to develop methods that can consider all the available information as the bioinformatics field is developing. Using large reference populations and Whole Genome Sequence data to detect QTLs combined with, for instance, Bayesian variable selection applied to all these interactions could be an exciting way forward. Other methods to consider for WGS data could be Machine learning methods (although Bayesian methods are a form of Machine Learning). With the large amount of data produced from WGS, it could be an option to detect patterns and connections that humans would not be able to detect but machine learning methods do detect.

9 General Conclusions

When predicting Genomic EBVs using high-density or WGS genomic data:

- A large reference population size can help accuracy for both within- and across-line predictions.

- For multi-line prediction, adding more within-line animals is more important than a larger number of across-line animals.

- Although GBLUP, BayesC and BayesGC all yielded similar genomic prediction accuracies, the accuracy of BayesGC was generally as high as or higher than that of GBLUP.

- Within the BayesGC method the accuracies could vary depending on the prior distributions and the genetic architecture of the trait.

- The BayesGC method benefited from a large reference population and was less dependent on the different genotype marker datasets to achieve a high prediction accuracy.

10 References

Andreescu, C., Avendano, S., Brown, S. R., Hassen, A., Lamont, S. J., & Dekkers, J. C. M. (2007). Linkage Disequilibrium in Related Breeding Lines of Chickens. *Genetics*, *177*(4), 2161–2169. https://doi.org/10.1534/genetics.107.082206

Brøndum, R. F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D., & Lund, M. S. (2015). Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of Dairy Science*, *98*(6), 4107–4116. https://doi.org/10.3168/jds.2014-9005

Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution 2010 42:1*, *42*(1), 1–8. https://doi.org/10.1186/1297-9686-42-2

Clark, S. A., Hickey, J. M., Daetwyler, H. D., & van der Werf, J. H. J. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genetics Selection Evolution*, *44*(1), 4. https://doi.org/10.1186/1297-9686-44-4

Clark, S. A., Hickey, J. M., & Van Der Werf, J. H. (2011). Different models of genetic variation and their effect on genomic evaluation. In *Genetics Selection Evolution* (Vol. 43). https://doi.org/10.1186/1297-9686-43-18

D. Habier, R.L. Fernando, & J.C.M. Dekkers. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, *177*(4), 2389–2397.

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., De Los Campos, G., & Hickey, J. M. (2013). *Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking*. https://doi.org/10.1534/genetics.112.147983

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, *185*(3), 1021–1031. https://doi.org/10.1534/genetics.110.116855

de Koning, D.-J. (2016). Meuwissen et al. on Genomic Selection . *Genetics*, *203*(1), 5–7. https://doi.org/10.1534/genetics.116.189795

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

de Roos, A. P. W., Hayes, B. J., Spelman, R. J., & Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics*, *179*(3), 1503–1512. https://doi.org/10.1534/genetics.107.084301

Dekkers, J. C. M. (2007). Prediction of response to marker-assisted and genomic selection using selection index theory. *Journal of Animal Breeding and Genetics = Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie*, *124*(6), 331–341. https://doi.org/10.1111/j.1439-0388.2007.00701.x

Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., … Goddard, M. E. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, *95*(7), 4114–4129. https://doi.org/10.3168/jds.2011-5019

Fernando, R. L., Dekkers, J. C., & Garrick, D. J. (2014). A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genetics, Selection, Evolution : GSE*, *46*(1). https://doi.org/10.1186/1297-9686-46-50

Gautier, M., Faraut, T., Moazami-Goudarzi, K., Navratil, V., Foglio, M., Grohs, C., … Eggen, A. (2007). Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics*, *177*(2), 1059–1070. https://doi.org/10.1534/GENETICS.107.075804

Gianola, D. (2013). Priors in whole-genome regression: The Bayesian alphabet returns.

Genetics, *194*(3), 573–596. https://doi.org/10.1534/genetics.113.151753

Gianola, D., De Los Campos, G., Hill, W. G., Manfredi, E., & Fernando, R. (2009). Additive Genetic Variability and the Bayesian Alphabet. *Genetics*, *183*(1), 347–363. https://doi.org/10.1534/genetics.109.103952

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, *136*(2), 245–257. https://doi.org/10.1007/s10709-008-9308-0

Goddard, M. E., & Hayes, B. J. (2009, June). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, Vol. 10, pp. 381–391. https://doi.org/10.1038/nrg2575

Groß, C., Derks, M., Megens, H.-J., Bosse, M., Groenen, M. A. M., Reinders, M., & de Ridder, D. (2020). pCADD: SNV prioritisation in Sus scrofa. *Genetics Selection Evolution*, *52*(1), 4. https://doi.org/10.1186/s12711-020-0528-9

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-186

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., & Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution*, *42*(1), 5. https://doi.org/10.1186/1297-9686-42-5

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, *92*(2), 433–443. https://doi.org/10.3168/JDS.2008-1646

Henry, D. (2006). *Aristotle on the Mechanism of Inheritance*. 425–455. https://doi.org/10.1007/s10739-005-3058-y

Hozé, C., Fritz, S., Phocas, F., Boichard, D., Ducrocq, V., & Croiseau, P. (2014). *Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population*. https://doi.org/10.3168/jds.2013-7761

Karaman, E., Lund, M. S., & Su, G. (2019). Multi-trait single-step genomic prediction accounting for heterogeneous (co)variances over the genome. *Heredity 2019 124:2*, *124*(2), 274–287. https://doi.org/10.1038/s41437-019-0273-4

Kemper, K. E., Bowman, P. J., Hayes, B. J., Visscher, P. M., & Goddard, M. E. (2018). A multi-trait Bayesian method for mapping QTL and genomic prediction. *Genetics Selection Evolution 2018 50:1*, *50*(1), 1–13. https://doi.org/10.1186/S12711-018-0377-Y

Kriaridou, C., Tsairidou, S., Houston, R. D., & Robledo, D. (2019). Genomic prediction using low density marker panels in aquaculture: performance across species, traits, and genotyping platforms. *BioRxiv*, 869628. https://doi.org/10.1101/869628

Larmer, S., Sargolzaei, M., Ventura, R., & Schenkel, F. (2011). Imputation accuracy from low to high density using within and across breed reference populations in Holstein, Guernsey and Ayrshire cattle. *Cgil.Uoguelph.Ca*. Retrieved from http://cgil.uoguelph.ca/dcbgc/Agenda1203/Imputation from low to high density using within and across breed reference populations in Holstein, Guernsey and Ayrshire cattle.pdf

Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, *92*(9), 4656–4663. https://doi.org/10.3168/JDS.2009-2061

Meuwissen, T., Berg, I. van den, & Goddard, M. (2021). On the use of whole-genome sequence data for across-breed genomic prediction and fine-scale mapping of QTL. *Genetics, Selection, Evolution : GSE*, *53*(1), 19. https://doi.org/10.1186/S12711-021-00607-4

Meuwissen, T., & Goddard, M. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics*, *185*(2), 623–631. https://doi.org/10.1534/genetics.110.116590

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829. Retrieved from http://www.genetics.org/content/157/4/1819.short

Meuwissen, T., van den Berg, I., & Goddard, M. (2021). On the use of whole-genome sequence data for across-breed genomic prediction and fine-scale mapping of QTL. *Genetics Selection Evolution*, *53*(1), 19. https://doi.org/10.1186/s12711-021-00607-4

Nordbø, Ø., Gjuvsland, A. B., Eikje, L. S., & Meuwissen, T. (2019). Level-biases in estimated breeding values due to the use of different SNP panels over time in ssGBLUP. *Genetics Selection Evolution*, *51*(1), 76. https://doi.org/10.1186/s12711-019-0517-z

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). *Frontiers in Genetics*, *5*(NOV), 1–8. https://doi.org/10.3389/fgene.2014.00402

Oldenbroek, K., & van der Waiij, L. (2015). Textbook Animal Breeding and Genetics for BSc students. Retrieved April 5, 2022, from Centre for Genetic Resources The Netherlands and Animal Breeding and Genomics Centre, Groen Kennisnet website: https://wiki.groenkennisnet.nl/display/TAB/

Pérez-Enciso, M., Rincón, J. C., & Legarra, A. (2015). Sequence- vs. chip-assisted genomic selection: Accurate biological information is advised. *Genetics Selection Evolution*, *47*(1), 1–14. https://doi.org/10.1186/S12711-015-0117-5/FIGURES/8

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Schaeffer, L. R. (2006). Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, *123*(4), 218–223. https://doi.org/https://doi.org/10.1111/j.1439-0388.2006.00595.x

Su, G., Brøndum, R. F., Ma, P., Guldbrandtsen, B., Aamand, G. P., & Lund, M. S. (2012). Comparison of genomic predictions using medium-density (∼54,000) and high-density (∼777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. *Journal of Dairy Science*, *95*(8), 4657–4665. https://doi.org/https://doi.org/10.3168/jds.2012-5379

Taylor, J. F., Whitacre, L. K., Hoff, J. L., Tizioto, P. C., Kim, J., Decker, J. E., & Schnabel, R. D. (2016). Lessons for livestock genomics from genome and transcriptome sequencing in cattle and other mammals. *Genetics, Selection, Evolution : GSE*, *48*(1), 59. https://doi.org/10.1186/s12711-016-0237-6

van Binsbergen, R., Bink, M. C. A. M., Calus, M. P. L., van Eeuwijk, F. A., Hayes, B. J., Hulsegge, I., & Veerkamp, R. F. (2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution*, *46*(1), 41. https://doi.org/10.1186/1297-9686-46-41

Van Binsbergen, R., Calus, M. P. L., Bink, M. C. A. M., Van Eeuwijk, F. A., Schrooten, C., & Veerkamp, R. F. (2015). Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution*, *47*(1), 1–13. https://doi.org/10.1186/s12711-015-0149-x

van den Berg, I., Meuwissen, T. H. E., MacLeod, I. M., & Goddard, M. E. (2019). Predicting the effect of reference population on the accuracy of within, across, and multibreed genomic prediction. *Journal of Dairy Science*, *102*(4), 3155–3174. https://doi.org/10.3168/jds.2018-15231

van den Berg, Irene, Fritz, S., & Boichard, D. (2013). QTL fine mapping with Bayes C(π): a simulation study. *Genetics, Selection, Evolution : GSE*, *45*(1), 19. https://doi.org/10.1186/1297-9686-45-19

VanRaden, P. M., Null, D. J., Sargolzaei, M., Wiggans, G. R., Tooker, M. E., Cole, J. B., … Doak, G. A. (2012). Genomic imputation and evaluation using high-density Holstein genotypes. *Journal of Dairy Science*, *96*(1), 668–678. https://doi.org/10.3168/jds.2012-5702

Ventura, R. V, Lu, D., Schenkel, F. S., Wang, Z., Li, C., & Miller, S. P. (2014). Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle1. *Journal of Animal Science*, *92*(4), 1433–1444. https://doi.org/10.2527/jas.2013-6638

Verbyla, K. L., Bowman, P. J., Hayes, B. J., & Goddard, M. E. (2010). Sensitivity of genomic selection to using different prior distributions. *BMC Proceedings*, *4*(S1), 2–5. https://doi.org/10.1186/1753-6561-4-s1-s5

Visscher, P. M., & Goddard, M. E. (2019). From R.A. Fisher's 1918 Paper to GWAS a Century Later. *Genetics*, *211*(4), 1125–1130. https://doi.org/10.1534/genetics.118.301594

Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., & Losick, R. (2008). *Molecular Biology of the Gene* (6th ed.). Pearson.

Wetterstrand KA. (2021). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 2016-09-05. Retrieved March 27, 2022, from www.genome.gov/sequencingcostsdata/

Wientjes, Y. C. J., Veerkamp, R. F., & Calus, M. P. L. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, *193*(2), 621–631. https://doi.org/10.1534/genetics.112.146290

Zhang, C., Kemp, R. A., Stothard, P., Wang, Z., Boddicker, N., Krivushin, K., … Plastow, G. (2018). Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. *Genetics, Selection, Evolution : GSE*, *50*(1), 14. https://doi.org/10.1186/s12711-018-0387-9

Zhou, L., Heringstad, B., Su, G., Guldbrandtsen, B., Meuwissen, T., Svendsen, M., … Lund, M. (2014). *Genomic predictions based on a joint reference population for the Nordic Red cattle breeds*. https://doi.org/10.3168/jds.2013-7580

Zhou, L., Lund, M. S., Wang, Y., & Su, G. (2014). Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. *Journal of Animal Breeding and Genetics*, *131*(4), 249–257. https://doi.org/10.1111/jbg.12089

Paper I

M.H. Kjetså, J. Ødegård, T.H.E. Meuwissen

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

Aquaculture, 526, 735415.

Aquaculture 526 (2020) 735415

Contents lists available at ScienceDirect

Aquaculture

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

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

M.H. Kjetså[a,*], J. Ødegård[b], T.H.E. Meuwissen[a]

[a] *Norwegian University of Life Sciences, Faculty of Biosciences, PO Box 5003, 1432 Ås, Norway*
[b] *Breeding and Genetics, AquaGen, PO Box 1240, 7462 Trondheim, Norway*

ARTICLE INFO

Keywords:
Genomic prediction
Atlantic salmon
Salmon lice
Imputation
Prediction accuracy

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 (215 K and imputed 750 K). 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 215 K to 750 K genotype density. However, for across-family prediction a slight improvement in predictive ability was observed at the higher density compared to the lower.

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 et al., 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 and Veerkamp, 2007; Meuwissen et al., 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 et al., 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

[*] Corresponding author.
E-mail address: maria.kjetsa@nmbu.no (M.H. Kjetså).

https://doi.org/10.1016/j.aquaculture.2020.735415
Received 21 January 2020; Received in revised form 23 April 2020; Accepted 24 April 2020
Available online 25 April 2020

(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 and Meuwissen (2009) found in their simulation study that the accuracy of selection dropped when the challenge test was done only every other generation or only in one generation when using the GBLUP method. This implies that it would be necessary to challenge test every generation to get accurate predictions.

The accuracy of genomic predictions increases with the number of phenotypes relative to the effective number of genomic segments of the population (Daetwyler et al., 2010). Bayesian variable selection methods (Meuwissen et al., 2001; Verbyla et al., 2010) attempt to increase the relative weight of markers being in LD with casual mutation and remove markers that are not linked to causal loci (i.e., not useful for prediction), and thereby reduce the number of marker effects to estimate.

Bayesian selection approaches such as Bayes (A/B/C/R) have been found to have a higher predictive ability in simulation studies, but differences were smaller in studies using real data (Neves et al., 2012). One of the biggest differences between the Bayesian methods and GBLUP is that GBLUP assumes that genetic variance is evenly distributed over SNPs, whilst the Bayesian methods try to differentiate SNPs with respect to their relative importance. In the current study we investigate the BayesC (Habier et al., 2011), and BayesGC models (Iheshiulor et al., 2017). In BayesGC, a polygenic effect and a Bayesian term are fitted simultaneously, so that we account for both numerous SNPs of small effect, as well as a smaller group of SNPs with a potentially larger effect. In contrast to Iheshiulor et al. (2017), who used an iterative conditional expectation (ICE) algorithm for the BayesGC model, we fitted this model using a Gibbs-sampling approach.

The aim of this study was to compare three methods of genomic prediction: Genomic Best Linear Unbiased Prediction (GBLUP), using a genomic relationship matrix, two Bayesian variable selection methods BayesGC 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 215 K SNP genotypes and imputed 750 K 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 220 K Affymetrix genome-wide SNP-chip. The total number of SNPs after quality control was 215,610. A group of parents ($n = 59$) was genotyped with a high-density SNP-chip with 990 K 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 750 K 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 215 K and the 750 K imputed genotypes were used to construct two genomic relationship matrices (**G**-matrix; one using 215 K and one using 750 K), using own software based on VanRaden method 1 (VanRaden, 2008);

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

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:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

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 ($\mathbf{u}$~N(0,$\mathbf{A}\sigma_a^2$) where **A** is the pedigree relationship matrix; **e** is the residual effect, where ($\mathbf{e}$~N(0,$\mathbf{I}\sigma_e^2$), 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 analyzed 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 ($\boldsymbol{u}$~N(0,$\mathbf{G}\sigma_u^2$), 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:

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

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

2.4. BayesGC

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

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

$$\sigma_m^2 = \frac{Fr * \sigma_u^2}{\overline{HET}}$$

where σ_m^2 is the genetic variance explained by a single SNP,

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

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

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

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

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

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,

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

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:

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

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($\boldsymbol{u}$) = $\mathbf{G}\sigma_{pol}^2$. $\mathbf{X}_i$ = 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_m^2). **e** is the residual with variance **e** ~N(0, $\mathbf{I}\sigma_e^2$) where **I** is an identity matrix. The MCMC – chain was run for 4000 burn-in cycles and a total of 20,000 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 (215 K and imputed 750 K). 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:

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

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

2.6. Significance test

To test the models for significant differences in prediction accuracy

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

	acc	SE(acc)	b	π	σ_{pol}^2	σ_m^2	n_{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

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.
σ_{pol}^2 is the variance attributed to the polygenic effect.
σ_m^2 is the variance assumed for a single SNP effect (if fitted in the model).
n_{mrk} is the estimated number of markers fitted in the model based on the π value multiplied by the total number of markers.

we used a bootstrapping procedure (Efron and 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 215 K SNPchip 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 215 K SNPchip and across-family scenario (Table 2), 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 215 K genotypes neither within- nor across-family. For the 750 K 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 750 K 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.

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

	acc	SE(acc)	b	π	σ_{pol}^2	σ_m^2	n_{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

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.
σ_{pol}^2 is the variance attributed to the polygenic effect.
σ_m^2 is the variance assumed for a single SNP effect (if fitted in the model).
n_{mrk} is the estimated number of markers fitted in the model based on the π value multiplied by the total number of markers.

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

	acc	SE(acc)	b	π	σ_{pol}^2	σ_m^2	n_{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.
σ_{pol}^2 is the variance attributed to the polygenic effect.
σ_m^2 is the variance assumed for a single SNP effect (if fitted in the model).
n_{mrk} is the estimated number of markers fitted in the model based on the π value multiplied by the total number of markers.

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

	acc	SE(acc)	b	π	σ_{pol}^2	σ_m^2	n_{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

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.
σ_{pol}^2 is the variance attributed to the polygenic effect.
σ_m^2 is the variance assumed for a single SNP effect (if fitted in the model).
n_{mrk} is the estimated number of markers fitted in the model based on the π value multiplied by the total number of markers.

Increasing genotype density from 215 K to 750 K within family (Tables 1 and 3) had no effect on the accuracy of prediction. However, between the 215 K and 750 K 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 215 K and 750 K 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 and 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 divided between the G-matrix or the SNP-markers. This is one of the reasons we had hoped to see a bigger difference between the models for the across-family prediction scenario.

Other studies showed promising results for a BayesGC type of method. Solberg et al. (2009) fit a polygenic effect using pedigree information and the Bayes B method from Meuwissen et al. (2001) to fit SNP effects. They conclude that fitting a polygenic effect has a small impact on the accuracy of genome-wide EBVs in the generation immediately following phenotyping, but as the generations progress, the predictions with a polygenic effect retain a higher accuracy, and that this persistence in accuracy is significant for higher marker densities. Calus and Veerkamp (2007) found an increase in the prediction accuracy when including a polygenic effect when the SNP density and heritability was high. Calus et al. did not predict over generations and generally had a smaller genome size and lower marker densities than Solberg et al. (2009). Hence, it is expected that including a BayesC and polygenic term increases prediction accuracies, especially as the genetic relationships between the training and evaluation animals decrease. However, both these studies are simulation studies. We found from our study with real data, that there was no significant difference between our models in the across-family scenario compared to the within-family scenario at either genotypic densities.

Ma et al. (2019) found that using a Bayesian model including known QTLs increased the reliability of prediction accuracy regardless of the genetic distance between the reference population and the predicted population. They found that the Bayesian methods had a larger advantage for traits linked to major genes such as milk yield and fat compared to fertility and mastitis that had almost no effect. They also saw that a small reference population ($<$ 1000 individuals) could affect the reliability of the prediction. As we have both a relatively small reference population (~1000 individuals) in addition to a highly polygenic trait, this might have had an impact on why the Bayesian methods did not outperform GBLUP.

Iheshiulor et al. (2017) compared the Bayes GC method with GBLUP and BayesC on real data from cattle. Their BayesGC method used an iterative conditional expectation (ICE) algorithm to fit their BayesC term while we used a Gibbs sampling algorithm. They found that the BayesGC performed marginally better than GBLUP and BayesC for all their traits and for one trait the difference was significant. Iheshiulor et al. (2017) finds that BayesC always performs between GBLUP and BayesGC. Our results showed that the BayesC method performed either the same or worse than BayesGC and the same or slightly better than GBLUP. In other words, the BayesC term did not add prediction accuracy compared with the GBLUP model, which may explain why the BayesGC model did not have an advantage over GBLUP. Moreover, the performance of the Bayesian methods may be affected by the assumption that each SNP explains 0.1% of the genetic variance, which limits the number of SNPs fitted. However, fitting more SNPs would make the use of fitting both a polygenic trait and a Bayes C term redundant, as fitting many small SNPs would be practically the same as fitting polygenic effects. On the other hand, fitting fewer and larger SNPs would not agree with the polygenic nature of the trait. We did, however, test different assumptions for the BayesC method, assuming that each SNP explain $\frac{1}{500}$, $\frac{1}{2000}$ and $\frac{1}{10000}$ of genetic variance. None of these assumptions yielded a significantly different accuracy for the BayesC prediction accuracy and thus the results were not included here.

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 215 K SNPs at our lowest density and so the relationships are expected to be accurately fitted by a 215 K 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 750 K genotypes were imputed, whereas the 215 K 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 215 K 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.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aquaculture.2020.735415.

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.

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.

Cáceres, P., Barría, A., Christensen, K.A., Bassini, L.N., Correa, K., Lhorente, J.P., Yáñez, J.M., 2019. Genome-scale comparative analysis for host resistance against sea lice between Atlantic salmon and rainbow trout. bioRxiv 624031. https://doi.org/10.1101/624031.

Calus, M.P.L., Veerkamp, R.F., 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J. Anim. Breed. Genet. 124, 362–368. https://doi.org/10.1111/j.1439-0388.2007.00691.x.

Daetwyler, H.D., Pong-Wong, R., Villanueva, B., Woolliams, J.A., 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185, 1021–1031. https://doi.org/10.1534/genetics.110.116855.

Efron, B., Tibishirani, R.J., 1994. An Introduction to the Bootstrap [WWW Document]. CRC Press LLC, Boca Rat URL. https://books.google.no/books?hl=en&lr=&id=gLlpIUxRntoC&oi=fnd&pg=PR14&ots=A9BvU8J7F2&sig=rU1bHQeofAkRYvjRIucY5ei_XkQ&redir_esc=y#v=onepage&q&f=false (accessed 11.19.19).

Gjerde, B., Ødegård, J., Thorland, I., 2011. Estimates of genetic variation in the susceptibility of Atlantic salmon (*Salmo salar*) to the salmon louse *Lepeophtheirus salmonis*. Aquaculture 314, 66–72. https://doi.org/10.1016/j.aquaculture.2011.01.026.

Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J., 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12. https://doi.org/10.1186/1471-2105-12-186.

Iheshiulor, O.O.M., Woolliams, J.A., Svendsen, M., Solberg, T., Meuwissen, T.H.E., 2017. Simultaneous fitting of genomic-BLUP and Bayes-C components in a genomic prediction 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, 1819–1829.

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.

Overton, K., Dempster, T., Oppedal, F., Kristiansen, T.S., Gismervik, K., Stien, L.H., 2018. Salmon lice treatments and salmon mortality in Norwegian aquaculture: a review. Rev. Aquac. https://doi.org/10.1111/raq.12299.

Robledo, D., Gutiérrez, A.P., Barría, A., Lhorente, J.P., Houston, R.D., Yáñez, J.M., 2019. Discovery and functional annotation of quantitative trait loci affecting resistance to sea lice in Atlantic Salmon. Front. Genet. 10. https://doi.org/10.3389/fgene.2019.00056.

Rochus, C.M., Holborn, M.K., Ang, K.P., Elliott, J.A.K., Glebe, B.D., Leadbeater, S., Tosh, J.J., Boulding, E.G., 2018. Genome-wide association analysis of salmon lice (*Lepeophtheirus salmonis*) resistance in a North American Atlantic salmon population. Aquac. Res. 49, 1329–1338. https://doi.org/10.1111/are.13592.

Sargolzaei, M., Chesnais, J.P., Schenkel, F.S., 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15, 478. https://doi.org/10.1186/1471-2164-15-478.

Solberg, T., Sonesson, A., Woolliams, J., Degard, J., Meuwissen, T., 2009. Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. Genet. Sel. Evol. 41, 1–8. https://doi.org/10.1186/1297-9686-41-53.

Sonesson, A.K., 2007. Within-family marker-assisted selection for aquaculture species. Genet. Sel. Evol. 39, 301. https://doi.org/10.1186/1297-9686-39-3-301.

Sonesson, A.K., Meuwissen, T.H., 2009. Testing strategies for genomic selection in aquaculture breeding programs. Genet. Sel. Evol. 41, 1–9. https://doi.org/10.1186/1297-9686-41-37.

Torrissen, O., Jones, S., Asche, F., Guttormsen, A., Skilbrei, O.T., Nilsen, F., Horsberg, T.E., Jackson, D., 2013. Salmon lice - impact on wild salmonids and salmon aquaculture. J. Fish 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.

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

Verbyla, K.L., 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.

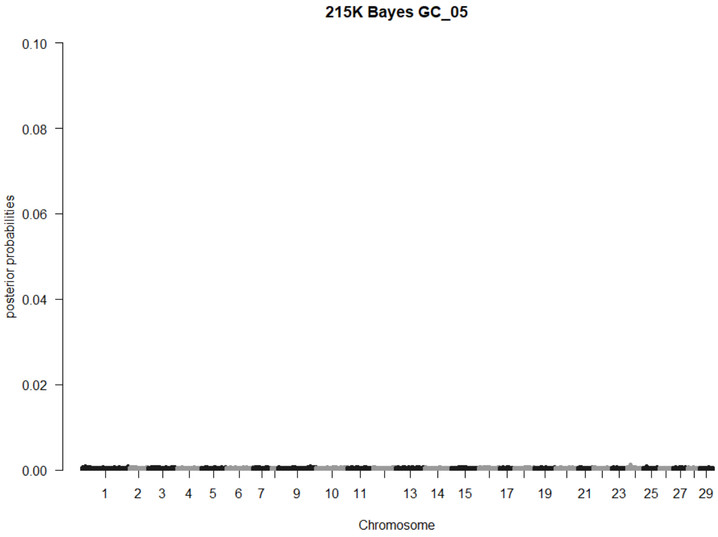
Wang, Y., Lin, G., Li, C., Stothard, P., 2016. Genotype imputation methods and their effects on genomic predictions in cattle. Springer Sci. Rev. 4, 79–98. https://doi.org/10.1007/s40362-017-0041-x.

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.

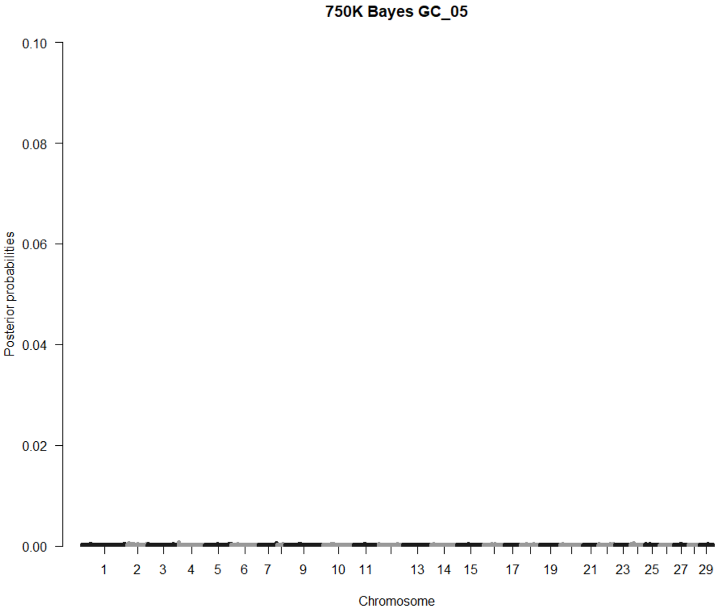
Xu, S., 2003. Theoretical Basis of the Beavis Effect. Genetics 165, 2259–2268.

Appendix A

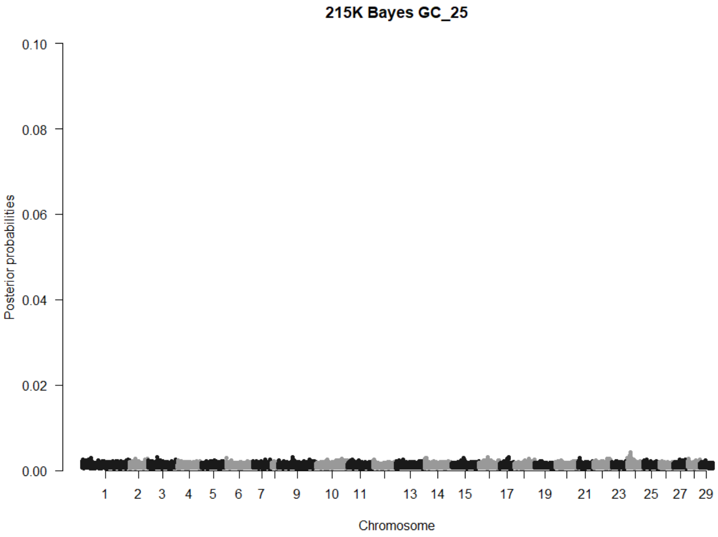
Supplementary Figure 1. Posterior probability distributions of SNPs from method BayesGC_05 at 215K genotype density.



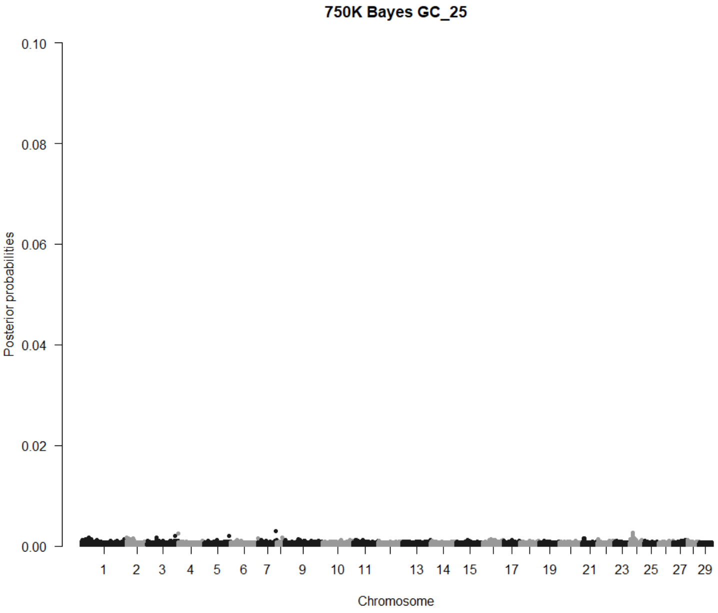
Supplementary Figure 2. Posterior probability distributions of SNPs from method BayesGC_05 at 750K genotype density.



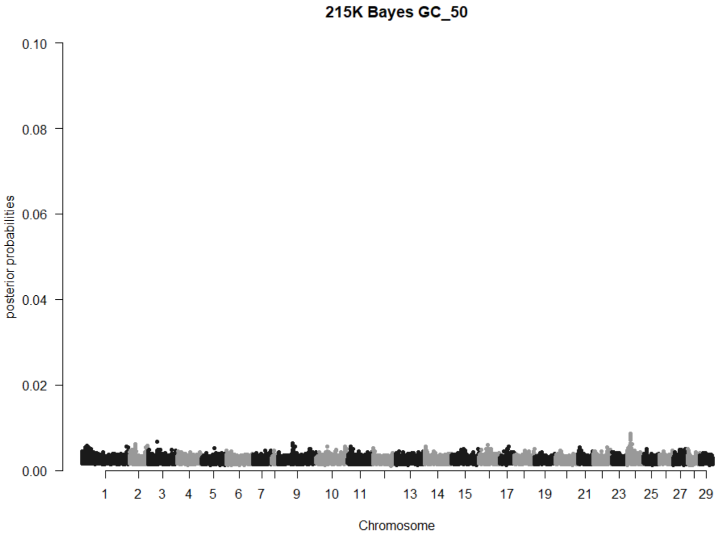
Supplementary Figure 3. Posterior probability distributions of SNPs from method BayesGC_25 at 215K genotype density.



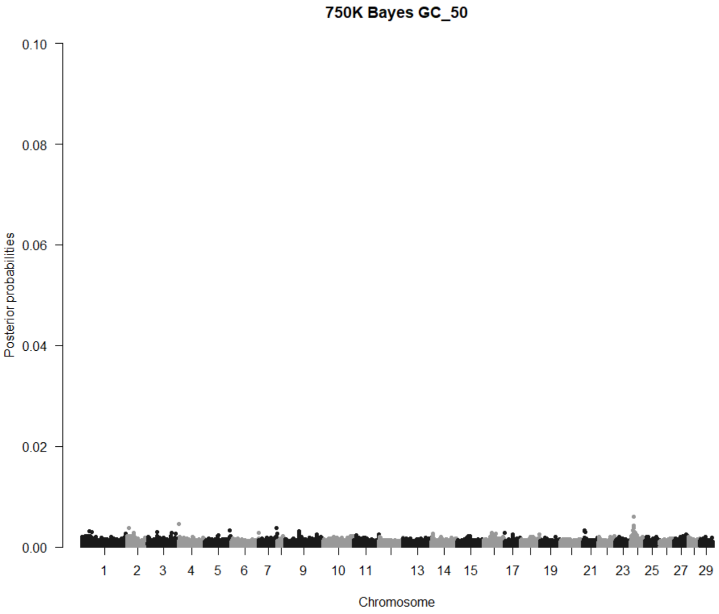
Supplementary Figure 4. Posterior probability distributions of SNPs from method BayesGC_25 at 750K genotype density



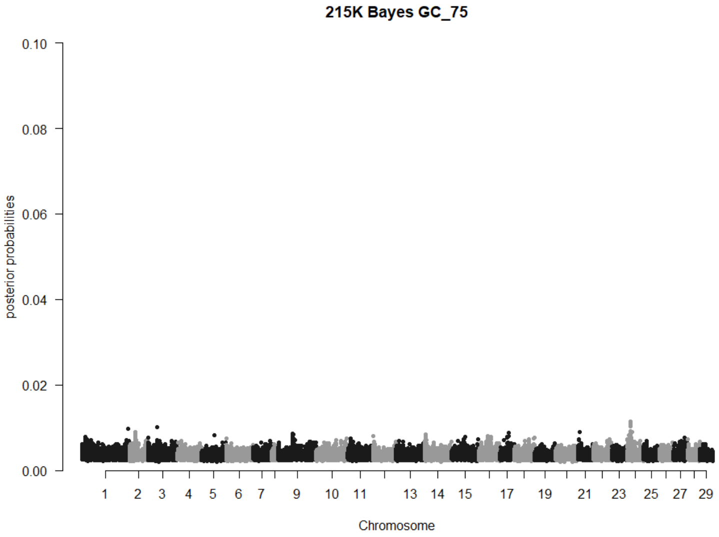
Supplementary Figure 5. Posterior probability distributions of SNPs from method BayesGC_50 at 215K genotype density.



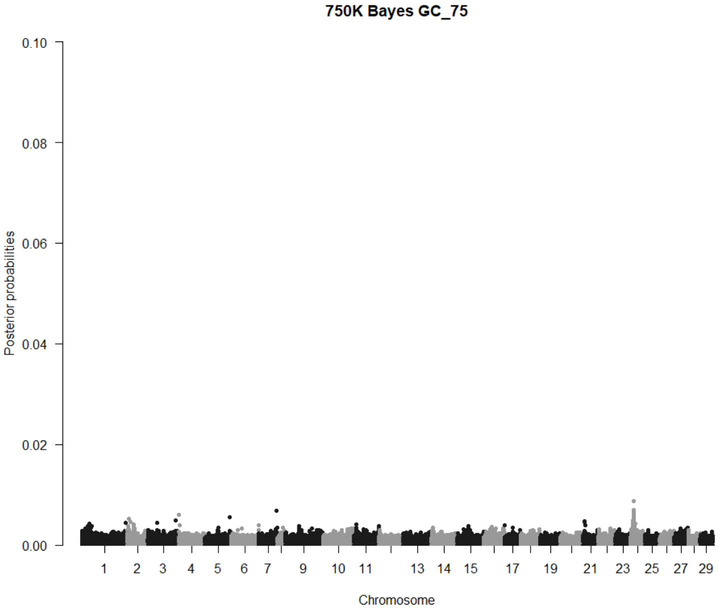
Supplementary Figure 6. Posterior probability distributions of SNPs from method BayesGC_50 using 750K genotype density.



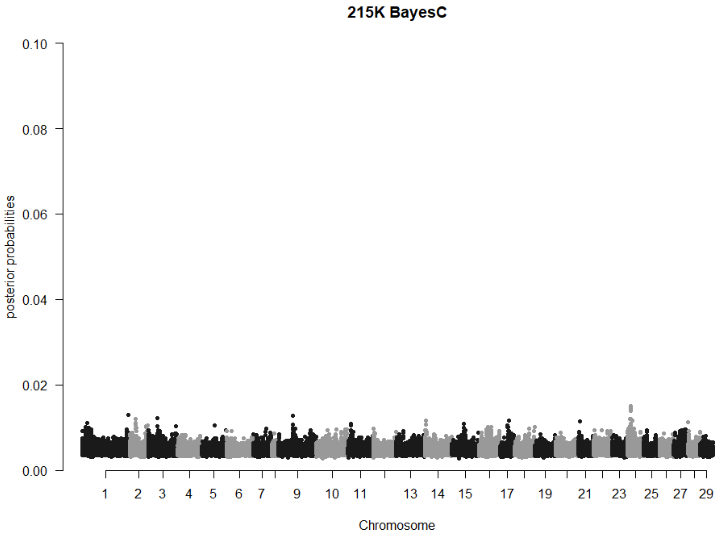
Supplementary Figure 7. Posterior probability distributions of SNPs from method BayesGC_75 using 215K genotype density.



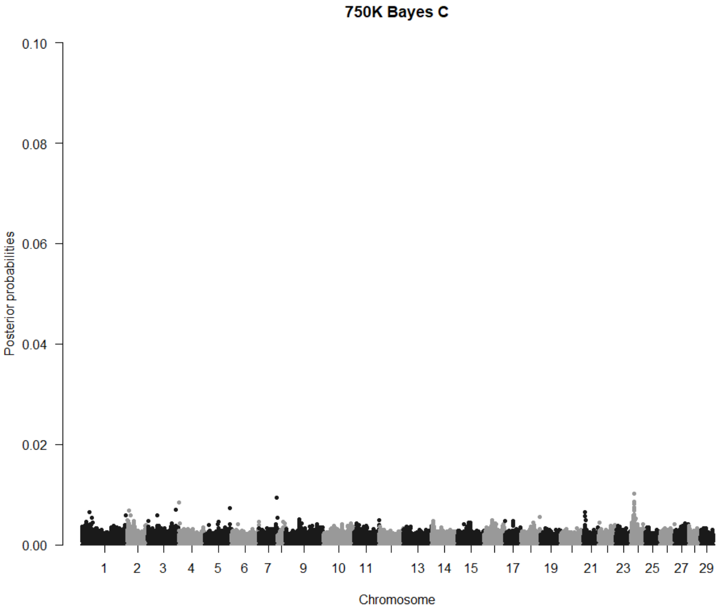
Supplementary Figure 8. Posterior probability distributions of SNPs from method BayesGC_75 using 215K genotype density.



Supplementary Figure 9. Posterior probability distributions of SNPs from method BayesC using 215K genotype density.



Supplementary Figure 10. Posterior probability distributions of SNPs from method BayesC using 750K genotype density.



Paper II

M. V. Kjetså, A. B. Gjuvsland, Ø. Nordbø, E. Grindflek, T.H.E. Meuwissen

Accuracy of genomic prediction of maternal traits in pigs using Bayesian variable selection methods.

Manuscript submitted to Journal of Animal Breeding and Genetics

Accuracy of genomic prediction of maternal traits in pigs using Bayesian variable selection methods

Maria V. Kjetså[1*], Arne B. Gjuvsland[2], Øyvind Nordbø[2], Eli Grindflek[2] and Theo Meuwissen[1]

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

[2] Norsvin SA, Storhamargata 44, 2317 Hamar, Norway

*Corresponding author

Corresponding E-mail address: maria.kjetsa@nmbu.no

Summary

The aim of this study was to compare three methods of genomic prediction: GBLUP, BayesC and BayesGC for genomic prediction of six maternal traits in Landrace sows using a panel of 660K SNPs. The effect of different priors for the Bayesian methods were also investigated. GBLUP does not take the genetic architecture into account as all SNPs are assumed to have equally sized effects and relies heavily on the relationships between the animals for accurate predictions. Bayesian approaches rely on both fitting SNPs that describe relationships between animals in addition to fitting single SNP effects directly. Both the relationship between the animals and single SNP effects are important for accurate predictions. Maternal traits in sows are often more difficult to record and have lower heritabilities. BayesGC was generally the method with the higher accuracy, although its accuracy was for some traits matched by that of GBLUP and for others by that of BayesC. For piglet mortality within 3 weeks, BayesGC achieved up to 9.2% higher accuracy. For many of the traits however, the methods did not show significant differences in accuracies.

KEYWORDS

Bayesian genomic prediction, genomic prediction accuracy, genomic selection, maternal traits

1 INTRODUCTION

Genomic Prediction (GP) (Meuwissen, Hayes, & Goddard, 2001) is a method to predict breeding values (GEBVs) in animal and plant breeding. GP predicts the GEBVs by using a reference population of animals with both phenotype and marker information to estimate marker effects. Meuwissen et al. proposed three methods for genomic prediction: two Bayesian variable selection methods (BayesA and B) and a linear marker effects model estimating marker effects from Single Nucleotide Polymorphisms (SNPs) using Best Linear Unbiased Prediction (BLUP), referred to as SNP-BLUP. An alternative method of SNP-BLUP is to use a marker-derived genomic relationship matrix (often called a **G**-matrix) as a covariance matrix when solving Mixed Model Equations (MME) (VanRaden, 2008) referred to as Genomic Best Linear Unbiased Predictions (GBLUP). The two methods, SNP-BLUP and GBLUP, are mathematically equivalent (Strandén & Garrick, 2009; VanRaden, 2008).

All markers are assumed to have equal weight in the prediction for the linear models, while Bayesian methods try to differentiate SNPs relative to their importance. Markers associated with causal mutations get a higher relative weight, and markers not linked to causal loci are down-weighted, thus only giving weights to the most important SNPs (Meuwissen, Hayes, & Goddard, 2001; Verbyla, Bowman, Hayes, & Goddard, 2010). Several alternative Bayesian variable selection methods are proposed, often referred to as the “Bayesian Alphabet” (Gianola, De Los Campos, Hill, Manfredi, & Fernando, 2009). Differences between the methods are the prior distributions used for the estimation of SNP effects. For example, BayesA uses one t-distribution for SNP-effects, while BayesB has a mixture of a t-distribution with probability π, and a null-effect with probability 1-π. BayesC (Habier, Fernando, Kizilkaya, & Garrick, 2011). is similar to BayesB, as both have a mixture distribution prior, where one has a null effect. However, BayesC uses a normal distribution instead of a t-distribution and assumes a common variance for all SNPs, while BayesB assumes SNP-specific variances. BayesR uses four normal distributions, where one of them has a null effect (Erbe et al., 2012). The recently proposed BayesGC method (Meuwissen, van den Berg, & Goddard, 2021), fits a polygenic effect through a **G**-matrix in addition to a BayesC term. Hence, BayesGC fits many SNPs with a small effect through the **G**-matrix and a group of SNPs selected by the model with more significant effects through the BayesC term.

In this paper we look at genomic prediction of maternal traits in landrace pigs which are considered complex traits with a low to moderate heritability and explore the effect of the genetic architecture on the prediction accuracy. Specifically, we look at the traits; total number of born piglets (**TNB**), number of stillborn piglets (**STB**), piglet mortality within 3 weeks, i.e., number of piglets dead after birth and until 3 weeks (**M3W**), total litter weight at 3 weeks (**LW3W**), sow shoulder lesions (**SHL**) and the sow's body condition score (**BCS**). These maternal traits were included in the breeding goal for Topigs Norsvin at the time of recording (Eriksen, 2018).

Maternal traits in pigs are related to the sow's ability to produce and raise offspring. Maternal traits are essential for efficiency in pig production, the economy, and animal welfare (Ocepek & Andersen, 2017). An ideal sow produces a litter of piglets corresponding to the number of functional teats available, and all the piglets born survive until weaning. Furthermore, the piglets should grow evenly, and the sow should not spend all her resources on the litter, implying that she maintains a good body condition score and does not develop shoulder lesions.

Simulation studies have shown great potential of using genomic prediction methods to predict maternal traits in pigs (Lillehammer, Meuwissen, & Sonesson, 2011, 2013). Although few studies have reported genomic prediction accuracies for maternal traits in pigs (Tan et al., 2017), there are very few that have reported prediction accuracies for Bayesian genomic prediction methods for maternal traits. Some have looked at Bayesian methods in growth and reproduction traits (Song et al., 2017) and slaughter traits (Salek Ardestani, Jafarikia, Sargolzaei, Sullivan, & Miar, 2021). Although the basis of inheritance and breeding is the same across livestock species, their differences in breeding structure, genetic architecture and trait biology make it important to study the different prediction methods across the species (Samorè & Fontanesi, 2016).

This study aimed to determine the prediction accuracy of six maternal traits in Landrace sows using a panel of 660k SNP markers and a large reference population (9-15 thousand reference animals). The study also compares three methods of genomic prediction: GBLUP (VanRaden, 2008), BayesC (Habier et al., 2011) and BayesGC (Meuwissen, van den Berg, & Goddard, 2021).

2 MATERIAL AND METHODS

2.1 Phenotypic data

The phenotypic data consisted of records from 15,703 unique individual Landrace sows with at least one record for one of the six traits; total number of born piglets (**TNB**), number of stillborn piglets (**STB**), piglet mortality within 3 weeks, i.e., number of piglets dead after birth and until 3 weeks (**M3W**), total litter weight at 3 weeks (**LW3W**), sow shoulder lesions (**SHL**) and the sow's body condition score (**BCS**). Of the 15,703 sows, 10,306 had records for all six traits. The traits were recorded between 2008 and 2019. Each of the different traits had between 10,611-15,690 phenotypic records (see Table 1).

Table 1. Number (n) of animals with records for each trait and partition into the reference and validation population, and mean (m) number of parity records in each trait.

Trait[1]	Total n	n reference	n validation	m parity
TNB	15,690	14,513	1,177	2.5
STB	15,690	14,513	1,177	2.5
M3W	10,611	9,466	1,145	1.7
LW3W	10,804	9,656	1,148	1.7
SHL	15,084	13,934	1,150	2.2
BCS	15,084	13,933	1,151	2.2

[1] Total number born (TNB), number of stillborn piglets (STB), piglet mortality within 3 weeks, i.e., number of piglets dead after birth and until 3 weeks (M3W), total litter weight at 3 weeks (LW3W), sow shoulder lesions (SHL) and the sow's body condition score (BCS).

Yield Deviations (VanRaden & Wiggans, 1991) for the six traits were derived from the commercial breeding value evaluations from Topigs Norsvin. There were multiple records for each trait, as we had one YD for each parity. The maximum number of parities recorded for each trait were 6, and the mean number of parities recorded for each trait is shown in Table 1. Because the software used for the Bayesian variable selection models (Meuwissen et al., 2021) could not handle multiple records per animals, we used the average YD for each sow, with a weighting of each record corresponding to the effective number of records calculated by the formula $\frac{n*(1+\text{[redacted]}}{(n+\text{[redacted]}}$ where [redacted] is σ_e^2/σ_{pe}^2 and *n* is the number of records for each individual, σ_e^2 is the residual variance and σ_{pe}^2 is the permanent environmental variance ([redacted] was obtained from Topigs Norsvin's breeding value evaluation).

2.2 Genotype data

The sows were genotyped with varying SNP densities and imputed to a 660K-genotype density. Of the 15703 sows, 526 were genotyped on a 10K chip (GGP Porcine LD), and the rest were genotyped on medium density chips: Illumina PorcineSNP60 (60K) and two Illumina GeneSeek custom chips (80K and 50K). All genotypes were imputed using Fimpute v2.2 (Sargolzaei, Chesnais, & Schenkel, 2014), first to the 50K chip, and then to the 660K Axiom Porcine Genotyping Array with reference genotypes from 467 Landrace animals. After quality control, the 660K High-Density genotype data had a total of 429, 403 SNPs with MAF>0.01.

2.3 Validation and reference data

The ~1,000 youngest sows were used for validation of the predictions, in order to imitate a typical genomic breeding program where one wishes to predict the breeding values of young animals before they have their own recorded traits. This was done by masking their phenotypic records in the analysis and using them for validation. The number of validation sows was between 1,145 and 1,177 (Table 1). The rest of the animals was used as the reference data with both phenotypic and genotype records. Our smallest reference dataset consisted of 9,466 animals for the trait M3W, and the largest of 14,513 animals for traits TNB and STB (Table 1).

2.4 Prediction accuracy and regression coefficients

The accuracy of prediction for all methods was estimated as:

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

And the bias (coefficient of regression) was the calculated slope (*b*) of the linear regression Y = *a* + *b*X where a is the intercept, Y is the Yield Deviation (YD) and X is the Genomic Estimated Breeding Value (GEBV) of the sows in the validation datasets, estimated with only marker information and not phenotypic records. h^2 is the heritability of the trait and was estimated on the full dataset.

2.5 Variance components and GBLUP

We estimated variance components for each trait using the pedigree relationship matrix and the DMUAI package from the DMU software (Madsen & Jensen, 2013). The variance components were estimated on the full dataset (i.e., both the reference and validation animals). The model for the variance component estimations were as follows:

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

where **y** is a vector of the average YD of a sow, **1** is a vector of ones corresponding to the size of **y**, μ is the mean, **Z** is a design matrix linking individuals to the phenotype, ***u*** is the random effect of the individual animal (**u**~N(0, **A**σ_u^2), where **A** is the pedigree relationship matrix and **e** = the residual effect (**e**~N(0, **D**σ_e^2)), where **D** is a diagonal matrix where the diagonals are the inverses of the effective number of records. The same model was used for the GBLUP analyses except that the individual animal effect was modelled as (**u**~N(0, **G**σ_u^2)). The variance components used were from the above pedigree-based estimates. The **G**-matrix was calculated using the VanRaden method 1 (VanRaden, 2007).

2.6 BayesC

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

$\mathbf{y} = \mathbf{1}\mu + \sum_i \mathrm{I_i}\mathbf{x}_i\mathrm{s_i} + \mathbf{e}$

where **y** = a vector of Yield Deviations, **1** is a vector of ones, μ is overall mean, $\mathbf{x}_i$ is a vector of genotypes for SNP *i* containing -2p_i for homozygote individuals, 1-2p_i for heterozygotes, and 2-2p_i for the alternative homozygote genotype with p_i being the allele-frequency of SNP *i*, and I_i is an indicator of whether the SNP *i* is in the model in a particular MCMC-cycle or not (0/1), where the prior probability of I_i being equal to 1 is denoted by π (values in Table 2), $\mathrm{s_i}$ is the SNP effect, where if the SNP *i* is in the model: $\mathrm{s_i}$ ~N(0, σ_m^2), **e** is the residual with variance **e** ~N(0 **D**σ_e^2)) where **D** is an diagonal matrix where the diagonals are the inverses of the effective number of records and σ_e^2 is the residual variance estimated from the variance component estimations (Table 5). The Markov Chain Monte Carlo (MCMC) – chain was run for 20,000 Gibbs-cycles using 4,000 burn-in cycles, in two distinct chains.

We used the same variance components as for the GBLUP analyses, however the total genetic variance σ_u^2 was partitioned. In the following, we describe how the total genetic variance σ_u^2 (see Table 4) is partitioned over the fitted SNPs for the Bayes C method:

$\sigma_m^2 = \frac{Fr * \sigma_u^2}{\overline{HET}}$

where σ_m^2 is the genetic variance explained by a single SNP,

Fr is the fraction of the total genetic variance explained by a single fitted SNP, i.e., 1/1,000 when we assume each SNP explains 1/1,000th of the genetic variance. We test different values of Fr, namely 1/100, 1/500, 1/1,000, 1/5,000 and 1/10,000 respectively.

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

Where p_i is the allele frequency of locus i and N_{loci} is the total number of loci.

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

$\pi_c = \frac{1/Fr}{N_{loci}}$

Such that the total genetic variance is $\sigma_u^2 = \pi_c \cdot N_{loci} \cdot \overline{HET} \cdot \sigma_m^2$.

Table 2. π values used for BayesC and BayesGC methods at different fraction of total genetic variance explained by a single fitted SNP (Fr).

Fr	BayesGC_10[1]	BayesGC_50[2]	BayesGC_90[3]	BayesC
1/100	0.00002	0.00012	0.00021	0.0002
1/500	0.00012	0.00058	0.00105	0.0012
1/1,000	0.00023	0.00116	0.00210	0.0023
1/5,000	0.00116	0.00582	0.01048	0.0116
1/10,000	0.00233	0.01164	0.02096	0.0233

[1]BayesGC_10 is Bayes_GC with 10% marker variance and 90% polygenic variance.
[2]BayesGC_50 has 50% marker variance and 50% polygenic variance.
[3]BayesGC_90 has 90% marker variance and 10% polygenic variance.

2.7 BayesGC

The BayesGC model is as follows:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \sum_i \mathrm{I}_i \mathbf{x}_i s_i + \mathbf{e}$$

where **y** is a vector of the Yield Deviations, **1** is a vector of ones, μ is overall mean, **Z** is a design matrix that links individuals to the **y**, *u* is a vector of random polygenic effects with variance V(*u*) = $\mathbf{G}\sigma^2_{\mathrm{pol}}$, $\mathbf{x}_i$ is the vector of genotypes for SNP i coded as for BayesC. $\mathrm{I_i}$ 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 being equal to 1 is π (listed in Table 2), $\mathrm{s_i}$ is the SNP effect, where if the SNP *i* is in the model: $\mathrm{s_i} \sim N(0, \sigma^2_{\mathrm{m}})$, **e** is the residual with variance **e** $\sim N(0, \mathbf{D}\sigma_e{}^2)$ where **D** is an diagonal matrix where the diagonals are the inverses of the effective number of records and $\sigma_e{}^2$ is the residual variance estimated from the variance component estimations (Table 5). The MCMC – chain was run for 4,000 burn-in cycles and a total of 20,000 Gibbs-cycles for two independent chains. The EBVs from the two Gibbs-chains for both BayesC and BayesGC 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.

Table 3. priors of variance of a single marker (σ^2_{m}) used in the BayesC and BayesGC methods under the different priors for the fraction of total genetic variance explained by a single fitted SNP (Fr) where $\sigma^2_m = \frac{Fr*\sigma^2_u}{HET}$ for each trait.

Fr	σ^2_{m}					
	TNB[1]	STB[1]	M3W[1]	LW3W[1]	SHL[1]	BCS[1]
1/100	0.03196	0.00509	0.00518	0.53802	0.00259	0.00274
1/500	0.00639	0.00102	0.00104	0.10760	0.00052	0.00055
1/1,000	0.00320	0.00051	0.00052	0.05380	0.00026	0.00027
1/5,000	0.00064	0.00010	0.00010	0.01076	0.00005	0.00005
1/10,000	0.00032	0.00005	0.00005	0.00538	0.00003	0.00003

[1] Total number born (TNB), number of stillborn piglets (STB), piglet mortality within 3 weeks, i.e., number of piglets dead after birth and until 3 weeks (M3W), total litter weight at 3 weeks (LW3W), sow shoulder lesions (SHL) and the sow's body condition score (BCS).

The BayesGC model basically fits the previous two models (GBLUP and BayesC) simultaneously, i.e., it fits a polygenic and a BayesC term. The polygenic effect is fitted using the genomic relationship matrix (**G**) as in the GBLUP model. The BayesC term assumes SNPs to have normally distributed effects with probability (π) or an effect of 0 with probability (1-π).

Table 4. priors for variance attributed to the polygenic effect for the different traits for the different BayesGC methods.

	Trait	BayesGC_10[1]	BayesGC_50[2]	BayesGC_90[3]
	TNB	0.944	0.525	0.105
	STB	0.150	0.084	0.017
	M3W	0.153	0.085	0.017
σ^2_{pol}	LW3W	15.89	8.830	1.766
	SHL	0.077	0.043	0.009
	BCS	0.081	0.045	0.009

[1]BayesGC_10 is BayesGC with 10% marker variance and 90% polygenic variance.
[2]BayesGC_50 has 50% marker variance and 50% polygenic variance.
[3]BayesGC_90 has 90% marker variance and 10% polygenic variance.

In the following we describe how the total genetic variance σ_u^2 is partitioned over the fitted SNPs and the polygenic effect. For BayesGC, we need an assumption on the fraction of the variance that is explained by the individually fitted SNPs in the BayesC term of the model. 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 the BayesC term, 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 method. Four different values of q were tested for BayesGC, q = 0.1, 0.5 and 0.9 corresponding to the BayesC term with fitted marker effects explaining 10%, 50% or 90% of the total genetic variance (denoted BayesGC_10, BayesGC_50, BayesGC_90, respectively), with the rest of the variance 1-q explained by the polygenic effect through the **G**-matrix. The values of σ_{m}^2 used are shown in table 3 and the values of σ_{pol}^2 are shown in Table 4.

3 RESULTS

The heritabilities of the traits ranged from 0.09 (M3W) to 0.34 (SHL) (Table 5). M3W had the lowest heritability of 0.09, followed by STB and TNB with moderate heritabilities of 0.13 and 0.19. LW3W, BCS, and SHL had the highest heritabilities with 0.31, 0.31 and 0.34 respectively (Table 5).

Table 5. The estimated total genetic variance ($\boldsymbol{\sigma_u^2}$) , residual variance ($\boldsymbol{\sigma_e^2}$) and heritabilities ($\boldsymbol{h^2}$) for the six maternal traits.

Trait[1]	$\boldsymbol{\sigma_u^2}$	$\boldsymbol{\sigma_e^2}$	$\boldsymbol{h^2}$
TNB	1.049	4.490	0.189
STB	0.167	1.125	0.130
M3W	0.170	1.791	0.087
LW3W	17.66	39.99	0.306
SHL	0.085	0.163	0.343
BCS	0.090	0.203	0.307

[1] total number born (TNB), number of stillborn piglets (STB), piglet mortality within 3 weeks, i.e., number of piglets dead after birth and until 3 weeks (M3W), total litter weight at 3 weeks (LW3W), sow shoulder lesions (SHL) and the sow's body condition score (BCS).

For the trait Total Number Born (TNB) the highest accuracy was achieved at 0.610-0.611 for GBLUP and BayesGC_10 (Table 6) and the method giving the lowest prediction accuracy is BayesC (Fr 1/100) (Figure 1) which achieved an accuracy of 0.515 for TNB. For all the Bayesian methods (BayesGC_10, BayesGC_50, BayesGC_90 and BayesC), fitting more SNPs (Fr = 1/10000) gave the highest accuracy of prediction for the trait TNB. The accuracy of prediction for Stillborn (STB) (Figure 2), is lower than the other traits.

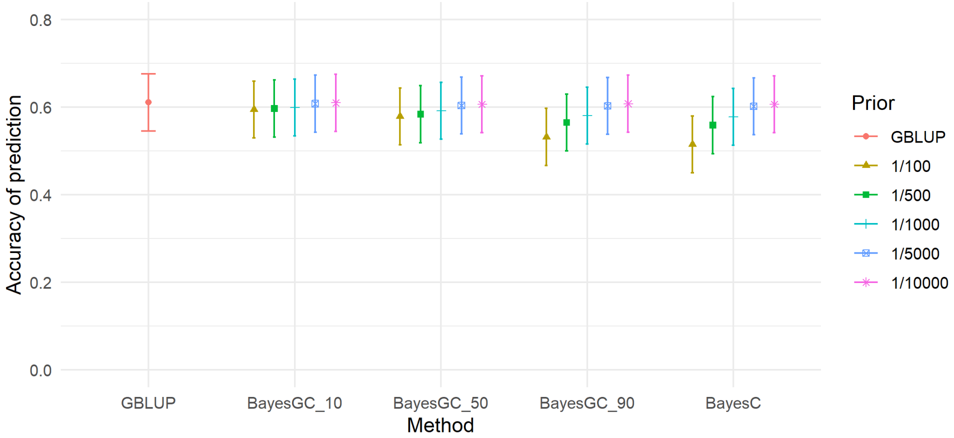


Figure 1. The accuracy of prediction for the trait TNB, from the different prediction methods at the different priors for fraction of variance explained by a single SNP (bars denote standard errors).

For STB, there were also minor, but not significant differences in accuracy between the methods, with the highest accuracy achieved by BayesGC_10 (Fr 1/5,000) and GBLUP at 0.318 and the lowest accuracy for STB was BayesC (Fr 1/100) at 0.272 (see Figure 2). M3W (Figure 3) is the trait with the largest differences between the methods. GBLUP and BayesC (Fr 1/500) had an accuracy of 0.441 and 0.464 respectively, while the highest accuracy from the BayesGC methods was achieved by BayesGC_50 (Fr 1/100) with an accuracy of 0.484 (Table 6), making a difference of 9.8% between GBLUP and BayesGC.

However, the difference was not significant. For the trait LW3W (Figure 4) all the methods had high accuracies of 0.717, 0.722 and 0.718 for the methods GBLUP, BayesGC_90 (Fr 1/10,000) and BayesC (Fr 1/10,000) respectively.

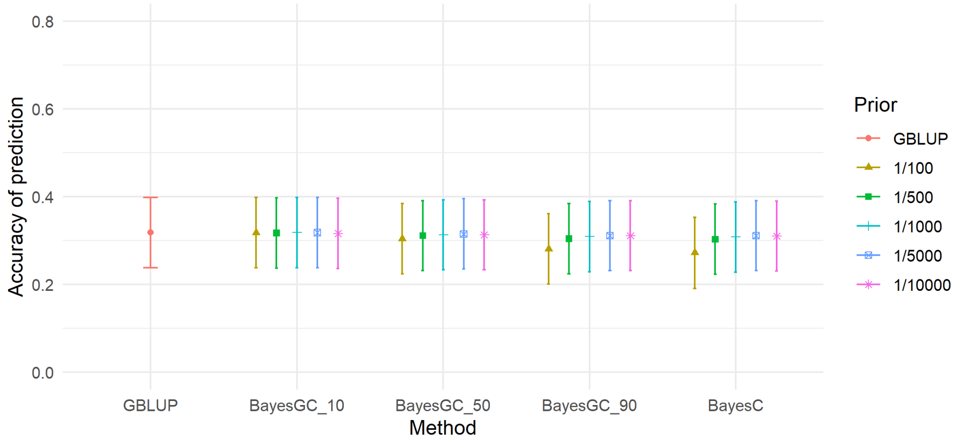


Figure 2. The accuracy of prediction for the trait STB, from the different prediction methods at the different priors for fraction of variance explained by a single SNP (bars denote standard errors).

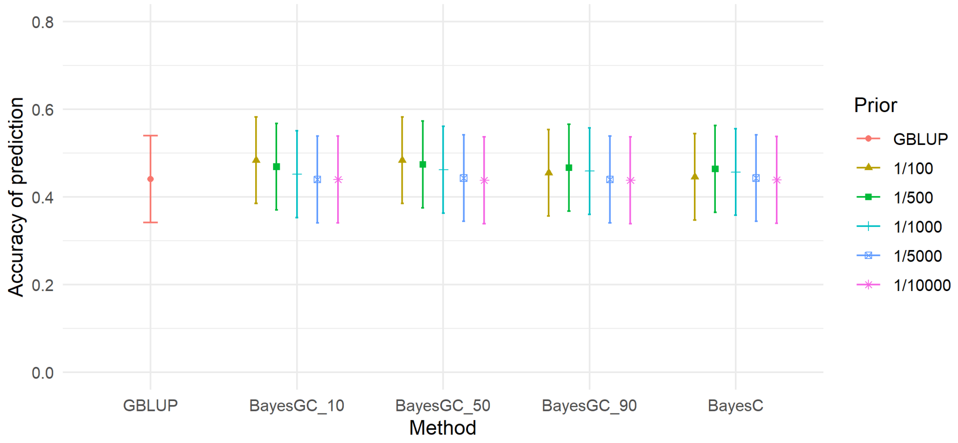


Figure 3. The accuracy of prediction for the trait M3W, from the different prediction methods at the different priors for fraction of variance explained by a single SNP (bars denote standard errors).

Shoulder Lesions (SHL) (Figure 5) showed a prediction accuracy of 0.406 and 0.409 for GBLUP and BayesC while the highest accuracy for BayesGC was 0.418 for BayesGC_50

(Fr1/100). Trait BCS (Figure 6) also had minor differences between the methods and obtained the highest accuracy from the BayesC and BayesGC_90 (Fr 1/10,000) methods with an accuracy of 0.518 for both methods, while GBLUP obtained an accuracy of 0.511.

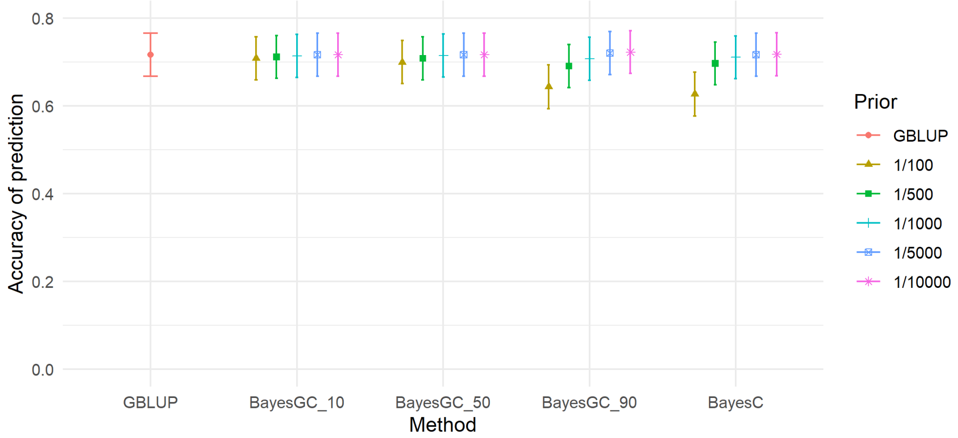


Figure 4. The accuracy of prediction for the trait LW3W, from the different prediction methods at the different priors for fraction of variance explained by a single SNP (bars denote standard errors).

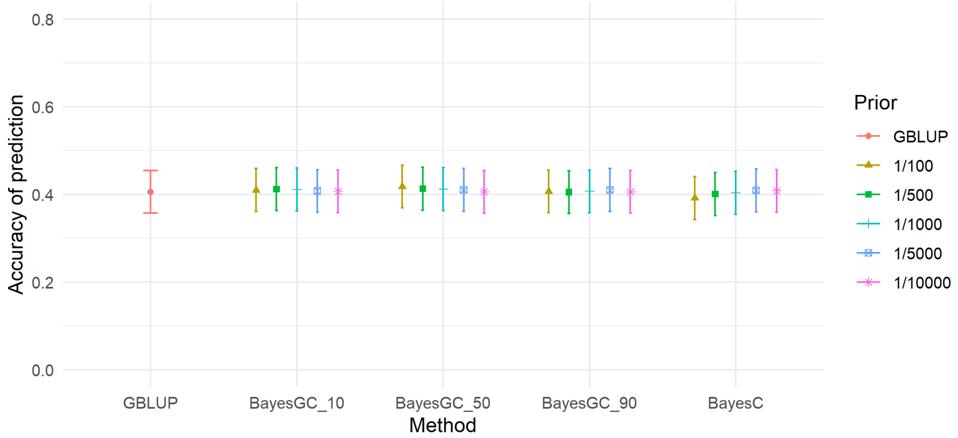


Figure 5. The accuracy of prediction for the trait SHL, from the different prediction methods at the different priors for fraction of variance explained by a single SNP (bars denote standard errors).

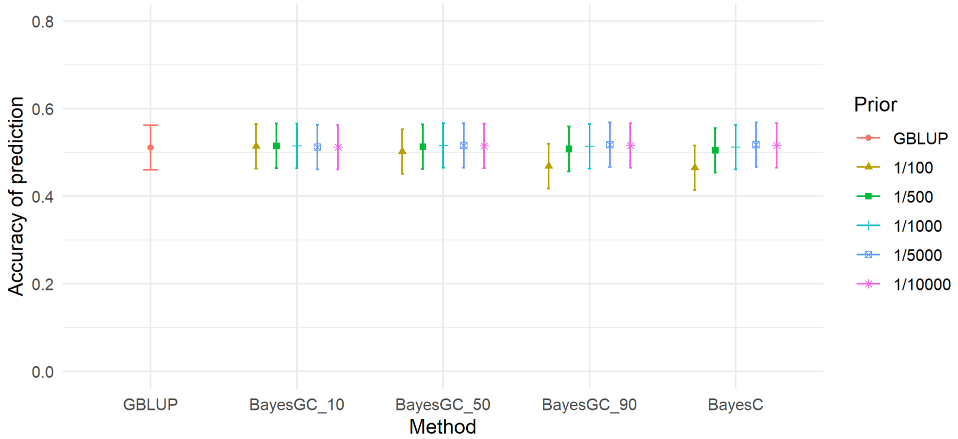


Figure 6. The accuracy of prediction for the trait BCS, from the different prediction methods at the different priors for fraction of variance explained by a single SNP (bars denote standard errors).

The regression coefficients for the method yielding the highest accuracy were also the highest regression coefficient for all the traits. L3W3 was the only trait with a regression coefficient above 1. TNB was the trait with a regression coefficient closest to 1 with a regression coefficient of 0.966 while SHL was the trait with a regression coefficient furthest from 1 at 0.436 (see Table 6), implying that the variance of the GEBV for SHL was inflated.

Table 6. Accuracy, standard error (SE) and regression coefficients (b) for each trait from the method and fraction of the total genetic variance explained by a single fitted SNP (Fr) yielding the highest accuracy for each trait.

Trait[1]	**Method**	**Fr**	**Accuracy**	**SE**	**b**
TNB	BayesGC_10[2]	1/10,000	0.610	0.07	0.97
	BayesC	1/10,000	0.607	0.07	0.95
	GBLUP	-	0.611	0.07	0.98
STB	BayesGC_10[2]	1/5,000	0.318	0.08	0.50
	BayesC	1/5,000	0.311	0.08	0.49
	GBLUP	-	0.318	0.08	0.50
M3W	BayesGC_50[3]	1/100	0.484	0.10	0.79
	BayesC	1/500	0.464	0.10	0.74
	GBLUP	-	0.441	0.10	0.74
LW3W	BayesGC_90[4]	1/10,000	0.722	0.05	1.17
	BayesC	1/10,000	0.718	0.05	1.04
	GBLUP	-	0.717	0.05	1.04
SHL	BayesGC_50[3]	1/100	0.418	0.05	0.44
	BayesC	1/5,000	0.409	0.05	0.41
	GBLUP	-	0.406	0.05	0.41
BCS	BayesGC_50[3]	1/5,000	0.518	0.05	0.70
	BayesC	1/5,000	0.518	0.05	0.70
	GBLUP	-	0.511	0.05	0.70

[1] Total number born (**TNB**), number of stillborn piglets (**STB**), piglet mortality within 3 weeks, i.e., number of piglets dead after birth and until 3 weeks (**M3W**), total litter weight at 3 weeks (**LW3W**), sow shoulder lesions (**SHL**) and the sow's body condition score (**BCS**).

[2] BayesGC_10 is BayesGC with 10% marker variance and 90% polygenic variance.

[3] BayesGC_50 has 50% marker variance and 50% polygenic variance.

[4] BayesGC_90 has 90% marker variance and 10% polygenic variance.

4 DISCUSSION

4.1 Genomic prediction methods

We have compared GBLUP, BayesGC and BayesC for six maternal traits with different priors for BayesC and BayesGC. In general, for all traits, one of the BayesGC methods yielded the highest prediction accuracy (Table 6), although its accuracy was often, but not always, matched by GBLUP and for one trait (BCS) by BayesC. This implies that fitting a combination of individual SNPs with large effects and a polygenic effect often yielded the highest prediction accuracy, however the differences were not significant. The traits M3W and SHL yielded a 9.8% and 3.0% increase in accuracy when moving from GBLUP to BayesGC_50 (Fr1/100) (Table 6). The trait BCS had a somewhat increased accuracy of prediction (1.4% higher than GBLUP) when fitting either BayesC (Fr 1/5,000) or BayesGC_50 (Fr 1/5,000). The trait LW3W had a 0.7% higher accuracy for BayesGC_90 (Fr 1/10,000) than GBLUP. The traits TNB and STB showed no benefit of fitting Bayesian variable selection methods compared to GBLUP.

A limited increase in accuracy when going from GBLUP to BayesGC could be because the accuracy of prediction for the trait using GBLUP already is quite high. Our reference population was quite large (9-15,000 animals). A reference population of 7-11,000 animals were sufficient to obtain GEBV prediction accuracies comparable to the EBVs obtained with progeny testing for Japanese Black cattle (Takeda et al., 2021). TNB and LW3W with an accuracy of ~0.6 and ~0.7 for GBLUP respectively, might not have as much potential for increasing their accuracy as M3W, with a much lower general accuracy of prediction (~0.44 for GBLUP). However, the trait STB showing the least benefit of fitting a Bayesian model also has the lowest general prediction accuracy of ~0.3. This could mean that there are other factors impacting the possible prediction accuracy of STB. For example, there could be fewer or no major QTL for the trait STB, lower linkage disequilibrium between markers and QTL, or low minor allele frequency of QTL for STB.

4.2 Genetic architecture

The accuracy of GP depends on the proportion of genetic variance captured by the markers, the size of the reference population, the additive genetic relationship between the animals in the reference and the validation population, the heritability of the trait, the number of independent QTL and the effective number of chromosome segments (Habier, Fernando, & Dekkers, 2007; Daetwyler, Calus, Pong-Wong, de los Campos, & Hickey, 2013; Daetwyler, Pong-Wong, Villanueva, & Woolliams, 2010; Habier, Tetens, Seefried, Lichtner, & Thaller, 2010; Wientjes, Veerkamp, & Calus, 2013). Most individuals have records for all traits in our current data, which implies that the genetic relationships between the reference and validation populations are approximately the same over the six traits. However, some individuals have missing records for some traits, resulting in reduced reference population size. M3W and LW3W have ~10K reference animals, while the other traits have ~15K reference populations. LW3W, SHL, and BCS have the highest heritability, implying more informative reference data (Table 1). Thus, it seems that the main differences between the traits in our study are the genetic architectures of the traits. I.e., how much genetic variance is captured by the markers and the size and number of major QTLs present for each trait.

While the traits are all considered to be complex and polygenic, some of the traits might have major genes and SNPs in close linkage disequilibrium that explain a substantial part of the genetic variance. However, if there happen to be many SNPs with substantial LD to a major gene, e.g., due to high genetic drift in the region, the GBLUP method may still perform well, since it can use many SNPs to explain the major gene effect. Also, for some traits, genomic predictions may have been over larger genetic distances, i.e., reduced relationships between reference and validation animals, which favors variable selection genomic prediction methods since they focus on SNPs that are in close LD with the QTL (Meuwissen and Goddard, 2010; Solberg et al., 2009).

The QTL database (Pig QTLDdb; Hu, Park, & Reecy, 2022) for each trait shows that there were 228 detected QTL for “Total number born alive” (TNB) and 138 QTL for “Number of stillborn” (STB). The trait “Piglet mortality within 3 weeks” (M3W) did not exist in the database. However, 10 QTL were found for the trait “Piglet Mortality”. There was also no trait in the database defined as “Total litter weight within 3 weeks” (LW3W), but 1 QTL was listed for “Total litter weight at weaning” (He et al., 2021). There was also no QTL listed for Body Condition Score (BCS) or Shoulder Lesions (SHL). However, the published QTL listed in the database do not only reflect the genetic architecture of the traits but serve also as indicators for which traits that are more or less investigated.

QTL markers identified by GWAS on sequence data may be included in 50k marker panels for genomic prediction. In Holstein cattle (Brøndum et al., 2015), this method showed increased reliability of genomic prediction, especially when the QTL are included as a separate variance component, as it allows for extra emphasis on the QTL. If large QTL included in the prediction model can help increase the prediction accuracy, why not just include the QTL directly in the linear model? This, however, requires a two-step approach, where one first finds the QTL associated with the trait, and then includes them into the genomic prediction model. The BayesGC method fits both the polygenic trait and the important SNPs in one analysis. Both approaches do however show that there is room for improvement in prediction accuracy by including important SNPs with higher emphasis in a genomic prediction model. Bayesian variable selection methods also have the potential to find the functional SNPs to include in a linear model (Meuwissen et al., 2021; van den Berg, Fritz, & Boichard, 2013).

4.3 Prior distributions

Bayesian variable selection methods use priors, which need to be carefully chosen or hyper-parameters of the priors estimated as part of the prediction method. The latter would extend the number of MCMC cycles substantially, as these hyper-parameters converge much slower to their equilibrium distribution than GEBVs. In our study we tried a range of different priors, varying both the number of SNPs to be included in the model through Fr , the emphasis of each SNP through the variance explained by markers (σ_{m}^2) and the ratio between variance explained by markers and variance explained by the polygenic effect (σ_{pol}^2), where with BayesGC_10, 10% of the total genetic variance is fitted with markers ($\sigma_{\mathrm{m}}{}^2$) and 90% with the polygenic effect (σ_{pol}^2), BayesGC_50 the variance is split 50/50 and with BayesGC_90, 90% of the total genetic variance is fitted with markers and 10% with the polygenic effect. For the Bayesian methods, the priors on the fraction of variance explained by a single SNP (Fr) seems more important than how much genetic variance is explained by either polygenic effect ($\sigma_{\mathrm{pol}}{}^2$) or the marker effects ($\sigma_{\mathrm{m}}{}^2$), i.e., there are more differences within the methods BayesGC_10, BayesGC_50 or BayesGC_90 than between them.

M3W (Figure 3) showed the largest differences in accuracy between GBLUP and BayesGC and it seems the accuracy increases gradually as Fr becomes larger (fewer SNPs fitted) but only when the ratio between σ_{m}^2 and σ_{pol}^2 is favoring σ_{pol}^2 in such a way that the model fits 50-90% of the genetic variance as σ_{pol}^2 and the remaining variance is fitted with very few SNPs that in turn get fitted with a relatively high emphasis through Fr. The traits SHL and BCS (Figure 5 and 6) show a similar pattern, i.e., fitting a few SNPs is not improving the overall prediction accuracy unless it is also accompanied by a high emphasis on σ_{pol}^2. This could indicate that finding QTL and fitting them on their own is not sufficient to obtain high prediction accuracy. One also needs the support of a polygenic effect through e.g., a genomic relationship matrix. However, when fitting many SNPs through the BayesC term, the Bayesian variable selection method would also fit many SNPs with a small effect – similar to GBLUP. The benefit of a Bayesian variable selection method compared to GBLUP is thus expected to be lower for the methods with a higher π-value, like the Fr 1/5,000 and 1/10,000.

4.4 Further developments

A further development of the BayesGC would be to expand to the model to include - non-genotyped animals in the estimation of breeding values through e.g. single-step genomic prediction (Legarra et al., 2009; Christensen and Lund, 2010; Fernando et al. 2014). The challenge of including non-genotyped animals with genomic prediction is the need to impute genotypes. With linear methods there are methods such as ssGBLUP (Aguilar et al., 2010; Christensen & Lund, 2010; Legarra, Aguilar, & Misztal, 2009; Misztal, Legarra, & Aguilar, 2009) where an additive relationship matrix H is combining information from both pedigree and SNP data. Bayesian methods for combining genotyped and non-genotyped animals have been developed that could be adapted to this model by imputing genotypes for non-genotyped animals using MCMC methods that could be used with whole-genome data (Fernando, Dekkers, & Garrick, 2014). BayesGC includes a polygenic effect in the form of a **G**-matrix which could also be exchanged with an **H**-matrix to include non-genotyped animals, and the marker-model-based single step approach of Fernando et al. (2014) could be used for the additional SNPs fitted by the BayesGC model. Other options for using BayesGC results in routine genomic evaluations would be to use the analysis of genotyped animals to find SNPs that need extra weight. In a regular GBLUP /ssGBLUP analysis these SNPs would thus attain extra weights when constructing **G**, and implicitly the **H**-matrix. The information on SNP variance from a Bayesian analysis could thus be used to improve the genomic relationship matrix for GBLUP or ssGBLUP analyses.

Another way to improve BayesGC could be to expand the software towards multi trait analyses as many routine breeding evaluations today are based on multi-trait models. Expanding the BayesGC model towards multi-trait analyses is relatively straightforward if one assumes that a SNP with a large effect, is affecting all the (related) traits (Karaman, Lund, & Su, 2019; Kemper, Bowman, Hayes, Visscher, & Goddard, 2018). In situations where we cannot assume this, multi-trait variable selection modelling requires to sample which combination of traits is affected by each of the SNPs. If there are many traits, there are many such combinations. Applying the BayesGC results to multi-trait routine evaluations may be by giving extra weight to some SNP genotypes, resulting in a different **G** matrix for each of the traits, and consequently also for different pairs of traits (since the **G** matrix modelling covariances between traits i and j is constructed as the cross-product of the SNP genotypes weighted for trait i and those weighted for trait j). Modifications of routine software packages may be needed to accommodate these per trait alternative **G** matrices.

Bayesian variable selection methods have a lot of potential for further development to be used in routine breeding value estimations. One of the biggest drawbacks today is the high computational costs of running the MCMC chains. However, computational power has historically increased and will most likely continue to increase, in addition to further research developing into more efficient algorithms using parallel computations.

5 CONCLUSIONS

The accuracy of genomic prediction on six maternal traits in landrace pigs varied greatly ranging from 0.31 to 0.61. The prediction accuracies did not vary much between the different genomic prediction methods. The two traits M3W and BCS could benefit from using a BayesGC approach with a 9.8 and 3.0% increase in accuracy respectively, while TNB, STB, LW3W, and SHL showed only minor improvements. Although GBLUP, BayesC and BayesGC all yielded similar genomic prediction accuracies, the accuracy of BayesGC was always as high as or higher than that of GBLUP. Within the BayesGC method the accuracies could vary depending on the prior distributions. The models were more sensitive to how many markers were fitted in the model through varying the fraction of the total genetic variance explained by a single marker (Fr) compared to the amount of total genetic variance explained by marker effects as a whole (BayesGC_10, BayesGC_50 or BayesGC_90), but overall, most traits were robust against varying the prior distributions.

ACKNOWLEDGEMENTS

We gratefully acknowledge the funding from the Norwegian Research Council (project 255297). Topigs Norsvin SA is acknowledged for providing data and genotype information. The helpful comments of two anonymous reviewers are also gratefully acknowledged.

REFERENCES

Aguilar, I., Misztal, I., Johnson, D., Legarra, A., Tsuruta, S., & Lawlor, T. (2010). Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science*, *93*(2), 743–752. https://doi.org/10.3168/JDS.2009-2730

Brøndum, R. F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D., & Lund, M. S. (2015). Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of Dairy Science*, *98*(6), 4107–4116. https://doi.org/10.3168/jds.2014-9005

Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution 2010 42:1*, *42*(1), 1–8. https://doi.org/10.1186/1297-9686-42-2

D. Habier, R.L. Fernando, & J.C.M. Dekkers. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, *177*(4), 2389–2397.

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., & Hickey, J. M. (2013). Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics*, *193*(2), 347–365. https://doi.org/10.1534/genetics.112.147983

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, *185*(3), 1021–1031. https://doi.org/10.1534/genetics.110.116855

Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., … Goddard, M. E. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, *95*(7), 4114–4129. https://doi.org/10.3168/jds.2011-5019

Eriksen, E. M. (2018). *Effects of changes in the breeding goal on genetic improvement for maternal traits in Landrace pigs* (Norwegian University of Life Science (NMBU)). Retrieved from https://nmbu.brage.unit.no/nmbu-xmlui/bitstream/handle/11250/2572879/Masteroppgave.pdf?sequence=1&isAllowed=y

Fernando, R. L., Dekkers, J. C., & Garrick, D. J. (2014). A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genetics, Selection, Evolution : GSE*, *46*(1). https://doi.org/10.1186/1297-9686-46-50

Gianola, D., De Los Campos, G., Hill, W. G., Manfredi, E., & Fernando, R. (2009). Additive Genetic Variability and the Bayesian Alphabet. *Genetics*, *183*(1), 347–363. https://doi.org/10.1534/genetics.109.103952

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-186

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., & Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution*, *42*(1), 5. https://doi.org/10.1186/1297-9686-42-5

He, Y., Zhou, X., Zheng, R., Jiang, Y., Yao, Z., Wang, X., … Yuan, X. (2021). The Association of an SNP in the EXOC4 Gene and Reproductive Traits Suggests Its Use as a Breeding Marker in Pigs. *Animals : An Open Access Journal from MDPI*, *11*(2). https://doi.org/10.3390/ani11020521

Hu, Z.-L., Park, C. A., & Reecy, J. M. (2022). Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. *Nucleic Acids Research*, *50*(D1), D956–D961. https://doi.org/10.1093/nar/gkab1116

Karaman, E., Lund, M. S., & Su, G. (2019). Multi-trait single-step genomic prediction accounting for heterogeneous (co)variances over the genome. *Heredity 2019 124:2*, *124*(2), 274–287. https://doi.org/10.1038/s41437-019-0273-4

Kemper, K. E., Bowman, P. J., Hayes, B. J., Visscher, P. M., & Goddard, M. E. (2018). A multi-trait Bayesian method for mapping QTL and genomic prediction. *Genetics Selection Evolution 2018 50:1*, *50*(1), 1–13. https://doi.org/10.1186/S12711-018-0377-Y

Kjetså, M. H., Ødegård, J., & Meuwissen, T. H. E. (2020). Accuracy of genomic prediction of host resistance to salmon lice in Atlantic salmon (Salmo salar) using imputed high-density genotypes. *Aquaculture*, *526*, 735415. https://doi.org/10.1016/j.aquaculture.2020.735415

Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, *92*(9), 4656–4663. https://doi.org/10.3168/JDS.2009-2061

Lillehammer, M., Meuwissen, T. H. E., & Sonesson, A. K. (2011). Genomic selection for maternal traits in pigs. *Journal of Animal Science*, *89*(12), 3908–3916. https://doi.org/10.2527/jas.2011-4044

Lillehammer, M., Meuwissen, T. H. E., & Sonesson, A. K. (2013). Genomic selection for two traits in a maternal pig breeding scheme. *Journal of Animal Science*, *91*(7), 3079–3087. https://doi.org/10.2527/jas.2012-5113

Madsen, P., & Jensen, J. (2013). *A User's Guide to DMU A Package for Analysing Multivariate Mixed Models*. Retrieved from http://dmu.agrsci.dk

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829. Retrieved from http://www.genetics.org/content/157/4/1819.short

Meuwissen, T., van den Berg, I., & Goddard, M. (2021). On the use of whole-genome sequence data for across-breed genomic prediction and fine-scale mapping of QTL. *Genetics Selection Evolution*, *53*(1), 19. https://doi.org/10.1186/s12711-021-00607-4

Misztal, I., Legarra, A., & Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science*, *92*(9), 4648–4655. https://doi.org/10.3168/JDS.2009-2064

Ocepek, M., & Andersen, I. L. (2017). What makes a good mother? Maternal behavioural traits important for piglet survival. *Applied Animal Behaviour Science*, *193*, 29–36. https://doi.org/10.1016/j.applanim.2017.03.010

Salek Ardestani, S., Jafarikia, M., Sargolzaei, M., Sullivan, B., & Miar, Y. (2021). Genomic Prediction of Average Daily Gain, Back-Fat Thickness, and Loin Muscle Depth Using Different Genomic Tools in Canadian Swine Populations . *Frontiers in Genetics* , Vol. 12, p. 735. Retrieved from https://www.frontiersin.org/article/10.3389/fgene.2021.665344

Samorè, A. B., & Fontanesi, L. (2016). Genomic selection in pigs: State of the art and perspectives. *Italian Journal of Animal Science*, *15*(2), 211–232. https://doi.org/10.1080/1828051X.2016.1172034

Sargolzaei, M., Chesnais, J. P., & Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*, *15*, 478. https://doi.org/10.1186/1471-2164-15-478

Song, H., Zhang, J., Jiang, Y., Gao, H., Tang, S., Mi, S., … Ding, X. (2017). Genomic prediction for growth and reproduction traits in pig using an admixed reference population1. *Journal of Animal Science*, *95*(8), 3415–3424. https://doi.org/10.2527/jas.2017.1656

Strandén, I., & Garrick, D. J. (2009). Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of Dairy Science*, *92*(6), 2971–2975. https://doi.org/10.3168/jds.2008-1929

Takeda, M., Inoue, K., Oyama, H., Uchiyama, K., Yoshinari, K., Sasago, N., … Uemoto, Y. (2021). Exploring the size of reference population for expected accuracy of genomic prediction using simulated and real data in Japanese Black cattle. *BMC Genomics*, *22*(1), 1–11. https://doi.org/10.1186/S12864-021-08121-Z/FIGURES/3

Tan, C., Wu, Z., Ren, J., Huang, Z., Liu, D., He, X., & Prakapenka, D. (2017). Genome - wide association study and accuracy of genomic prediction for teat number in Duroc pigs using genotyping - by - sequencing. *Genetics Selection Evolution*, 1–13. https://doi.org/10.1186/s12711-017-0311-8

van den Berg, I., Fritz, S., & Boichard, D. (2013). QTL fine mapping with Bayes C(π): a simulation study. *Genetics, Selection, Evolution : GSE*, *45*(1), 19. https://doi.org/10.1186/1297-9686-45-19

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

VanRaden, P. M., & Wiggans, G. R. (1991). Derivation, Calculation, and Use of National Animal Model Information. *Journal of Dairy Science*, *74*(8), 2737–2746. https://doi.org/10.3168/jds.S0022-0302(91)78453-1

VanRaden, Paul M. (2007). Genomic Measures of Relationship and Inbreeding. *Interbull Bulletin*, *25*(37), 111–114. https://doi.org/10.1007/s13398-014-0173-7.2

Verbyla, K. L., Bowman, P. J., Hayes, B. J., & Goddard, M. E. (2010). Sensitivity of genomic selection to using different prior distributions. *BMC Proceedings*, *4*(S1), 2–5. https://doi.org/10.1186/1753-6561-4-s1-s5

Wientjes, Y. C. J., Veerkamp, R. F., & Calus, M. P. L. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, *193*(2), 621–631. https://doi.org/10.1534/genetics.112.146290

Paper III

M. V. Kjetså, A. B. Gjuvsland, E. Grindflek, T.H.E. Meuwissen

Whole-genome sequence and pCADD marker-based genomic prediction for maternal traits in two pig lines.

Manuscript

Whole-genome sequence and pCADD marker-based genomic prediction for maternal traits in two pig lines

Maria V. Kjetså[1*], Arne B. Gjuvsland· [2], Eli Grindflek[2] and Theo Meuwissen[1]

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

[2] Norsvin SA, Storhamargata 44, 2317 Hamar, Norway

*Corresponding author

Corresponding E-mail address: maria.kjetsa@nmbu.no

Abstract

Background
In this study, we investigate several factors affecting the prediction accuracy of genomic selection. The data consists of two nucleus pig populations, one pure-bred Landrace (L) and one Synthetic (S) Yorkshire/Large White line. All animals have records on maternal traits and are genotyped, with up to 30K animals in each line. Our aim was to investigate the reference population size required to obtain substantial prediction accuracy within- and across-lines and the effect of using a multi-line reference population with both a high ratio of within-line and a high ratio of across-line animals in the reference population. Prediction accuracy was tested with three different marker data sets: High-Density, pCADD, and Whole Genome Sequence (WGS). Also, two different genomic prediction methods (GBLUP and BayesGC) were compared for four maternal traits in pigs.

Results
A reference population of 3K-6K animals for within-line prediction was generally sufficient to achieve high prediction accuracy. However, increasing to 30K animals in the reference population significantly increased prediction accuracy for two traits. A reference population of 30K across-line animals achieved a similar accuracy to 1K within-line animals. For multi-line prediction accuracy, the accuracy was most dependent on the number of within-line animals in the reference data. The S-line provided a generally higher prediction accuracy compared to the L-line. Using pCADD scores to reduce the number of markers from WGS data in combination with the GBLUP method generally reduced prediction accuracies relative to GBLUP_HD analyses. When using BayesGC, prediction accuracies were generally similar when using HD, pCADD, or WGS marker data, suggesting that the BayesGC method selects a suitable set of markers irrespective of the markers provided (HD, pCADD, or WGS).

Conclusions
A large reference population size can help accuracy for both within- and across-line predictions. For multi-line prediction, adding more within-line animals is more important than a larger number of across-line animals. There were few differences between the different marker data sets and methods regarding prediction accuracy. The BayesGC method benefited from a large reference population and was less dependent on the different genotype marker datasets to achieve a high prediction accuracy.

Background

Maternal traits are related to the sow's ability to produce large litters of healthy piglets that survive past weaning (Ocepek and Andersen 2017). For pig production, the maternal traits are of high economic importance. One disadvantage of breeding for maternal traits is that the traits can only be recorded directly on sows. For the selection of boars for maternal traits, we depend on records from relatives, which can be challenging for achieving a high prediction accuracy. New methods to increase prediction accuracy are required to improve the prediction accuracy for maternal traits in pigs.

Genomic Selection (GS) (Meuwissen, Hayes, & Goddard, 2001) is a method for predicting breeding values that is beneficial for traits where one cannot produce records directly on selection candidates. Many factors could affect the prediction accuracy of GS. Some of the factors are the size of the reference population, the relationship between the reference and the validation population, the heritability of the trait, the effective population size (Ne), the marker density, the number of QTL, the linkage disequilibrium (LD) between markers and QTL, and the minor allele frequencies of causative mutations (Clark et al. 2012; Clark, Hickey, and Van Der Werf 2011; Daetwyler et al. 2010; Druet, Macleod, and Hayes 2014; Goddard 2009; Habier et al. 2010; Hayes et al. 2009; Wientjes, Veerkamp, and Calus 2013).

For smaller breeds or lines, or breeding systems where the budgets are small, it could be difficult to obtain a large enough reference population to achieve high prediction accuracy. Using a reference population with animals from a different population could be tempting. However, the results so far report zero or close to zero accuracies (Kachman et al. 2013). Several studies have also attempted to combine genotyped animals from different populations in the reference data (Erbe et al. 2012; Hozé et al. 2014; L Zhou et al. 2014; L. Zhou et al. 2014). The accuracies using e.g., 50K or 770K marker density genotype has only slightly improved prediction accuracies so far, with the biggest advantage coming from large reference populations.

A low across population prediction accuracy is consistent with a large effective number of genomic segments (M_e) to estimate (van den Berg et al. 2019). For across population predictions, the Linkage Disequilibrium (LD) between markers and QTL may not persist. The relationships between the animals across populations are tiny, there could be differences in the allele substitution effects across populations, and QTL that segregate in one population may not segregate in the other population (De Roos, Hayes, and Goddard 2009).

The advantage of WGS is that it has a high chance of including SNPs that are causative mutations or are in high linkage disequilibrium (LD) with causative mutations which is consistent across populations. There have been some successful applications using WGS data for multi-breed genomic prediction (Van Binsbergen et al. 2015). However, to maximise the effect of using WGS, the size of the reference population should be large (Iheshiulor et al. 2016; De Roos et al. 2009). WGS may contain information about Structural Variants (SVs) or Copy Number Variants (CNVs) in addition to the Single Nucleotide Polymorphisms (SNPs), in contrast to the more common SNP arrays ("SNP-chips"), which only include SNPs. The costs for WGS continue to reduce (Wetterstrand, 2021), which will make WGS data even more commonly available in the future.

One of the major drawbacks of utilising WGS is the sheer size of the genotype datasets. With millions of markers, the computational power needed to process a large dataset with many animals is very demanding. One option is to derive functional markers from WGS and use a reduced set of markers in the prediction. Many methods for reducing the number of markers has been used, such as removing markers in high LD with each other and removing markers with low Minor Allele Frequencies (MAF). However, by doing this, we risk losing one of the most significant advantages of WGS, namely that WGS includes the causative mutations.

Combined annotation dependent depletion for pigs (pCADD) is a method for prioritizing single nucleotide variants (SNVs) in the pig genome for their putative deleteriousness by the biological significance of the genomic location (Groß et al. 2020). The combined annotation dependent depletion model (CADD) was first developed for human populations and is designed to capture signals of evolutionary selection across many generations and combines this with genomic features, epigenetic data, and other predictors to estimate a deleteriousness score for a given variant. pCADD was developed to help researchers and breeders to evaluate newly observed SNVs, and rank potentially harmful SNVs that are propagated by breeding. The pCADD score is a log-rank score that ranges from ~95 to 0, with the higher scores indicating a higher probability for a deleterious SNV. The top 1% and 0.1% highest scored SNVs have a pCADD score higher than 20 and 30, respectively, and thus the more deleterious variants are differentiated from the likely neutral ones. This study investigates the prediction accuracy of a marker data set derived from WGS, based on their pCADD score.

Genomic Best Linear Unbiased Prediction (GBLUP) is a widely used and accepted method for predicting genomic breeding values, which does not account for the variation from markers that have a greater effect on the trait, such as the Quantitative Trait Loci (QTL). GBLUP assumes that all markers explain the same variance across the genome. Bayesian variable selection methods are a way to differentiate between individual SNPs so that markers with larger effects are fitted in the model, and others are down-weighted. Many different Bayesian methods have been proposed, such as A, B, C and R, and are often referred to as the "Bayesian alphabet" (Gianola 2013). The downside of the Bayesian methods is that they are computationally costly.

In this paper, we use the BayesGC method (Meuwissen, van den Berg, and Goddard 2021), which fits both a BayesC marker term with a prior distribution that is a mixture of the normal distribution and a zero-effects distribution, in addition to fitting a polygenic term through a Genomic Relationship Matrix ("**G**-matrix"). The method was to utilise high-density and whole-genome sequence data, where the polygenic term accounted for the many SNPs with small effects, and the BayesC term will fit the few SNPs with a large effect (Meuwissen et al. 2021). Thus, BayesGC combines the positives of a GBLUP model using a **G**-matrix to fit all the markers with a small effect computationally efficiently, with the positive effect of adding individual markers with large effects through the BayesC term.

This study aims to, firstly, compare a pCADD derived marker panel, a high density (HD) SNP-chip set, and Whole Genome Sequence (WGS) marker data, using both a linear prediction method (GBLUP) and a Bayesian variable selection method (BayesGC). Secondly, we compare the effect of within-, across- and multi-breed genomic predictions at different sizes of reference populations, using alternative prediction methods (GBLUP and BayesGC) and genome marker sets (HD, pCADD and WGS).

Methods

Dataset

The data consisted of two commercial pig populations; a pure-bred Landrace and a Synthetic Large White/Yorkshire line, denoted as the L- and S-line respectively. All data were obtained from herds owned by Norsvin and Topigs Norsvin (www.topigsnorsvin.com). There were phenotypic records of four traits: Total Number Born piglets (TNB), Total Number of Stillborn piglets (STB), Shoulder Lesion Score (SHL) and Body Condition Score (BCS). All traits are measured on sows. Because a lot of animals had records on more than one of the traits, the traits were grouped into two trait groups, where all animals had records on both traits in their group. The two trait groups are denoted as *A-traits*; consisting of animals with records on both TNB and STB, and *B-Traits*; consisting of animals with records on both SHL and BCS. Only 332 A-trait and 329 B-trait animals were removed because they did not have records for both traits. Combining the traits highly reduced computer times for the analyses as the two traits in the trait group share genotype datasets. The A-traits set consisted of 31751 animals of the L-line and 30356 animals of the S-line. The B-traits consisted of 27456 and 6840 animals from the L- and S-line respectively (Table 1).

Table 1. Number of animals for each trait group and line.

Line	Trait	Total n	Val n	Ref 1K	Ref 3K	Ref 6K	Ref 15K	Ref 30K
L	A	31751	1247	1436	3691	6114	15054	30504
S	A	30256	1259	1444	3321	5874	15533	29097
L	B	27456	1089	1286	3681	6523	15392	26367
S	B	6804	1202	1088	3277	5638	-	-

Line indicates which line of pig the animals are from, and Trait indicates which trait group the animals in the dataset has phenotypic records of. Total n is the total number of animals for that line and trait group. Val n is the number of animals used for validation.

Ref 1K, Ref 3K, Ref 6K, Ref 15K and Ref 30K is the number of animals used for within-line reference data in the sub-datasets corresponding to the number of animals in the reference data set, where K indicates numbers in 1,000.

Three genotype marker data sets were used in the study; High-Density (HD), Whole Genome Sequence (WGS) and pig combined annotation dependent depletion (pCADD) markers derived from WGS. The animals were genotyped with varying SNP densities, but all were imputed to a HD 660K genotype density using FImpute v2.2 (Sargolzaei, Chesnais, and Schenkel 2014) through the routine imputation process of Topigs Norsvin. After quality control, the 660K High-Density genotype data had a total of 433,451 SNPs with MAF>0.01 in both breeds.

The 660K genotypes were imputed to a reference panel of 756 sequenced animals for 25,9 million sequence variants, using Eagle 2.4.1 (Loh et al. 2016) for phasing and Minimac4 (Das et al. 2016) for imputation. After imputation we filtered out variants with low MAF in one or both breeds, keeping 11.3 million variants with MAF>0.01 in both breeds which were used to create the WGS and pCADD datasets. The WGS dataset was created by LD-pruning with the option "–indep-pairwise" in PLINK 1.9 (Chang et al. 2015; Purcell and Chang 2019) using an r^2 threshold of >0.99 and a window size of 10K variants. After pruning we were left with a total of 1,946,188 variants. For the pCADD dataset the variant were ranked by pCADD-score. We used markers with a pCADD score >11 giving a total of 416,828 SNPs, which is comparable to the number of markers on the High-Density SNP-chip.

Table 2. The designs of the sub-datasets that are compared for the different traits and lines.

Traits	A-Traits		B-Traits	
Validation data	L-Line	S-Line	L-Line	S-Line
Across Reference data	S1	L1	S1	L1
	S3	L3	S3	L3
	S6	L6	S6	L6
	S15	L15	-	L15
	S30	L30	-	L30
Within reference data	L1	S1	L1	S1

	L3	S3	L3	S3
	L6	S6	L6	S6
	L15	S15	L15	-
	L30	S30	L30	-
Multi-line reference data	L1_S1	S1_L1	L1_S1	S1_L1
	L1_S3	S1_L3	L1_S3	S1_L3
	L1_S6	S1_L6	L1_S6	S1_L6
	L3_S1	S3_L1	L3_S1	S3_L1
	L3_S3	S3_L3	L3_S3	S3_L3
	L3_S6	S3_L6	L3_S6	S3_L6
	L6_S1	S6_L1	L6_S1	S6_L1
	L6_S3	S6_L3	L6_S3	S6_L3
	L6_S6	S6_L6	L6_S6	S6_L6

Traits indicates which trait group the dataset is for; A or B. Validation data indicates which line the validation is from, L or S. Across Reference data indicates that the reference data is a different line than the validation data. Within reference data is the datasets of which the validation and reference data are from the same line. The Multi-line reference data contains animals from both lines in the reference data. The first letter indicates the line of which the data is from (L or S). The number indicates the number of animals in the reference data. For the multi-line reference data, it is indicated the size of each reference data from each line, e.g., L1_S3 indicates 1000 animals from the L-line and 3000 animals from the S-line.

Validation and sub-datasets

To compare the methods for different sizes and types of reference populations, the animals were divided into sub-datasets. The sub-datasets were set up both for pure within-, across-, and multi-line reference sets. The animals were divided into reference animals and validation animals. The reference animals consisted of animals with both phenotypic and genotypic records. For the validation animals, the phenotypic records were masked to validate the prediction accuracy of the genomic prediction method.

The validation animals were the same for all analyses: approximately the 1000 youngest animals for each Trait and Line. In-total there were 4 validation sets (Table 1). For the A-traits we had sub-datasets with reference sets of approximately 1K, 3K, 6K, 15K and full 30K animals for both the L- and S-line, denoted Lx or Sx where x denotes the size of the reference population (in thousands).

For the B-traits of the S-line there were reference sets consisting of 1K, 3K and 6K animals, while the L-line had reference sets up to 30K animals. An overview of the sub-datasets is shown in Table 2. Note that for the within-line predictions, the youngest animals were chosen in the reference data, which avoids the inclusion of progeny of the validation animals in the reference data. Moreover, these forward predictions are most relevant for breeding schemes. To avoid splitting animals that were born at the same day (potential siblings), we found a cut-off date that gave a reference population size close to the desired size. However, for the across-line predictions forward prediction was not seen as an important factor and thus the animals for the reference population were chosen at random. Table 1 shows an overview of the exact number of animals for each Reference set for within-line prediction. For the across-line reference data sets, as random selection was used to select the individuals, the number of animals is exact (i.e., for a 3K size reference data set there are exactly 3,000 animals). The accuracy of prediction for all methods were corrected for the trait-heritability as:

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

Table 3. Estimated variance components and variance priors used in the BayesGC method for the traits TNB, STB, SHL and BCS.

	TNB	STB	SHL	BCS
σ_u^2	0.644	0.020	0.104	0.167
σ_e^2	6.870	0.227	0.073	0.188
h^2	0.086	0.080	0.590	0.471
σ_{pol}^2	0.322	0.010	0.052	0.084
σ_m^2 HD	0.00079	0.00002	0.00013	0.00020
σ_m^2 pCADD	0.00088	0.00003	0.00014	0.00023
σ_m^2 WGS	0.00088	0.00003	0.00014	0.00023

TNB is Total Number Born. STB is Stillborn. SHL is Shoulder Lesions. BCS is Body Condition Score. σ_u^2 is the total genetic variance. σ_e^2 is the error variance. h^2 is the heritability. σ_{pol}^2 is the variance prior for the polygenic term in the BayesGC method. σ_m^2 HD is the marker variance prior for the high density (HD) genotype data. σ_m^2 pCADD is the marker variance prior for the pCADD genotype data. σ_m^2 WGS is the marker variance prior for the whole genome sequence (WGS) genotype data.

Model for analysis

Yield Deviations (VanRaden & Wiggans, 1991) for the four traits were derived from the commercial breeding value evaluations from Topigs Norsvin, using a traditional (pedigree-based) animal model. There were multiple records for each trait, as we had one YD for each parity. Because the software used for the Bayesian variable selection models (Meuwissen, van den Berg, & Goddard, 2021) could not handle multiple records per animals, we used the average YD for each sow, with a weighting of each record corresponding to the effective number of records calculated as $\frac{n*(1+\text{[redacted]}}{(n+\text{[redacted]}}$ where [redacted] is σ_e^2/σ_{pe}^2 and *n* is the number of records for each individual, σ_e^2 is the residual variance and σ_{pe}^2 is the permanent environmental variance. These variance components were obtained from commercial Topigs Norsvin breeding value evaluations.

The variance components of the yield deviations (incl. their weights) were estimated for each trait using the pedigree relationship matrix and the DMUAI package from the DMU software (Madsen and Jensen 2013). The following model was used:

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

where **Y** is the phenotypic record of a sow. **1** is a vector of ones corresponding to the size of **Y**. μ is the mean, **Z** is a design matrix linking individuals to the phenotype. ***u*** is the random effect of the individual animal (**u**~N(0, **A**□$_u^2$) where **A** is the pedigree relationship matrix. **e** = residual effect (**e**~N(0, **D**□$_e^2$)), and **D** is a diagonal matrix where the diagonals are the inverses of the weights of the records. The resulting estimated variance components are presented in Table 3.

Two methods were used for genomic prediction: GBLUP and BayesGC.

The GBLUP method used the model:

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

where **Y** is a vector of the average YD of a sow. **1** is a vector of ones corresponding to the size of **Y**. μ is the mean, **Z** is a design matrix linking individuals to the phenotype. ***u*** is the random effect of the individual animal (**u**~N(0, $\mathbf{G}\square_{u}^{2}$) where **G** is the genomic relationship matrix. **e** = residual effect (**e**~N(0, $\mathbf{D}\square_{e}^{2}$)), and **D** is a diagonal matrix where the diagonals are the inverses of the weights of the records.

And the BayesGC model was:

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

where **Y** is a vector of the Yield Deviations. **1** is a vector of ones. μ is an overall mean. **Z** is a design matrix that links individuals to the **Y**. ***u*** = random polygenic effect with variance V(***u***) = $\mathbf{G}\sigma^2_{\mathrm{pol}}$ where **G** is a genomic relationship matrix. $\mathbf{X_i}$ = vector of genotypes for SNP i containing 0 for homozygote individuals, 1 for heterozygous, and 2 for the alternative homozygote genotype. $\mathrm{I_i}$ is an indicator of whether SNP *i* is in the model in a MCMC-cycle or not (0/1). The prior probability of $\mathrm{I}_i = 1$ is π. $\mathrm{s_i}$ is the SNP effect, where if the SNP *i* is in the model: $\mathrm{s_i}$ ~N(0, σ^2_{m}). **e** is the residual with variance **e** ~N(0, $\mathbf{D}\sigma^2_{\mathrm{e}}$), and **D** is an diagonal matrix where the diagonals are the inverses of the weights of the records. The MCMC – chain was run for 2000 burn-in cycles and a total of 12000 Gibbs-cycles for twenty independent chains. The EBVs from the twenty Gibbs-chains had correlations of ~0.99 and thus the EBVs were assumed to be converged, and the results presented is the average of twenty Gibbs-chains.

The Genomic Relationship Matrices (**G**-matrix) were calculated using the VanRaden method 1 (VanRaden 2007) for both the HD and pCADD SNP-data sets. One **G**-matrix was calculated for each sub-dataset so that the relationships were based only on the genotypes of the animals present in each sub-dataset. For the WGS analyses using BayesGC, the **G**-matrix used for the variance of the polygenic effect V($\boldsymbol{u}$) was made using the HD genotypes.

For the BayesGC model, the total genetic variance σ_u^2 was partitioned into a variance over the markers σ_m^2 explained by the BayesC term of the model and variance over the polygenic effect σ_{pol}^2.

Let q be the fraction of σ_u^2 explained by the BayesC term, then the variance explained by the polygenic effect is σ_{pol}^2 = (1-q) σ_u^2. Hence,

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

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

p is the allele frequency of a single loci and N_{loci} is the total number of loci.

$$\sigma_m^2 = \frac{Fr*\sigma_u^2}{\overline{HET}}$$

Where σ_m^2 is the genetic variance explained by a single SNP,

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

π is the prior probability of a SNP I_i = 1, indicating whether a SNP i is in the model in a particular MCMC-cycle or not (0/1).

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

$$\pi_c = \frac{1/Fr}{N_{loci}}$$

Such that the total genetic variance is $\sigma_u^2 = \pi_c \cdot N_{loci} \cdot \overline{HET} \cdot \sigma_m^2$.

For the partitioned model of BayesGC, it follows that

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

Based on a previous study (Kjetså et al. 2022, submitted manuscript) the fraction of the partitioning (q) does not have a great effect on the accuracy and the optimum seem to be to split the variance 50/50 between the polygenic σ_{pol}^2 and the marker σ_m^2 variance, giving a q=0.5.

The Fr could have a greater effect on the accuracy and thus the π in our study was estimated, using a π based on an Fr = 1/1000 as the starting value.

For the different genotypes used this meant a starting value for π of 0.0023, 0.0024 and 0.0005 for HD, pCADD and Sequence genotypes respectively.

The priors used for BayesGC for each trait can be found in Table 3 in addition to the estimated variance components. The $\overline{HET}$ depended on the genotypes and was not trait specific and was calculated to be 0.41, 0.37 and 0.37 for the HD, pCADD and WGS genotypes respectively.

Significance testing

The accuracy differences between two alternative methods were tested for their significance using a bootstrapping procedure (Efron, B. Tibishirani 1994). From datasets consisting of triplets (two EBVs (one from each of the models compared) and the corresponding YD of a validation animal), bootstrap samples were randomly sampled by sampling these triplets with replacement (following Iversen et al., 2019). 10,000 bootstrap samples were constructed for each pairwise comparison of models, and the model which yielded the higher correlation with the YD for each bootstrap sample was determined. The models were considered significantly different if one of the models had a higher correlation in at least 97.5% of the bootstrap samples (resulting in a p-value of 0.05 due to the two-sidedness of the test).

Results

For the L-line animals in Figure 1a, the accuracy of prediction for the different reference populations for within population (L1, L3, L6, L15 and L30) ranged from 0.32 (L3 BGC_HD and BGC_WGS) to 0.73 (L30 BGC_HD, BGC_PCADD and BGC_WGS). For across-line predictions of L-Line validation animals for TNB (Figure 1a), BGC_WGS had an accuracy of 0.19, 0.21, 0.0, 0.26 and 0.29 for S1, S3, S15 and S30 respectively, giving a gradual increase in prediction accuracy with a larger reference population size (except for S6 where predictions did not achieve any accuracy). For the largest reference population across-line with 30,000 animals (S30) the accuracies were 0.20, 0.28, 0.29, 0.29 and 0,32 for the GBLUP_PCADD, GBLUP_HD, BGC_PCADD, BGC_WGS and BGC_HD respectively. This is comparable to accuracies obtained with a small within-line reference populations of 1,000-3,000 animals where accuracies ranged from 0.33-0.36.

The prediction accuracy of the S-line animals for the TNB trait (Figure 1b) seems to give slightly higher prediction accuracy compared to the general accuracies of the L-line animals (Figure 1a). For within-line predictions the accuracies ranged from 0.47 for S1 (BGC_WGS and BGC_HD) to 0.83 for S30 (BGC_WGS). For the across-line predictions (Figure 1b) the accuracies ranged from 0.08 for L3 (GBLUP_PCADD) to 0.34 for L30 (BGC_WGS and GBLUP_HD).

For the multi-breed reference population predicting on L-line (Figure 1a), L1_S1 containing 1,000 L-line animals and 1,000 S-line animals obtained accuracies of 0.36-0.37 for all methods. For L1_S3 the accuracies increased ranging between 0.37-0.39 across the methods. For L1_S6 however, where the number of S-line animals were increased to 6,000, the accuracies decreased again and ranged from 0.34-0.36. L3 and S6 seem to be outlier data sets, resulting in poor predictions.

The multi-line prediction accuracies (Figure 1b) yielded the same or slightly higher accuracies than those of within-line predictions with the same amount of within-line reference animals. In contrast to the L-line (Figure 1a), the multi-line prediction in the S-line had a consistent increase of prediction accuracy when adding more animals from the opposite line. Adding just 1000 animals to the reference data of L-animals seems to give the same or slightly higher prediction accuracies compared to the pure within S-line predictions.

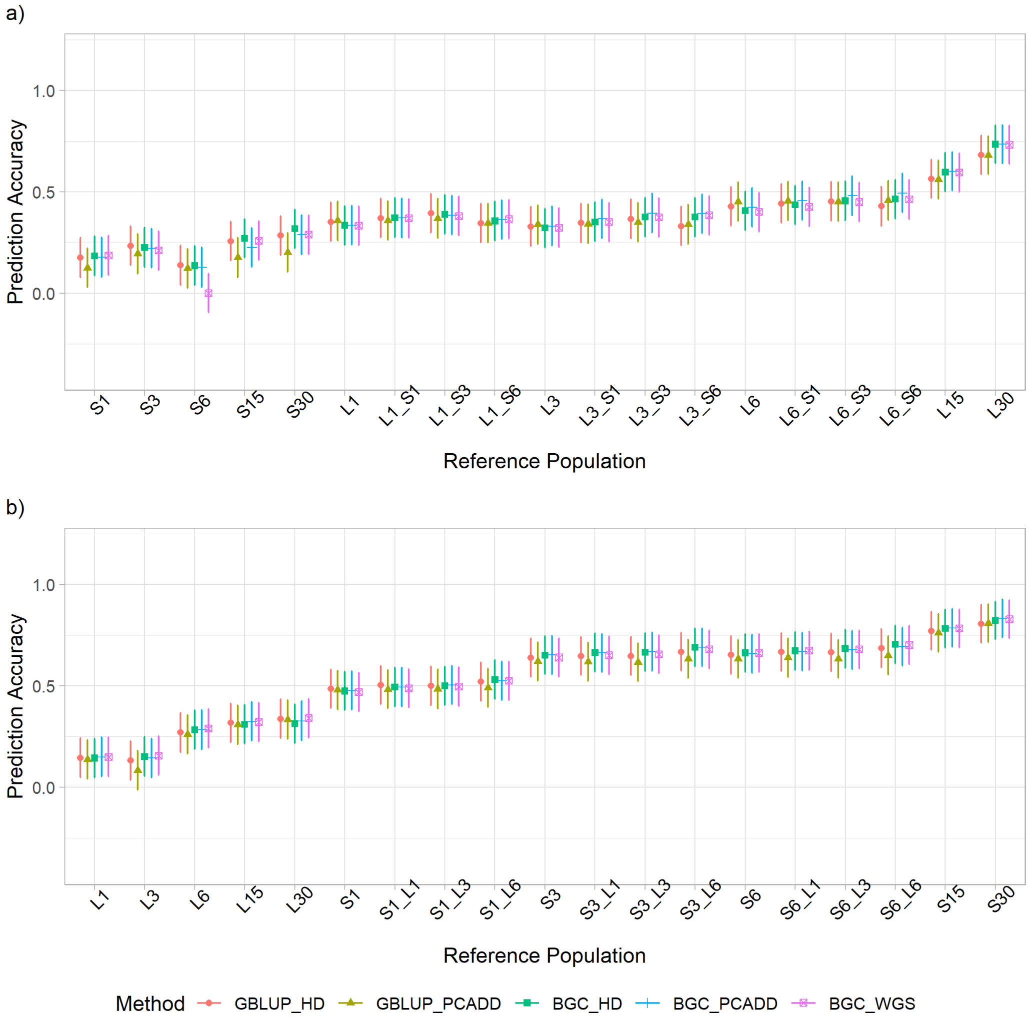


Figure 1. Accuracy of prediction for the trait Total Number Born (TNB) with validation animals from a) L-line animals and b) S-line animals, for different reference populations[1] and Methods[2]. Bars denote the Standard Error of the accuracy (SE).

[1]The name of the reference population refers to the line of the animals from the reference population, and the number of animals of that line, e.g., L1 is with a reference population with 1000 L-line animals, and S1_L3 is a reference population of 1000 S-line and 3000 L-line animals.

[2]GBLUP_HD is the results for the GBLUP method using the HD genotype data, GBLUP_PCADD is the GBLUP method using the pCADD genotype data, BGC_HD is the BayesGC method using HD genotype data, BGC_PCADD is the BayesGC method using pCADD data and BGC_WGS is the BayesGC method using Whole Genome Sequence data.

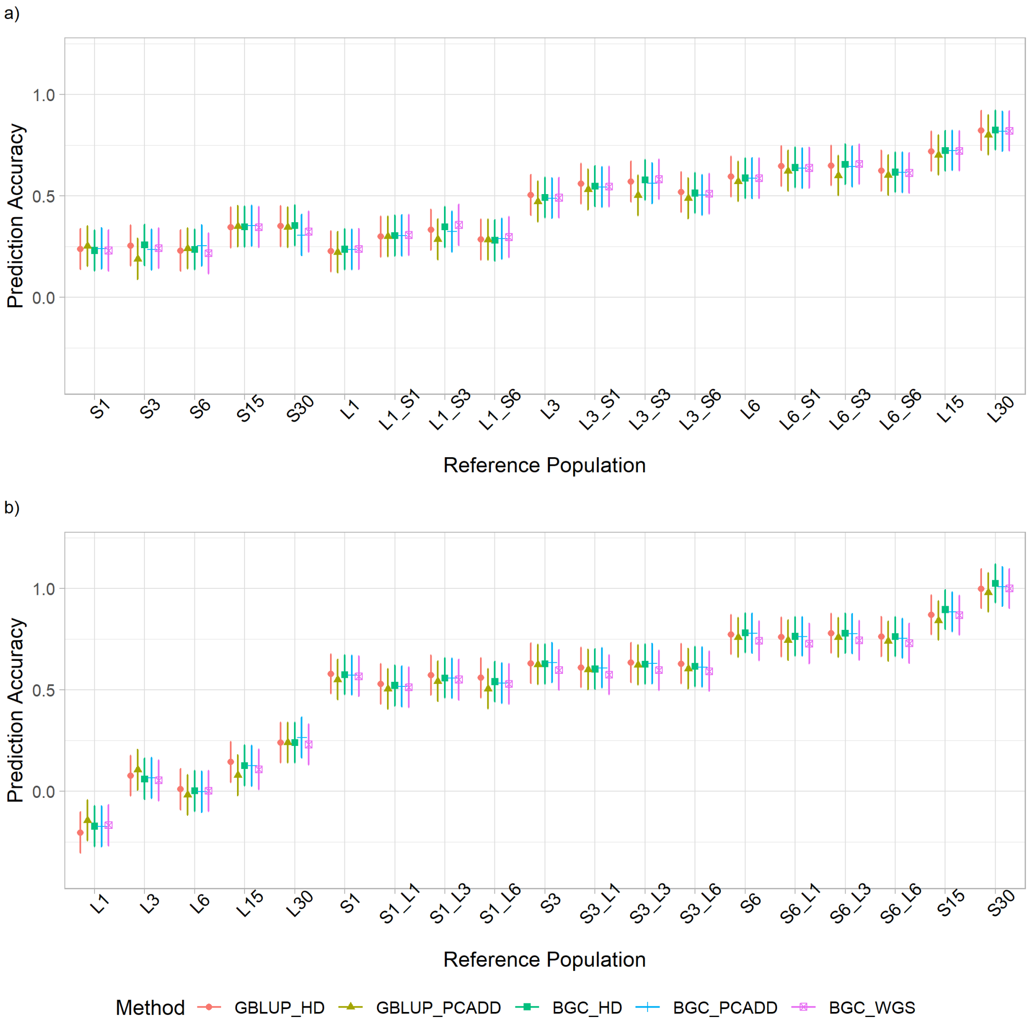


Figure 2. Accuracy of prediction for the trait Number of Stillborn (STB) with validation animals from a) L-line animals and b) S-line animals, for different reference populations[1] and Methods[2]. Bars denote the Standard Error of the accuracy (SE).

[1] The name of the reference population refers to the line of the animals from the reference population, and the number of animals of that line, e.g., L1 is with a reference population with 1000 L-line animals, and S1_L3 is a reference population of 1000 S-line and 3000 L-line animals.
[2] GBLUP_HD is the results for the GBLUP method using the HD genotype data, GBLUP_PCADD is the GBLUP method using the pCADD genotype data, BGC_HD is the BayesGC method using HD genotype data, BGC_PCADD is the BayesGC method using pCADD data and BGC_WGS is the BayesGC method using Whole Genome Sequence data.

Stillborn (STB) prediction accuracies ranged from 0.22 for L1 (GBLUP_PCADD) to 0.82 for L30 (GBLUP_HD, BGC_HD, BGC_PCADD, PGC_WGS) for the within-line prediction of L-line animals (Figure 2a). For across-line predictions the accuracies ranged from 0.19 for S1 (GBLUP_PCADD) to 0.35 for S15 (GBLUP_PXADD, BGC_HD and BGC_PCADD and BGC_WGS) and S30 (GBLUP_HD and BGC_HD).

Prediction accuracies for S-line validation animals for the STB trait (Figure 2b) ranged from 0.55 (S1, GBLUP_PCADD) to 1.03 for L30 (BGC_HD) for within-line predictions, from -0.20 (L1, GBLUP_HD) to 0.27 (L30, BGC_PCADD) for across-line predictions and from 0.50 (S1_L1, GBLUP_PCADD) to 0.78 (S6_L3, BGC_HD and BGC_PCADD) for multi-line predictions. An accuracy greater than 1 denotes a high prediction accuracy that is somewhat over-estimated, which is possible here since our estimator of the prediction accuracy is not bounded to the 0 – 1 interval.

In general, the accuracies for across-line prediction of S-animals for STB were low. For L1 all accuracies were negative, which showed again that accuracy estimates are not bound to the 0-1 interval. The highest prediction accuracies were obtained with the L30 dataset with accuracies ranging between 0.24 and 0.27. In comparison the prediction accuracy for within-line predictions were between 0.55-0.57 for S1.

The prediction accuracies for the multi-line reference data for L-line animals (Figure 2a) range from 0.28 for L1_S3 (BGC_PCADD and L1_S6 (BGC_PCADD and BGC_HD) to 0.66 for L6_S6 (BGC_WGS). Accuracies are slightly higher when including 1,000-3,000 animals from the S-line compared to within-line predictions. However, increasing the within-line proportion increases the accuracy more. When predicting on S-line animals using multi-line reference populations (Figure 2b), the accuracies range from 0.51-0.57 for S1_L1-L6, from 0.58-0.63 for S3_L1-L6 and from 0.73-0.78 for S6_L1-L6.

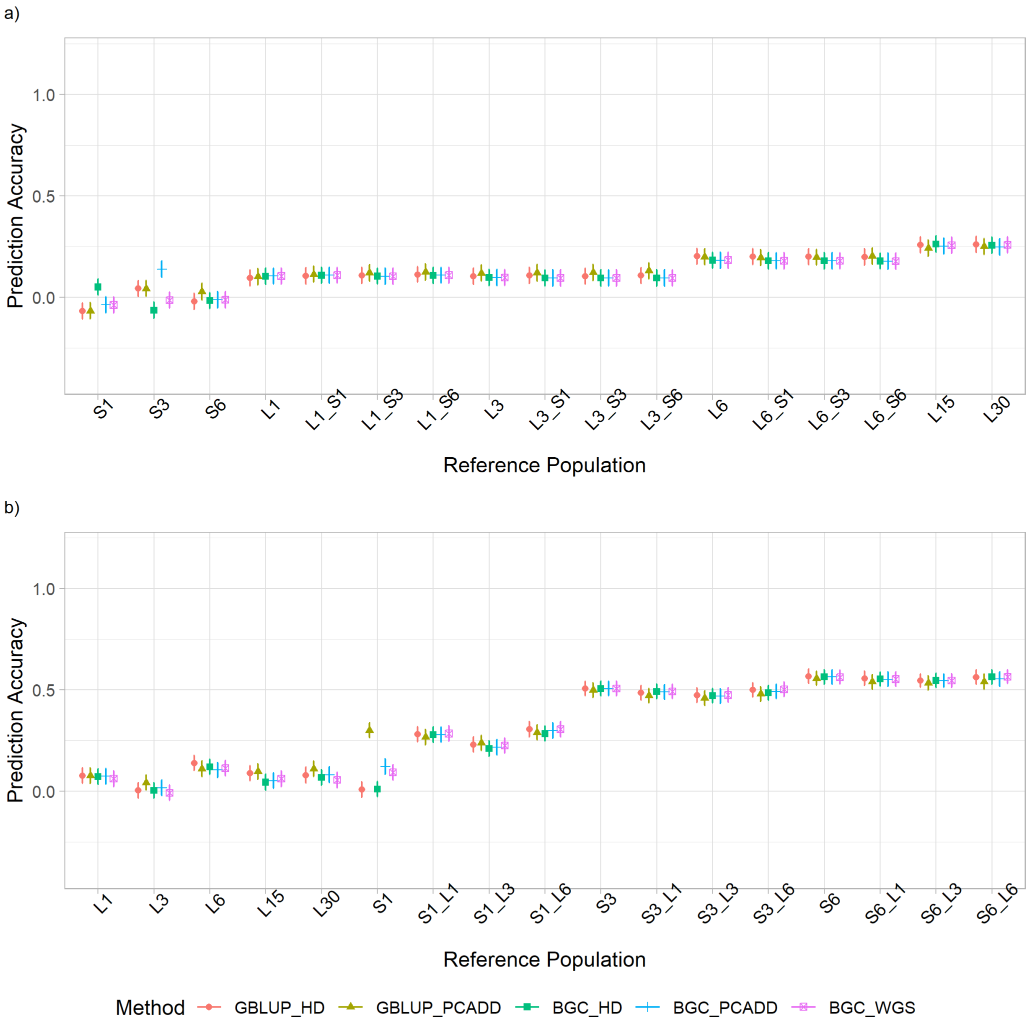


Figure 3. Accuracy of prediction for the trait Shoulder Lesions (SHL) with validation animals from a) L-line animals and b) S-line animals, for different reference populations[1] and Methods[2]. Bars denote the Standard Error of the accuracy (SE).

[1] The name of the reference population refers to the line of the animals from the reference population, and the number of animals of that line, e.g., L1 is with a reference population with 1000 L-line animals, and S1_L3 is a reference population of 1000 S-line and 3000 L-line animals.

[2] GBLUP_HD is the results for the GBLUP method using the HD genotype data, GBLUP_PCADD is the GBLUP method using the pCADD genotype data, BGC_HD is the BayesGC method using HD genotype data, BGC_PCADD is the BayesGC method using pCADD data and BGC_WGS is the BayesGC method using Whole Genome Sequence data.

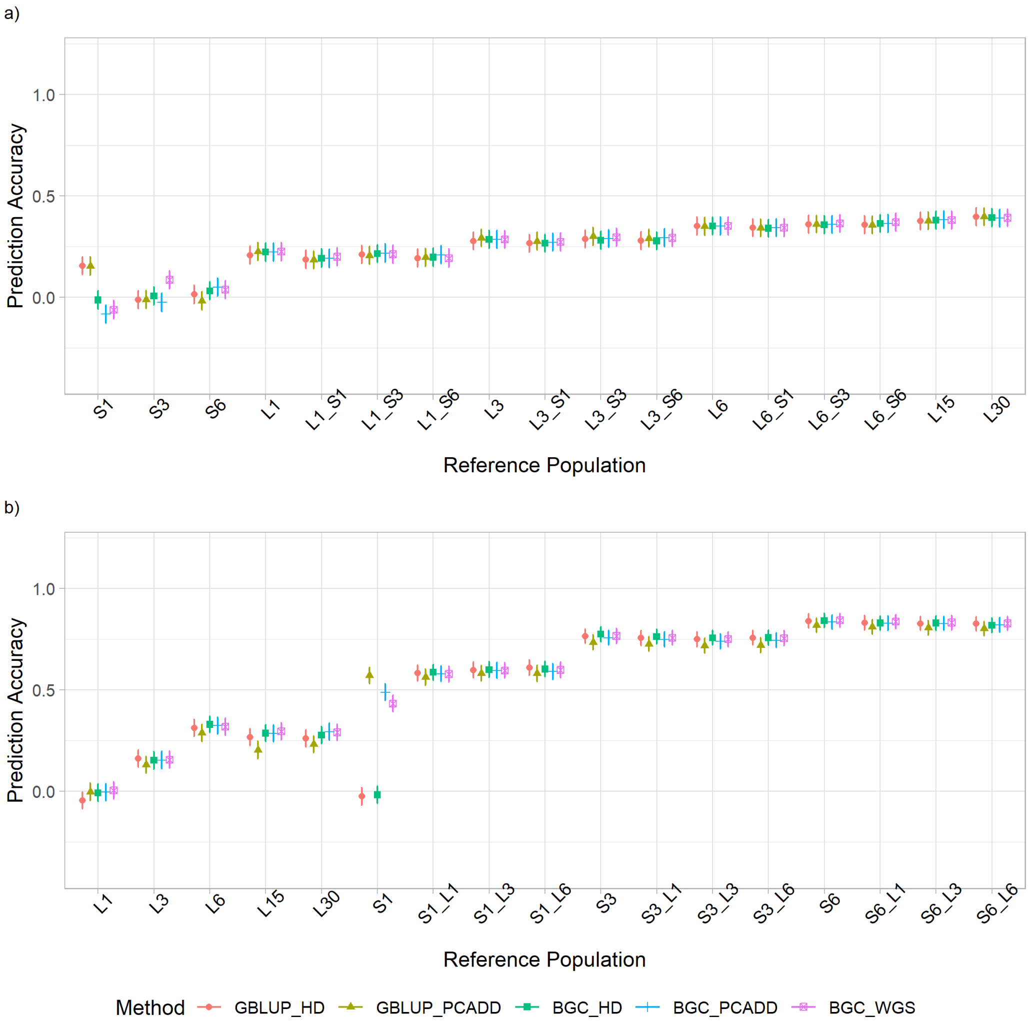


Figure 4. Accuracy of prediction for the trait Body Condition Score (BCS) with validation animals from a) L-line animals and b) S-line animals, for different reference populations[1] and Methods[2]. Bars denote the Standard Error of the accuracy (SE).

[1]The name of the reference population refers to the line of the animals from the reference population, and the number of animals of that line, e.g., L1 is with a reference population with 1000 L-line animals, and S1_L3 is a reference population of 1000 S-line and 3000 L-line animals.
[2]GBLUP_HD is the results for the GBLUP method using the HD genotype data, GBLUP_PCADD is the GBLUP method using the pCADD genotype data, BGC_HD is the BayesGC method using HD genotype data, BGC_PCADD is the BayesGC method using pCADD data and BGC_WGS is the BayesGC method using Whole Genome Sequence data.

The accuracies for within-line predictions for shoulder lesions (SHL) for L-line ranged from 0.09 for L1 (GBLUP_HD) to 0.26 (L15 GBLUP_HD and BGC_HD and L30 GBLUP_HD, BGC_HD and BGC_WGS) (Figure 3a). For across-line predictions, the accuracies ranged from -0.07 (S1 GBLUP_HD and GBLUP_PCADD) to 0.14 (BGC_PCADD) and for multi-line prediction the accuracies ranged from 0.10 (L1_S1 GBLUP_HD) to 0.20 (L6_S1 GBLUP_HD and L6_S3+L6_S6 GBLUP_HD and GBLUP_PCADD). Within-line prediction accuracies for SHL on S-line animals ranged from 0.01 (L1 GBLUP_HD and BGC_HD) to 0.57 (S6 GBLUP_HD) (Figure 3b). For across-line predictions the accuracies ranged from -0.01 (L3 BGC_WGS) to 0.14 (L6 GBLUP_HD). The multi-line prediction accuracies ranged from 0.21 (S1_L3 BGC_HD) to 0.56 (S6_L6 GBLUP_HD, BGC_HD and BGC_WGS, S6_L1 GBLUP_HD). The accuracies for within-line predictions with only 1,000 animals (S1) were rather low (0.01-0.30), but it improved greatly for S3 and S6 with a prediction accuracy of 0.5-0.56.

Prediction accuracies for Body Condition Score (BCS) for L-line animals (Figure 4a) ranged from 0.21 (L1 GBLUP_HD) to 0.4 (L30 GBLUP_HD) for within-line prediction, from -0.08 (S1 BGC_PCADD) to 0.15 (S1 GBLUP_HD and GBLUP_PCADD) for across-line prediction and from 0.18 (L1_S1 GBLUP_PCADD) to 0.37 (L6_S6 BGC_WGS) for multi-line prediction. The prediction accuracies for S-line animals for BCS (Figure 4b) ranged from -0.03 (S1 GBLUP_HD to 0.84 (S6 GBLUP_HD, BGC_HD, BGC_PCADD and BGC_WGS) for within-line predictions, from -0.05 (L1 GBLUP_HD) to 0.33 (L6 BGC_HD) for across-line predictions and from 0.56 (S1_L1 GBLUP_PCADD) to 0.84 (S6_L1 BGC_WGS) for multi-line predictions.

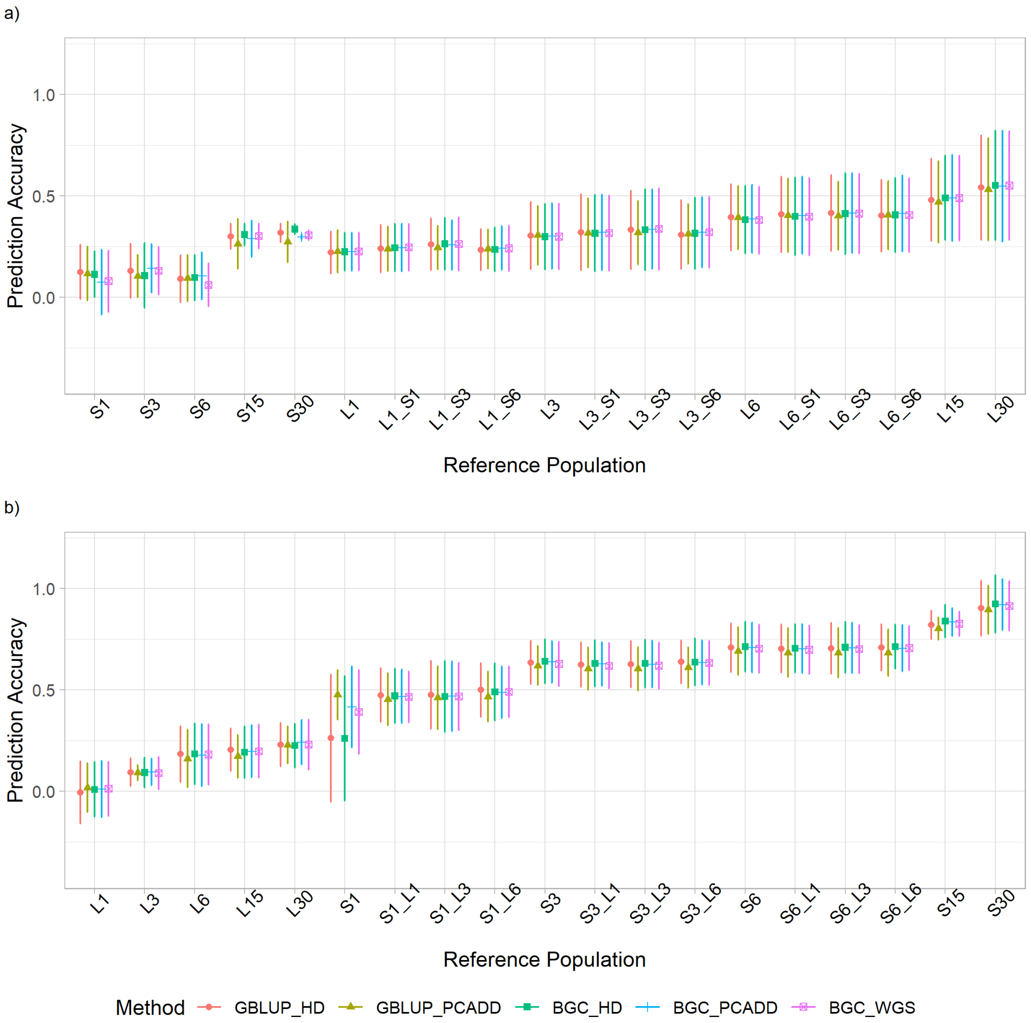


Figure 5. Mean prediction accuracy for all four traits, with validation animals from a) L-line animals and b) S-line animals, for different reference populations[1] and Methods[2]. Bars denote Standard Deviation (SD) of the accuracy of the four traits.

[1] The name of the reference population refers to the line of the animals from the reference population, and the number of animals of that line, e.g., L1 is with a reference population with 1000 L-line animals, and S1_L3 is a reference population of 1000 S-line and 3000 L-line animals.

[2] GBLUP_HD is the results for the GBLUP method using the HD genotype data, GBLUP_PCADD is the GBLUP method using the pCADD genotype data, BGC_HD is the BayesGC method using HD genotype data, BGC_PCADD is the BayesGC method using pCADD data and BGC_WGS is the BayesGC method using Whole Genome Sequence data.

Comparing genomic prediction methods

Significant differences between methods are shown in the supplementary tables 1-8. In general, for TNB, the GBLUP methods performed numerically but not significantly better than the BayesGC method for L1 and L3. For L6, GBLUP_HD, GBLUP_PCADD and BGC_PCADD performed significantly better than BGC_HD and BGC_WGS. However, for the larger reference data, L15 and L30, the BayesGC methods had a significantly higher accuracy compared to the GBLUP methods (Supplementary Table 1). For S15, (across) the accuracy obtained with BGC_HD was significantly higher than GBLUP_PCADD, and for S30, both BGC_HD and BGC_PCADD were significantly higher than GBLUP_PCADD. For prediction on TNB on S-line animals, for all except the S1 data set, the BayesGC methods were slightly but not significantly more accurate than the GBLUP methods for within-line prediction. For across-line prediction: the BGC_WGS analysis yielded consistently higher accuracies except for L15 where BGC_PCADD yielded the highest accuracy. However, none of the differences between methods for the across-line prediction analysis were significant. For multi-line prediction, the BayesGC methods perform better than both GBLUP methods for S3_L1-L6 and S6_L1-L6 (Supplementary Table 2).

For prediction on STB on L-line there was no significant difference between the methods for within-line or across-line predictions. For multi-line however, the methods BGC_HD, BGC_PCADD and BGC_WGS all had a significantly higher accuracy compared to GBLUP_PCADD for L1_S3. BGC_HD had also a higher prediction accuracy compared to GBLUP_HD and BGC_PCADD, while BGC_WGS also had a significantly higher accuracy than BGC_PCADD. For L3_S3 and L6_S3 all methods were significantly more accurate than GBLUP_PCADD. For L3_S3 BGC_HD and BGC_WGS had a significantly higher accuracy compared to BGC_PCADD (Supplementary Table 3).

There is no obvious difference between the methods for STB prediction of S-line animals. For prediction on SHL L-line animals, there was not a consistent difference between the methods, but the GBLUP methods tended to have a slightly higher accuracy compared to the BayesGC methods, and for example for L3_S1-S6 GBLUP_PCADD performed significantly better than the three BayesGC methods and for L6_S1-S6 GBLUP_HD performed significantly better than the three BayesGC methods (Supplementary Table 5). BGC_PCADD also had a significantly higher accuracy for S1 compared to GBLUP_HD, BGC_HD and BGC_WGS while BGC_WGS had a significantly higher accuracy compared to GBLUP_HD and BGC_HD for SHL predicted on S-line animals (Supplementary Table 6).

For BCS predicted on L-line, there were no differences between the methods except for across-line S1 where both GBLUP_HD and GBLUP_PCADD had a significantly higher prediction accuracy compared to BGC_PCADD and BGC_WGS and for S3 BGC_WGS had a significantly higher accuracy compared to BGC_PCADD (Supplementary Table 7). For BCS predicted on S-line, GBLUP_HD and BGC_HD performed significantly worse compared to the other methods for within-line S1 predictions (Supplementary Table 8).

When comparing the genomic prediction methods performance across traits (Figure 5) one cannot find a significant difference between the methods. The standard deviation (SD) seemed to increase when the reference population size was larger when predicting on L-line (Figure 5a). BGC_PCADD and BGC_WGS seemed to give slightly lower prediction accuracies across line for 1K animals (S1), but all methods performed better than L1 (whin-line) when number of across-line animals were 15-30K (Figure 5a). This was not the case for across-line prediction for S-line animals (Figure 5b). Averaging the prediction accuracy across traits shows a slight improvement for the three

BayesGC methods compared to the two GBLUP methods in both lines for reference population size >15,000 animals.

Discussion

Size of reference population

Across all the traits and lines, the size of the reference population and the relationship between the reference and validation animals influences the prediction accuracy most. For 3,000 within-line reference animals, the accuracies were consistently high across all the traits and methods. For SHL and BCS on S-line animals (Figures 6 and 8), the within-line prediction accuracy with a reference population of 1,000 animals (S1) did not seem large enough to be stable for all methods and genotypes. For example, GBLUP_HD and BGC_HD yielded zero accuracies. In contrast, GBLUP_PCADD acquired an accuracy similar to the results obtained for the multi-breed reference populations with added L-line animals (S1_L1). For smaller reference population sizes, the constitution of the animals in the reference population and their relationship with the validation animals become more critical compared to when the reference population is larger.

A larger reference population size generally always seemed to increase the prediction accuracy. However, there could be a limit beyond which a larger reference population does not significantly contribute to prediction accuracy. For Total Number Born (TNB) and Total Number of Stillborn (STB), there was a slight increase in accuracy even when increasing the size of the reference population from 15,000 to 30,000 individuals for both the L and S-line (Figures 1-4). However, for Shoulder Lesions (SHL) and Body Condition Score (BCS) for L-line animals, there was a difference increasing from 6- to 15,000 animals but not a significant increase in accuracy when going from 15- to 30,000 (Figures 5 and 7). Takeda et al. (2021) found a reference population of

7,000-11,000 animals to be sufficient for genomic prediction methods to reach an accuracy similar to that of progeny testing for within breed prediction, which is in line with our results.

Across and multi-line reference population

The across-line prediction accuracy was generally low, with values close to zero. When having as many as 30,000 animals in the across-line reference population, the accuracy was for several traits close to the accuracy of 1,000 within-line reference animals. For STB with S-line reference predicting on L-line, the across-line prediction was higher for S15 and S30 than for L1. When adding animals from a different line for multi-line prediction, prediction accuracies generally improved compared to the accuracy of the pure within-line prediction with the same number of within-line animals in the reference data. However, there were some differences between the different traits and lines, and in some cases, adding more animals from a different line did not improve the prediction accuracy compared to pure within-line predictions.

Adding S-line to predict L-line generally improved prediction accuracy, while adding L-line animals to predict S-line generally lowered the prediction accuracy slightly. However, adding too many S-line animals could also decrease the prediction accuracy of L-line animals. For instance, the multi-line reference population containing 3K S-line animals often performed better than the prediction accuracy that included 6K S-line animals.

It is not always an advantage to increase the reference population of the "foreign" breed in a multi-breed reference population when the breed of interest has a small number of animals, as the SNP effects estimates become dominated by the larger breed. Other studies have looked for solutions to this issue. For example, Karaman et al. (2021) found that adding an admixed population could be a way for the prediction method to account for the breed origin of the alleles. Hayes et al. (2009)

found that a Bayesian variable selection method yields the best accuracy for multi-breed prediction, maybe because the Bayesian variable selection methods yield more accurate estimates of individual SNP effects.

Differences between lines

In addition to the reference population's size, the reference population's genetic constitution affects prediction accuracies. The S-line had a generally higher prediction accuracy than the L-line for all traits. The differences are likely due to the relatedness between reference and validation animals and the effective number of chromosomal segments, M_e (van den Berg et al. 2019). The L-line is a closed pure-bred line with no new genetics introduced in the last 60-70 years, while the S-line is a synthetic line with at least two breeds or lines admixed. The synthetic line will have more long-range LD which helps in prediction accuracy, especially when reference population size is small, as there will not be so many independent segments that need to be evaluated. When predicting across-lines, the chromosomal segments in common are smaller, and M_e increases. One needs more records to estimate the increased number of chromosomal segments and thus a large reference population to get a high prediction accuracy. Our study has a large reference population and many markers, which should be sufficient to estimate the larger number of chromosomal segments in the S-line.

Differences between lines could also be due to a difference in heritability between the lines. If one of the lines has a higher heritability, it would be more valuable for within- and across-line prediction. Our study estimated the variance component for the two lines as if they were the same for both populations. Heritabilities estimated for pure L-line animals were 0.19, 0.13, 0.34 and 0.31 for TNB, STB, SHL and BCS, respectively (Kjetså et al. 2022). For the combined L and S- lines, heritabilities were 0.09, 0.08, 0.59 and 0.47 for TNB, STB, SHL and BCS. To compare the results of within, across, and multi-line reference populations, the decision was made to use the variance components estimated from the combined population, which may also explain why some prediction accuracies were estimated above 1, as the accuracies may be scaled with a too low (across-line) heritability.

Genotypes

One of the aims of this study was to see if using marker data based on pCADD values would improve prediction accuracies. There was little difference between the genotype marker sets of pCADD, High Density (HD) and Whole Genome Sequence (WGS.) However, the pCADD data tended to give a slightly lower prediction accuracy compared to the others, especially when combined with GBLUP. For TNB predicted on L and S-line (Tables 4 and 5) and BCS predicted on L line (Supplementary Table 7), BGC_PCADD had a significantly higher accuracy than GBLUP_PCADD for many of the reference populations, especially for a sizeable within-line reference population and multi-line reference populations. The HD and pCADD marker sets had approximately the same number of markers (433K and 417K, respectively). However, the average heterozygosity for pCADD, HD and WGS were 0.37, 0.40 and 0.37, respectively. Hence, the average heterozygosity and thus marker information was lower for pCADD and WGS compared to HD. In addition, pCADD marker genotypes and WGS data relied heavily on genotype imputation, whereas HD genotypes were primarily obtained from direct SNP-chip genotyping, which may have impaired the prediction accuracies obtained by pCADD and WGS data.

Imputation and genotyping errors

Most studies using high-density SNP panels and WGS data, including the current, use genotype imputation to increase the size of the reference population with similar (high) density of marker genotypes. Studies on the accuracy of imputation do generally report high imputation accuracy.

Larmer et al. (2011) reported accuracies of 0.89-0.99 for cattle imputing from 6K to HD (~777K SNP panel) and from 50K to HD, where the smaller breed generally gave a lower accuracy compared to the larger breed and the accuracy of imputation was higher when imputing from 50K to HD. Van Binsbergen et al. (2014) reported higher accuracies when imputing stepwise from 50K to HD and then to WGS, compared to a direct imputation from 50K to WGS. The studies on imputation accuracies report that imputation accuracies can vary for different SNPs and the location on the genome. Many factors can affect the accuracy of imputation, such as the size and the composition of the reference population and the relationship between the reference population and the imputed population. When comparing different marker genotypes, the lack of increase in accuracy when going to higher density was likely partly due to imputation errors. WGS data is also a lot more prone to genotyping errors than SNP-chip genotypes (Pérez-Enciso, Rincón, and Legarra 2015), which could also affect the accuracy of GEBVs when using WGS data.

Prediction methods

A linear model Genomic Best Linear Unbiased Prediction (GBLUP) and a Bayesian method (BayesGC) were tested for both the high density (HD) marker set and the pCADD marker set. The WGS data was only tested with BayesGC since a variable selection genomic prediction analysis is required to make best use of WGS data (Meuwissen et al. 2021). GBLUP does not generally improve when WGS data is used (Van Binsbergen et al. 2015), probably because a lower marker density is enough to construct the genomic relationship matrix accurately. The differences between the GBLUP and BayesGC analyses were minor in our data, although BayesGC tended to have slightly higher accuracy, although often not significant. In some of the data sets, GBLUP yielded even higher accuracy than BayesGC, but this was only seen for small reference populations, indicating that one needs a reasonably large reference population to benefit from Bayesian variable selection.

Further developments

Future studies are needed to optimise the reference populations for multi-line predictions. It may be possible to optimise the number of reference animals from a different breed. Our study also shows that one line could be more valuable as a reference than another. Hence, it is possible to choose a suitable line for boosting prediction accuracy, such as S-line animals to predict L-line animals but not the other way around. Further investigations would be needed to confirm this.

Furthermore, we could utilise WGS data to extract markers that aid HD-based prediction. It has been shown that including pre-selected markers in high LD with QTL derived from WGS can improve the GBLUP prediction (Brøndum et al. 2015; Warburton et al. 2020). Further development could be to combine a top 1% or 0.1% pCADD markers with markers known to have high LD with QTL, and neutral markers derived from WGS, covering the genome densely and evenly across the genome to account for markers in LD with potentially unknown QTL.

Conclusions

The main contributor to prediction accuracy is the size of the within-line reference population, where 3,000-6,000 animals were sufficient to get a high prediction accuracy of >0.5 for all traits except TNB predicted on L-line and BCS. However, increasing to 30,000 animals in the reference population further increases prediction accuracy for traits. A reference population of 30,000 animals for across-line prediction could achieve similar accuracy as 1,000 within-line animals. For multi-line prediction accuracy, the accuracy was most dependent on the number of within-line animals in the reference data. Adding S-line was more beneficial for multi-breed prediction on L-line than vice versa. Using pCADD scores to reduce the number of markers from WGS data combined with GBLUP generally reduced prediction accuracies relative to GBLUP_HD analyses, probably due to the lower information content of the pCADD markers. When using BayesGC, prediction accuracies were generally similar when using HD, pCADD or WGS marker data, which suggests that the variable selection method selects a suitable set of markers irrespective of the marker set provided (HD, pCADD or WGS).

Acknowledgements

We are grateful to acknowledge funding from the Norwegian Research Council (project 255297). Topigs Norsvin is gratefully acknowledged for providing data and genotype information.

List of abbreviations

BCS – Body Condition Score

BLUP – Best Linear Unbiased Prediction

GBLUP – Genomic Best Linear Unbiased Prediction

GEBV – Genomic Estimated Breeding Values

GS – Genomic Selection

HD – High Density

LD – Linkage Disequilibrium

MCMC – Markov Chain Monte Carlo

pCADD – pig Combined Annotation Dependent Depletion

QTL – Quantitative Trait Loci

STB – total number of Still Born piglets

SHL – Shoulder Lesions

SNP – Single Nucleotide Polymorphism

SNV – Single Nucleotide Variants

TNB – Total Number of Born piglets

WGS – Whole Genome Sequence

YD – Yield Deviations

References

van den Berg, I., T. H. E. Meuwissen, I. M. MacLeod, and M. E. Goddard. 2019. "Predicting the Effect of Reference Population on the Accuracy of within, across, and Multibreed Genomic Prediction." *Journal of Dairy Science* 102(4):3155–74.

van Binsbergen, Rianne, Marco C. A. M. Bink, Mario P. L. Calus, Fred A. van Eeuwijk, Ben J. Hayes, Ina Hulsegge, and Roel F. Veerkamp. 2014. "Accuracy of Imputation to Whole-Genome Sequence Data in Holstein Friesian Cattle." *Genetics Selection Evolution* 46(1):41.

Van Binsbergen, Rianne, Mario P. L. Calus, Marco C. A. M. Bink, Fred A. Van Eeuwijk, Chris Schrooten, and Roel F. Veerkamp. 2015. "Genomic Prediction Using Imputed Whole-Genome Sequence Data in Holstein Friesian Cattle." *Genetics Selection Evolution* 47(1):1–13.

Brøndum, RF, G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and MS Lund. 2015. "Quantitative Trait Loci Markers Derived from Whole Genome Sequence Data Increases the Reliability of Genomic Prediction." *Journal of Dairy Science* 98:4107–16.

Chang, Christopher C., Carson C. Chow, Laurent C. A. M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4(1):s13742-015-0047–0048.

Clark, Samuel A., John M. Hickey, Hans D. Daetwyler, and Julius H. J. van der Werf. 2012. "The Importance of Information on Relatives for the Prediction of Genomic Breeding Values and the Implications for the Makeup of Reference Data Sets in Livestock Breeding Schemes." *Genetics Selection Evolution* 44(1):4.

Clark, Samuel A., John M. Hickey, and Julius Hj Van Der Werf. 2011. *Different Models of Genetic Variation and Their Effect on Genomic Evaluation*. Vol. 43.

Daetwyler, Hans D., Ricardo Pong-Wong, Beatriz Villanueva, and John A. Woolliams. 2010. "The Impact of Genetic Architecture on Genome-Wide Evaluation Methods." *Genetics*

185(3):1021–31.

Das, Sayantan, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E. Locke, Alan Kwong, Scott I. Vrieze, Emily Y. Chew, Shawn Levy, Matt McGue, David Schlessinger, Dwight Stambolian, Po-Ru Loh, William G. Iacono, Anand Swaroop, Laura J. Scott, Francesco Cucca, Florian Kronenberg, Michael Boehnke, Gonçalo R. Abecasis, and Christian Fuchsberger. 2016. "Next-Generation Genotype Imputation Service and Methods." *Nature Genetics* 48(10):1284–87.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. "Toward Genomic Prediction from Whole-Genome Sequence Data: Impact of Sequencing Design on Genotype Imputation and Accuracy of Predictions." *Heredity* 112:39–47.

Efron, B. Tibishirani, R. J. 1994. "An Introduction to the Bootstrap." *Boca Raton: CRC Press LLC*. Retrieved November 19, 2019 (https://books.google.no/books?hl=en&lr=&id=gLlpIUxRntoC&oi=fnd&pg=PR14&ots=A9BvU8J7F2&sig=rU1bHQeofAkRYvjRIucY5ei_XkQ&redir_esc=y#v=onepage&q&f=false).

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. "Improving Accuracy of Genomic Predictions within and between Dairy Cattle Breeds with Imputed High-Density Single Nucleotide Polymorphism Panels." *Journal of Dairy Science* 95(7):4114–29.

Gianola, Daniel. 2013. "Priors in Whole-Genome Regression: The Bayesian Alphabet Returns." *Genetics* 194(3):573–96.

Goddard, Mike. 2009. "Genomic Selection: Prediction of Accuracy and Maximisation of Long Term Response." *Genetica* 136(2):245–57.

Groß, Christian, Martijn Derks, Hendrik-Jan Megens, Mirte Bosse, Martien A. M. Groenen, Marcel Reinders, and Dick de Ridder. 2020. "PCADD: SNV Prioritisation in Sus Scrofa." *Genetics*

Selection Evolution 52(1):4.

Habier, David, Jens Tetens, Franz-Reinhold Seefried, Peter Lichtner, and Georg Thaller. 2010. “The Impact of Genetic Relationship Information on Genomic Breeding Values in German Holstein Cattle.” *Genetics Selection Evolution* 42(1):5.

Hayes, Ben J., Phillip J. Bowman, Amanda C. Chamberlain, Klara Verbyla, and Mike E. Goddard. 2009. “Accuracy of Genomic Breeding Values in Multi-Breed Dairy Cattle Populations.” *Genetics Selection Evolution* 41(1):1–9.

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, and P. Croiseau. 2014. “Efficiency of Multi-Breed Genomic Selection for Dairy Cattle Breeds with Different Sizes of Reference Population.”

Iheshiulor, Oscar O. M., John A. Woolliams, Xijiang Yu, Robin Wellmann, and Theo H. E. Meuwissen. 2016. “Within- and across-Breed Genomic Prediction Using Whole-Genome Sequence and Single Nucleotide Polymorphism Panels.” *Genetics Selection Evolution* 48(1):15.

Iversen, Maja Winther, Øyvind Nordbø, Eli Gjerlaug-Enger, Eli Grindflek, Marcos Soares Lopes, and Theo Meuwissen. 2019. “Effects of Heterozygosity on Performance of Purebred and Crossbred Pigs.” *Genetics Selection Evolution* 51(8).

Kachman, Stephen D., Matthew L. Spangler, Gary L. Bennett, Kathryn J. Hanford, Larry A. Kuehn, Warren M. Snelling, R. Mark Thallman, Mahdi Saatchi, Dorian J. Garrick, Robert D. Schnabel, Jeremy F. Taylor, and E. John Pollak. 2013. “Comparison of Molecular Breeding Values Based on Within- and across-Breed Training in Beef Cattle.” *Genetics Selection Evolution* 45(1):30.

Karaman, Emre, Guosheng Su, Iola Croue, and Mogens S. Lund. 2021. “Genomic Prediction Using a Reference Population of Multiple Pure Breeds and Admixed Individuals.” *Genetics Selection*

Evolution 53(1):46.

Kjetså, M. V., A. B. Gjuvsland, Ø. Nordbø, E. Grindflek, and T. Meuwissen. 2022. *Accuracy of Genomic Prediction of Maternal Traits in Pigs Using Bayesian Variable Selection Methods*.

Larmer, S., M. Sargolzaei, R. Ventura, and F. Schenkel. 2011. “Imputation Accuracy from Low to High Density Using within and across Breed Reference Populations in Holstein, Guernsey and Ayrshire Cattle.” *Cgil.Uoguelph.Ca*.

Loh, Po-Ru, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R. Abecasis, Richard Durbin, and Alkes L Price. 2016. “Reference-Based Phasing Using the Haplotype Reference Consortium Panel.” *Nature Genetics* 48(11):1443–48.

Madsen, Per and Just Jensen. 2013. *A User’s Guide to DMU A Package for Analysing Multivariate Mixed Models*.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. “Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps.” *Genetics* 157(4):1819–29.

Meuwissen, Theo, Irene van den Berg, and Mike Goddard. 2021. “On the Use of Whole-Genome Sequence Data for across-Breed Genomic Prediction and Fine-Scale Mapping of QTL.” *Genetics Selection Evolution* 53(1):19.

Ocepek, Marko and Inger Lise Andersen. 2017. “What Makes a Good Mother? Maternal Behavioural Traits Important for Piglet Survival.” *Applied Animal Behaviour Science* 193:29–36.

Pérez-Enciso, Miguel, Juan C. Rincón, and Andrés Legarra. 2015. “Sequence- vs. Chip-Assisted Genomic Selection: Accurate Biological Information Is Advised.” *Genetics, Selection, Evolution : GSE* 47(1):43.

Purcell, Shaun and Christopher Chang. 2019. “PLINK 1.9.” Retrieved (www.coggenomics.org/plink/1.9/).

De Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. “Reliability of Genomic Predictions across Multiple Populations.” *Genetics* 183(4):1545–53.

Sargolzaei, Mehdi, Jacques P. Chesnais, and Flavio S. Schenkel. 2014. “A New Approach for Efficient Genotype Imputation Using Information from Relatives.” *BMC Genomics 2014 15:1* 15(1):1–12.

Takeda, Masayuki, Keiichi Inoue, Hidemi Oyama, Katsuo Uchiyama, Kanako Yoshinari, Nanae Sasago, Takatoshi Kojima, Masashi Kashima, Hiromi Suzuki, Takehiro Kamata, Masahiro Kumagai, Wataru Takasugi, Tatsuya Aonuma, Yuusuke Soma, Sachi Konno, Takaaki Saito, Mana Ishida, Eiji Muraki, Yoshinobu Inoue, Megumi Takayama, Shota Nariai, Ryoya Hideshima, Ryoichi Nakamura, Sayuri Nishikawa, Hiroshi Kobayashi, Eri Shibata, Koji Yamamoto, Kenichi Yoshimura, Hironori Matsuda, Tetsuro Inoue, Atsumi Fujita, Shohei Terayama, Kazuya Inoue, Sayuri Morita, Ryotaro Nakashima, Ryohei Suezawa, Takeshi Hanamure, Atsushi Zoda, and Yoshinobu Uemoto. 2021. “Exploring the Size of Reference Population for Expected Accuracy of Genomic Prediction Using Simulated and Real Data in Japanese Black Cattle.” *BMC Genomics* 22(1):1–11.

VanRaden, Paul M. 2007. “Genomic Measures of Relationship and Inbreeding.” *Interbull Bulletin* 25(37):111–14.

Warburton, Christie L., Bailey N. Engle, Elizabeth M. Ross, Roy Costilla, Stephen S. Moore, Nicholas J. Corbet, Jack M. Allen, Alan R. Laing, Geoffry Fordyce, Russell E. Lyons, Michael R. McGowan, Brian M. Burns, and Ben J. Hayes. 2020. “Use of Whole-Genome Sequence Data and Novel Genomic Selection Strategies to Improve Selection for Age at Puberty in Tropically-Adapted Beef Heifers.” *Genetics Selection Evolution* 52(1):1–13.

Wetterstrand KA. 2021. “DNA Sequencing Costs: Data from the NHGRI Genome Sequencing

Program (GSP) Available at: Www.Genome.Gov/Sequencingcostsdata. Accessed 2016-09-05." Retrieved March 27, 2022 (www.genome.gov/sequencingcostsdata/).

Wientjes, Yvonne C. J., Roel F. Veerkamp, and Mario P. L. Calus. 2013. "The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction." *Genetics* 193(2):621–31.

Zhou, L, B. Heringstad, G. Su, B. Guldbrandtsen, The Meuwissen, M. Svendsen, H. Grove, Us Nielsen, and MS Lund. 2014. "Genomic Predictions Based on a Joint Reference Population for the Nordic Red Cattle Breeds."

Zhou, L., M. S. Lund, Y. Wang, and G. Su. 2014. "Genomic Predictions across Nordic Holstein and Nordic Red Using the Genomic Best Linear Unbiased Prediction Model with Different Genomic Relationship Matrices." *Journal of Animal Breeding and Genetics* 131(4):249–57.

Supplementary Information

Supplementary Table 1. Accuracy of prediction for L-line animals for the trait Total Number Born (TNB).

	Ref. Pop.	**GBLUP HD**	**GBLUP PCADD**	**BGC HD**	**BGC PCADD**	**BGC WGS**
Within-line	L1	0.35±0.10	0.36±0.10	0.33±0.10	0.33±0.10	0.33±0.10
	L3	0.33±0.10	0.34±0.10	0.32±0.10	0.33^{c}±0.10	0.32±0.10
	L6	0.43^{ce}±0.10	0.45^{ce}±0.10	0.41±0.10	0.42^{ce}±0.10	0.40±0.10
	L15	0.56±0.10	0.56±0.10	0.60^{a}±0.10	0.60^{ab}±0.10	0.60±0.10
	L30	0.68±0.09	0.68±0.09	0.73^{ab}±0.09	0.73^{ab}±0.09	0.73^{ab}±0.09
Across-Line	S1	0.17±0.10	0.12±0.10	0.18±0.10	0.18±0.10	0.19±0.10
	S3	0.23±0.10	0.19±0.10	0.23±0.10	0.22±0.10	0.21±0.10
	S6	0.14±0.10	0.12±0.10	0.14±0.10	0.13±0.10	0.00±0.10
	S15	0.26±0.10	0.18±0.10	0.27^{b}±0.10	0.23±0.10	0.26±0.10
	S30	0.28±0.10	0.20±0.10	0.32^{b}±0.10	0.29^{b}±0.10	0.29±0.10
Multi-Line	L1_S1	0.37±0.10	0.36±0.10	0.37±0.10	0.37±0.10	0.37±0.10
	L1_S3	0.39±0.10	0.37±0.10	0.39±0.10	0.38±0.10	0.38±0.10
	L1_S6	0.34±0.10	0.35±0.10	0.36±0.10	0.36±0.10	0.36±0.10
	L3_S1	0.35±0.10	0.34±0.10	0.35±0.10	0.37^{ac}±0.10	0.35±0.10
	L3_S3	0.37±0.10	0.35±0.10	0.37±0.10	0.40^{b}±0.10	0.37±0.10
	L3_S6	0.33±0.10	0.34±0.10	0.37^{a}±0.10	0.39^{ab}±0.10	0.38^{a}±0.10
	L6_S1	0.44±0.10	0.45±0.10	0.44^{e}±0.10	0.46^{ce}±0.10	0.42±0.10
	L6_S3	0.45±0.10	0.45±0.10	0.46±0.10	0.48^{ce}±0.10	0.45±0.10
	L6_S6	0.43±0.10	0.46±0.10	0.46^{a}±0.10	0.49^{abc}±0.10	0.46^{a}±0.10

Ref. Pop. Is the reference population, Letters L and S indicates animals belonging to either L- or S- line and the number indicating number of animals in reference population in number of 1,000. **GBLUP** and **BGC** is the statistical method (GBLUP or BayesGC), **HD** indicates the High Density marker data, **PCADD** indicates the pCADD marker data and **WGS** indicates the Whole Genome Sequence marker data.

[a] indicates a significantly larger prediction accuracy compared to GBLUP_HD
[b] indicates a significantly larger prediction accuracy compared to GBLUP_PCADD
[c] indicates a significantly larger prediction accuracy compared to BGC_HD
[d] indicates a significantly larger prediction accuracy compared to BGC_PCADD
[e] indicates a significantly larger prediction accuracy compared to BGC_WGS

Supplementary Table 2. Accuracy of prediction for S-line animals for the trait Total Number Born (TNB).

	Ref. Pop.	**GBLUP HD**	**GBLUP PCADD**	**BGC HD**	**BGC PCADD**	**BGC WGS**
Within-line	S1	0.49±0.10	0.48±0.10	0.47±0.10	0.48±0.10	0.47±0.10
	S3	0.64±0.09	0.62±0.09	0.65±0.09	0.65[b]±0.09	0.64±0.09
	S6	0.65±0.09	0.63±0.09	0.66±0.09	0.66±0.09	0.66±0.09
	S15	0.77±0.09	0.76±0.09	0.78±0.09	0.79±0.09	0.78±0.09
	S30	0.81±0.09	0.81±0.09	0.82±0.09	0.83±0.09	0.83±0.09
Across-Line	L1	0.15±0.10	0.14±0.10	0.14±0.10	0.15±0.10	0.15±0.10
	L3	0.13±0.10	0.08±0.10	0.15±0.10	0.15±0.10	0.16±0.10
	L6	0.27±0.10	0.26±0.10	0.28±0.10	0.28±0.10	0.29±0.10
	L15	0.32±0.10	0.31±0.10	0.31±0.10	0.33±0.10	0.32±0.10
	L30	0.34±0.10	0.33±0.10	0.31±0.10	0.33±0.10	0.34±0.10
Multi-Line	S1_L1	0.50±0.10	0.48±0.10	0.49±0.10	0.49±0.10	0.49±0.10
	S1_L3	0.50±0.10	0.49±0.10	0.50±0.10	0.50±0.10	0.50±0.10
	S1_L6	0.52±0.10	0.49±0.10	0.53±0.10	0.53±0.10	0.53±0.10
	S3_L1	0.65±0.09	0.62±0.09	0.66[b]±0.09	0.66[b]±0.09	0.65±0.09
	S3_L3	0.65±0.09	0.62±0.09	0.67[ab]±0.09	0.67[b]±0.09	0.66±0.09
	S3_L6	0.67±0.09	0.63±0.09	0.69[ab]±0.09	0.69[b]±0.09	0.68±0.09
	S6_L1	0.67±0.09	0.64±0.09	0.67±0.09	0.67[b]±0.09	0.67±0.09
	S6_L3	0.67±0.09	0.63±0.09	0.68[ab]±0.09	0.68[b]±0.09	0.68[b]±0.09
	S6_L6	0.69±0.09	0.65±0.09	0.70[ab]±0.09	0.69[b]±0.09	0.70[b]±0.09

Ref. Pop. Is the reference population, Letters L and S indicates animals belonging to either L- or S- line and the number indicating number of animals in reference population in number of 1,000. **GBLUP** and **BGC** is the statistical method (GBLUP or BayesGC), **HD** indicates the High Density marker data, **PCADD** indicates the pCADD marker data and **WGS** indicates the Whole Genome Sequence marker data.

[a] indicates a significant difference from GBLUP_HD
[b] indicates a significant difference from GBLUP_PCADD
[c] indicates a significant difference from BGC_HD
[d] indicates a significant difference from BGC_PCADD
[e] indicates a significant difference from BGC_WGS

Supplementary Table 3. Accuracy of prediction for L-line animals for the trait Number of Stillborn (STB).

	Ref. Pop.	GBLUP HD	GBLUP PCADD	BGC HD	BGC PCADD	BGC WGS
Within-line	L1	0.23±0.10	0.22±0.10	0.24±0.10	0.24±0.10	0.24±0.10
	L3	0.50±0.10	0.47±0.10	0.49[b]±0.10	0.49±0.10	0.49±0.10
	L6	0.60±0.10	0.57±0.10	0.59±0.10	0.59±0.10	0.59±0.10
	L15	0.72±0.10	0.70±0.10	0.72±0.10	0.72±0.10	0.72±0.10
	L30	0.82±0.10	0.80±0.10	0.82±0.10	0.82±0.10	0.82±0.10
Across-Line	S1	0.24±0.10	0.25±0.10	0.23±0.10	0.24±0.10	0.23±0.10
	S3	0.25±0.10	0.19±0.10	0.26±0.10	0.23±0.10	0.24±0.10
	S6	0.23±0.10	0.24±0.10	0.23±0.10	0.25±0.10	0.22±0.10
	S15	0.34±0.10	0.35±0.10	0.35±0.10	0.35±0.10	0.35±0.10
	S30	0.35±0.10	0.34±0.10	0.35±0.10	0.31±0.10	0.32±0.10
Multi-Line	L1_S1	0.30±0.10	0.30±0.10	0.30±0.10	0.30±0.10	0.31±0.10
	L1_S3	0.33±0.10	0.28±0.10	0.35[abd]±0.10	0.32[b]±0.10	0.36[bd]±0.10
	L1_S6	0.28±0.10	0.28±0.10	0.28±0.10	0.29±0.10	0.30±0.10
	L3_S1	0.56±0.10	0.53±0.10	0.55±0.10	0.54±0.10	0.55±0.10
	L3_S3	0.57[b]±0.10	0.50±0.10	0.58[bd]±0.10	0.56[b]±0.10	0.58[bd]±0.10
	L3_S6	0.52±0.10	0.49±0.10	0.51±0.10	0.50±0.10	0.51±0.10
	L6_S1	0.65±0.10	0.62±0.10	0.64±0.10	0.64±0.10	0.64±0.10
	L6_S3	0.65[b]±0.10	0.60±0.10	0.65[b]±0.10	0.64[b]±0.10	0.66[b]±0.10
	L6_S6	0.62±0.10	0.60±0.10	0.62±0.10	0.61±0.10	0.61±0.10

Ref. Pop. Is the reference population, Letters L and S indicates animals belonging to either L- or S- line and the number indicating number of animals in reference population in number of 1,000. **GBLUP** and **BGC** is the statistical method (GBLUP or BayesGC), **HD** indicates the High Density marker data, **PCADD** indicates the pCADD marker data and **WGS** indicates the Whole Genome Sequence marker data.

[a] indicates a significant difference from GBLUP_HD
[b] indicates a significant difference from GBLUP_PCADD
[c] indicates a significant difference from BGC_HD
[d] indicates a significant difference from BGC_PCADD
[e] indicates a significant difference from BGC_WGS

Supplementary Table 4. Accuracy of prediction for S-line animals for the trait Number of Stillborn (STB).

	Ref. Pop.	**GBLUP HD**	**GBLUP PCADD**	**BGC HD**	**BGC PCADD**	**BGC WGS**
Within-line	S1	0.58±0.10	0.55±0.10	0.57±0.10	0.57±0.10	0.57±0.10
	S3	0.63[e]±0.10	0.63±0.10	0.63[e]±0.10	0.64[e]±0.10	0.60±0.10
	S6	0.77[e]±0.10	0.76±0.10	0.78[ae]±0.10	0.78[e]±0.10	0.74±0.10
	S15	0.87[b]±0.10	0.84±0.10	0.90[abe]±0.10	0.88[b]±0.10	0.87±0.10
	S30	1.00±0.10	0.98±0.10	1.03[abe]±0.10	1.01[b]±0.10	1.00±0.10
Across-Line	L1	-0.20±0.10	-0.14±0.10	-0.17[a]±0.10	-0.17[a]±0.10	-0.17[ac]±0.10
	L3	0.08±0.10	0.11±0.10	0.06[e]±0.10	0.07[e]±0.10	0.05±0.10
	L6	0.01±0.10	-0.02±0.10	0.00±0.10	0.00±0.10	0.00±0.10
	L15	0.14[e]±0.10	0.08±0.10	0.13[e]±0.10	0.13±0.10	0.11±0.10
	L30	0.24±0.10	0.24±0.10	0.24±0.10	0.27±0.10	0.23±0.10
Multi-Line	S1_L1	0.53±0.10	0.50±0.10	0.52±0.10	0.52±0.10	0.51±0.10
	S1_L3	0.57±0.10	0.54±0.10	0.56±0.10	0.56±0.10	0.55±0.10
	S1_L6	0.56[bcde]±0.10	0.51±0.10	0.54[e]±0.10	0.53±0.10	0.53±0.10
	S3_L1	0.61[ce]±0.10	0.60±0.10	0.60[e]±0.10	0.61[e]±0.10	0.58±0.10
	S3_L3	0.63[ce]±0.10	0.62±0.10	0.63[e]±0.10	0.63[e]±0.10	0.60±0.10
	S3_L6	0.63[ce]±0.10	0.61±0.10	0.62[e]±0.10	0.61[e]±0.10	0.59±0.10
	S6_L1	0.76[e]±0.10	0.75±0.10	0.76±0.10	0.76±0.10	0.73±0.10
	S6_L3	0.78[e]±0.10	0.76±0.10	0.78[e]±0.10	0.78[e]±0.10	0.75±0.10
	S6_L6	0.76[e]±0.10	0.74±0.10	0.76[e]±0.10	0.75[e]±0.10	0.73±0.10

Ref. Pop. Is the reference population, Letters L and S indicates animals belonging to either L- or S- line and the number indicating number of animals in reference population in number of 1,000. **GBLUP** and **BGC** is the statistical method (GBLUP or BayesGC), **HD** indicates the High Density marker data, **PCADD** indicates the pCADD marker data and **WGS** indicates the Whole Genome Sequence marker data.

[a] indicates a significant difference from GBLUP_HD
[b] indicates a significant difference from GBLUP_PCADD
[c] indicates a significant difference from BGC_HD
[d] indicates a significant difference from BGC_PCADD
[e] indicates a significant difference from BGC_WGS

Supplementary Table 5. Accuracy of prediction for L-line animals for the trait Shoulder Lesions (SHL)

	Ref.	**GBLUP**	**GBLUP**	**BGC**	**BGC**	**BGC**
	Pop.	**HD**	**PCADD**	**HD**	**PCADD**	**WGS**
Within-line	L1	0.09±0.04	0.10±0.04	0.10±0.04	0.10[c]±0.04	0.10±0.04
	L3	0.10±0.04	0.12±0.04	0.10±0.04	0.10[c]±0.04	0.10±0.04
	L6	0.20[cde]±0.04	0.20±0.04	0.18±0.04	0.18±0.04	0.18±0.04
	L15	0.26[b]±0.04	0.24±0.04	0.26±0.04	0.25±0.04	0.25±0.04
	L30	0.26±0.04	0.25±0.04	0.26±0.04	0.25±0.04	0.26±0.04
Across-Line	S1	-0.07±0.04	-0.07±0.04	0.05±0.04	-0.04±0.04	-0.04±0.04
	S3	0.04±0.04	0.04±0.04	-0.06±0.04	0.14[c]±0.04	-0.01±0.04
	S6	-0.02±0.04	0.03±0.04	-0.02±0.04	-0.01±0.04	-0.01±0.04
Multi-Line	L1_S1	0.10±0.04	0.11±0.04	0.11±0.04	0.11±0.04	0.11±0.04
	L1_S3	0.11±0.04	0.12±0.04	0.10±0.04	0.10±0.04	0.10±0.04
	L1_S6	0.11±0.04	0.12±0.04	0.11±0.04	0.11[c]±0.04	0.11[c]±0.04
	L3_S1	0.11±0.04	0.12[cde]±0.04	0.09±0.04	0.09±0.04	0.09±0.04
	L3_S3	0.10±0.04	0.12[cde]±0.04	0.09±0.04	0.09±0.04	0.09±0.04
	L3_S6	0.11±0.04	0.13[acde]±0.04	0.09±0.04	0.09±0.04	0.09±0.04
	L6_S1	0.20[cde]±0.04	0.19±0.04	0.18±0.04	0.18±0.04	0.18±0.04
	L6_S3	0.20[cde]±0.04	0.20±0.04	0.18±0.04	0.18±0.04	0.18±0.04
	L6_S6	0.20[cde]±0.04	0.20[cde]±0.04	0.18±0.04	0.18±0.04	0.18±0.04

Ref. Pop. Is the reference population, Letters L and S indicates animals belonging to either L- or S- line and the number indicating number of animals in reference population in number of 1,000. **GBLUP** and **BGC** is the statistical method (GBLUP or BayesGC), **HD** indicates the High Density marker data, **PCADD** indicates the pCADD marker data and **WGS** indicates the Whole Genome Sequence marker data.

[a] indicates a significant difference from GBLUP_HD
[b] indicates a significant difference from GBLUP_PCADD
[c] indicates a significant difference from BGC_HD
[d] indicates a significant difference from BGC_PCADD
[e] indicates a significant difference from BGC_WGS

Supplementary Table 6. Accuracy of prediction for S-line animals for the trait Shoulder Lesions (SHL)

	Ref.	**GBLUP**	**GBLUP**	**BGC**	**BGC**	**BGC**
	Pop.	**HD**	**PCADD**	**HD**	**PCADD**	**WGS**
Within-line	S1	0.01±0.04	0.30[acde]±0.04	0.01±0.04	0.12[ace]±0.04	0.09[ac]±0.04
	S3	0.51±0.03	0.50±0.03	0.51±0.03	0.51[e]±0.03	0.51±0.03
	S6	0.57±0.03	0.56±0.03	0.56±0.03	0.56[e]±0.03	0.56±0.03
Across-Line	L1	0.08[e]±0.04	0.08±0.04	0.07±0.04	0.07±0.04	0.06±0.04
	L3	0.00±0.04	0.04[acde]±0.04	0.01±0.04	0.02±0.04	-0.01±0.04
	L6	0.14[d]±0.04	0.11±0.04	0.12±0.04	0.11±0.04	0.11±0.04
	L15	0.09[cd]±0.04	0.10[cd]±0.04	0.04±0.04	0.05±0.04	0.06±0.04
	L30	0.08±0.04	0.11±0.04	0.07±0.04	0.08±0.04	0.06±0.04
Multi-Line	S1_L1	0.28±0.04	0.27±0.04	0.28±0.04	0.28±0.04	0.28[c]±0.04
	S1_L3	0.23[c]±0.04	0.24[d]±0.04	0.21±0.04	0.22±0.04	0.23[c]±0.04
	S1_L6	0.31±0.04	0.29±0.04	0.29±0.04	0.30±0.04	0.31±0.04
	S3_L1	0.49±0.03	0.47±0.04	0.49±0.03	0.49±0.03	0.49±0.03
	S3_L3	0.47±0.04	0.46±0.04	0.47±0.04	0.47±0.04	0.47±0.03
	S3_L6	0.50[b]±0.03	0.48±0.03	0.49±0.03	0.49±0.03	0.50[bc]±0.03
	S6_L1	0.56±0.03	0.54±0.03	0.55±0.03	0.55±0.03	0.55±0.03
	S6_L3	0.55±0.03	0.53±0.03	0.55±0.03	0.55±0.03	0.55±0.03
	S6_L6	0.56[b]±0.03	0.54±0.04	0.56[b]±0.03	0.55±0.03	0.56[b]±0.03

Ref. Pop. Is the reference population, Letters L and S indicates animals belonging to either L- or S- line and the number indicating number of animals in reference population in number of 1,000. **GBLUP** and **BGC** is the statistical method (GBLUP or BayesGC), **HD** indicates the High Density marker data, **PCADD** indicates the pCADD marker data and **WGS** indicates the Whole Genome Sequence marker data.

[a] indicates a significant difference from GBLUP_HD
[b] indicates a significant difference from GBLUP_PCADD
[c] indicates a significant difference from BGC_HD
[d] indicates a significant difference from BGC_PCADD
[e] indicates a significant difference from BGC_WGS

Supplementary Table 7. Accuracy of prediction for L-line animals for the trait Body Condition Score (BCS).

	Ref. Pop.	**GBLUP HD**	**GBLUP PCADD**	**BGC HD**	**BGC PCADD**	**BGC WGS**
Within-line	L1	0.21±0.04	0.22±0.04	0.22[a]±0.04	0.22[a]±0.04	0.22[acd]±0.04
	L3	0.28±0.04	0.29[ce]±0.04	0.28±0.04	0.28±0.04	0.28±0.04
	L6	0.35±0.04	0.35±0.04	0.35±0.04	0.35±0.04	0.35±0.04
	L15	0.38±0.04	0.38±0.04	0.38±0.04	0.38[c]±0.04	0.38±0.04
	L30	0.40±0.04	0.40±0.04	0.39±0.04	0.39±0.04	0.39±0.04
Across-Line	S1	0.15[de]±0.04	0.15[de]±0.04	-0.01±0.04	-0.08±0.04	-0.06±0.04
	S3	-0.01±0.04	-0.01±0.04	0.01±0.04	-0.03±0.04	0.09[d]±0.04
	S6	0.01±0.04	-0.02±0.04	0.03±0.04	0.05±0.04	0.04±0.04
Multi-Line	L1_S1	0.19±0.04	0.18±0.04	0.19±0.04	0.19±0.04	0.20[a]±0.04
	L1_S3	0.21±0.04	0.21±0.04	0.21±0.04	0.22±0.04	0.21±0.04
	L1_S6	0.19±0.04	0.20±0.04	0.20±0.04	0.21±0.04	0.19±0.04
	L3_S1	0.27±0.04	0.28±0.04	0.27±0.04	0.27±0.04	0.27±0.04
	L3_S3	0.29±0.04	0.30±0.04	0.28±0.04	0.29±0.04	0.30±0.04
	L3_S6	0.28±0.04	0.29±0.04	0.28±0.04	0.29±0.04	0.29±0.04
	L6_S1	0.34±0.04	0.34±0.04	0.34±0.04	0.34±0.04	0.34±0.04
	L6_S3	0.36±0.04	0.36±0.04	0.36±0.04	0.36±0.04	0.36±0.04
	L6_S6	0.36±0.04	0.36±0.04	0.36±0.04	0.36±0.04	0.37±0.04

Ref. Pop. Is the reference population, Letters L and S indicates animals belonging to either L- or S- line and the number indicating number of animals in reference population in number of 1,000. **GBLUP** and **BGC** is the statistical method (GBLUP or BayesGC), **HD** indicates the High Density marker data, **PCADD** indicates the pCADD marker data and **WGS** indicates the Whole Genome Sequence marker data.

[a] indicates a significant difference from GBLUP_HD
[b] indicates a significant difference from GBLUP_PCADD
[c] indicates a significant difference from BGC_HD
[d] indicates a significant difference from BGC_PCADD
[e] indicates a significant difference from BGC_WGS

Supplementary Table 8. Accuracy of prediction for S-line animals for the trait Body Condition Score (BCS).

	Ref. Pop.	**GBLUP HD**	**GBLUP PCADD**	**BGC HD**	**BGC PCADD**	**BGC WGS**
Within-line	S1	-0.03±0.04	0.57^{acde}±0.04	-0.02±0.04	0.49^{ace}±0.04	0.43^{ac}±0.04
	S3	0.77^{b}±0.04	0.74±0.04	0.77^{bde}±0.04	0.76^{b}±0.04	0.77^{b}±0.04
	S6	0.84^{b}±0.03	0.82±0.03	0.84^{b}±0.03	0.84^{b}±0.03	0.84^{b}±0.03
Across-Line-	L1	-0.05±0.04	0.00^{acde}±0.04	-0.01±0.04	0.00^{ace}±0.04	0.01^{ac}±0.04
	L3	0.16±0.04	0.13±0.04	0.15±0.04	0.15±0.04	0.16±0.04
	L6	0.31±0.04	0.29±0.04	0.33^{abe}±0.04	0.32^{b}±0.04	0.32±0.04
	L15	0.27^{b}±0.04	0.20±0.04	0.29^{ab}±0.04	0.29^{ab}±0.04	0.30^{ab}±0.04
	L30	0.26±0.04	0.23±0.04	0.28^{b}±0.04	0.29^{ab}±0.04	0.29^{ab}±0.04
Multi-Line	S1_L1	0.58^{b}±0.04	0.56±0.04	0.59^{be}±0.04	0.58^{b}±0.04	0.58±0.04
	S1_L3	0.60±0.04	0.58±0.04	0.60±0.04	0.60^{b}±0.04	0.60±0.04
	S1_L6	0.61^{bde}±0.04	0.58±0.04	0.60^{d}±0.04	0.59±0.04	0.60±0.04
	S3_L1	0.76^{b}±0.04	0.73±0.04	0.76^{bd}±0.04	0.75^{b}±0.04	0.76^{b}±0.04
	S3_L3	0.75^{b}±0.04	0.72±0.04	0.76^{bd}±0.04	0.74^{b}±0.04	0.75^{b}±0.04
	S3_L6	0.76^{bd}±0.04	0.72±0.04	0.76^{bd}±0.04	0.74^{b}±0.04	0.76^{b}±0.04
	S6_L1	0.83^{b}±0.03	0.81±0.03	0.83^{b}±0.03	0.83^{b}±0.03	0.84^{b}±0.03
	S6_L3	0.83^{b}±0.03	0.81±0.04	0.83^{b}±0.03	0.83^{b}±0.03	0.83^{b}±0.03
	S6_L6	0.83^{b}±0.03	0.80±0.04	0.82^{b}±0.03	0.82^{b}±0.03	0.83^{bc}±0.03

Ref. Pop. Is the reference population, Letters L and S indicates animals belonging to either L- or S- line and the number indicating number of animals in reference population in number of 1,000. **GBLUP** and **BGC** is the statistical method (GBLUP or BayesGC), **HD** indicates the High Density marker data, **PCADD** indicates the pCADD marker data and **WGS** indicates the Whole Genome Sequence marker data.

[a] indicates a significant difference from GBLUP_HD
[b] indicates a significant difference from GBLUP_PCADD
[c] indicates a significant difference from BGC_HD
[d] indicates a significant difference from BGC_PCADD
[e] indicates a significant difference from BGC_WGS

ISBN: 978-82-575-1906-6
ISSN: 1894-6402

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no

11968 / ANDVORDGRAFISK.NO