WHOLE-GENOME SEQUENCE DATA IN FARM ANIMALS: FROM SNP SELECTION TO GENOMIC PREDICTIONS

by

SUNGBONG JANG

(Under the Direction of Daniela Lourenco)

ABSTRACT

Using whole-genome sequence (WGS) data to identify the causative variants and improve genomic prediction is of current research interest. However, single nucleotide polymorphisms (SNP) chips are still the primary source for genomic predictions. Regular SNP chips only include a small number of SNP. Therefore, more accurate genomic predictions would be expected with WGS data. The objective of first study was to investigate the impact of using preselected variants from WGS for large-scale single-step GBLUP (ssGBLUP) genomic predictions in maternal and terminal pig lines separately. Genomic predictions with regular SNP chip data were compared with preselected SNP sets. Preselection of SNP relied on genome-wide association studies (GWAS) and linkage disequilibrium (LD) pruning. A second study aimed to explore the use of selected WGS variants in a multi-line ssGBLUP genomic evaluation (MLE), which comprised over 200,000 sequenced/imputed animals. A multi-line GWAS was conducted to preselect WGS variants, and unknown parent groups (UPGs) or metafounders (MFs) accounted for genetic differences among lines in a joint evaluation. Those first two studies reported small to no gain in accuracy of genomic prediction with WGS data. To explore the possible reasons for the limited gain in accuracy of genomic prediction with WGS data, a simulation study with different effective

population sizes (*Ne*) was carried out in the third study. We investigated different discovery set sizes in GWAS, relating them to the limited dimensionality of genomic information. The selected variants based on different GWAS sample sizes were then added to simulated SNP panels that mimicked regular chips used commercially. Populations with smaller effective sizes (*Ne* = 20) require more data to capture causative variants, whereas for large populations (*Ne* = 200), using the number of genotyped animals equal to that of the largest eigenvalues explaining 98% of the variance of the genomic relationship matrix suffices. However, only a small proportion of the causative variants can be discovered if those genotyped animals do not have many progeny records. Even when several causative variants are preselected, their impact on ssGBLUP genomic predictions is minimal because medium-density commercial SNP chips already account for most of the information added.

INDEX WORDS: whole-genome sequence, genome-wide association study, variants selection, multi-line genomic evaluation, limited dimensionality of genomic information

WHOLE-GENOME SEQUENCE DATA IN FARM ANIMALS: FROM SNP SELECTION TO GENOMIC PREDICTIONS

by

SUNGBONG JANG

B.S., Chungnam National University, South Korea, 2017

M.S., Chungnam National University, South Korea, 2019

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2022

WHOLE-GENOME SEQUENCE DATA IN FARM ANIMALS: FROM SNP SELECTION TO GENOMIC PREDICTIONS

by

SUNGBONG JANG

Major Professor:	Daniela Lourenco
Committee:	Ignacy Misztal
	Romdhane Rekaya
	Ching-Yi Chen

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2022

DEDICATION

To my parents and people who I love

ACKNOWLEDGEMENTS

I would like to show my sincere appreciation to **Dr. Daniela Lourenco** for having me as her student. Without her commitment, guidance, advice, and encouragement, I could not make it. I have learned a lot from her over the last three years. **Dr. Ignacy Misztal**, thanks for all the discussions about my projects and for always making our group one team. It was my great honor to be a part of this group. **Dr. Romdhane Rekaya**, I really enjoyed talking to him, and thanks for all the great teaching and advice that he gave to me. I will remember and keep his messages in my mind. **Dr. Ching-Yi Chen**, it was my great fortune to work and pursue my project with her. Thanks for being on my committee and always encouraging and answering me with a smile and positivity. **Dr. Shogo Tsuruta**, thanks for always having great talks with me not only about research but also about non-research topics. Time flew so quickly whenever I talked with him. **Dr. Steve Miller**, thanks for the supervision during my internship at American Angus Association and even after that. I am very lucky to meet him and learn from him during my Ph.D. I would like to also thank **Dr. Seunghwan Lee**, my master's supervisor. He is a very special professor to me who gave me all the support, advice, respect, and encouragement since I met him.

Thank you to all my colleagues and friends in our group and all the visitors that I have met. Especially, **Andre**, he is always there to help me and answer my questions since I joined this group. I hope we could work out together and talk about many random topics again. **Natalia**, often, I am astonished by the fact that we have very similar opinions and feelings about many things although we are from completely different countries. Feel free to reach out to me whenever need someone to talk to. **Fiona**, she always brings positive energy and good laughter to us. It really

added something special to my graduate school life. **Evan**, I appreciate his care for our friends and his way of never saying ‘no’ to any help and suggestions. **Jorge**, he is one of the greatest teammates with a nice team spirit that I’ve ever met. He has many things that I should learn. **Ashley**, thanks for accepting my mean joke and teasing every time. I hope she knows that I did it to only a few people. **Taylor**, thanks for all the kind answers to my straightforward questions about American culture, people, and way of living. It helps me to understand and respect them as well as broaden my perspective. **Enrico**, my very honest friend. The travel with him to his hometown still comes to me as unreal. I will definitely invite him to South Korea. **Jiayi**, whenever she would like to talk to me, I will be always there for her as a Korean big brother. Lastly, thank and love my family and friends in South Korea, especially my father and mother, **Yongsam Jang** and **Bokyung Cheon** for all their infinite and warm love and support. Now, it is my turn to give it back.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Using whole-genome sequence (WGS) data in the genomic prediction of farm animals is becoming feasible as the cost of sequencing decreases. The WGS data covers the entire genome with abundant linkage disequilibrium (LD) information between the single nucleotide polymorphisms (SNP) and causative variants or possibly includes causative variants that may not be present in the regular SNP chip data. However, the benefit of using WGS data on genomic prediction could be limited as many SNP are redundant, and strongly correlated to each other with a high extent of LD when their physical distance is small. Therefore, identification of only significant variants and utilization of them could be an efficient strategy to improve the performance of genomic predictions. Pinpointing causative variants through genome-wide association studies (GWAS) has been a common choice when WGS data is used. Although many genotyped animals may have WGS data, non-redundant information is finite. In other words, genomic information has a limited dimensionality, which means that additive genetic information in a population is contained in a limited number of independent chromosome segments (*Me*). Thus, the limited dimensionality of the genomic information might give insights into the number of genotyped animals to use in variant preselection through GWAS for genomic prediction.

The main objective of this dissertation was to 1) investigate the impact of using preselected variants selected from WGS data on genomic prediction in pigs for both single-line and multi-line scenarios and 2) scrutinize the effects of different data set sizes in GWAS and genomic prediction, where the sizes depended on the limited dimensionality in the genomic information, as well as

explore a change in accuracy by adding preselected variants from WGS to regular SNP chip data through a simulation study. In Chapter 2, a literature review is addressed. Afterward, objective 1) is discussed in Chapters 3 and 4 for single-line and multi-line genomic evaluations, respectively. Finally, objective 2) is extensively discussed in Chapter 5 of this dissertation.

CHAPTER 2

LITERATURE REVIEW

Single nucleotide polymorphisms (SNP) have been extensively used as a source of genomic information to estimate the genetic merit of farm animals. The large adoption of SNP is due to (1) abundance, (2) informativity, (3) efficiency, and (4) affordability, among others. SNP are evenly distributed across the entire genome and may be in linkage disequilibrium (LD) with quantitative trait loci (QTL). In that regard, Meuwissen et al. (2001) proposed the Bayesian models to calculate SNP effects and genomic breeding values (GEBV) using phenotypic and genomic information. In line with that, genomic BLUP (GBLUP) was proposed afterward, which calculates GEBV assuming the covariance structure among animals is given by the additive genetic variance and the genomic relationship matrix (**G**) (VanRaden, 2008). However, GBLUP only considers information on genotyped animals, which does not reflect the structure of animal populations because most of the animals are not genotyped. To handle all available information on both genotyped and non-genotyped animals in a sole evaluation, a method called single-step GBLUP (ssGBLUP) was developed by Aguilar et al. (2010); Christensen and Lund (2010). This method uses all available phenotypes, pedigree, and genomic information in the mixed model equations (MME). In the ssGBLUP, the covariance structure among animals is given by the additive genetic variance and the realized relationship matrix (**H**), which its inverse is given by (Aguilar et al., 2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix},$$

Where $\mathbf{A}^{-1}$ is the inverse of the pedigree relationship matrix among all animals ($\mathbf{A}$), $\mathbf{A}_{22}^{-1}$ is the inverse of the pedigree relationship matrix only for the genotyped animals ($\mathbf{A}_{22}$), and $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix ($\mathbf{G}$), which is often constructed as (VanRaden, 2008):

$$\mathbf{G} = \frac{\mathbf{MDM'}}{2\sum p_j(1-p_j)},$$

where $\mathbf{M}$ is the matrix of centered genotypes for current allele frequencies, p_j is the minor allele frequency of SNP_j, $\mathbf{D}$ is the diagonal matrix of SNP weights. All the SNP were presumed to have homogeneous weights in ssGBLUP, meaning that $\mathbf{D}$ is an identity matrix ($\mathbf{I}$). To avoid singularity issues, $\mathbf{G}$ is often blended with a small proportion of $\mathbf{A}_{22}$ (e.g., 5%).

Implementation of ssGBLUP has made a considerable impact in animal breeding and has been applied for routine genomic evaluation in many farm animals, such as cattle, pigs, chickens, and sheep (Chen et al., 2011; Forni et al., 2011; Lourenco et al., 2015; Brown et al., 2018). For routine genomic evaluation, 50k SNP chips have been generally used under the assumption that SNP in the chip are in LD with existing QTL. Therefore, if SNP are in strong LD with QTL, they are likely inherited together, even though the SNP chips may not contain QTL in practice. In that sense, using whole-genome sequence (WGS) data has the potential to improve the accuracy of genomic predictions because they cover the entire genome with a large number of SNP. For example, WGS data in cattle consists of about 20 to 30 million SNP. By spanning the entire genome, WGS possibly includes causative variants that primarily affect the traits of interest but are not present in the SNP chip data. However, there was limited or no improvement in genomic predictions with WGS data in cattle and pigs (Van Binsbergen et al., 2015; Zhang et al., 2018; Song et al., 2019). A procedure of choice to overcome this limitation when using WGS data is the preselection of significant variants through genome-wide association studies (GWAS). The

selected variants may be combined with regular SNP chips for genomic predictions or not, which has been investigated in several studies (Brøndum et al., 2015; VanRaden et al., 2017; Fragomeni et al., 2019).

The primary purpose of GWAS is to identify genetic variants affecting the trait of interest, e.g., mostly disease-related traits in humans and economic traits in livestock. The first GWAS was conducted in humans to identify a susceptibility gene for myocardial infarction trait (Ozaki et al., 2002). Afterward, GWAS became a common procedure also in livestock and plants. The initial GWAS considered all the genetic variants identically and independently distributed. This was because the primary method aimed to explore the populations with unrelated individuals (Risch and Merikangas, 1996). However, that consideration does not hold in reality because of the relatedness between individuals in the population. In humans, individuals in the population are distantly related as no selection is conducted, and the effective population size (*Ne*) is larger compared to livestock. In livestock, individuals in a population are more closely related to each other, especially for populations under selection that have small *Ne*. This relatedness can hinder the identification of causative variants and induce false-positive results (Sul et al., 2018). To resolve this issue, population structure should be accounted for by the GWAS model. Several studies (Kang et al., 2008; Kang et al., 2010) attempted to correct for population structure through the mixed-effects model (Henderson, 1975) by fitting either **A**, **G**, or principal components in the model.

Besides accounting for population structure, several factors would affect the performance of GWAS. The number of causative variants and SNP, sample size, heritability of the trait, statistical methods, *Ne*, and the number of independent chromosome segments (*Me*) are the major ones (Berisa and Pickrell, 2016; Visscher et al., 2017; Jang et al., 2022). Among those, the

investigation of the sample size associated with *Ne* and *Me* is still questionable. Theoretically, increasing the sample size for GWAS improves the resolution and statistical power to identify the significant variants, avoiding spurious results. The sample size is critical when investigating a large number of SNP, as is the case for WGS data. As of June 2022, the number of genotyped Holsteins in the US is 5.4 million (https://queries.uscdcb.com/Genotype/cur_freq.html), and Angus is 1.2 million (K. Retallick, American Angus Association, Saint Joseph, MO, personal communication). However, computations with such a number are challenging, and not all genotyped animals are informative. Therefore, it is crucial to understand how many genotyped animals are effectively needed to be used as a variant discovery set for GWAS. Understanding the efficient sample size for GWAS could help alleviate economic and computational costs in practical applications for populations with small and large effective sizes.

Misztal (2016) hypothesized that the additive genetic information in a population is contained in a limited *Me*. Stam (1980) showed that given a species with a genome length equal to *L* Morgans, the *Me* segregating in a population could be expressed as a function of *Ne* and *L*, which is 4*NeL*. Pocrnic (2016) showed that *Me*, and therefore, *Ne* and *L*, is a function of the number of largest eigenvalues (EIG) explaining a certain proportion of variance in **G**, such that EIG90 ≈ *NeL*, EIG95 ≈ 2*NeL*, and EIG98 ≈ 4*NeL*. Based on that, the optimal number of animals that carry all the independent chromosome segments, representing the non-redundant genomic information, is approximated by the number EIG explaining 98% of the variance in **G**. This number varies from 4k to 6k in pigs and chickens and from 10k to 15k in cattle (Pocrnic et al., 2016).

The accuracy of genomic prediction is affected by many factors. Among them, the size of reference population and the relationship between the reference and validation individuals are important ones. Thus, the gain in accuracy of genomic prediction in populations with a small

reference size could be limited. In such a case combining multiple populations (breeds or lines) could help boost the accuracy of genomic predictions for the populations with fewer individuals (Calus et al., 2014; Rolf et al., 2015; Song et al., 2017; Cesarani et al., 2022). Combining multiple populations can be challenging, especially with genomic information. Several studies explored adjusting **G** for breed-specific allele frequencies in multi-breed evaluations in cattle and pigs, but no benefits were observed in prediction accuracy (Makgahlela et al., 2013; Lourenco et al., 2016). Cesarani et al. (2022) showed that accurate multibreed cattle predictions are obtained when all fixed effects in the model are breed-specific and unknown parent groups (UPGs) are fitted to account for genetic differences at the breed level, where UPGs can be used to model non-zero breeding values for missing parents (Westell et al. 1988). Macedo et al. (2020) also observed that better multibreed genomic predictions are possible if differences in base populations are correctly modeled. This difference can be modeled by UPGs, which ssGBLUP can handle in two ways based on the extension of the QP-transformed equation (Quaas, 1988) to all elements of $\mathbf{H}^{-1}$ (Misztal et al. 2013):

$$\mathbf{H}^{*} = \mathbf{A}^{*} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & -(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q} \\ 0 & -\mathbf{Q}'(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}) & \mathbf{Q}'(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q} \end{bmatrix},$$

or only the pedigree relationship matrices (Tsuruta et al., 2019):

$$\mathbf{H}^{*} = \mathbf{A}^{*} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & \mathbf{A}_{22}^{-1}\mathbf{Q} \\ 0 & -\mathbf{Q}'\mathbf{A}_{22}^{-1} & \mathbf{Q}'\mathbf{A}_{22}^{-1}\mathbf{Q} \end{bmatrix},$$

where $\mathbf{H}^{*}$is the inverse of the realized relationship matrix with UPGs added to the **A**, $\mathbf{A_{22}}$, and **G** or without **G**. **Q** is an incidence matrix relating animals in vector **u** to UPGs in vector **g,** $\mathbf{A}^{*}$ is the inverse of **A** with UPGs constructed with the QP transformation. However, UPGs assume that the

ancestors in base populations are neither inbred nor related. Therefore, Legarra et al. (2015) proposed metafounders (MFs), which assume the individuals in the base populations are related and inbred. In this method, **A** was modified to be compatible with **G** centered with allele frequencies of 0.5 ($\mathbf{G}_{0.5}$), so $\mathbf{H}^{-1}$ with metafounders is computed as:

$$\mathbf{H}^{\Gamma-1} = \mathbf{A}^{\Gamma-1} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}_{05}^{-1} - \mathbf{A}_{22}^{\Gamma-1} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where $\mathbf{A}^{\Gamma-1}$ and $\mathbf{A}_{22}^{\Gamma-1}$ are the altered $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}^{-1}$ with the parameter $\mathbf{\Gamma}$. The $\mathbf{\Gamma}$ matrix is computed using SNP markers under a generalized least square approach (Garcia-Baccino et al., 2017). Although the MFs theory is well defined, its application in multibreed or multi-line evaluations is still under investigation.

REFERENCES

Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta, and T. Lawlor. 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. doi:10.3168/jds.2009-2730

Berisa, T., and J. K. Pickrell. 2016. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics 32(2):283. doi:10.1093/bioinformatics/btv546

Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 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. doi: 10.3168/jds.2014-9005

Brown, D., A. Swan, V. Boerner, L. Li, P. Gurman, A. McMillan, J. Van der Werf, H. Chandler, B. Tier, and R. Banks. 2018. Single-step genetic evaluations in the Australian sheep industry. In: Proceedings of the world congress on genetics applied to livestock production. p 460

Calus, M. P., H. Huang, A. Vereijken, J. Visscher, J. Ten Napel, and J. J. Windig. 2014. Genomic prediction based on data from three layer lines: a comparison between linear methods. Genetics Selection Evolution 46(1):1-13. doi:10.1186/s12711-014-0057-5

Cesarani, A., D. Lourenco, S. Tsuruta, A. Legarra, E. Nicolazzi, P. VanRaden, and I. Misztal. 2022. Multibreed genomic evaluation for production traits of dairy cattle in the United States using single-step genomic best linear unbiased predictor. Journal of Dairy Science. doi:10.3168/jds.2021-21505

Chen, C., I. Misztal, I. Aguilar, S. Tsuruta, T. Meuwissen, S. Aggrey, T. Wing, and W. Muir. 2011. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: an example using broiler chickens. Journal of animal science 89(1):23-28. doi:10.2527/jas.2010-3071

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

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43(1):1. doi:10.1186/1297-9686-43-1

Fragomeni, B., D. Lourenco, A. Legarra, P. VanRaden, and I. Misztal. 2019. Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in

US Holsteins in the presence of selected sequence variants. Journal of dairy science 102(11):10012-10019. doi:10.3168/jds.2019-16262

Garcia-Baccino, C. A., A. Legarra, O. F. Christensen, I. Misztal, I. Pocrnic, Z. G. Vitezica, and R. J. Cantet. 2017. Metafounders are related to F st fixation indices and reduce bias in single-step genomic evaluations. Genetics Selection Evolution 49(1):1-14. doi:10.1186/s12711-017-0309-2

Henderson, C. R. 1975. Comparison of alternative sire evaluation methods. Journal of Animal Science 41(3):760-770. doi:10.2527/jas1975.413760x

Jang, S., S. Tsuruta, N. G. Leite, I. Misztal, and D. Lourenco. 2022. Dimensionality of genomic information and its impact on GWA and variant selection: a simulation study. bioRxiv. doi:10.1101/2022.04.13.488175

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. Nature genetics 42(4):348-354. doi:10.1038/ng.548

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178(3):1709-1723. doi:10.1534/genetics.107.080101

Legarra, A., J. Bertrand, T. Strabel, R. Sapp, J. Sanchez, and I. Misztal. 2007. Multi-breed genetic evaluation in a Gelbvieh population. Journal of Animal Breeding and Genetics 124(5):286-295. doi:10.1111/j.1439-0388.2007.00671.x

Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal. 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. Genetics 200(2):455-468. doi:10.1534/genetics.115.177014

Lourenco, D., S. Tsuruta, B. Fragomeni, C. Chen, W. Herring, and I. Misztal. 2016. Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. Journal of animal science 94(3):909-919. doi:10.2527/jas.2015-9748

Lourenco, D., S. Tsuruta, B. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. Bertrand, T. Amen, L. Wang, and D. Moser. 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. Journal of animal science 93(6):2653-2662. doi:10.2527/jas.2014-8836

Macedo, F. L., O. F. Christensen, J.-M. Astruc, I. Aguilar, Y. Masuda, and A. Legarra. 2020. Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. Genetics Selection Evolution 52(1):1-10. doi:10.1186/s12711-020-00567-1

Makgahlela, M., I. Strandén, U. Nielsen, M. Sillanpää, and E. Mäntysaari. 2013. The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. Journal of dairy science 96(8):5364-5375. doi:10.3168/jds.2012-6523

Mancin, E., D. L. Lourenco, M. Bermann, R. Mantovani, and I. Misztal. 2021. Accounting for population structure and phenotypes from relatives in association mapping. Frontiers in Genetics 12:658. doi:10.3389/fgene.2021.642065

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819-1829. doi:10.1093/genetics/157.4.1819

Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. Genetics 202(2):401-409. doi:10.1534/genetics.115.182089

Misztal, I., Z.-G. Vitezica, A. Legarra, I. Aguilar, and A. Swan. 2013. Unknown-parent groups in single-step genomic evaluation. Journal of Animal Breeding and Genetics 130(4):252-258. doi:10.1111/jbg.12025

Ozaki, K., Y. Ohnishi, A. Iida, A. Sekine, R. Yamada, T. Tsunoda, H. Sato, H. Sato, M. Hori, and Y. Nakamura. 2002. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nature genetics 32(4):650-654. doi:10.1038/ng1047

Pocrnic, I., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016. The dimensionality of genomic information and its effect on genomic prediction. Genetics 203(1):573-581. doi:10.1534/genetics.116.187013

Quaas, R. 1988. Additive genetic model with groups and relationships. Journal of Dairy Science 71(5):1338-1345. doi:10.1016/S0022-0302(88)79986-5

Risch, N., and K. Merikangas. 1996. The future of genetic studies of complex human diseases. Science 273(5281):1516-1517. doi:10.1126/science.273.5281.1516

Rolf, M. M., D. J. Garrick, T. Fountain, H. R. Ramey, R. L. Weaber, J. E. Decker, E. J. Pollak, R. D. Schnabel, and J. F. Taylor. 2015. Comparison of Bayesian models to estimate direct genomic values in multi-breed commercial beef cattle. Genetics Selection Evolution 47(1):1-14. doi:10.1186/s12711-015-0106-8

Song, H., S. Ye, Y. Jiang, Z. Zhang, Q. Zhang, and X. Ding. 2019. Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. Genetics Selection Evolution 51(1):58. doi:10.1186/s12711-019-0500-8

Song, H., J. Zhang, Y. Jiang, H. Gao, S. Tang, S. Mi, F. Yu, Q. Meng, W. Xiao, and Q. Zhang. 2017. Genomic prediction for growth and reproduction traits in pig using an admixed reference population. Journal of Animal Science 95(8):3415-3424. doi:10.2527/jas.2017.1656

Stam, P. 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genetics Research 35(2):131-155. doi:10.1017/S0016672300014002

Sul, J. H., L. S. Martin, and E. Eskin. 2018. Population structure in genetic studies: Confounding factors and mixed models. PLoS genetics 14(12):e1007309. doi:10.1371/journal.pgen.1007309

Tsuruta, S., D. Lourenco, Y. Masuda, I. Misztal, and T. Lawlor. 2019. Controlling bias in genomic breeding values for young genotyped bulls. Journal of dairy science 102(11):9956-9970. doi:10.3168/jds.2019-16789

Van Binsbergen, R., M. P. Calus, M. C. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 47(1):71. doi:10.1186/s12711-015-0149-x

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. Journal of dairy science 91(11):4414-4423. doi:10.3168/jds.2007-0980

VanRaden, P. M., M. E. Tooker, J. R. O'connell, J. B. Cole, and D. M. Bickhart. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. Genetics Selection Evolution 49(1):32. doi:10.1186/s12711-017-0307-4

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 2017. 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics 101(1):5-22. doi:10.1016/j.ajhg.2017.06.005

Westell, R. A., Quaas, R. L, and Van Vleck, L. D. 1988. Genetic groups in an animal model. Journal of dairy science 71(5), 1310-1318. doi:10.3168/jds.S0022-0302(88)79688-5

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price. 2014. Advantages and pitfalls in the application of mixed-model association methods. Nature genetics 46(2):100-106. doi:10.1038/ng.2876

Yin, T., and S. König. 2019. Genome-wide associations and detection of potential candidate genes for direct genetic and maternal genetic effects influencing dairy cattle body weight at different ages. Genetics Selection Evolution 51(1):1-14. doi:10.1186/s12711-018-0444-4

Zhang, C., R. A. Kemp, P. Stothard, Z. Wang, N. Boddicker, K. Krivushin, J. Dekkers, and G. Plastow. 2018. Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. Genetics Selection Evolution 50(1):1-13. doi:10.1186/s12711-018-0387-9

CHAPTER 3

USING WHOLE-GENOME SEQUENCE DATA FOR SINGLE-STEP GENOMIC PREDICTIONS IN MATERNAL AND TERMINAL PIG LINES[1]

[1]Sungbong Jang, Roger Ros-Freixedes, John M. Hickey, Ching-Yi Chen, William O. Herring, Ignacy Misztal, and Daniela Lourenco. To be submitted to *Genetics Selection Evolution*.

ABSTRACT

Whole-genome sequence (WGS) data harbor causative variants that may not be present in the regular SNP chip data. The objective of this study was to investigate the impact of using preselected variants from WGS for single-step genomic predictions in maternal and terminal pig lines with up to 1.8k sequenced and 100k imputed sequenced animals. Two maternal and four terminal lines were investigated for eight and seven traits, respectively. The number of sequenced animals ranged from 1,365 to 1,491 in maternal lines and 381 to 1,865 in terminal lines. Imputation occurred within each line, and the number of animals imputed to sequence ranged from 66k to 76k in maternal lines and 29k to 104k in terminal lines. Three preselected SNP sets were generated based on genome-wide association studies (GWAS) or linkage disequilibrium (LD) pruning. Top40k had the top SNP in each 40k genomic window; ChipPlusSign included significant variants integrated into the regular porcine SNP chip, whereas LDTags were tag variants retained after pruning SNP with LD r^2>0.1 in 10-Mb windows. Those SNP sets were compared to the regular porcine SNP chip using single-step GBLUP with equal or different SNP variances. In maternal lines, ChipPlusSign, Top40k, and LDTags showed, on average, 0.62%, 4.9%, and 1.56% increased accuracy compared to the regular porcine SNP chip. The greatest changes were for fertility traits, with an average gain of 31.1%. However, for sire lines, most of the SNP sets resulted in a loss of accuracy ranging from 1% to 12.4%. Only ChipPlusSign provided a positive, albeit small, gain (0.85%). Assigning different variances for SNP slightly improved accuracies when using variances obtained from BayesR; however, the increase was not consistent across lines and traits. The benefit of using sequence data depends on the line, genetic architecture of the traits, size of the genotyped population, and how the WGS variants are preselected. When WGS is available on hundreds of

thousands of animals, the advantage of sequence data is present but could be limited in maternal and terminal pig lines.

INTRODUCTION

Using SNP chip data for genomic prediction relies on the linkage disequilibrium (LD) between SNP and causative variants (De Roos et al., 2008). Because of the initial high cost of SNP genotyping, most of the SNP chips utilized in farm animals are still limited to less than 100k markers, which could restrict the information available for genomic predictions. Whole-genome sequence (WGS) data harbor millions of variants, possibly including causative variants that primarily affect the traits of interest but are not present in regular SNP chips. As sequencing is becoming cheaper, WGS data is becoming available for some agricultural species. Whether this data can help increase the accuracy of genomic predictions beyond that already achieved by SNP chips is still questionable because no or marginal gains were reported by several studies (Brøndum et al., 2015; van den Berg et al., 2016; VanRaden et al., 2017; Fragomeni et al., 2019). Specifically, in pigs, Zhang et al. (2018) showed that the 80k SNP chip outperformed the 650k SNP chip and WGS data for genomic predictions of average daily feed intake and backfat traits. In contrast, Song et al. (2019) reported a marginal gain in prediction accuracy when WGS data was used. The absence of benefits reported in those studies could be due to the small number of sequenced animals (maximum of 289 animals), poor imputation accuracy, statistical methods, and sequence SNP redundant with the ones in the chip.

Imputation is an inevitable step when working with WGS data because sequencing a large number of individuals is still unfeasible. So far, the most efficient approach is to sequence the

important animals in a population and impute the sequence to other genotyped animals (Ros-Freixedes et al., 2020). Not all variants might be causative or in high LD with the causative ones; thus, using the entire WGS data would not benefit genomic predictions (Van Binsbergen et al., 2015). Hence, the preselection of variants helps narrow down the WGS data to only significant ones. Several approaches have been investigated to select significant or causative variants for genomic prediction, such as genome-wide association studies (GWAS) (VanRaden et al., 2017), SNP functional annotation (Lopez et al., 2021), and gene expression (de Las Heras-Saldana et al., 2020). Among these approaches, GWAS has been used to preselect WGS variants in pig populations (Zhang et al., 2018; Song et al., 2019); however, the number of individuals initially sequenced was small, and the imputation applied to less than 7k animals, indicating that only a small number of animals were used for variant selection and genomic prediction.

Fragomeni et al. (2017) used simulated sequence data to show that once all causative variants are known, together with their position and percentage of additive genetic variance explained, prediction accuracy is maximized. Conversely, if only neighboring SNP are identified, the accuracy is inversely proportional to the distance between the causative variants and the neighbor SNP. If only a small proportion of the causative variants were known, the increase in accuracy was also proportional. When a few causative variants were known from a real beef cattle data, Gualdrón-Duarte et al. (2020) showed an increase in prediction accuracy for carcass traits of up to seven points when using single-step genomic BLUP (ssGBLUP) with BayesR SNP weights, but no improvements with non-linear weights (VanRaden, 2008; Fragomeni et al., 2019).

Using simulated sequence data, Jang et al. (2022) looked at the dimensionality of the genomic information (Pocrnic et al., 2016a) to assess the number of genotyped animals needed to maximize the percentage of discoveries in GWAS. The authors showed that populations with

smaller effective size (*Ne* = 20) require more data to capture causative variants, whereas for large populations (*Ne* = 200), using the number of genotyped animals equal to that of the largest eigenvalues explaining 98% of the variance of the genomic relationship matrix suffices. Still, only a small proportion of the causative variants can be discovered if those genotyped animals do not have many progeny records.

In pigs, the *Ne* varies from 30 to 50, and the dimensionality of the genomic information or the number of independent chromosome segments segregating in the population ranges from 4k to 6k (Pocrnic et al., 2016b). Based on Jang et al. (2022), using a sample size for GWAS of 7k in a population with *Ne* of 20 allowed identifying causative variants explaining 20% of the additive genetic variance. Moreover, larger sample sizes resulted in larger prediction accuracies with selected variants. Recently, Ros-Freixedes et al. (2020) proposed an approach to accurately impute sequence to hundreds of thousands of pigs, which resulted in WGS data for over 300k animals across maternal and terminal lines. The present study investigates the impact of using preselected variants from WGS data for genomic prediction in maternal and terminal pig lines with up to 1.8k sequenced and 100k imputed sequenced animals. We explored different ways to preselect variants and the changes in accuracy when using single-step GBLUP (ssGBLUP) and ssGBLUP weighted with BayesR SNP variances (WssGBLUP).

MATERIALS AND METHODS

Animal Care and Use Committee approval was not needed as data were obtained from preexisting databases.

Data

Datasets provided by the Pig Improvement Company (PIC; Hendersonville, TN) comprised two maternal lines (ML1, ML2) and four terminal lines (TL1, TL2, TL3, and TL4) with diverse genetic backgrounds. We investigated average daily feed intake (ADFI), average daily gain (ADG), backfat thickness (BF), loin depth (LDP), total born (BTP), number of stillborn (NSB), return to oestrus seven days after weaning (RET), and litter weaning weight (WWT) for maternal lines. For terminal lines, we investigated ADFI in purebreds, ADG, BF, and LDP, in purebreds and crossbreds (ADGX, BFPX, and LDPX). In maternal lines, two-trait models were considered for ADG and ADFI (ADFI model), ADG and BF (GROWTH model), ADG and LDP (LOIN model), BTP and NSB (REPROD model), whereas single-trait models were used for RET (RET model) and WWT (WWT model). In terminal lines, the described ADFI model was applied, but four-trait models were used for GROWTH (ADG, BF, ADGX, and BFX) and LOIN (ADG, LDP, ADGX, and LDPX) model. The total number of animals in the pedigree and records for each trait are in Table 3.1. Pigs were genotyped with the GGP-Porcine HD BeadChip (GeneSeek, Lincoln, NE). We filtered out the monomorphic SNP as well as SNP with a call rate lower than 0.90, minor allele frequency lower than 0.01, and the difference between observed and expected genotype frequencies greater than 0.15. Individuals with more than 10% missing genotypes were also removed. Table 3.2 depicts the number of genotyped animals and SNP per line after quality control.

Whole-genome sequencing and imputation

The whole-genome sequence data used in this study were generated by Ros-Freixedes et al. (2020); Ros-Freixedes et al. (2022). In summary, a low-coverage sequencing strategy was followed by joint calling, phasing, and imputation of the whole-genome genotypes using the 'hybrid peeling' method implemented in AlphaPeel (Whalen et al., 2018). The number of

sequenced individuals for each line is provided in Table 3.2. The 'hybrid peeling' method used all the GGP-Porcine HD and WGS data available across the pedigrees. Imputation was carried out separately for each line. Individuals with low predicted imputation accuracy were excluded, as described by Ros-Freixedes et al. (2020). A total of 76,230 (ML1), 66,608 (ML2), 60,474 (TL1), 41,573 (TL2), 29,330 (TL3), and 104,661 (TL4) sequenced/imputed individuals remained in each line after quality control. These individuals were predicted to have an average dosage correlation of 0.97 (median: 0.98 based on the imputation accuracy of 284 pigs that had both WGS (high coverage) and marker array data. All SNP with a minor allele frequency lower than 0.023 were removed since their estimated dosage correlations were lower than 0.90 (Ros-Freixedes et al., 2020).

Training and test sets

Before the GWAS, all animals with WGS data were separated into training and test, which were defined as in Ros-Freixedes et al. (2022). Test sets were generated by extracting entire litters from the last generation of the pedigree. The training sets were created by establishing a threshold on the pedigree relationship coefficients between training and test sets. We removed individuals with a relationship coefficient equal to or greater than 0.5 to test animals to resemble the selection candidate evaluation done by pig breeding companies. Ros-Freixedes et al. (2022) investigated different training sets for variant discovery and genomic prediction and reported no improvement or drop in accuracy. Therefore, the same training dataset was adopted for both analyses.

Pre-selected SNP panels

Three different pre-selected SNP panels were created from WGS for the genomic prediction as described in Ros-Freixedes et al. (2022): (1) LDTags (2), Top40k, and (3) ChipPlusSign. The LDTags were tag variants retained after pruning SNP with LD with an $r^2>0.1$

in any 10-Mb window so that, on average, 30k variants remained (range: 5k to 80k). Top40k were the variants with the lowest p-value (not necessarily below the significance threshold) in each consecutive non-overlapping 55-kb window along the genome, based on GWAS analyses. ChipPlusSign combined the GGP-Porcine HD SNP and significant variants ($p \leq 10^{-6}$) from the GWAS in a way that when a 55-kb window contained more than one significant variant, only that with the lowest p-value was selected. Genomic predictions with the three sets were compared against the GGP-Porcine HD chip (Chip). For scenarios that used multi-trait models, the preselected variants for each trait were combined for the traits included in each model. For example, those pre-selected variants for each ADFI and ADG were combined for the ADFI model and used for genomic prediction. As sequence information was available only on purebred animals, no variants were selected for ADGX, BFPX, and LDPX. After constructing all the pre-selected SNP panels, quality control was done to remove SNP with the difference between observed and expected genotype frequencies greater than 0.15 and to exclude individuals with parent-progeny Mendelian conflicts. Because the number of animals available for each SNP panel (Chip, LDtag, Top40k, and ChipPlusSign) was different after quality control, we only used the common animals that passed quality control for all the SNP panels for a fair comparison of genomic prediction (Table 3.3).

Genomic prediction

Single-trait, two-trait, or four-trait models were used for genomic predictions, depending on the traits. Herein, only a four-trait GROWTH model (ADG, BF, ADGX, and BFX) of terminal lines is described:

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

where **y** is the vector of phenotypes; **X** is an incidence matrix for fixed effects (contemporary group as a cross-classified effect for all traits, offtest weight and carcass weight as a covariate only for BF and BFX, respectively) contained in **b**; **W** is an incidence matrix for the random, diagonal litter effect contained in **c** (**c**~ MVN(0, $\mathbf{I}\otimes\mathbf{L}_0$)); **Z** is an incidence matrix for the random additive genetic effect contained in **u** (**u**~ MVN(0, $\mathbf{H}\otimes\mathbf{\Sigma}_0$)); and **e** (**e**~MVN(0, $\mathbf{I}\otimes\mathbf{R}_0$)) is a vector of residual effects. Matrices $\mathbf{L}_0$, $\mathbf{\Sigma}_0$, and $\mathbf{R}_0$ are as follows:

$$\mathbf{L}_0 = \begin{bmatrix} \sigma^2_{l_{ADG}} & \sigma_{l_{ADG},l_{BF}} & 0 & 0 \\ \sigma_{l_{BF},l_{ADG}} & \sigma^2_{l_{BF}} & 0 & 0 \\ 0 & 0 & \sigma^2_{l_{ADGX}} & \sigma_{l_{ADGX},l_{BFX}} \\ 0 & 0 & \sigma_{l_{BFX},l_{ADGX}} & \sigma^2_{l_{BFX}} \end{bmatrix},$$

$$\mathbf{\Sigma}_0 = \begin{bmatrix} \sigma^2_{a_{ADG}} & \sigma_{a_{ADG},a_{BF}} & \sigma_{a_{ADG},a_{ADGX}} & \sigma_{a_{ADG},a_{BFX}} \\ \sigma_{a_{BF},a_{ADG}} & \sigma^2_{a_{BF}} & \sigma_{a_{BF},a_{ADGX}} & \sigma_{a_{BF},a_{BFX}} \\ \sigma_{a_{ADGX},a_{ADG}} & \sigma_{a_{ADGX},a_{BF}} & \sigma^2_{a_{ADGX}} & \sigma_{a_{ADGX},a_{BFX}} \\ \sigma_{a_{BFX},a_{ADG}} & \sigma_{a_{BFX},a_{BF}} & \sigma_{a_{BFX},a_{ADGX}} & \sigma^2_{a_{BFX}} \end{bmatrix},$$

$$\mathbf{R}_0 = \begin{bmatrix} \sigma^2_{e_{ADG}} & \sigma_{e_{ADG},e_{BF}} & 0 & 0 \\ \sigma_{e_{BF},e_{ADG}} & \sigma^2_{e_{BF}} & 0 & 0 \\ 0 & 0 & \sigma^2_{e_{ADGX}} & \sigma_{e_{ADGX},e_{BFX}} \\ 0 & 0 & \sigma_{e_{BFX},e_{ADGX}} & \sigma^2_{e_{BFX}} \end{bmatrix},$$

where σ^2_l is the litter variance, σ^2_a is the additive genetic variance, and σ^2_e is the residual variance. **I** is an identity matrix and **H** is the realized relationship matrix that combines pedigree and genomic relationships in ssGBLUP. The genomic prediction was performed with both ssGBLUP and WssGBLUP using BLUPF90 family of programs (Misztal et al., 2014b), which used the inverse of **H** ($\mathbf{H}^{-1}$) as follows (Aguilar et al., 2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}^{-1}_{22} \end{bmatrix},$$

where $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix and $\mathbf{A}_{22}^{-1}$ is the inverse of the pedigree relationship matrix for the genotyped individuals. The **G** was created using the first method of VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{MDM'}}{2\sum p_j(1-p_j)},$$

where **M** is a matrix of genotypes centered or current allele frequencies, p_j is the minor allele frequency of SNP j, and **D** is the diagonal matrix of SNP weights. All the SNP were presumed to have homogeneous weights in ssGBLUP, meaning that **D** is an identity matrix (**I**). To ensure compatibility between **G** and $\mathbf{A}_{22}$ and circumvent singularity issues, **G** was tuned and then blended with 5% of $\mathbf{A}_{22}$.

The algorithm for proven and young (APY) was applied to obtain $\mathbf{G^{-1}}$ while avoiding the direct inversion of **G** (2014a) for the lines with more than 50k genotyped animals, i.e., ML1, ML2, TL1, and TL4. To ensure reliable estimation of GEBV, the number of core animals corresponded to the number of largest eigenvalues explaining 98% of the total variation in **G** assessed with regular chip data (Pocrnic et al., 2016a). Therefore, the number of core animals in each line was: 4,200, 5,400, 3,400, and 5,500 for ML1, ML2, TL1, and TL4, respectively.

For the WssGBLUP, we calculated SNP variances from BayesR (Erbe et al., 2012) and assigned those variances as weights for SNP in an iterative way. Each iteration stored individual SNP variances, and posterior SNP variance was calculated as the average variance across all the iterations. Afterward, the weights were re-scaled to make the trace of **D** equal to the number of SNP. More details about BayesR weighting are described in Gualdrón-Duarte et al. (2020). This approach was only applied to the four largest lines, ML1, ML2, TL1, and TL4 for growth-related traits (ADFI, ADG, BF, and LD) with the Top40k and ChipPlusSign data.

Validation

The accuracy of genomic prediction was calculated by correlating genomic EBV (GEBV) with deregressed EBV (dEBV) (VanRaden and Wiggans, 1991) for the animals in the test sets. Inflation or deflation levels were assessed as the b_1 of the regression of dEBV on GEBV. The b_1 values lower than 1 indicated inflation of GEBV and greater than 1 indicated deflation. All animals with no dEBV were removed from the test sets.

RESULTS

Genomic prediction accuracy of maternal lines using ssGBLUP

Fig. 3.1 shows the changes in prediction accuracy (%) of ChipPlusSign, Top40k, and LDtags compared to Chip for the two maternal lines. To compare the results between lines by the size of the genotyped population rather than the SNP panels, the x-axis was sorted by the number of genotyped animals in each line (small to large; M2 and M1 by order), but the gain range of y-axis in each figure did not equalize to other figures to show the gain or reduction in each genotype scenario. For many traits, ChipPlusSign and Top40k showed greater accuracy than Chip (Fig. 3.1). Using ChipPlusSign instead of Chip resulted in a maximum gain of 1.49% for ADFI in ML2 and 1.61% for ADG in ML1, and the greatest decrease of -0.74% and -0.26% for WWT in ML2 and ML1, respectively. The mean across all eight traits increased as the number of genotyped animals increased from ML2 to ML1, although the percentage of gain was very small (0.49% to 0.75%) compared to Chip. Accuracy gains by using Top40k were greater than the ones from ChipPlusSign. The average gain using Top40k for ML2 and ML1 was 4.34% and 5.54%, respectively, following the increasing number of animals with WGS data. The largest gains in each line with Top40k were 22.87% and 34.77% for RET in ML2 and ML1, whereas the largest reduction was -4.25% and -5.19% for ADFI in ML2 and WWT in ML1, respectively.

LDTags showed inconsistent results among traits. In ML1, the gains compared to Chip were 12.36% for BTP, 13.67% for NSB, and 94.06% for RET. Although a gain of 94.06% is impressive, the accuracy for RET using Chip was as low as 0.14. Likewise, in ML1, using LDTags in ML2 resulted in greater accuracy than Chip for BTP (2.82%) and RET (35.04%), but not for NSB (-8.80%). Therefore, pre-selection of variants from GWAS (ChipPlusSign and Top40k) could improve prediction accuracy for most traits in maternal lines, although those gains were small to modest. The pre-selection of SNP based on LD pruning (LDTags) also increased prediction accuracy for some reproduction-related traits; however, meaningful reductions were observed in most growth-related traits.

Genomic prediction accuracy of terminal lines using ssGBLUP

Changes in prediction accuracy (%) of ChipPlusSign, Top40k, and LDTags compared to Chip for four terminal lines are described in Fig. 3.2. The x-axis was sorted by the number of genotyped animals in each line as in maternal lines (small to large; TL3, TL2, TL1, and TL4 by order). ChipPlusSign reported a consistent gain in accuracy among all traits and lines, except for LDPX in TL2 (-5.26%) and LDPX in TL3 (-0.45%). As the results of ChipPlusSign in maternal lines, the result in terminal lines also showed an increasing pattern as the number of genotyped animals increased following the order, TL3, TL2, TL1, and TL4. The average gain for all seven traits was 0.37%, 0.56%, 0.95%, and 1.50% and the maximum gain was 0.94% (TL3-ADFI), 2.27% (TL2-ADGX), 1.32% (TL1-BF), and 2.61% (TL4-ADGX), following the previous order. Contrary to ChipPlusSign results, Top40k reported inconsistent results among the traits and lines. Although TL3 and TL4 showed accuracy gains for most traits, except ADFI in TL4 (-0.49%), TL1 and TL2 reported a reduction in accuracy for many traits (six traits for TL1 and four traits for TL2). On average, TL3 showed the second greatest accuracy gain for all seven traits (2.35%), with

the smallest number of genotyped animals among all terminal lines. TL2, TL1, and TL4 reported -3.32%, -6.28%, and 3.30% accuracy changes, meaning that the number of genotyped animals did not affect the gain with Top40k for terminal lines. However, the largest genotyped line (TL4) showed the most notable average gain (3.30%). The maximum gains in each line were 3.74% (LDP), 16.16% (ADGX), 1.55% (BFX), and 7.91% (ADGX) for TL3, TL2, TL1, and TL4, respectively. For both ChipPlusSign and Top40k, TL2 showed the greatest standard deviation among the traits, 2.67 and 18.37, respectively, meaning that accuracy changes highly depend on the traits in TL2.

Genomic prediction results using LDTags reported accuracy reduction for all the traits among the models but gain for BFX in TL2 (2.81%). In general, the use of ChipPlusSign reported improved accuracy for most of the traits in the terminal lines. However, those gains were limited (maximum 2.61% for ADGX in TL4). Results of Top40k showed decreased accuracy in most of the traits for TL1 and TL2, on the contrary, increased accuracy was outlined in TL3 and TL4 for almost all traits. LDTags indicated lower accuracy than Chip data for all terminal lines.

Inflation/deflation of GEBV

Fig. 3.3 describes the b1 values for all genotyped scenarios for maternal (a) and terminal lines (b). All the values in each genotype scenario were averaged across all traits in each line. Fig. 3.3-(a) described the results of maternal lines. When the number of genotyped animals increased (ML2 to ML1), b1 departed from 1.0 (0.86 in ML2 and 0.79 in ML1). More specifically, LDTags showed more notable deflation of GEBV from ML2 to ML1 (1.17 to 1.23, respectively). However, other genotype panels had greater inflation of GEBV from ML2 to ML1. The results of Chip, ChipPlusSign, and Top40k were similar within the lines.

The results of terminal lines are in Fig. 3.3-(b). Compared to the results of maternal lines, inconsistent patterns were observed in terminal lines among the traits and lines. Overall, all the terminal lines showed inflated GEBVs for all traits and genotype panels. LDTags showed less inflation for TL3, TL1, and TL4 (0.80, 0.79, and 0.88, respectively), whereas the greatest inflation was observed in TL2 (0.43). On average, TL3 reported the best result (0.76), followed by TL4, TL1, and TL2 (0.72, 0.70, and 0.62, respectively). The Chip, ChipPlusSign, and Top40k yielded similar results in TL3 and TL4. In TL1 and TL2, however, Top40k showed either less inflation (+0.05) or more inflation (-0.13) compared to Chip and ChipPlusSign.

Genomic prediction using WssGBLUP

WssGBLUP using BW was only applied to the four largest lines (ML1, ML2, TL1, and TL4) for growth-related traits (ADFI, ADG, BF, and LD) with Top40k and ChipPlusSign because those sets showed the best performance among the preselected genotype panels. Comparisons were made between Top40k and Top40k_BW and between ChipPlusSign and ChipPlusSign_BW. Results of prediction accuracy are in Table 3.4. For ML1, no gain was observed with WssGBLUP. In ML2, using Top40k_BW and ChipPlusSign_BW showed 0.02 accuracy gain in BF, whereas -0.09 reduction in LD. TL1 results also showed no improvement in accuracy. The greatest gains (~0.06) were outlined in the results of TL4 when Top40k_BW or ChipPlusSign_BW were used. No gain was observed for ADFI. However, 0.06, 0.03, and 0.04 accuracy gains were observed for ADG, BF, and LD when Top40k_BW was used instead of Top40k. Similarly, 0.03, 0.02, and 0.04 increases in accuracy were reported with ChipPlusSign_BW for ADG, BF, and LD. Overall, WssGBLUP showed similar b1 values to the regular ssGBLUP except for a few scenarios (results not shown).

DISCUSSION

The current study investigated the impact of using large-scale WGS data for genomic prediction through ssGBLUP and WssGBLUP in maternal and terminal pig lines. This is the first study applying ssGBLUP to over 1.8k sequenced and 100k imputed sequenced animals in pigs. Several preselected genotype scenarios were created to compare genomic predictions with WGS and the regular SNP chip. Our results showed that preselected variants could outperform the regular SNP chip for genomic prediction, although not consistently across the lines and traits and with relatively limited gains. In addition, we observed the potential to improve prediction accuracy through WssGBLUP using posterior variance from the BayesR as SNP weight, especially for the largest genotyped populations. Our results suggest effective scenarios to construct preselected variant sets depending on traits and population sizes for maternal and terminal pig lines. In the discussion section, we will address the three topics: (1) Impact of preselected variants on genomic prediction (2) Using WGS data for genomic prediction in pigs (3) Comparison of different weighting methods in ssGBLUP.

Impact of preselected variants on genomic prediction

Theoretically, using WGS data can improve genomic predictions because they cover the entire genome, assuming that causative variants are likely presented in the data. Therefore, genomic prediction does not rely on LD between SNP and causative variants but directly uses causative variants (Meuwissen and Goddard, 2010). A common assumption is that all the variants can explain a large proportion of genetic variance in the WGS data. However, several studies reported that using WGS data did not improve prediction accuracies (Van Binsbergen et al., 2015; Zhang et al., 2018; Song et al., 2019). A plausible reason for that would be the use of redundant SNP. As the WGS data have millions of SNP across the entire genome, adjacent SNP likely have

a strong LD with causative variants or other SNP in certain genome blocks, indicating that most SNP are correlated, providing the same information. Therefore, fitting all the WGS data in the prediction model could lead to biased GEBV. To avoid bias, many studies have investigated the preselection of significant variants for genomic prediction (Brøndum et al., 2015; VanRaden et al., 2017; Fragomeni et al., 2019). Besides preselection, removing correlated SNP based on LD extent has been also explored (Calus et al., 2016; Song et al., 2019).

Thus, in the current study, three different preselected genotype panels were designed and compared to the regular chip data. Those three panels were constructed following different assumptions. For ChipPlusSign, significant variants ($p \leq 10^{-6}$) based on GWAS were added to the regular SNP chip with an expectation of better prediction accuracy if the significant SNP had large effects or were causative and not presented in Chip. Incorporating preselected, significant SNP into the regular SNP chip has been investigated in many studies with WGS data (Fragomeni et al., 2017; VanRaden et al., 2017; Moghaddar et al., 2019). Top40k was created to mimic the number of SNP in the regular medium-density SNP chips used for routine genomic evaluation in many farm animals (e.g., pigs, cattle, and chickens). As most of the regular SNP chips contain evenly spaced SNP, Top40k also consisted of 40k SNP selected from each consecutive non-overlapping 55-kb window of WGS with the lowest p-value (i.e., from GWAS) in each window. Therefore, gains in prediction accuracy were expected if those preselected 40k SNP from WGS data were more informative and explained a more notable proportion of genetic variation than the SNP in the regular chip. As WGS data had many SNP with strong LD extent to each other or causative variants, removing highly correlated SNP was considered a reasonable strategy to preselect SNP, which was applied in the LDTags scenario. Song et al. (2019) reported that using LD-pruned WGS data outperformed an 80k SNP panel in pigs. Therefore, we expected that LDTags could

outperform Chip if those redundant SNP were correctly removed and SNP with considerable effects retained.

Among the preselected genotype sets, ChipPlusSign showed small to moderate accuracy gain for many traits in maternal and terminal lines. In this study, ChipPlusSign showed the most consistent results across lines and traits, with accuracy gains in most of them; however, within a limited range (from 0.12% to 2.61%). Several studies have been conducted to investigate genomic prediction by adding selected variants to regular chip data through either real or simulated data sets (VanRaden et al., 2017; Fragomeni et al., 2019; Moghaddar et al., 2019; Jang et al., 2022). In US Holstein, VanRaden et al. (2017) investigated the reliability of GEBV for 33 traits when preselected SNP (N = 16k) from WGS were added to a 60k SNP chip. They reported up to a 4.8 percentage point increase in reliability (15.35%) with an average of 2.7 extra points (9.15%) compared to the reliability obtained from a 60k SNP chip. However, when Fragomeni et al. (2019) investigated the performance of ssGBLUP using the same preselected variants set created by VanRaden et al. (2017), almost no gain in reliability (0.92%) was observed, although reliabilities were greater in Fragomeni et al. (2019). One major difference between those two studies was the method of genomic prediction, multistep in VanRaden et al. (2017) and ssGBLUP in Fragomeni et al. (2019). When ssGBLUP was used, it combined all the information from genotyped and non-genotyped animals, allowing the inclusion of a massive amount of data. In such a scenario, gains in reliability are less likely if the selected variants are redundant, not truly causative, or have a small effect on the traits of interest. Our results agree with the ones from Fragomeni et al. (2019), especially with ChipPlusSign. In a simulated study, Jang et al. (2022) investigated the dimensionality of genomic information for variant selection and genomic prediction with sequence data. Their results showed that populations with small *Ne* obtained a maximum accuracy gain of

0.86% to 1.98% when either significant variants or hundreds of variants with high effect sizes preselected from GWAS were added to a 50k SNP chip. The small *Ne* scenario they simulated was 20, close to the *Ne* in pig populations (32~48) (Pocrnic et al., 2016b).

In our study, the results of Top40k highly depended on the traits and lines. Top40k showed the greatest gain for RET across all maternal lines (22.87% to 34.77%), but relatively marginal gains or reductions for other traits. In the terminal lines, results of Top40k fluctuated more among lines, with increased or decreased accuracies. The possible reason for the large improvement observed for reproduction or fertility traits in maternal lines might be the nature of the traits and the lack of informative SNP in the regular SNP chip. The lack of informative SNP for fertility traits led to a recent change in the SNP chip for beef cattle evaluations (https://www.angus.org/AGI/global/AngusGS.pdf).

Heritabilities for RET were relatively lower than other traits. Consequently, this trait had the lowest prediction accuracies (0.14 for ML1 and 0.20 for ML2 with Chip) among other traits in maternal lines. Thus, there would be a greater room for improvement in genomic predictions through using preselected genotype data if the SNP in Chip could not explain a large proportion of the genetic variance. Therefore, we speculated that there would be more informative SNP in Top40k for RET, which were not included in the regular SNP chip. Likewise, small-scale differences in accuracy observed between Chip and Top40k were possibly due to variants capturing similar proportions of genetic variance and having similar LD patterns across the genome. In terminal lines, the maximum gain was observed for ADGX (16.16% in TL2), recorded in crossbred animals. The ADGX was investigated in the GROWTH model along with three correlated traits (ADG, BF, and BFX), and the Top40k was created based on GWAS for ADG and BF, individually, and then each Top40k was combined for genomic prediction. Thus, this result

showed the potential for prediction improvement for the traits recorded in crossbred animals if many phenotypes and WGS animals are available although those traits were not directly used for preselection of variants.

Those marginal gains for most traits with ChipPlusSign and Top40k raised a question about the amount of information that has been used for preselecting the variants and performing genomic predictions. Examining the dimensionality of the genomic information can help assess the efficient number of genotyped animals needed to maximize the percentage of discoveries in GWAS and prediction accuracy gains (Jang et al., 2022). According to Jang et al. (2022), in populations with a larger effective size (N_e = 200), using the number of genotyped animals equal to the number of largest eigenvalues explaining 98% of the variance of **G** sufficed to capture the most informative variants, although only a tiny proportion of the causative variants was discovered for highly polygenic traits. However, their study showed that populations with a smaller effective size (N_e = 20) required much more data to capture causative variants. For example, when 30k genotyped animals were used in GWAS for highly polygenic traits, only three causative variants were identified, explaining 3.9% of genetic variation. In addition, incorporating preselected variants to regular chip data reported a nearly 2% maximum gain in accuracy for the scenarios with N_e = 20. In the current study, the number of WGS animals used for GWAS was 29k to 104k, which is the largest WGS data in pigs by far. However, the fine-mapping of causative variants was still challenging and the benefits for genomic predictions were limited. Since pig populations have small N_e and most of the traits are highly polygenic, to capture the most informative variants, a very large number of WGS animals having lots of progeny records would be required (Jang et al., 2022).

In an initial batch of analyses (results not shown), we used only significant variants (TopSign) for genomic predictions, which showed no benefits compared to Chip. In fact, accuracies were reduced for most of the traits and lines. The number of variants in TopSign ranged from 6 to 1,705 depending on the lines and traits. Fragomeni et al. (2017) outlined that the maximum accuracy of GEBV could be obtained if the true causative variants were identified with their exact substitution effects, position in the genome, and genetic variance explained by each variant assigned as weight. Therefore, our results revealed that the variants in TopSign might not be the true causative variants, so the use of those variants underperformed regular SNP chip information.

In addition to using GWAS to preselect variants, LD pruning was also performed to reduce the number of variants in WGS. The LD extent between variants in WGS data is greater than in the chip data because of the dense distribution of many SNP across the genome. Therefore, several studies investigated the impact of LD pruning of WGS data on genomic prediction (Calus et al., 2016; Song et al., 2019). According to Song et al. (2019), using WGS after LD pruning showed the greatest prediction accuracy for both reproduction and production traits in pigs. Calus et al. (2016) concluded that pruning variants based on LD is an important step as those many variants having strong LD reduced the prediction performance. Our results showed decreased accuracy in terminal lines for almost all traits, and some gains in maternal lines, specifically for the reproduction traits (BTP, NSB, and RET). The possible reason for the discrepancy between previous studies and the current study could be the different criteria for LD pruning. The previous studies by Calus et al. (2016); Song et al. (2019), pruned the variants for $r^2 > 0.9$, whereas we pruned the variants for $r^2 > 0.1$ to maintain only a small number of variants. However, the chances of removing the true causative variants with stringent than lenient criteria are higher because LD

pruning only considered the LD extent between the variants, not the substitution effects of variants on traits of interest.

Using WGS data for genomic prediction in pigs

The cost of sequencing is getting cheaper, so using sequence data for genomic prediction of farm animals (e.g., sheep, beef cattle, dairy cattle, and pigs) has been more approachable than in the past. Several studies have been carried out and reported no or marginal benefits of using WGS on genomic prediction in sheep, beef, and dairy cattle (VanRaden et al., 2017; Fragomeni et al., 2019; Moghaddar et al., 2019; de Las Heras-Saldana et al., 2020; Lopez et al., 2021). Compared to other farm animals, using WGS data for genomic prediction in pigs has been barely investigated and only small-scale data sets were used (Zhang et al., 2018; Song et al., 2019). The number of WGS pigs in those studies was less than 7k. As the number of variants in WGS increases, more samples are required to resolve the well-known issue, '$N << p$', where N is the sample size and p is the number of variants. If the sample size is not sufficient, estimation of SNP effects and identification of causative SNP could be troublesome, especially for populations with small N_e and highly polygenic traits.

In the current study, the number of sequenced animals was around 380 to 1.8k, which represented nearly 2% of the population in each line (Ros-Freixedes et al., 2020). However, depending on the line, the WGS information was imputed to be about 29k to 104k animals. Applying large-scale WGS data to preselect the variants through GWAS and using those variants for genomic prediction in this study showed limited improvement as found in Zhang et al. (2018); Song et al. (2019), in which only a small number of animals had WGS. Theoretically, increasing the sample size enhanced the power to detect causative variants and improved genomic predictions (Daetwyler et al., 2010; Meuwissen and Goddard, 2010). However, the pig populations are highly

structured and have a small N_e. Therefore, increasing only the sample size might not help improve the performance of both variant selection and genomic prediction. Jang et al. (2022) reported that using animals with greater EBV reliability (more data available) helped better identify the causative variants than using animals that had lower EBV reliability. Therefore, selecting high-reliability animals and using them could be a possible strategy. Another possible reason for limited benefit could be the imputation accuracy (Van Den Berg et al., 2017; Ros-Freixedes et al., 2020). The ideal situation to use WGS data is to sequence all the animals in the population without imputation from genotype to sequence level. However, as sequencing the entire population is still impossible, imputation is an inevitable procedure for dealing with WGS data. Since only a limited number of WGS animals are used as a reference for imputation, sequencing more animals and using robust statistical tools to impute alleles accurately are required.

Comparison of different weighting methods in ssGBLUP

WssGBLUP was investigated in addition to ssGBLUP. A major assumption of GBLUP-based methods is that all markers have homogeneous variance. Those methods have been extensively applied for most of the traits in farm animals due to their highly polygenic nature (Aguilar et al., 2010; Lourenco et al., 2015). However, that assumption does not biologically hold because not all markers in the genome explain the same proportion of variance (Meuwissen et al., 2001). Therefore, assigning heterogeneous variance per marker for genomic prediction has been investigated in several studies (Wang et al., 2012; Zhang et al., 2016; Gualdrón-Duarte et al., 2020). Weighting SNP in ssGBLUP was initially proposed by Wang et al. (2012) by assigning unequal SNP variance through squared SNP effects weighted by allele frequencies. However, this method caused a reduction in GEBV accuracy and extra bias over iteration due to the extreme values of SNP variance, especially for the polygenic traits (Lee et al., 2017; Lourenco et al., 2017).

Following the increased accuracy reported by Gualdrón-Duarte et al. (2020), we used the posterior SNP variances from BayesR as SNP weights. In BayesR, SNP effects are sampled from a mixture of four normal distributions with mean zero and variances equivalent to the following classes: 0, $0.0001\sigma_g^2$, $0.001\sigma_g^2$, and $0.01\sigma_g^2$ (Moser et al., 2015). Therefore, we assumed that this strategy would construct a better weighting matrix close to the true variance of SNP. Our results showed that BW outperformed ssGBLUP for ADG, BF, and LD in TL4 for both Top40k and ChipPlusSign up to 0.06. However, other traits in ML1, ML2, and TL1 showed similar results as in ssGBLUP. Gualdrón-Duarte et al. (2020) compared the performances of weighting strategies in Belgian Blue beef cattle. In their study, the average reliability of genomic prediction for 14 traits using GBLUP and BayesR weighted GBLUP showed no differences. However, applying posterior variance of marker effect from the Bayesian mixture model (similar to BayesR) as the weighting factor showed the best performance among other weighting strategies and regular GBLUP in the Nordic Holstein population (Su et al., 2014).

The current study showed absent or modest gains in prediction accuracy depending on the lines and traits. We expected almost no gain with WssGBLUP, especially for the largest genotyped population (TL4), as the SNP effects were likely dominated by a large amount of data in the single-step system. However, we observed potential room for improvements in predictions when using the posterior variance of BayesR even with a large amount of data. In other words, although the volume of data could overwhelm *a priori* assumption of SNP effects, improvements can still occur if the variances used as SNP weights are accurate enough.

CONCLUSIONS

Preselection of significant variants from whole-genome sequence data and their utilization could help to improve genomic predictions in both maternal and terminal pig lines with tens of thousands of sequenced/imputed animals. However, a limited gain is expected even in large populations. Greater gains are observed when selected variants for some traits are not already present in the commercial SNP chips. Weighting SNP using BayesR variances slightly boosts prediction accuracies. The results of genomic prediction using several preselected variant sets highly depend on the genetic architecture of traits, population structure, number of genotyped animals, and method to select variants.

REFERENCES

Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta, and T. Lawlor. 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. doi:10.3168/jds.2009-2730

Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 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. doi:10.3168/jds.2014-9005

Calus, M. P., A. C. Bouwman, C. Schrooten, and R. F. Veerkamp. 2016. Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection. Genetics Selection Evolution 48(1):1-19. doi:10.1186/s12711-016-0225-x

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

de Las Heras-Saldana, S., B. I. Lopez, N. Moghaddar, W. Park, J.-e. Park, K. Y. Chung, D. Lim, S. H. Lee, D. Shin, and J. H. van der Werf. 2020. Use of gene expression and whole-genome sequence information to improve the accuracy of genomic prediction for carcass traits in Hanwoo cattle. Genetics Selection Evolution 52(1):1-16. doi:10.1186/s12711-020-00574-2

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

Erbe, M., B. Hayes, L. Matukumalli, S. Goswami, P. Bowman, C. Reich, B. Mason, and M. 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):4114-4129. doi:10.3168/jds.2011-5019

Fragomeni, B., D. Lourenco, A. Legarra, P. VanRaden, and I. Misztal. 2019. Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants. Journal of dairy science 102(11):10012-10019. doi:10.3168/jds.2019-16262

Fragomeni, B. O., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2017. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. Genetics Selection Evolution 49(1):59. doi:10.1186/s12711-017-0335-0

Gualdrón-Duarte, J. L., A.-S. Gori, X. Hubin, D. Lourenco, C. Charlier, I. Misztal, and T. Druet. 2020. Performances of Adaptive MultiBLUP, Bayesian regressions, and weighted-GBLUP approaches for genomic predictions in Belgian Blue beef cattle. BMC genomics 21(1):1-18. doi:10.1186/s12864-020-06921-3

Jang, S., S. Tsuruta, N. G. Leite, I. Misztal, and D. Lourenco. 2022. Dimensionality of genomic information and its impact on GWA and variant selection: a simulation study. bioRxiv. doi:10.1101/2022.04.13.488175

Lee, J., H. Cheng, D. Garrick, B. Golden, J. Dekkers, K. Park, D. Lee, and R. Fernando. 2017. Comparison of alternative approaches to single-trait genomic prediction using genotyped and non-genotyped Hanwoo beef cattle. Genetics Selection Evolution 49(1):1-9. doi:10.1186/s12711-016-0279-9

Lopez, B. I. M., N. An, K. Srikanth, S. Lee, J.-D. Oh, D.-H. Shin, W. Park, H.-H. Chai, J.-E. Park, and D. Lim. 2021. Genomic Prediction Based on SNP Functional Annotation Using Imputed Whole-Genome Sequence Data in Korean Hanwoo Cattle. Frontiers in genetics:1523. doi:10.3389/fgene.2020.603822

Lourenco, D., B. Fragomeni, H. Bradford, I. Menezes, J. Ferraz, I. Aguilar, S. Tsuruta, and I. Misztal. 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. Journal of Animal Breeding and Genetics 134(6):463-471. doi:10.1111/jbg.12288

Lourenco, D., S. Tsuruta, B. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. Bertrand, T. Amen, L. Wang, and D. Moser. 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. Journal of animal science 93(6):2653-2662. doi:10.2527/jas.2014-8836

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

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819-1829. doi:10.1093/genetics/157.4.1819

Misztal, I., A. Legarra, and I. Aguilar. 2014a. Using recursion to compute the inverse of the genomic relationship matrix. Journal of dairy science 97(6):3943-3952. doi:10.3168/jds.2013-7752

Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, and Z. Vitezica. 2014b. Manual for BLUPF90 family of programs. Athens: University of Georgia

Moghaddar, N., M. Khansefid, J. H. van der Werf, S. Bolormaa, N. Duijvesteijn, S. A. Clark, A. A. Swan, H. D. Daetwyler, and I. M. MacLeod. 2019. Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. Genetics Selection Evolution 51(1):72. doi:10.1186/s12711-019-0514-2

Moser, G., S. H. Lee, B. J. Hayes, M. E. Goddard, N. R. Wray, and P. M. Visscher. 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS genetics 11(4):e1004969. doi:10.1371/journal.pgen.1004969

Pocrnic, I., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016a. The dimensionality of genomic information and its effect on genomic prediction. Genetics 203(1):573-581. doi:10.1534/genetics.116.187013

Pocrnic, I., D. A. Lourenco, Y. Masuda, and I. Misztal. 2016b. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. Genetics Selection Evolution 48(1):1-9. doi:10.1186/s12711-016-0261-6

Ros-Freixedes, R., M. Johnsson, A. Whalen, C.-Y. Chen, B. D. Valente, W. O. Herring, G. Gorjanc, and J. M. Hickey. 2022. Genomic prediction with whole-genome sequence data in intensely selected pig lines. bioRxiv:2022.2002.2002.478838. doi: 10.1101/2022.02.02.478838

Ros-Freixedes, R., A. Whalen, C.-Y. Chen, G. Gorjanc, W. O. Herring, A. J. Mileham, and J. M. Hickey. 2020. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. Genetics Selection Evolution 52:1-15. doi:10.1186/s12711-020-00536-8

Song, H., S. Ye, Y. Jiang, Z. Zhang, Q. Zhang, and X. Ding. 2019. Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. Genetics Selection Evolution 51(1):58. doi:10.1186/s12711-019-0500-8

Su, G., O. Christensen, L. Janss, and M. Lund. 2014. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. Journal of dairy science 97(10):6547-6559. doi:10.3168/jds.2014-8210

Van Binsbergen, R., M. P. Calus, M. C. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 47(1):71. doi:10.1186/s12711-015-0149-x

van den Berg, I., D. Boichard, B. Guldbrandtsen, and M. S. Lund. 2016. Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: a simulation study. G3: Genes, Genomes, Genetics 6(8):2553-2561. doi:10.1534/g3.116.027730

Van Den Berg, I., P. J. Bowman, I. M. MacLeod, B. J. Hayes, T. Wang, S. Bolormaa, and M. E. Goddard. 2017. Multi-breed genomic prediction using Bayes R with sequence data and dropping variants with a small effect. Genetics Selection Evolution 49(1):70. doi:10.1186/s12711-017-0347-9

VanRaden, P., and G. Wiggans. 1991. Derivation, calculation, and use of national animal model information. Journal of Dairy Science 74(8):2737-2746. doi:10.3168/jds.S0022-0302(91)78453-1

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. Journal of dairy science 91(11):4414-4423. doi:10.3168/jds.2007-0980

VanRaden, P. M., M. E. Tooker, J. R. O'connell, J. B. Cole, and D. M. Bickhart. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. Genetics Selection Evolution 49(1):32. doi:10.1186/s12711-017-0307-4

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics Research 94(2):73-83. doi:10.1017/S0016672312000274

Whalen, A., R. Ros-Freixedes, D. L. Wilson, G. Gorjanc, and J. M. Hickey. 2018. Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees. Genetics Selection Evolution 50(1):1-15. doi:10.1186/s12711-018-0438-2

Zhang, C., R. A. Kemp, P. Stothard, Z. Wang, N. Boddicker, K. Krivushin, J. Dekkers, and G. Plastow. 2018. Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. Genetics Selection Evolution 50(1):1-13. doi:10.1186/s12711-018-0387-9

Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal. 2016. Weighting strategies for single-step genomic BLUP: an iterative approach for accurate calculation of GEBV and GWAS. Frontiers in genetics 7:151. doi:10.3389/fgene.2016.00151

TABLES

Table 3.1. Number of records and animals in the pedigree

Line	ADFI	ADG	BF	LDP	BTP	NSB	RET	WWT	Pedigree
ML1	35k	1.06M	820k	604k	1.08M	1.13M	0.86M	34k	3.75M
ML2	34k	1.52M	936k	631k	5.11M	5.28M	4.10M	29k	9.18M
Line	ADFI	ADG	BF	ADGX	BFX	LDP		LDPX	Pedigree
TL1	35k	356k	339k	150k	149k	305k		148k	1.13M
TL2	40k	298k	295k	158k	156k	294k		156k	0.84M
TL3	16k	233k	226k	155k	153k	212k		152k	1.30M
TL4	64k	937k	859k	299k	247k	753k		243k	3.14M

*ADFI: Average daily feed intake; ADG: Average daily gain; BF: Backfat thickness; LDP: Loin depth; BTP: Total born; NSB: Number of stillborn; RET: Return to oestrus seven days after weaning; WWT: Litter weaning weight; ADGX: ADG recorded in crossbred; BFX: BF recorded in crossbred; LDPX: LDP recorded in crossbred

*ML1: Maternal line1; ML2: Maternal line2; TL1: Terminal line1; TL2: Terminal line2; TL3: Terminal line3; TL4: Terminal line4

Table 3.2. Number of genotyped individuals, SNP, and sequenced animals in six lines

Line	Number of genotyped individuals	Number of SNP	Number of sequenced individuals
ML1	76,227	40,592	1,365
ML2	66,608	42,746	1,491
TL1	60,467	35,786	731
TL2	41,572	40,311	760
TL3	29,328	39,999	381
TL4	104,644	43,032	1,865

*ML1: Maternal line1; ML2: Maternal line2; TL1: Terminal line1; TL2: Terminal line2; TL3: Terminal line3; TL4: Terminal line4

Table 3.3. Number of animals with genomic information that was retained after quality control and used in the analyses with all SNP panels

Line	ADFI	GROWTH	LOIN	REPROD	RET	WWT
ML1	74,148	74,153	74,152	73,919	73,891	74,058
ML2	64,654	64,655	64,659	64,599	64,653	63,456
TL1	56,423	56,424	56,422	-	-	-
TL2	38,477	38,475	38,477	-	-	-
TL3	27,671	27,671	27,671	-	-	-
TL4	102,586	102,590	102,588	-	-	-

*In ML1 and ML2; ADFI: two-trait ADFI model (ADG and ADFI); GROWTH: two-trait GROWTH model (ADG and BF); LOIN: two-trait LOIN model (ADG and LDP); REPROD: two-trait REPROD model (BTP and NSB); RET: single-trait RET model (RET); WWT: single-trait WWT model (WWT)
*In TL1, TL2, TL3, and TL4; ADFI: two-trait GROWTH model (ADG and ADFI); GROWTH: four-trait GROWTH model (ADG, BF, ADGX, and BFX); LOIN: four-trait LOIN model (ADG, LDP, ADGX, and LDPX)

Table 3.4. Prediction accuracy of WssGBLUP compared to ssGBLUP

Line	Description	ADFI	ADG	BF	LDP
ML1	Top40k	0.37	0.49	0.51	0.53
	Top40k_BW	0.37	0.49	0.51	0.53
	ChipPlusSign	0.37	0.47	0.51	0.52
	ChipPlusSign_BW	0.37	0.47	0.51	0.51
ML2	Top40k	0.36	0.61	0.64	0.62
	Top40k_BW	0.35	0.62	0.66	0.53
	ChipPlusSign	0.37	0.61	0.63	0.62
	ChipPlusSign_BW	0.37	0.62	0.65	0.53
TL1	Top40k	0.34	0.45	0.59	0.56
	Top40k_BW	0.34	0.45	0.59	0.55
	ChipPlusSign	0.36	0.49	0.61	0.60
	ChipPlusSign_BW	0.36	0.50	0.60	0.60
TL4	Top40k	0.39	0.51	0.60	0.59
	Top40k_BW	0.39	0.57	0.63	0.63
	ChipPlusSign	0.40	0.51	0.60	0.57
	ChipPlusSign_BW	0.40	0.54	0.62	0.61

*ML1: Maternal line1; ML2: Maternal line2; TL1: Terminal line1; TL4: Terminal line4
*ADFI: Average daily feed intake; ADG: Average daily gain; BF: Backfat thickness; LDP: Loin depth
*Top40k: Top40k preselected genotype panel; Top40k_BW: Top40k using BayesR weighting
*ChipPlusSign: ChipPlusSign preselected genotype panel; ChipPlusSign_BW: ChipPlusSign using BayesR weighting

FIGURES

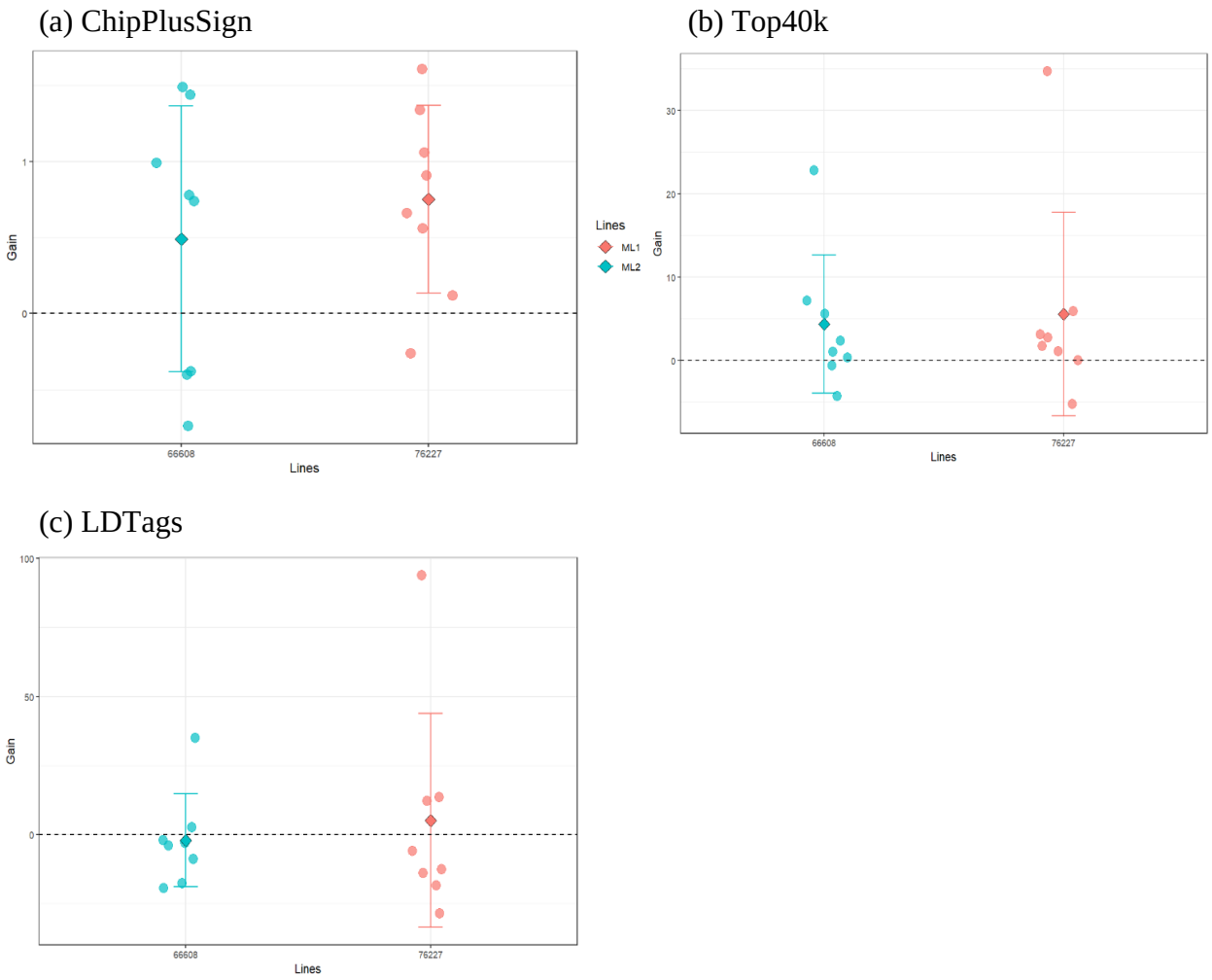
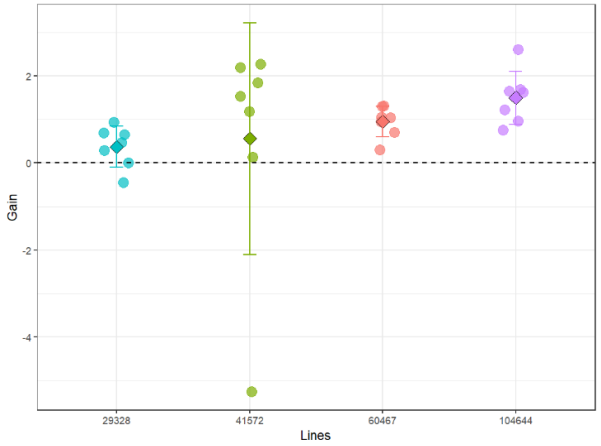
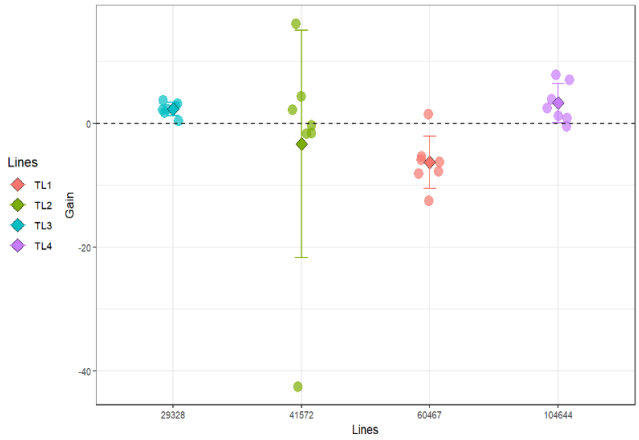


Figure 3.1. Accuracy changes (%) of ChipPlusSign, Top40k, and LDTags compared to Chip in maternal lines

(a) ChipPlusSign



(b) Top40k



(c) LDTags

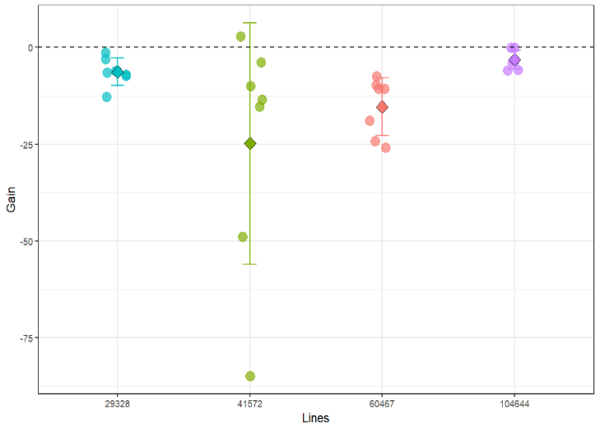


Figure 3.2. Accuracy changes (%) of ChipPlusSign, Top40k, and LDTags in terminal lines

(a) Maternal lines (b) Terminal lines

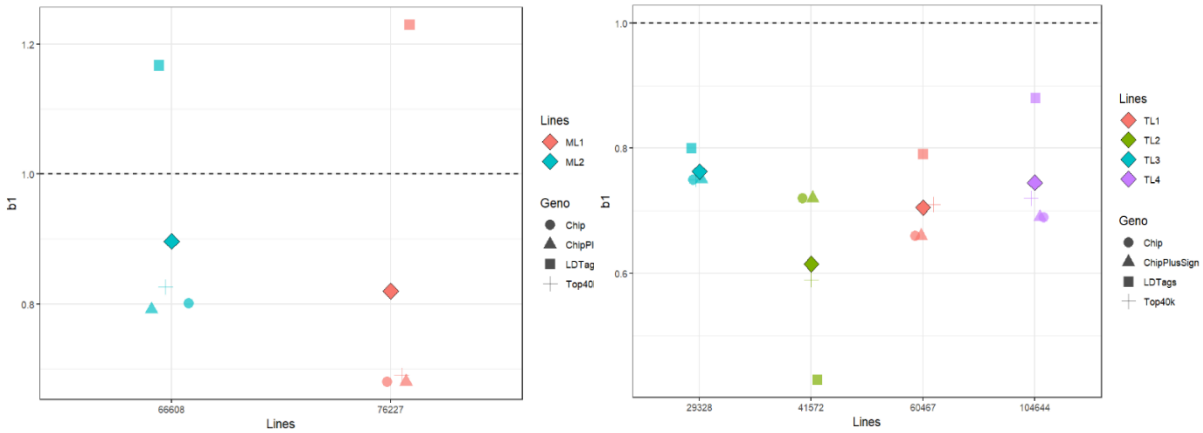


Figure 3.3. b1 values for all the genotype scenarios in both maternal and terminal lines

CHAPTER 4

MULTI-LINE SINGLE-STEP GENOMIC EVALUATION USING PRESELECTED MARKERS FROM WHOLE-GENOME SEQUENCE DATA IN PIGS[2]

[2]Sungbong Jang, Roger Ros-Freixedes, John M. Hickey, Ching-Yi Chen, William O. Herring, Ignacy Misztal, and Daniela Lourenco. To be submitted to *Genetics Selection Evolution*.

ABSTRACT

Genomic evaluations in pigs could benefit from using multi-line data along with whole-genome sequence (WGS) if data are large enough to represent the variability across populations. The objective of this study was to investigate strategies to combine large-scale data from different terminal pig lines in a multi-line genomic evaluation (MLE) through single-step GBLUP (ssGBLUP) models while including variants preselected from whole-genome sequence (WGS) data. We explored three terminal lines under single-line and multi-line evaluations for five traits. The number of sequenced animals in each line ranged from 731 to 1,865, with 60k to 104k imputed to WGS. Unknown parent groups (UPG) and metafounders (MF) were explored to account for genetic differences among lines and improve the compatibility between pedigree and genomic relationships in the MLE evaluations. Sequence SNP were preselected based on multi-line genome-wide association studies (GWAS) or linkage disequilibrium (LD) pruning. Those preselected variants sets were used for ssGBLUP predictions without and with weights from BayesR, and the performances were compared to that of a commercial porcine SNP chip. Using UPG and MF in MLE showed small to no gain in prediction accuracy (up to 0.02), depending on the lines and traits, compared to the single-line genomic evaluation (SLE). Likewise, adding selected variants from GWAS to the commercial SNP chip resulted in a maximum increase in prediction accuracy of 0.02, only for average daily feed intake in the most numerous lines. Besides, no benefits were observed when using preselected sequence variants in multi-line genomic predictions. Weights from BayesR did not help improve the performance of ssGBLUP. This study revealed limited benefits of using preselected whole-genome sequence variants for multi-line genomic predictions, even when tens of thousands of animals had imputed sequence data. Correctly accounting for line differences with unknown parent groups or metafounders in multi-

line evaluations is essential to obtain predictions similar to single-line; however, the only observed benefit of a multi-line evaluation is to have comparable predictions across lines. Further investigation on the amount of data and novel methods to preselect whole-genome causative variants in combined populations would be of great interest.

INTRODUCTION

Genomic evaluations have been successfully implemented in pig breeding programs to increase the accuracy of predicting genomic EBV (GEBV) and better identify the best animals to be parents of the next generation. However, lines with a small reference size may not experience the same benefits as large lines because the accuracy of GEBV could be limited by the size of the reference data set. Combining multiple lines could be a possible strategy for the small lines to benefit from genomics and the increased reference size. Several studies have investigated the impact of combining multiple lines or breeds in farm animals, such as dairy and beef cattle, chicken, and pigs (Calus et al., 2014; Rolf et al., 2015; Song et al., 2017; Cesarani et al., 2022). This would also allow the comparison of animals across lines and the identification of the best gene combinations. However, multi-line genomic evaluations (MLE) in pigs are still challenging because the main breeding objective is to improve pure lines for crossbred performance. In contrast, lines have heterogeneous genetic backgrounds and may be distantly related.

Single-step genomic BLUP (ssGBLUP) has been commonly used for genomic evaluation in pigs (Chen et al., 2011; Pocrnic et al., 2019; Song et al., 2019). The fundamental idea of this method is to use all available data, connecting genotyped and non-genotyped animals through a joint relationship matrix (**H**) (Legarra et al., 2009; Christensen and Lund, 2010). For that, ssGBLUP relies on the compatibility between pedigree (**A**) and genomic (**G**) relationship matrices

(Misztal et al., 2013). Two significant causes of incompatibility between **A** and **G** are missing pedigrees and heterogeneous base populations (Vitezica et al., 2011; Misztal et al., 2013). In theory, allele frequencies to construct **G** would correspond to the ones from the base population in the pedigree (VanRaden, 2008); however, base animals are seldom genotyped, making base allele frequencies unknown (Legarra et al., 2015). The incompatibility issue becomes critical in multi-line populations because of the heterogeneous base population across lines. Macedo et al. (2020) reported that better genomic predictions could be obtained for such scenarios if differences in base populations are correctly modeled. Unknown parent groups (UPG) could mitigate this issue by modeling the differences in genetic base across classes of missing parents and accounting for the differences among breeds or lines; however, UPG assume that the base populations are unrelated (Legarra et al., 2007). Legarra et al. (2015) proposed using metafounders (MF), which are pseudo-animals that act as proxies for the base individuals and can be related. A few studies investigated using ssGBLUP with MF in pigs, but only for crossbred (Xiang et al., 2017; van Grevenhof et al., 2019) and single-line (Fu et al., 2021) evaluations.

Another factor affecting the performance of genomic predictions in MLE would be the inconsistent linkage disequilibrium (LD) structure between single nucleotide polymorphisms (SNP) and quantitative trait loci (QTL) across the lines. Pig populations have a smaller effective size (*Ne*) than dairy and beef cattle, resulting in smaller numbers of independent chromosome segments (*Me*) (Pocrnic et al., 2016b). Therefore, if the lines are distantly related, they are not likely to share many chromosome segments. This could lead to no benefits from combining multiple lines in genomic predictions. Song et al. (2017) evaluated genomic predictions for growth and reproduction traits in pigs using a combined data set with genotypes for 80k SNP, but no benefits were observed over the single-line genomic evaluation (SLE). In another study, the same

authors (Song et al., 2019) showed that pruned whole-genome sequence (WGS) data outperformed the 80k SNP chip for genomic predictions in combined populations; however, no benefit was observed through the direct use of WGS data. This could be due to the redundancy of many SNP across the whole genome with strong LD extent to each other in certain genomic blocks. Therefore, preselection of significant SNP or removal of redundant SNP could be a possible strategy to improve the accuracy of genomic predictions when WGS data are used. In the case of MLE, a joint preselection of SNP from WGS can help identify variants segregating across lines, which may not be possible with commercial SNP chips because of the limited number of SNP (~ 40k to 80k).

In the current study, we aimed to (1) investigate strategies to combine different lines in a multi-line evaluation through the use of unknown parent groups or metafounders; (2) evaluate the impact of using jointly preselected SNP from WGS in multi-line evaluations under ssGBLUP without and with weights from BayesR.

MATERIALS AND METHODS

Animal Care and Use Committee approval was not needed because information was obtained from pre-existing databases.

Data

All datasets were provided by Pig Improvement Company (PIC; Hendersonville, TN). We investigated average daily feed intake (ADFI), average daily gain (ADG), backfat thickness (BF), ADG recorded in crossbred animals (ADGX), and BF recorded in crossbred animals (BFX) in three terminal pig lines named TL1, TL2, and TL3. A two-trait model was considered for ADFI and ADG (ADFI model), whereas a four-trait model was used for ADG, BF, ADGX, and BFX (GROWTH model). Two scenarios were considered in this study: SLE and MLE. For the MLE scenar

io, all the data from every single line were combined. The total number of records and animals in the pedigree for each line and MLE are in Table 4.1. Individuals in each line were genotyped with the GGP-Porcine HD BeadChip (GeneSeek, Lincoln, NE) and jointly imputed for MLE. We filte red out the monomorphic SNP and SNP with a call rate lower than 0.90, minor allele frequency lo wer than 0.01, and the difference between observed and expected genotype frequencies greater th an 0.15. Genotyped pigs with greater than 0.10 missing genotypes were removed as well. For ML E, all genotyped individuals in the three terminal lines were combined. Identical quality control w as applied to the imputed MLE chip data (Chip). The total number of genotyped animals in all lin es and SNP after quality control is described in Table 4.2.

Whole-genome sequencing and imputation

The WGS data used in this study were generated by Ros-Freixedes et al. (2020); Ros-Freixedes et al. (2022). In brief, a low-coverage sequencing strategy was followed by joint calling, phasing, and imputing the whole-genome genotypes using the 'hybrid peeling' method implemented in AlphaPeel (Whalen et al., 2018). The number of WGS individuals for each line is provided in Table 4.2. The 'hybrid peeling' method used all marker array and WGS data that was available across the pedigrees. Imputation was carried out separately in each line. Individuals with low predicted imputation accuracy were excluded, as described by Ros-Freixedes et al. (2020). A total of 60,474 (TL1), 41,573 (TL2), and 104,661 (TL3) WGS individuals remained in each line after quality control. These individuals were predicted to have an average dosage correlation of 0.97 (median: 0.98). SNP with a minor allele frequency lower than 0.023 were removed since their estimated dosage correlations were lower than 0.90 (Ros-Freixedes et al., 2020).

Training and test sets

Before the SNP preselection, all animals with (imputed) WGS were split into two non-overlapping data sets: training and test. Test sets were generated by extracting entire genotyped individuals in the litters from the last generation of the pedigree. The training sets were created by establishing a threshold on the pedigree relationship coefficient between training and test sets. We removed individuals with a relationship coefficient equal to or greater than 0.5 to test animals. Through this, the relationship between training and test sets became minimized, which could be a potential advantage when commercial data is used in the training data set. Training data sets were also used as discovery sets for GWAS in this study. Although using the same sets for training and GWAS can reduce accuracy and increase the bias of genomic predictions (Veerkamp et al., 2016), Ros-Freixedes et al. (2022) showed that using different sets for each task helped alleviate the bias but reduced prediction accuracy.

Pre-selected SNP panels

Five different preselected SNP panels were created from WGS for genomic predictions: (1) LDTags (2) Top40k, (3) TopSign, (4) ChipPlusSign, and (5) AllComb. After imputation, all WGS individuals from each line were combined for preselecting SNP for the MLE as described in (Ros-Freixedes et al., 2022). The LDTags were tag variants retained after pruning based on LD with an $r^2>0.1$. Top40k were the variants with the lowest p-value (not necessarily below the significance threshold) in each consecutive non-overlapping 55-kb window along the genome, based on multi-line GWAS analyses. TopSign only included significant variants ($p \leq 10^{-6}$) from the multi-line GWAS in a way that, when a 55-kb window contained more than one significant variant, only that with the lowest p-value was selected. For the multi-line GWAS, seven lines (TL1, TL2, TL3, TL4, and three maternal lines) were jointly investigated as described in Ros-Freixedes et al. (2022). ChipPlusSign combined TopSign and Chip because sometimes the number

of significant variants is small. AllComb contained the variants from LDTags, Top40k, TopSign, and Chip. The preselected variants for each trait were combined according to the traits included in the ADFI and GROWTH models. In the ADFI model, preselected variants for ADFI and ADG were combined and used for the genomic predictions. Likewise, preselected variants from ADG and BF were merged and used for genomic predictions in the GROWTH model; variants were not selected for ADGX and BFX because crossbred animals were not sequenced. For a fair comparison to the commercial SNP chip data (i.e., Chip), 206,634 animals were extracted from each preselected SNP set. Afterward, quality control was done with the same criteria mentioned above, except that individuals with parent-progeny conflicts were also removed. Table 4.3 depicts the number of genotyped animals and SNP for all preselected SNP panels after quality control.

Single-line genomic prediction

To compare the performance of genomic prediction using single-line and multi-line, Chip data was tested for SLE. Linear mixed models were used to perform genomic predictions with two and four traits for the ADFI and GROWTH models, respectively. Only a four-trait GROWTH model (ADG, BF, ADGX, and BFX) of TL1, TL2, and TL3 is described:

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

where **y** is the vector of phenotypes; **b** is a vector of fixed effects; **c** is a vector of random litter effects; **u** is a vector of random additive genetic effects; and **e** is a vector of residual effects. Matrix **X** is an incidence matrix relating phenotypes in vector **y** to fixed effects (contemporary group as a fixed effect for all traits, offtest weight and carcass weight as a covariate only for BF and BFX, respectively) in vector **b**, matrix **W** is an incidence matrix for random litter effects in vector **c**, matrix **Z** is an incidence matrix for random additive genetic effect in vector **u**.

Multi-line genomic prediction with unknown parent groups and Metafoudners

Linear mixed models were used to carry out genomic predictions with two and for traits for the ADFI and GROWTH models, respectively. Two UPG or MF were used to model the heterogeneous base across the lines. Herein, only the four-trait GROWTH model (ADG, BF, ADGX, and BFX) is described:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wc} + \mathbf{Zu} + \mathbf{ZQg} + \mathbf{e},$$

where **y** is the vector of phenotypes; **b** is a vector of fixed effects; **c** is a vector of random litter effects; **u** is a vector of random additive genetic effects; **g** is a vector of UPG; and **e** is a vector of residual effects. Matrix **X** is an incidence matrix relating phenotypes in vector **y** to fixed effects (contemporary group as a line-specific fixed effect for all traits, offtest weight and carcass weight as a covariate only for BF and BFX, respectively) in vector **b**, matrix **W** is an incidence matrix for random litter effects (line-specific) in vector **c**, matrix **Z** is an incidence matrix for random additive genetic effect in vector **u**, and matrix **Q** is an incidence matrix relating animals in vector **u** to UPG in vector **g**.

The genomic predictions were performed with ssGBLUP without and with weights from BayesR (WssGBLUP) using the BLUPF90 family of programs (Misztal et al., 2014b). In ssGBLUP and WssGBLUP, the inverse of the realized relationship matrix ($\mathbf{H}^{-1}$), which combines pedigree and genomic relationships is represented by (Aguilar et al., 2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{\mathbf{22}}^{-1} \end{bmatrix},$$

where $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix and $\mathbf{A}_{\mathbf{22}}^{-1}$ is the inverse of the pedigree relationship matrix for the genotyped individuals. The G was created with method 1 of VanRaden (2008) as:

$$\mathbf{G} = \frac{(\mathbf{M} - 2\mathbf{P})\mathbf{D}(\mathbf{M} - 2\mathbf{P})'}{2\sum_{i=1}^{m} p_i(1 - p_i)}$$

where **M** contains genotypes coded as {0,1,2}, **D** is a matrix of weights (**D** = **I** in ssGBLUP and **D** ≠ **I** in WssGBLUP), and **P** is a matrix whose columns contain observed allele frequencies across the entire data set of the second allele at a locus p_i. To avoid singularity issues, **G** was blended with 5% of $\mathbf{A}_{22}$. The GEBV for UPG models were calculated as:

$$\text{GEBV} = \mathbf{Qg} + \mathbf{u},$$

We investigated two ways to fit UPG in ssGBLUP. The first considered UPG in **A**, $\mathbf{A}_{22}$, and **G** (Misztal et al., 2013), and was called UPG1. The $\mathbf{H}^{-1}$ with UPG1 is described as follows:

$$\mathbf{H}^{*}_{\text{UPG1}} = \mathbf{A}^{*} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & -(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q} \\ 0 & -\mathbf{Q}'(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}) & \mathbf{Q}'(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q} \end{bmatrix}$$

where $\mathbf{A}^{*}$ is the inverse of **A** with UPG constructed with the QP transformation (Quaas, 1988). The second model related UPG only to **A** and $\mathbf{A}_{22}$, was called UPG2 and had $\mathbf{H}^{-1}$ represented by:

$$\mathbf{H}^{*}_{\text{UPG2}} = \mathbf{A}^{*} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & -\mathbf{A}_{22}^{-1}\mathbf{Q} \\ 0 & -\mathbf{Q}'\mathbf{A}_{22}^{-1} & \mathbf{Q}'\mathbf{A}_{22}^{-1}\mathbf{Q} \end{bmatrix}$$

Alternatively to the UPG models, we used MF to fit the heterogeneous genetic base across different lines (Legarra et al., 2015). Based on the MF theory, the pedigree relationship matrices are modified to be compatible with **G** centered with allele frequencies of 0.5 ($\mathbf{G}_{0.5}$) (Christensen, 2012; Legarra et al., 2015). $\mathbf{H}^{-1}$ with MF is described as follows:

$$\mathbf{H}^{\Gamma-1} = \mathbf{A}^{\Gamma-1} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}_{05}^{-1} - \mathbf{A}_{22}^{\Gamma-1} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

where $\mathbf{A}^{\Gamma-1}$ and $\mathbf{A}_{22}^{\Gamma-1}$ are the altered $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}^{-1}$ with the parameter $\boldsymbol{\Gamma}$, which is a matrix of relationships among MF. The $\boldsymbol{\Gamma}$ matrix was computed using SNP markers under a generalized least squares approach (Garcia-Baccino et al., 2017) through the gammaf90 program of the BLUPF90 software suite (Misztal et al., 2014b).

For UPG and MF models, six groups of base animals were defined based on the lines of origin. The first three were assigned to TL1, TL2, and TL3; one was assigned to another terminal line (TL4), one represented a crossbred line (CL), and the last one represented the remaining base animals with unknown origin (UNK). TL4 is another important terminal line in the routine evaluation and might be connected to the TL1 to TL3 in the base population, so we defined it as one group of base animals. Due to an issue of estimating $\boldsymbol{\Gamma}$ with all animals far back in the pedigree, animals born before 2000 were removed. Therefore, the total number of pedigreed animals was 5.16M for the GROWTH model and 5.04M for the ADFI model after data truncation. Due to the unequal number of phenotypes between the two models, the total number of pedigreed animals differed. The description of groups of UPG and MF is described in Table 4.4.

To efficiently compute the $\mathbf{G^{-1}}$ without the direct inversion of $\mathbf{G}$, the algorithm for proven and young (APY) (Misztal et al., 2014a) was applied to SLE and MLE. The number of core animals in each line corresponded to the number of largest eigenvalues explaining 98% of the total variation in $\mathbf{G}$ constructed by Chip (Pocrnic et al., 2016a). Therefore, the number of core animals was 3,996, 5,739, and 6,848 for TL1, TL2, and TL3, respectively. The number of core animals for MLE was selected as mentioned above but after combining all three terminal lines. Therefore, the number of core animals in MLE was 8,574. To fairly select the core animals from each line, we sampled 30% (2,572), 20% (1,715), and 50% (4,287) from TL1, TL2, and TL3, respectively. Those numbers were equivalent to the proportion of genotyped animals in each line for the MLE scenario.

In WssGBLUP, BayesR (Erbe et al., 2012) was used to estimate individual SNP variances, which were considered as weights. Each iteration stored individual SNP variances, and posterior SNP variance was calculated as the average variance across all the iterations. Afterward, the weights were re-scaled so that the trace of **D** was equal to the number of SNP. More details about BayesR weighting are described in Gualdrón-Duarte et al. (2020).

Validation

The LR validation method (Legarra and Reverter, 2018) was used to evaluate model performance. A total of 5,970 (TL1), 3,750 (TL2), and 11,308 (TL3) youngest genotyped animals in the test sets had their phenotypes removed from the evaluation. In the MLE scenario, the total number of records in the ADFI model was 1,476,644 (TL1), 1,478,431 (TL2), and 1,472,021 (TL3). For the GROWTH model, the number of records was 2,187,538, 2,189,326, and 2,182,916 for TL1, TL2, and TL3, respectively. These will be referred to as the reduced data and will be represented by the subscript r. On the other hand, the whole data, with no phenotype truncation, will be represented by the subscript w. Under the LR method, the accuracy of GEBV was calculated as $\widehat{acc} = \sqrt{\frac{cov(\hat{\boldsymbol{u}}_w,\ \hat{\boldsymbol{u}}_r)}{(1-\bar{F})\hat{\sigma}_u^2}}$, where u is the vector of GEBV and $\bar{F}$ is the average inbreeding coefficient for validation animals; $\hat{\sigma}_u^2$ was additive genetic variance in each model. Bias was calculated as the difference between the mean of GEBV from the reduced and whole datasets, which is $\mu_{w,r} = \bar{\hat{u}}_r - \bar{\hat{u}}_w$, with an expected estimator of 0 if unbiased. Dispersion of GEBV was assessed as the deviation of the regression coefficient (b_1) from 1, where b_1 was obtained from the regression of $\hat{u}_w$ on $\hat{u}_r$: $\hat{u}_w = b_0 + b_1\hat{u}_r$. Under the condition of neither over nor under dispersion, the expectation of this estimator would be one.

RESULTS

Population structure and Metafounders (Γ)

Principal component analysis (PCA) was performed to investigate the population structure of the three terminal lines. Chip data were used for 33,714 genotyped animals (TL1: 9,282, TL2: 7,900, and TL3: 16,532). Those were selected among 206,634 genotyped animals for efficient computation and had at least one progeny. The PCA plot showed a clear separation among the lines, with the first two principal components explaining 22.1% of the genetic variation (Fig. 4.1). These results reinforce the need to account for different genetic bases when having a multi-line evaluation.

The relationships within and between MF (**Γ**) are described in Table 4.4. Relationships within MF (diagonal values of **Γ**) were greater than one for TL1, TL2, and TL4, indicating that the base populations for those lines are inbred (Legarra et al., 2015). Contrary, TL3, CL, and UNK had values lower than one, indicating a high frequency of heterozygous compared to the population average. All the relationships between MF (off-diagonal values of Γ) showed positive values between 0 and 1, suggesting an overlap between ancestral populations.

Accuracy of GEBV

Fig. 4.2 shows the accuracy of predicting GEBV in SLE and MLE with UPG1, UPG2, and MF for five traits using Chip. The difference in prediction accuracy between SLE and MLE was up to 0.04. The results for TL1 are in Fig. 4.2 - (a). Only ADFI showed greater accuracy with MLE than SLE (0.56) for UPG1 (0.58), UPG2 (0.57), and MF (0.58). On the other hand, SLE had similar or better performance than MLE for the other four traits in the growth model. Among MLE scenarios, UPG1 outperformed UPG2 and MF for many traits, but the differences were minimal. Results for TL2 followed similar patterns (Fig. 4.2 - (b)); that is, all MLE scenarios outperformed

SLE for ADFI (0.63 vs. 0.61). Additionally, prediction accuracies of MLE scenarios were very similar for TL2. In general, TL3 reported greater prediction accuracies than TL1 and TL2; however, using MLE was not favorable. Only MLE with UPG1 outperformed SLE, which was for BFX (0.76 vs. 0.74). A comparison of MLE scenarios showed that UPG1 performed best for ADGP, BFP, ADGX, and BFX, with an accuracy gain of up to 0.03 (ADGP in TL3).

In the current study, five preselected genotype panels made from WGS were compared to the Chip for genomic prediction. As UPG1 showed the best prediction accuracy with Chip among all MLE scenarios, only results with UPG1 are in Table 4.5. In the results of TL1, no benefits of using preselected genotype panels were observed, meaning that Chip performed the best. Among preselected panels, ChipPlusSign showed the greatest prediction accuracy for all five traits. Top40k, TopSign, and AllComb had very similar prediction accuracies. However, LDTags displayed the lowest prediction accuracy among all genotype panels. Similar patterns were observed for TL2. In TL3, ChipPlusSign reported the greatest prediction accuracy only for ADFI (0.81). Likewise for TL1 and TL2, Top40k, TopSign, and AllComb showed very similar results to each other, but lower accuracy was noticed with AllComb. LDTags underperformed all the other genotype panels.

Bias of GEBV

Fig. 4.3 shows the bias of GEBV when using SLE and MLE with UPG1, UPG2, and MF for five traits in each line. Overall, bias was not considerable across all lines and models, except for ADGP in SLE and MLE with UPG1. In the TL1, the bias in ADGP with UPG1 was -2.91, whereas for ADGX in the same UPG scenario was -0.82. In both ADGP and ADGX, SLE showed the second largest bias, whereas UPG2 and MF reported almost negligible bias in both traits. For the other three traits (ADFI, BFP, and BFX), virtually no bias (0.05) was observed in all four

models. In TL2, almost no bias (0.06) was displayed for ADFI, ADGX, BFP, and BFX for all four models. ADGP showed the greatest bias among the traits, but the value was still small (-0.82 in SLE). Similar patterns were observed in the result of TL3 for ADFI, ADGX, BFP, and BFX (-0.46). A large bias was also reported in ADGP with a UPG1 (-2.10) and SLE (0.75).

Bias was also evaluated when the preselected genotype panels from WGS were used for genomic predictions. Results with UPG1 are described in Table 4.6. For TL1, only AllComb showed a smaller bias than Chip for ADGP, ADGX, and BFX. Except for those, Chip reported the smallest bias for other traits. Among the preselected panels, Top40k displayed the greatest bias in ADGP, BFP, ADGX, and BFX. Especially, for ADGP and ADGX, the bias was very large (-21.82 in ADGP and -8.09 in ADGX) with Top40k. ADFI trait showed no bias. Only one case of smaller bias than Chip was identified in TL2, which is for ADGX with AllComb (-0.05). The Chip data showed the smallest bias for the other four traits. Among the preselected panels, AllComb reported the least bias, whereas other panels indicated inconsistent results varying according to the traits. Interestingly, for TL3, both ChipPlusSign and AllComb reported a smaller bias than Chip for ADGP, BFP, ADGC, and BFX. In addition, Top40k and TopSign also showed a smaller bias than Chip for ADGX. Contrary to the other preselected genotype panels, bias with LDTags was always greater than with Chip.

Dispersion (inflation/deflation) of GEBV

The regression coefficients (b1) of GEBV whole on GEBV reduced when using SLE and MLE with UPG1, UPG2, and MF for five traits of each line are depicted in Fig. 4.4. Values of b1 greater than one indicate deflation of reduced GEBV, and smaller than one indicate inflation. In general, negligible deflation (1.01) and slight inflation (0.92) of GEBV were observed. The results of TL1 indicate relatively greater inflation with SLE for ADFI (0.95) and ADGP (0.92) compared

to other models with UPG and MF, although the differences are minimal. For the other three traits (ADGX, BFP, and BFX), all four models reported very similar b1 values (0.97 − 1.00). A similar pattern was identified for ADFI in TL2, which showed the greatest inflation of GEBV with SLE (0.93). For the other four traits, no considerable differences were found between the models (0.96 − 1.01). TL3 reported the most consistent results in each trait. For ADFI, all four models showed b1 values equal to 0.96.

To compare the same scenarios with preselected genotype panels from WGS as was done for prediction accuracy and bias, only results with UPG1 are described in Table 4.7. The dispersion of GEBV in TL1 with Top40k (0.94 − 1.00), TopSign (0.95 − 1.01), and ChipPlusSign (0.94 − 0.99) was very close to the result of Chip (0.94 − 1.00) for all five traits. However, LDTags and AllComb showed greater inflation of GEBV than other genotype panels. Among those two, LDTags indicated the greatest inflation of GEBV across all the traits (0.77 − 0.97). Only Top40k and TopSign reported better b1 values (close to 1) than in the Chip for some traits (Top40k – ADFI and ADGX; TopSign – ADFI and ADGP). Likewise, Top40k (0.96 − 1.00), TopSign (0.99 − 1.01), and ChipPlusSign (0.95 − 1.00) displayed similar results to Chip (0.97 − 1.00) in TL2, but a large inflation of GEBV with LDTags (0.78 − 0.90) and AllComb (0.88 − 0.96). Same patterns were observed in TL3. Remarkably, TopSign showed less inflation of GEBV than Chip for all five traits, although the differences are small (0.02).

Genomic prediction using WssGBLUP

Top40k and ChipPlusSign were used for WssGBLUP with BayesR weighting (BW) because those two panels showed the best performance among the preselected genotype panels. Additionally, BayesR weights were considered in ssGBLUP to make the results of this study more comparable to those in Ros-Freixedes et al. (2022) who used the same data under BayesR. Results

of prediction accuracy using BW for all five traits are summarized in Table 4.8. No benefits of using WssGBLUP over ssGBLUP were observed in Top40k and ChipPlusSign. In fact, the maximum accuracy gains of WssGBLUP compared to ssGBLUP was 0.01; however, a loss in accuracy of up to 0.04 was observed in several traits and lines.

DISCUSSION

In this study, we investigated the impact of using UPG and MF when combining different pig populations in multi-line evaluations. Additionally, we explored the potential benefits of using preselected SNP from WGS in those joint evaluations when SNP received equal or different weights in ssGBLUP. The novel aspect of this study is the amount of sequence data used for multi-line genomic predictions (i.e., over 200k pigs). This study brought insights into how accounting for different genetic bases in three large pig populations could affect the performance of joint genomic evaluations. It also proved that the forecasting regarding the usefulness of sequence data for across-breed predictions does not hold (Meuwissen et al., 2016), at least with the current methods. Overall, we will address three major topics in this discussion: (1) the impact of fitting UPG or MF in MLE, (2) the usefulness of WGS data for MLE, and (3) the impact of applying different weights to SNP selected from WGS in MLE. In a nutshell, two UPG and one MF model were considered in MLE, and performances of genomic predictions (accuracy, bias, and dispersion) were compared to SLE. Although the results varied depending on the lines and the traits, the maximum changes in prediction accuracy when moving from SLE to MLE were not that large (0.04). In the line with the smallest number of genotyped individuals (TL2), all three MLE scenarios performed very similarly. Most of the differences among scenarios were in the most extensive line (TL3). Regarding the use of WGS data, almost no benefits were observed when

preselected genotype panels were used for genomic prediction compared to Chip. This was true even when different weights were assigned to SNP.

Multi-line genomic evaluation with UPGs and MFs

Combining populations with different genetic backgrounds in genomic evaluations has been actively investigated in cattle (De Roos et al., 2009; Hayes et al., 2009; Olson et al., 2012; Cesarani et al., 2022), where the primary purpose is to increase the training size for small populations to improve the accuracy of genomic predictions. This is true if there are connections across populations and the training and validation sets are related (Meuwissen et al., 2001; De Roos et al., 2009; Hayes et al., 2009; Zhou et al., 2014). However, combining different pig populations may be challenging even if the lines belong to the same breeding company because the divergence may have happened a long time ago, and breeding objectives are different across lines. For these reasons, only a few studies have investigated combining multiple lines, populations, or breeds for genomic predictions in pigs (Fangmann et al., 2015; Aliakbari et al., 2020). In our study, although the PCA showed a clear separation among the three lines, representing three different breeds, as TL2 is the line with the least number of genotyped animals and shows a close distance to TL3 based on the PCA (the largest number), we would expect some improvements for TL2 in a joint evaluation. However, only ADFI in TL2 benefitted from MLE instead of SLE, and the increase in accuracy was slight. Additionally, the performance of MF, UPG1, and UPG2 were similar.

In the MF theory, a matrix of relationships within and across metafounders (**Γ**) is used to make pedigree relationships compatible with genomic relationships. We observed relationships within MF lower than one for TL3, CL, and UNK but greater than one for TL1, TL2, and TL4. Values of **Γ** smaller than one indicate a base population with broad genetic diversity with a higher

frequency of heterozygotes relative to the population average (Kluska et al., 2021). On the other hand, a value greater than one indicates inbreed and related base populations with a lower frequency of heterozygotes relative to the population average (Legarra et al., 2015). Besides, positive **Γ** values between MF imply overlapping among individuals in the base populations (Kluska et al., 2021). Our results showed only positive **Γ** values between MF, meaning that there was overlap between ancestors in their base populations. Xiang et al. (2017) reported similar results using pig data. Those authors calculated **Γ** values between two MF, which were defined as Landrace and Yorkshire, showing a positive value (0.259). Therefore, we could speculate that although the lines in our study diverged from different breeds, they share ancestors in the base population. In addition, **Γ** value smaller than one for MF assigned CL is explained by the fact that this line is crossbred, indicating that a large amount of genetic variability existed in the base population compared to the purebred lines. Conversely, TL1, TL2, and TL4 represent single breeds. Therefore, **Γ** values greater than one for MF assigned to TL1, TL2, and TL4 agree with their historical development. However, a study by Xiang et al. (2017) reported 0.756 and 0.730 **Γ** values for Landrace and Yorkshire, respectively, although they were pure breeds. The possible reasons for different **Γ** values in purebred between the current study and the study by Xiang et al. (2017) could be the use of terminal breeds and maternal breeds as well as the use of different SNP data and lines from different companies.

Fangmann et al. (2015) investigated using multi-subpopulation reference sets to improve the predictive ability of genomic predictions in pigs; however, almost no benefit was reported, even though all the subpopulations diverged from German Large White. Predictive abilities were reduced when distantly related subpopulations were added to the training data. Although our results agree with those in Fangmann et al. (2015), comparing the two studies might be unfair

because of the different data sizes and genomic evaluation models. For example, Fangmann et al. (2015) used only 2,053 animals with genotypes and phenotypes under GBLUP. We used ssGBLUP with over 5M animals, of which 206,634 were genotyped, and 140k − 1.6M were phenotyped depending on the trait. Most recently, Cesarani et al. [1] performed large-scale multibreed ssGBLUP in dairy cattle using five different breeds with 4M genotyped animals and 29.5M pedigree records. They reported similar predictive abilities for cows and reliabilities for bulls in single-breed and multibreed evaluations, even though some breeds had less than 50k genotyped animals and some had more than 500k. This was attributed to the use of UPG (i.e., UPG2) to model genetic differences across breeds, the inclusion of breed-specific fixed effects in the model, and a fair representation of all the breeds in the APY core. In our study, line-specific fixed effects were modeled to account for nongenetic differences among lines, the APY core properly represented the three lines, and UPG or MF were fit to account for the genetic differences among lines.

A preliminary analysis was done to compare the performance of genomic predictions in MLE with and without UPG. Most of the traits reported better accuracy, less bias, and less dispersion with UPG1 and UPG2 compared to the MLE without UPG (results not shown). The major difference between UPG1 and UPG2 was that groups were assigned to **A**, **G**, and $\mathbf{A}_{22}$ in UPG1, but only to **A** and $\mathbf{A}_{22}$ in UPG2. Theoretically, UPG are not needed in **G** because genomic relationships are not affected by missing pedigrees. However, adding UPG to **G** (UPG1) in multi-line evaluations could help to adjust the genetic base for each line (Tsuruta et al., 2019). However, UPG1 and UPG2 assume that the base populations are not related. Therefore, MF were also applied to MLE, which considered that individuals in the base populations were related and could be inbred (Legarra et al., 2015). Among the three methods used for MLE, UPG1 had a slight advantage, although the differences between MLE and SLE models were minimal. Our findings agreed with

the results by Fangmann et al. (2015); Song et al. (2017); Aliakbari et al. (2020), which showed that combining lines or breeds had almost no benefits in the performance of genomic predictions compared to the within-line predictions. However, several studies of cattle found some benefits of using the multi-breed reference on genomic prediction (Hayes et al., 2009; Lund et al., 2014). More benefits are likely in populations with a small number of genotyped animals, which was not the case in our study. Although MLE were not advantageous for these pig populations in terms of increased accuracy, having MLE facilitates comparing animals across breeds because of a single base for breeding values. In such a case, having similar accuracy, bias, and dispersion as in SLE is somehow an advantage. It indicates the lines can be successfully combined to identify "super-boars" to be used across lines if needed.

Impact of preselected markers on MLE

In addition to the size of the reference population and the relationships between the reference and test animals, another key factor affecting the prediction accuracy would be the existence of causal variants or informative SNP in LD with them. This factor could be particularly important for MLE. This is because the LD between SNP and causal variants within lines may not be consistent across lines, especially for the distantly related ones, and causative variants for all the lines may not be present in the commercial SNP chips. One plausible approach to help improve genomic predictions in MLE could be using WGS. As WGS data covers the entire genome, it has abundant LD information and possibly harbors all the causal variants. This could increase the power of identifying LD structures and causal variants across lines. However, several studies showed no benefits of using WGS data on genomic prediction without variant preselection (Van Binsbergen et al., 2015; Zhang et al., 2018) because of the redundancy of SNP in WGS. As WGS

has highly dense SNP information compared to the regular chip data, SNP close to each other would be strongly correlated, providing the same information about nearby QTL. Several studies investigated the impact of SNP preselection using WGS data on genomic predictions, showing slight to no improvement (Brøndum et al., 2015; VanRaden et al., 2017; Fragomeni et al., 2019; Jang et al., 2022). In addition, several studies in cattle scrutinized the impact of WGS in multibreed or across-breed genomic predictions (Van Den Berg et al., 2017; Raymond et al., 2018; Meuwissen et al., 2021), but not much is available in pigs (Song et al., 2019).

Our study used multi-line GWAS or LD pruning to construct preselected genotype panels for different pig terminal lines. Consequently, five genotype sets were designed: ChipPlusSign, Top40k, TopSign, LDTags, and AllComb. Those sets were used for MLE to compare the performance of genomic predictions with Chip. Among all the scenarios, only ChipPlusSign reported greater accuracy than Chip, and this was for one trait (ADFI) only in TL3. On average, ChipPlusSign showed the greatest prediction accuracy among all preselected scenarios but was still smaller than Chip, although the difference was slight. For the multi-line GWAS, we used seven lines (TL1, TL2, TL3, TL4, and three maternal lines) as described in (Ros-Freixedes et al., 2022). However, we used only TL1 to TL3 because of the amount of data, completeness of pedigree, and different traits being measured in the terminal and maternal lines. Compared to single-line GWAS, multi-line GWAS could help identify significant SNP affecting the traits as it makes a long-distance LD short, allowing more accurate identification of significant SNP across the lines (Moghaddar et al., 2019). This can be especially helpful for species with small *Ne*, such as pigs and chickens with small *Me* and a strong LD extent, which makes the identification of causative variants more difficult (Jang et al., 2022). Several studies in dairy cattle reported up to a 7% increase in reliability when using variants selected from multi-breed GWAS for genomic

predictions (Brøndum et al., 2015; van den Berg et al., 2016); however, they used methods other than ssGBLUP. Using ssGBLUP, Fragomeni et al. (2019) reported no benefits of using preselected WGS variants but larger reliabilities than in GBLUP. This is because ssGBLUP allows for more data than in multi-step methods. With enough data, the effects of existing variants are well-captured, and chromosome segments are correctly estimated.

Adding significant SNP to the regular chip data or using only them could potentially improve prediction accuracies only if those are real causative variants with known effects, positions, and variance explained (Fragomeni et al., 2017; Jang et al., 2022). To accurately identify the significant ones, there should be a sufficient sample size, enough data connected to the genotyped samples, and a robust statistical model, among others. Jang et al. (2022) extensively investigated the impact of data quantity on the variant selection using simulated WGS and its effect on genomic prediction with populations having different *Ne*. They showed that identifying significant quantitative trait nucleotides (QTN) is more difficult in populations with smaller than larger *Ne* because of the strong LD extent across the genome in the former. Accordingly, improvements in the accuracy of genomic predictions using those selected QTN combined with a 50k SNP chip in the population with smaller *Ne* were very limited (~1.98%) compared to the population that had larger *Ne* (~9.01%). Therefore, although multi-line GWAS could make the long-distance LD to be shorter across lines, the benefits would be still limited when it comes to genomic prediction.

Impact of WssGBLUP on MLE

In the current study, we used WssGBLUP with weights computed from BayesR. We did not use the standard weights proposed by Wang et al. (2012) because several studies reported no

improvement in the accuracy and increased inflation of genomic predictions when using those weights (Wang et al., 2012; Zhang et al., 2016; Gualdrón-Duarte et al., 2020; Jang et al., 2022). Additionally, we wanted to make our results comparable to those in Ros-Freixedes et al. (2022) who applied BayesR to the same dataset. In BayesR, SNP effects are sampled from a mixture of four normal distributions with mean zero and variances equivalent to the four classes. Thus, we assumed that better prediction performance could be observed with BW if it assigned weights closer to the actual SNP variances; however, no advantages were observed. More details about BayesR and using its weights in GBLUP-based methods are in Moser et al. (2015); Gualdrón-Duarte et al. (2020).

This is the first study in large-scale pig lines using MLE with WGS selected variants using the WssGBLUP approach. In practice, the benefits of using WssGBLUP seemed very limited, especially with large data sets and many genotyped animals (Lourenco et al., 2017). Our MLE scenario used around 206k genotyped animals, and the total number of animals traced back through the combined pedigree data was more than 5M. In ssGBLUP, any prior information about SNP is overwhelmed by the data because this method allows the use of all sources of information simultaneously, making SNP weighting ineffective (Lourenco et al., 2020).

CONCLUSIONS

This study revealed limited benefits of using preselected whole-genome sequence variants for multi-line genomic predictions, even when hundreds of thousands of animals had imputed sequence data. Correctly accounting for line differences with unknown parent groups or metafounders in multi-line evaluations is essential to obtain predictions similar to single-line; however, the only observed benefit of a multi-line evaluation is to have comparable predictions

across lines. Further investigation on the amount of data and novel methods to preselect whole-genome causative variants in combined populations would be of great interest.

REFERENCES

Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta, and T. Lawlor. 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.

Aliakbari, A., E. Delpuech, Y. Labrune, J. Riquet, and H. Gilbert. 2020. The impact of training on data from genetically-related lines on the accuracy of genomic predictions for feed efficiency traits in pigs. Genetics Selection Evolution 52(1):1-15.

Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 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.

Calus, M. P., H. Huang, A. Vereijken, J. Visscher, J. Ten Napel, and J. J. Windig. 2014. Genomic prediction based on data from three layer lines: a comparison between linear methods. Genetics Selection Evolution 46(1):1-13.

Cesarani, A., D. Lourenco, S. Tsuruta, A. Legarra, E. Nicolazzi, P. VanRaden, and I. Misztal. 2022. Multibreed genomic evaluation for production traits of dairy cattle in the United States using single-step genomic best linear unbiased predictor. Journal of Dairy Science

Chen, C.-Y., I. Misztal, I. Aguilar, A. Legarra, and W. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. Journal of animal science 89(9):2673-2679.

Christensen, O. F. 2012. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genetics Selection Evolution 44(1):1-10.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genetics Selection Evolution 42(1):1-8.

De Roos, A., B. Hayes, and M. J. G. Goddard. 2009. Reliability of genomic predictions across multiple populations. 183(4):1545-1553.

Erbe, M., B. Hayes, L. Matukumalli, S. Goswami, P. Bowman, C. Reich, B. Mason, and M. 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):4114-4129.

Fangmann, A., S. Bergfelder-Drüing, E. Tholen, H. Simianer, and M. Erbe. 2015. Can multi-subpopulation reference sets improve the genomic predictive ability for pigs? Journal of Animal Science 93(12):5618-5630.

Fragomeni, B., D. Lourenco, A. Legarra, P. VanRaden, and I. Misztal. 2019. Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants. Journal of dairy science 102(11):10012-10019.

Fragomeni, B. O., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2017. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. Genetics Selection Evolution 49(1):59.

Fu, C., T. Ostersen, O. F. Christensen, and T. Xiang. 2021. Single-step genomic evaluation with metafounders for feed conversion ratio and average daily gain in Danish Landrace and Yorkshire pigs. Genetics Selection Evolution 53(1):1-11.

Garcia-Baccino, C. A., A. Legarra, O. F. Christensen, I. Misztal, I. Pocrnic, Z. G. Vitezica, and R. J. Cantet. 2017. Metafounders are related to F st fixation indices and reduce bias in single-

step genomic evaluations. Genetics Selection Evolution 49(1):1-14.

Gualdrón-Duarte, J. L., A.-S. Gori, X. Hubin, D. Lourenco, C. Charlier, I. Misztal, and T. Druet. 2020. Performances of Adaptive MultiBLUP, Bayesian regressions, and weighted-GBLUP approaches for genomic predictions in Belgian Blue beef cattle. BMC genomics 21(1):1-18.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41(1):1-9.

Jang, S., S. Tsuruta, N. G. Leite, I. Misztal, and D. Lourenco. 2022. Dimensionality of genomic information and its impact on GWA and variant selection: a simulation study. bioRxiv

Kluska, S., Y. Masuda, J. B. S. Ferraz, S. Tsuruta, J. P. Eler, F. Baldi, and D. Lourenco. 2021. Metafounders May Reduce Bias in Composite Cattle Genomic Predictions. Frontiers in Genetics:1440.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. Journal of dairy science 92(9):4656-4663.

Legarra, A., J. Bertrand, T. Strabel, R. Sapp, J. Sanchez, and I. Misztal. 2007. Multi-breed genetic evaluation in a Gelbvieh population. Journal of Animal Breeding and Genetics 124(5):286-295.

Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal. 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. Genetics 200(2):455-468.

Legarra, A., and A. Reverter. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Genetics

Selection Evolution 50(1):53.

Lourenco, D., B. Fragomeni, H. Bradford, I. Menezes, J. Ferraz, I. Aguilar, S. Tsuruta, and I. Misztal. 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. Journal of Animal Breeding and Genetics 134(6):463-471.

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

Lund, M. S., G. Su, L. Janss, B. Guldbrandtsen, and R. F. Brøndum. 2014. Genomic evaluation of cattle in a multi-breed context. Livestock Science 166:101-110.

Macedo, F. L., O. F. Christensen, J.-M. Astruc, I. Aguilar, Y. Masuda, and A. Legarra. 2020. Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. Genetics Selection Evolution 52(1):1-10.

Meuwissen, T., B. Hayes, and M. Goddard. 2016. Genomic selection: A paradigm shift in animal breeding. Animal frontiers 6(1):6-14.

Meuwissen, T., I. van den Berg, and M. 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):1-15.

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

Misztal, I., A. Legarra, and I. Aguilar. 2014a. Using recursion to compute the inverse of the genomic relationship matrix. Journal of dairy science 97(6):3943-3952.

Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, and Z. Vitezica. 2014b. Manual for

BLUPF90 family of programs. Athens: University of Georgia

Misztal, I., Z.-G. Vitezica, A. Legarra, I. Aguilar, and A. Swan. 2013. Unknown-parent groups in single-step genomic evaluation. Journal of Animal Breeding and Genetics 130(4):252-258.

Moghaddar, N., M. Khansefid, J. H. van der Werf, S. Bolormaa, N. Duijvesteijn, S. A. Clark, A. A. Swan, H. D. Daetwyler, and I. M. MacLeod. 2019. Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. Genetics Selection Evolution 51(1):72.

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

Olson, K., P. VanRaden, and M. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. Journal of Dairy Science 95(9):5378-5383.

Pocrnic, I., D. A. Lourenco, C.-Y. Chen, W. O. Herring, and I. Misztal. 2019. Crossbred evaluations using single-step genomic BLUP and algorithm for proven and young with different sources of data. Journal of animal science 97(4):1513-1522.

Pocrnic, I., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016a. The dimensionality of genomic information and its effect on genomic prediction. Genetics 203(1):573-581.

Pocrnic, I., D. A. Lourenco, Y. Masuda, and I. Misztal. 2016b. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. Genetics Selection Evolution 48(1):1-9.

Quaas, R. 1988. Additive genetic model with groups and relationships. Journal of Dairy Science 71(5):1338-1345.

Raymond, B., A. C. Bouwman, C. Schrooten, J. Houwing-Duistermaat, and R. F. Veerkamp. 2018.

Utility of whole-genome sequence data for across-breed genomic prediction. Genetics Selection Evolution 50(1):1-12.

Rolf, M. M., D. J. Garrick, T. Fountain, H. R. Ramey, R. L. Weaber, J. E. Decker, E. J. Pollak, R. D. Schnabel, and J. F. Taylor. 2015. Comparison of Bayesian models to estimate direct genomic values in multi-breed commercial beef cattle. Genetics Selection Evolution 47(1):1-14.

Ros-Freixedes, R., M. Johnsson, A. Whalen, C.-Y. Chen, B. D. Valente, W. O. Herring, G. Gorjanc, and J. M. Hickey. 2022. Genomic prediction with whole-genome sequence data in intensely selected pig lines. bioRxiv:2022.2002.2002.478838. doi: 10.1101/2022.02.02.478838

Ros-Freixedes, R., A. Whalen, C.-Y. Chen, G. Gorjanc, W. O. Herring, A. J. Mileham, and J. M. Hickey. 2020. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. Genetics Selection Evolution 52:1-15.

Song, H., S. Ye, Y. Jiang, Z. Zhang, Q. Zhang, and X. Ding. 2019. Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. Genetics Selection Evolution 51(1):58.

Song, H., J. Zhang, Y. Jiang, H. Gao, S. Tang, S. Mi, F. Yu, Q. Meng, W. Xiao, and Q. Zhang. 2017. Genomic prediction for growth and reproduction traits in pig using an admixed reference population. Journal of Animal Science 95(8):3415-3424.

Tsuruta, S., D. Lourenco, Y. Masuda, I. Misztal, and T. Lawlor. 2019. Controlling bias in genomic breeding values for young genotyped bulls. Journal of dairy science 102(11):9956-9970.

Van Binsbergen, R., M. P. Calus, M. C. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 47(1):71.

van den Berg, I., D. Boichard, and M. S. Lund. 2016. Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genetics Selection Evolution 48(1):1-18.

Van Den Berg, I., P. J. Bowman, I. M. MacLeod, B. J. Hayes, T. Wang, S. Bolormaa, and M. E. Goddard. 2017. Multi-breed genomic prediction using Bayes R with sequence data and dropping variants with a small effect. Genetics Selection Evolution 49(1):70.

van Grevenhof, E. M., J. Vandenplas, and M. P. Calus. 2019. Genomic prediction for crossbred performance using metafounders. Journal of animal science 97(2):548-558.

VanRaden, P. M., M. E. Tooker, J. R. O'connell, J. B. Cole, and D. M. Bickhart. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. Genetics Selection Evolution 49(1):32.

VanRaden, P. M. J. J. o. d. s. 2008. Efficient methods to compute genomic predictions. 91(11):4414-4423.

Veerkamp, R. F., A. C. Bouwman, C. Schrooten, and M. P. Calus. 2016. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. Genetics Selection Evolution 48(1):95.

Vitezica, Z.-G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genetics Research 93(5):357-366.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics Research 94(2):73-83.

Whalen, A., R. Ros-Freixedes, D. L. Wilson, G. Gorjanc, and J. M. Hickey. 2018. Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees. Genetics Selection Evolution 50(1):1-15.

Xiang, T., O. Christensen, and A. Legarra. 2017. Genomic evaluation for crossbred performance in a single-step approach with metafounders. Journal of animal science 95(4):1472-1480.

Zhang, C., R. A. Kemp, P. Stothard, Z. Wang, N. Boddicker, K. Krivushin, J. Dekkers, and G. Plastow. 2018. Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. Genetics Selection Evolution 50(1):1-13.

Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal. 2016. Weighting strategies for single-step genomic BLUP: an iterative approach for accurate calculation of GEBV and GWAS. Frontiers in genetics 7:151.

Zhou, L., M. 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):249-257.

TABLES

Table 4.1. Number of records and animals in the pedigree for single- and multi-line datasets

Lines	ADFI	ADG	BF	ADGX	BFX	Number of Animals in pedigree
TL1	35k	0.36M	0.34M	150k	149k	1.13M
TL2	40k	0.30M	0.30M	158k	156k	0.84M
TL3	64k	0.94M	0.86M	299k	247k	3.14M
MLE	140k	1.60M	1.50M	578k	525k	5.28M (5.17M)

*The number of animals in the pedigree for the ADFI model is shown within brackets

*ADFI: Average daily feed intake; ADG: Average daily gain; BF: Backfat thickness; ADGX: ADG recorded in crossbred; BFX: BF recorded in crossbred

*TL1: Terminal line1; TL2: Terminal line2; TL3: Terminal line3; MLE: Multi-line evaluation

Table 4.2. Number of genotyped individuals, SNP, and sequenced animals in single- and multi-line datasets

Lines	Number of genotyped individuals*	Number of SNP	Number of sequenced individuals
TL1	60,450	37,909	731
TL2	41,561	42,897	760
TL3	104,622	44,022	1,865
MLE	206,634	41,303	3,356

*All genotyped individuals were imputed to sequence using the sequence individuals as reference
*TL1: Terminal line1; TL2: Terminal line2; TL3: Terminal line3; MLE: Multi-line evaluation

Table 4.3. Number of animals and SNP in all preselected genotype panels used for multi-line evaluations

	Number of genotyped animals		Number of SNP	
SNP panels	ADFI	GROWTH	ADFI	GROWTH
Chip	206,634	206,634	41,303	41,303
ChipPlusSign	206,452	206,630	62,906	59,756
LDTags	202,891	202,891	105,720	105,720
AllComb	205,729	205,680	215,361	210,619
Top40k	206,232	206,238	51,297	49,738
TopSign	206,228	206,228	21,772	18,593

*ADFI: Two-traits ADFI model (ADFI and ADG)

*GROWTH: Four-traits GROWTH model (ADG, BF, ADGX, and BFX)

*Chip: Imputed chip data; ChipPlusSign: Preselected SNP panel combining TopSign to Chip; LDTags: Preselected SNP panel after LD pruning; AllComb: Preselected SNP panel combining Chip, LDTags, Top40k, and TopSign; Top40k: Preselected SNP panel consisted of the variants with the lowest p-value in each 40k window; TopSign: Preselected SNP panel consisted of only significant variants

Table 4.4. Number of individuals that related to each unknown parent groups or metafounders and Γ values

			Γ					
MFs	Males	Females	TL1	TL2	TL3	TL4	CL	Unknown
TL1	3,649	2,713	1.03	0.45	0.5	0.67	0.46	0.47
TL2	2,083	1,949	0.45	1.26	0.44	0.39	0.42	0.5
TL3	7,148	6,135	0.5	0.44	0.62	0.43	0.38	0.43
TL4	30,141	29,707	0.67	0.39	0.43	1.06	0.42	0.41
CL	37,150	41,741	0.46	0.42	0.38	0.42	0.58	0.41
Unknown	158,929 (45,031)	158,929 (45,031)	0.47	0.5	0.43	0.41	0.41	0.50

*The number of animals in the pedigree for the ADFI model is shown within brackets

Table 4.5. Accuracy of GEBV in each line with all preselected genotype panels using unknown parent groups1

		Genotype panels					
Lines	Traits	Chip	Top40k	TopSign	ChipPlusSign	LDTags	AllComb
TL1	ADFI	0.58	0.55	0.54	0.57	0.47	0.54
	ADGP	0.62	0.58	0.58	0.61	0.48	0.57
	BFP	0.71	0.67	0.67	0.70	0.60	0.68
	ADGX	0.53	0.53	0.53	0.53	0.45	0.50
	BFX	0.83	0.74	0.75	0.81	0.66	0.75
TL2	ADFI	0.63	0.55	0.55	0.57	0.44	0.52
	ADGP	0.69	0.63	0.64	0.67	0.53	0.61
	BFP	0.64	0.60	0.60	0.63	0.51	0.60
	ADGX	0.53	0.48	0.49	0.51	0.39	0.44
	BFX	0.58	0.52	0.53	0.56	0.41	0.49
TL3	ADFI	0.79	0.79	0.79	0.81	0.72	0.78
	ADGP	0.81	0.77	0.79	0.81	0.69	0.78
	BFP	0.75	0.71	0.73	0.75	0.61	0.71
	ADGX	0.62	0.56	0.58	0.60	0.46	0.53
	BFX	0.76	0.71	0.73	0.74	0.57	0.68

*TL1: Terminal line1; TL2: Terminal line2; TL3: Terminal line3

*ADFI: Average daily feed intake; ADG: Average daily gain; BF: Backfat thickness; ADGX: ADG recorded in crossbred; BFX: BF recorded in crossbred

*Chip: Imputed chip data; Top40k: Preselected SNP panel consisted of the variants with the lowest p-value in each 40k window; TopSign: Preselected SNP panel consisted of only significant variants; ChipPlusSign: Preselected SNP panel combining TopSign to Chip; LDTags: Preselected SNP panel after LD pruning; AllComb: Preselected SNP panel combining Chip, LDTags, Top40k, and TopSign

Table 4.6. Bias of GEBV with preselected genotype panels when assigning unknown parent groups1

		Genotype panels					
Lines	Traits	Chip	Top40k	TopSign	ChipPlusSign	LDTags	AllComb
TL1	ADFI	0.00	0.00	-0.01	-0.01	-0.01	-0.01
	ADGP	-2.91	-21.82	-8.93	-3.13	-5.54	-1.08
	BFP	0.02	0.34	0.15	0.03	0.17	0.06
	ADGX	-0.82	-8.09	-3.25	-0.92	-2.31	-0.29
	BFX	0.05	0.67	0.28	0.05	0.31	0.00
TL2	ADFI	0.00	-0.01	-0.08	0.00	0.00	0.00
	ADGP	-0.25	6.42	3.42	0.82	1.32	-0.41
	BFP	0.00	0.11	0.07	-0.03	0.14	0.00
	ADGX	-0.18	-0.37	-0.64	-0.92	0.74	-0.05
	BFX	0.00	0.06	0.06	-0.05	0.19	0.01
TL3	ADFI	0.00	0.00	0.00	0.00	0.00	0.00
	ADGP	-2.10	7.86	6.36	0.08	4.39	0.00
	BFP	0.04	0.08	0.07	0.00	0.11	0.01
	ADGX	-0.46	0.26	0.09	-0.03	-0.60	0.03
	BFX	0.04	0.05	0.05	0.02	0.11	0.02

*TL1: Terminal line1; TL2: Terminal line2; TL3: Terminal line3

*ADFI: Average daily feed intake; ADG: Average daily gain; BF: Backfat thickness; ADGX: ADG recorded in crossbred; BFX: BF recorded in crossbred

*Chip: Imputed chip data; Top40k: Preselected SNP panel consisted of the variants with the lowest p-value in each 40k window; TopSign: Preselected SNP panel consisted of only significant variants; ChipPlusSign: Preselected SNP panel combining TopSign to Chip; LDTags: Preselected SNP panel after LD pruning; AllComb: Preselected SNP panel combining Chip, LDTags, Top40k, and TopSign

Table 4.7. Dispersion (b1) of GEBV with preselected genotype panels when assigning unknown parent groups1

		Genotype panels					
Lines	Traits	Chip	Top40k	TopSign	ChipPlusSign	LDTags	AllComb
TL1	ADFI	0.98	1.00	1.01	0.98	0.84	0.92
	ADGP	0.94	0.94	0.95	0.94	0.77	0.88
	BFP	0.99	0.99	0.99	0.99	0.91	0.96
	ADGX	0.99	1.00	1.01	0.98	0.91	0.94
	BFX	1.00	1.00	1.01	0.99	0.97	0.98
TL2	ADFI	0.97	0.97	1.00	0.95	0.78	0.88
	ADGP	1.00	0.96	0.99	0.97	0.80	0.89
	BFP	1.00	1.00	1.01	1.00	0.89	0.96
	ADGX	1.00	0.96	0.99	0.99	0.86	0.92
	BFX	1.00	1.00	1.01	1.00	0.90	0.96
TL3	ADFI	0.96	0.95	0.97	0.96	0.85	0.91
	ADGP	0.95	0.94	0.96	0.95	0.81	0.89
	BFP	0.98	0.97	0.99	0.98	0.85	0.93
	ADGX	0.98	0.99	1.00	0.97	0.85	0.92
	BFX	0.99	0.98	1.00	0.99	0.90	0.96

*TL1: Terminal line1; TL2: Terminal line2; TL3: Terminal line3

*ADFI: Average daily feed intake; ADG: Average daily gain; BF: Backfat thickness; ADGX: ADG recorded in crossbred; BFX: BF recorded in crossbred

*Chip: Imputed chip data; Top40k: Preselected SNP panel consisted of the variants with the lowest p-value in each 40k window; TopSign: Preselected SNP panel consisted of only significant variants; ChipPlusSign: Preselected SNP panel combining TopSign to Chip; LDTags: Preselected SNP panel after LD pruning; AllComb: Preselected SNP panel combining Chip, LDTags, Top40k, and TopSign

Table 4.8. Accuracy of GEBV using WssGBLUP through BayesR-weighting

Line	Description	ADFI	ADG	BF	ADGX	BFX
TL1	Top40k	0.55	0.58	0.67	0.53	0.74
	Top40k_BW	0.54	0.57	0.67	0.53	0.74
	ChipPlusSign	0.57	0.61	0.70	0.53	0.81
	ChipPlusSign_BW	0.54	0.60	0.69	0.50	0.79
TL2	Top40k	0.55	0.63	0.60	0.48	0.52
	Top40k_BW	0.54	0.62	0.60	0.49	0.52
	ChipPlusSign	0.57	0.67	0.63	0.51	0.56
	ChipPlusSign_BW	0.53	0.65	0.61	0.47	0.54
TL3	Top40k	0.79	0.77	0.71	0.56	0.70
	Top40k_BW	0.78	0.76	0.71	0.57	0.70
	ChipPlusSign	0.81	0.81	0.75	0.60	0.74
	ChipPlusSign_BW	0.77	0.78	0.72	0.56	0.71

*TL1: Terminal line1; TL2: Terminal line2; TL3: Terminal line3; MLE: Multi-line evaluation
*ADFI: Average daily feed intake; ADG: Average daily gain; BF: Backfat thickness; ADGX: ADG recorded in crossbred; BFX: BF recorded in crossbred
*Top40k: Top40k preselected genotype panel; Top40k_BW: Top40k using BayesR weighting
*ChipPlusSign: ChipPlusSign preselected genotype panel; ChipPlusSign_BW: ChipPlusSign using BayesR weighting

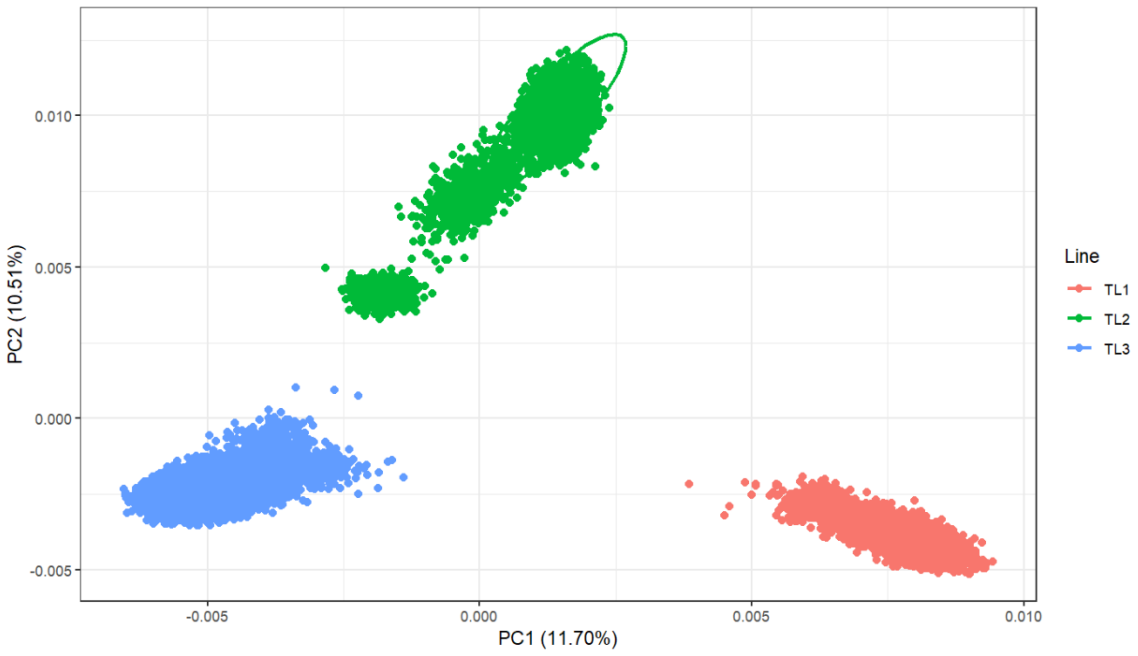
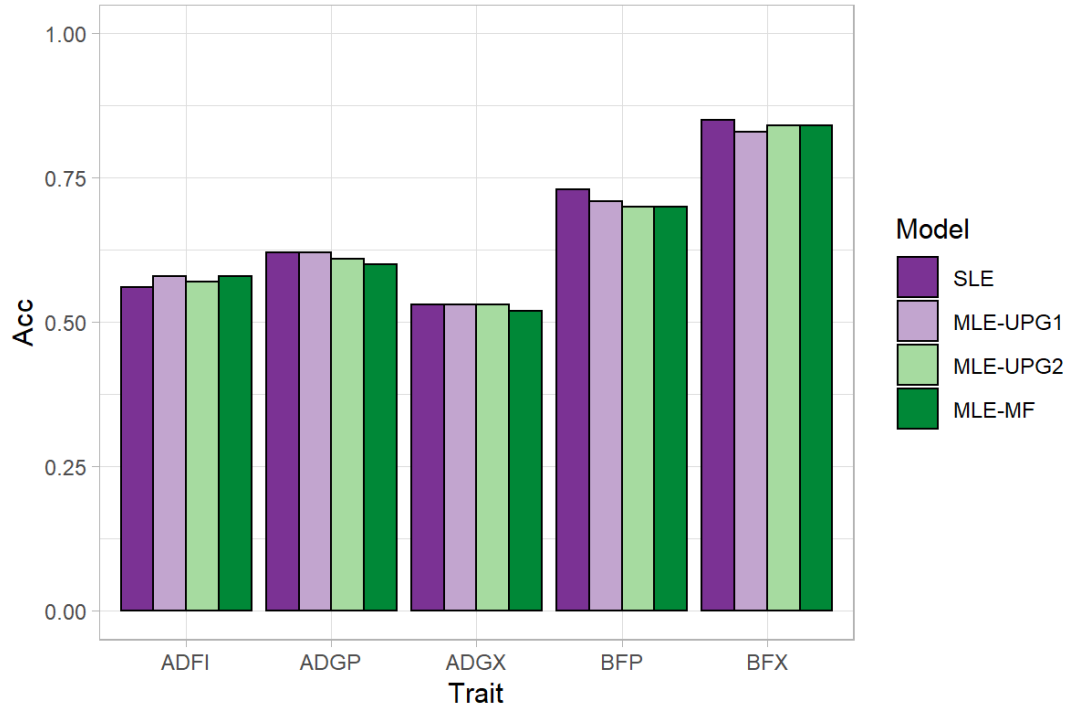
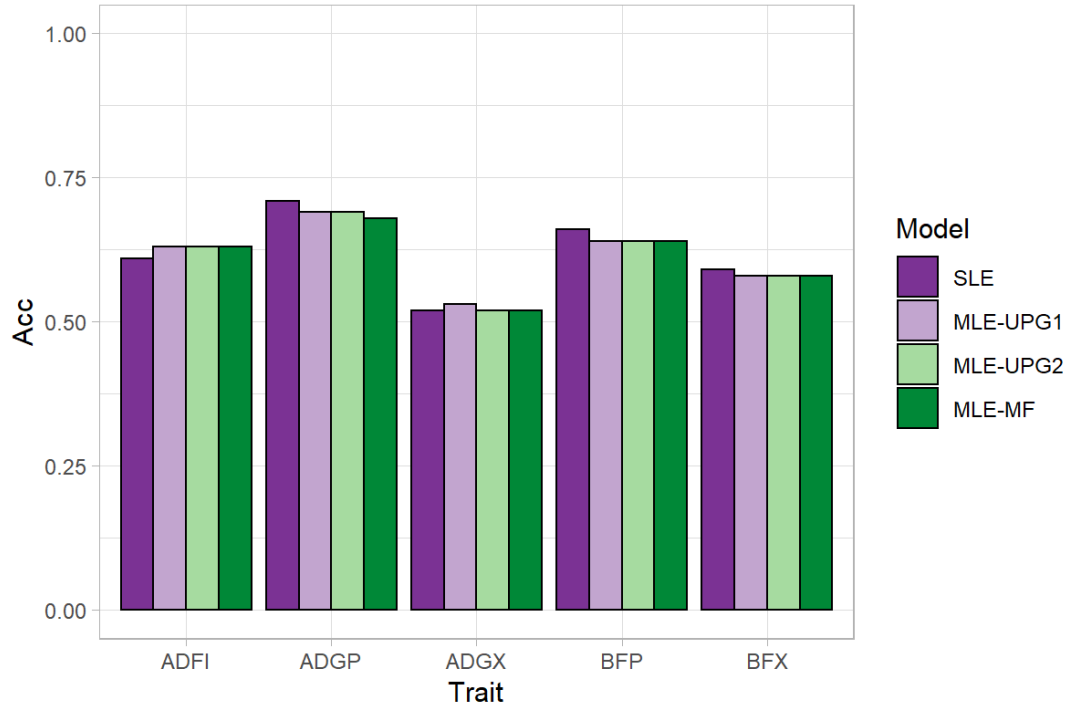


Figure 4.1. PCA plot for three terminal lines

(a) TL1


1.00
0.75
0.50
0.25
0.00
Acc
ADFI
ADGP
ADGX
BFP
BFX
Trait
Model
SLE
MLE-UPG1
MLE-UPG2
MLE-MF

(b) TL2


1.00
0.75
0.50
0.25
0.00
Acc
ADFI
ADGP
ADGX
BFP
BFX
Trait
Model
SLE
MLE-UPG1
MLE-UPG2
MLE-MF

(c) TL3

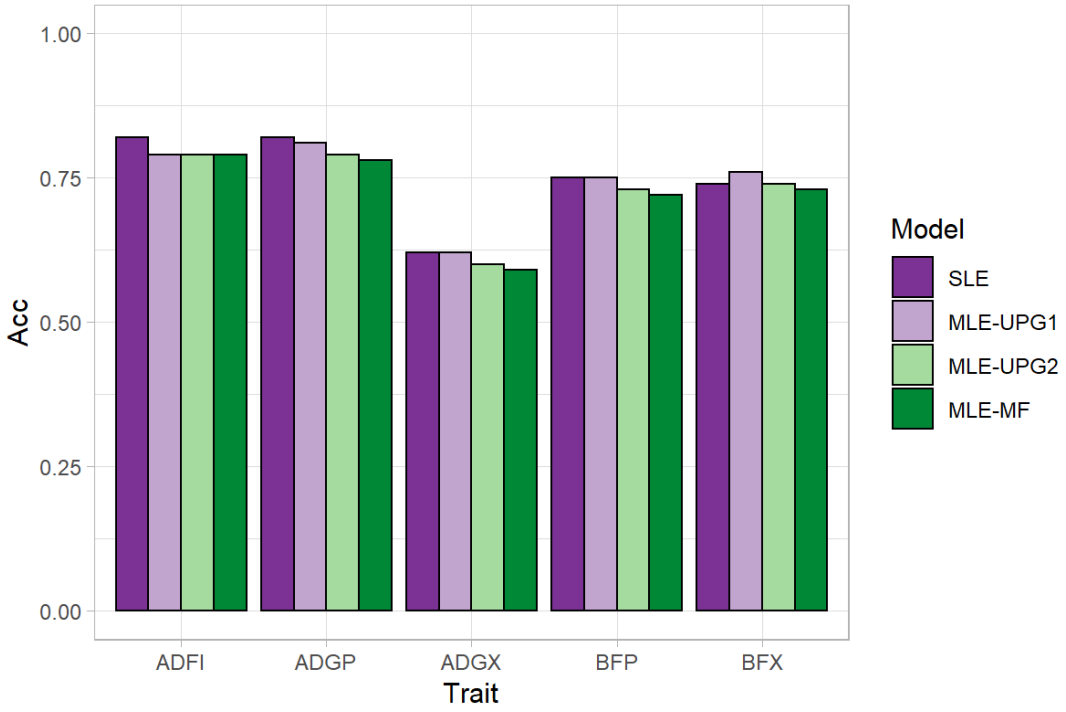
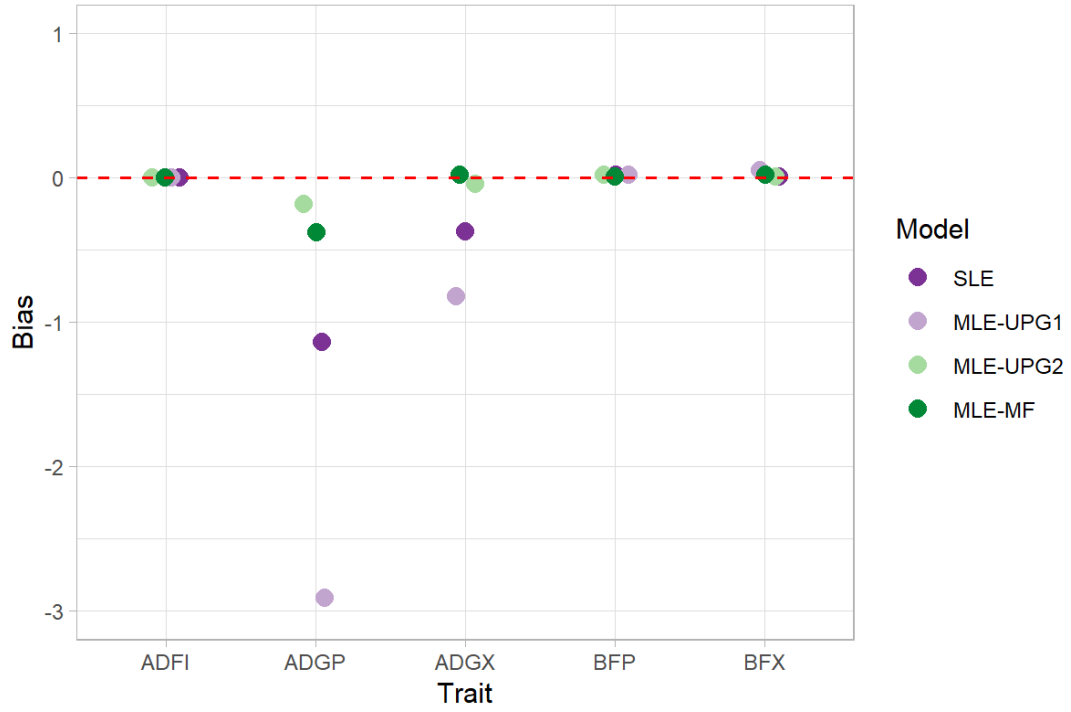
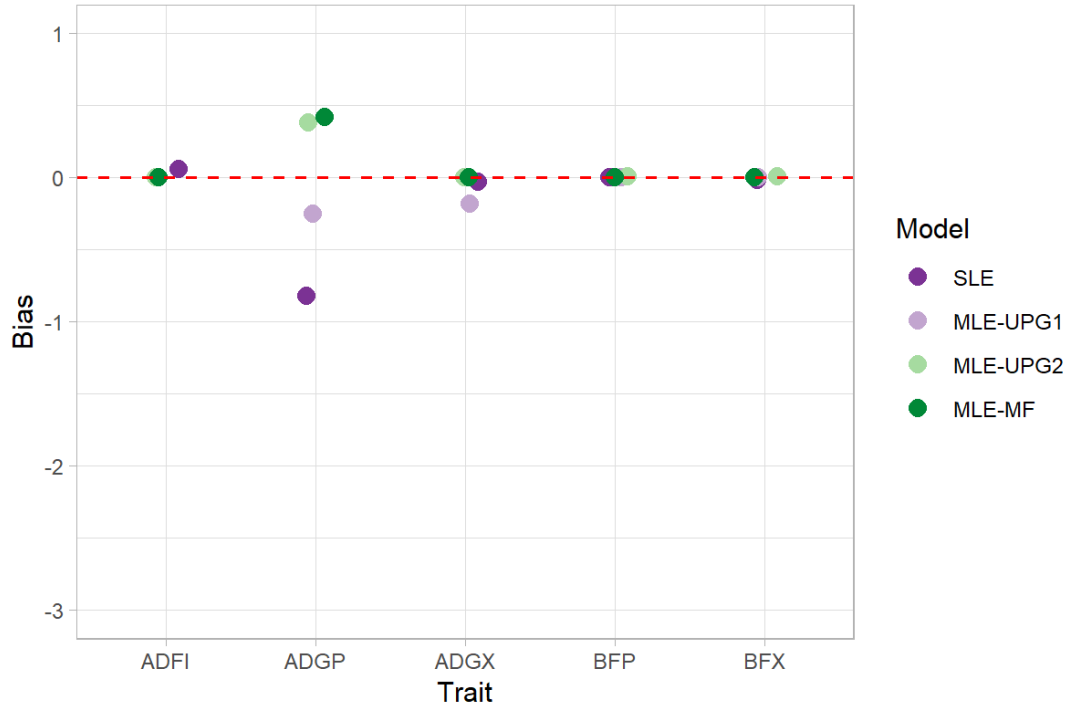


Figure 4.2. Accuracy of GEBV for single- and multi-line evaluations with unknown parent groups and metafounders using Chip data

(a) TL1


1
0
-1
-2
-3
Bias
ADFI
ADGP
ADGX
BFP
BFX
Trait
Model
SLE
MLE-UPG1
MLE-UPG2
MLE-MF

(b) TL2


1
0
-1
-2
-3
Bias
ADFI
ADGP
ADGX
BFP
BFX
Trait
Model
SLE
MLE-UPG1
MLE-UPG2
MLE-MF

(c) TL3

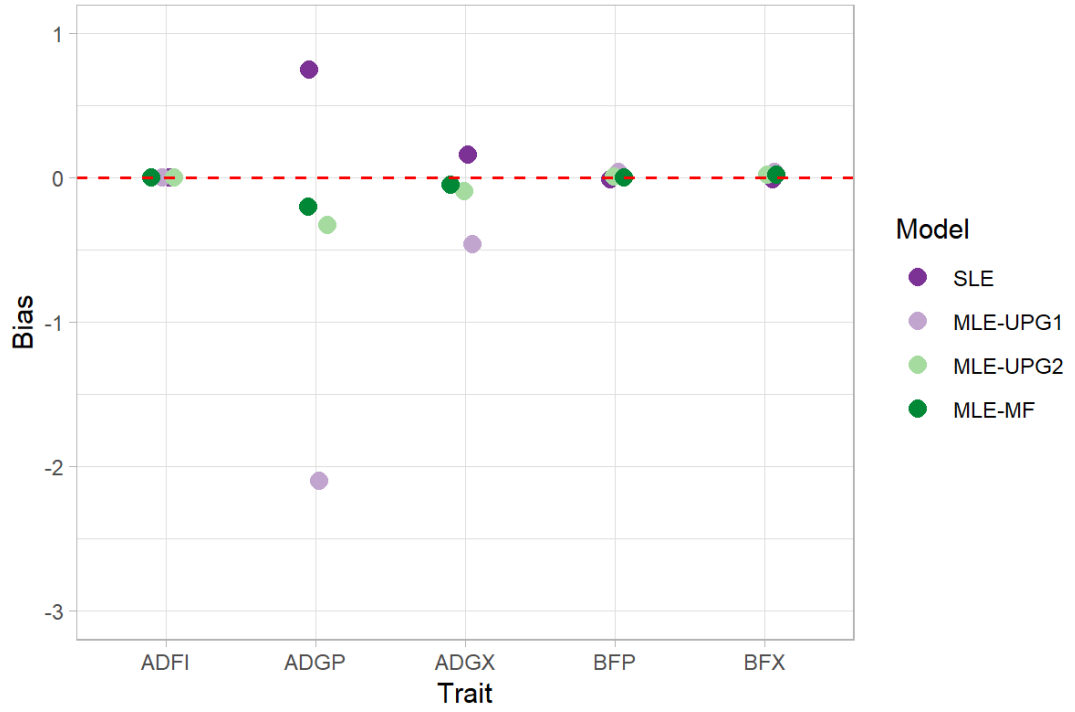
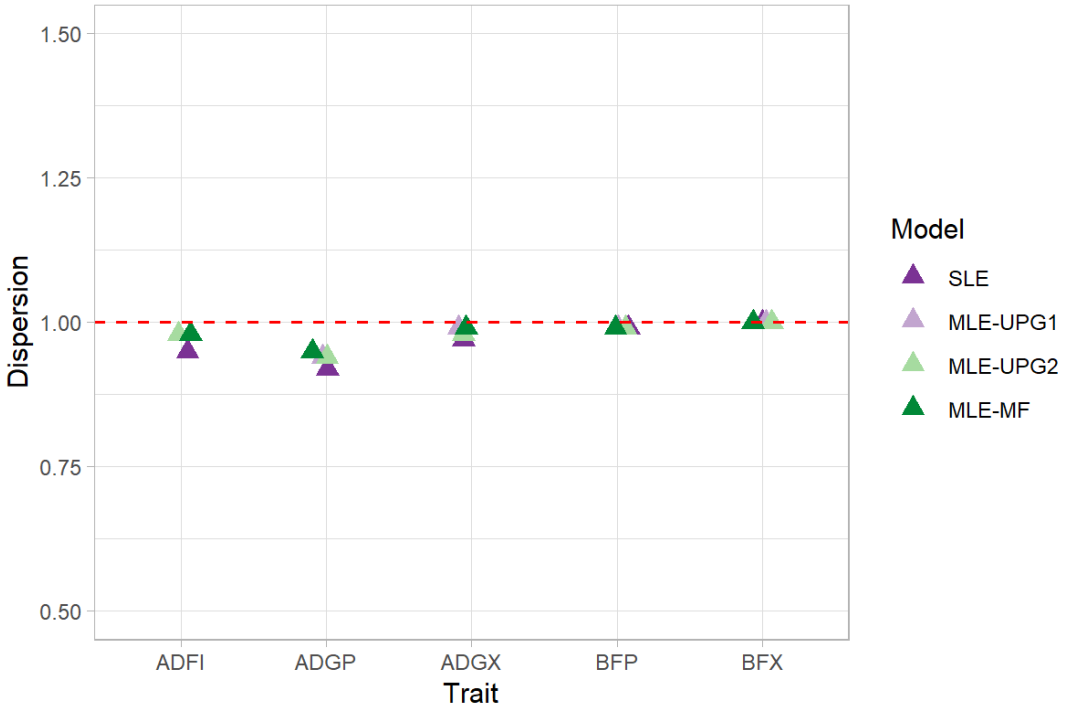
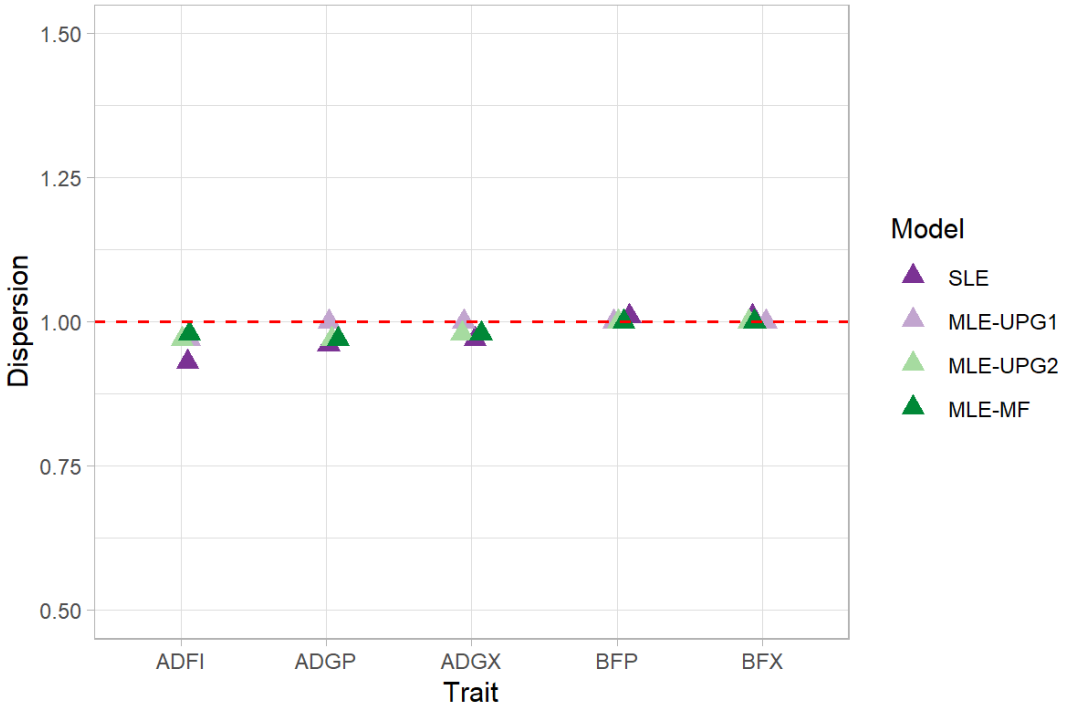


Figure 4.3. Bias of GEBV for single- and multi-line evaluations with unknown parent groups and metafounders using Chip data

(a) TL1


1.50
1.25
1.00
0.75
0.50
Dispersion
ADFI
ADGP
ADGX
BFP
BFX
Trait
Model
SLE
MLE-UPG1
MLE-UPG2
MLE-MF

(b) TL2


1.50
1.25
1.00
0.75
0.50
Dispersion
ADFI
ADGP
ADGX
BFP
BFX
Trait
Model
SLE
MLE-UPG1
MLE-UPG2
MLE-MF

(c) TL3

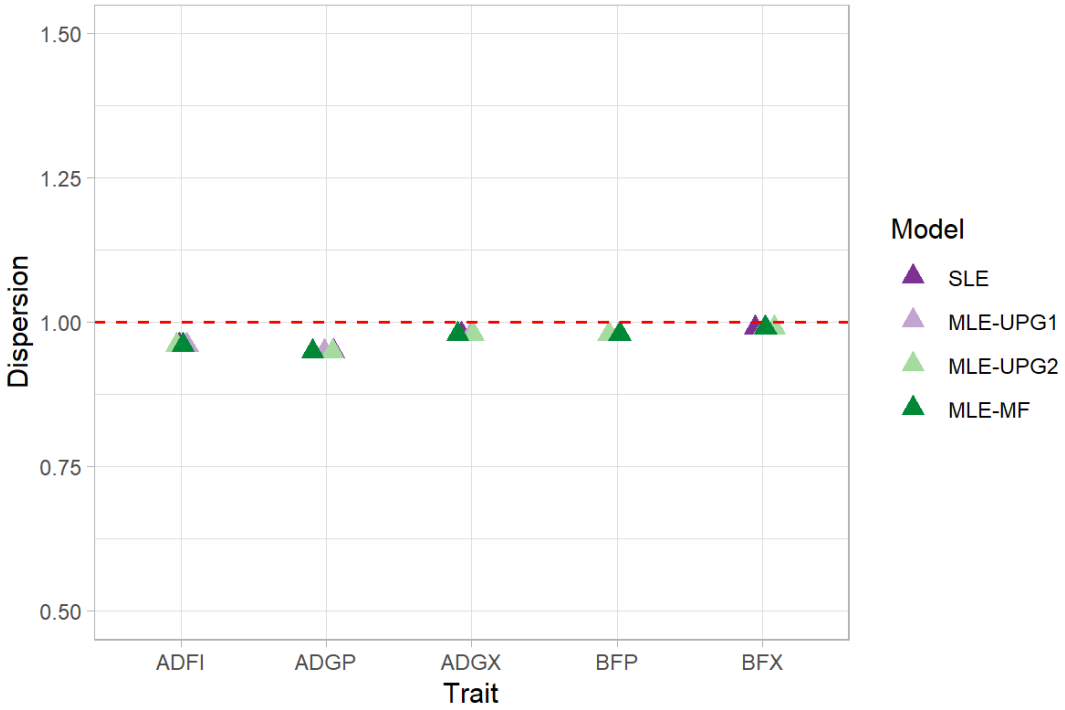


Figure 4.4. Dispersion (b1) of GEBV for single- and multi-line evaluations with unknown parent groups and metafounders using Chip data

CHAPTER 5

DIMENSIONALITY OF GENOMIC INFORMATION AND ITS IMPACT ON GWAS AND VARIANT SELECTION: A SIMULATION STUDY[3]

[3] Sungbong Jang, Shogo Tsuruta, Natalia Galoro Leite, Ignacy Misztal, and Daniela Lourenco. Submitted to *Genetics Selection Evolution*, 04/12/2022

ABSTRACT

Identifying true-positive variants in genome-wide association studies (GWAS) depends on several factors, including the number of genotyped individuals. The limited dimensionality of the genomic information may give insights into the optimal number of individuals to use in GWAS. This study investigated different discovery set sizes in GWAS based on the number of largest eigenvalues explaining a certain proportion of variance in the genomic relationship matrix (**G**). An additional investigation included the change in accuracy by adding variants, selected based on different set sizes, to the regular SNP chips used for genomic prediction. Sequence data were simulated containing 500k SNP with 200 or 2000 quantitative trait nucleotides (QTN). A regular 50k panel included one every ten simulated SNP. Effective population size (*Ne*) was 20 and 200. The GWAS was performed with the number of genotyped animals equivalent to the number of largest eigenvalues of **G** (EIG) explaining 50, 60, 70, 80, 90, 95, 98, and 99% of the variance. In addition, the largest discovery set consisted of 30k genotyped animals. Limited or extensive phenotypic information was mimicked by changing the trait heritability. Significant and high effect size SNP were added to the 50k panel and used for single-step GBLUP with and without weights. Using the number of genotyped animals corresponding to at least EIG98 enabled the identification of QTN with the largest effect sizes when *Ne* was large. Smaller populations required more than EIG98. Furthermore, using genotyped animals with higher reliability (i.e., higher trait heritability) helped better identify the most informative QTN. The greatest prediction accuracy was obtained when the significant or the high effect SNP representing twice the number of simulated QTN were added to the 50k panel. Weighting SNP differently did not increase prediction accuracy, mainly because of the size of the genotyped population. Accurately identifying causative variants from sequence data depends on the effective population size and, therefore, the dimensionality of

genomic information. This dimensionality can help identify the suitable sample size for GWAS and could be considered for variant selection. Even when variants are accurately identified, their inclusion in prediction models has limited implications.

INTRODUCTION

Several factors influence the statistical power to identify causative variants in genome-wide association studies (GWAS), including the number of quantitative trait nucleotides (QTN) affecting the trait, the number of single nucleotide polymorphisms (SNP) in the discovery panel, the number of genotyped individuals (Visscher et al., 2017), and the size of the genome blocks segregating in the population (Berisa and Pickrell, 2016), among others. Those genome blocks are chromosome segments inherited from founders and are subject to recombination every generation. Stam (1980) showed that segments are of different sizes but with a mean size of $1/4N_e$, where N_e is the effective population size. Given a species with a genome length equal to L Morgans, the number of independent chromosome segments (M_e) segregating in a population can be calculated as $4N_eL$.

Animal populations have lower N_e than human populations, implying smaller M_e. Pocrnic et al. (2016a) showed that although millions of individuals can be genotyped, non-redundant information is finite, which means the genomic information has a limited dimensionality; therefore, the additive genetic information in a population is contained in a limited M_e. The same authors related the limited dimensionality to $M_e = 4N_eL$ and observed that this quantity corresponds to the number of largest eigenvalues explaining 98% (EIG98) of the variance of the genomic relationship matrix (**G**). In cattle populations, EIG98 varies from 10K to 14K and is about 4K in

pigs and chickens. The minimum number of SNP needed to cover those segments is approximately $12M_e$ (MacLeod et al., 2005).

With the availability of sequence information, causal variants are expected to be in the data, generating more opportunities for discovery than with mid-density SNP panels (Meuwissen and Goddard, 2010). When the causal variants are known and included in the usual SNP panels, the accuracy of predicting genomic estimated breeding values (GEBV) should increase. This is clearly observed in simulated studies where the QTN and their effects are known (Pérez-Enciso et al., 2015; Fragomeni et al., 2017). However, the accuracy increased by using significant variants from the sequence in real populations is almost inexistent (Veerkamp et al., 2016; Zhang et al., 2018; Fragomeni et al., 2019). This raises a question on the effectiveness of GWAS in real populations. Although most traits of economic importance in farm animal populations are polygenic, very few peaks are usually statistically associated with traits of interest.

Misztal et al. (2021) investigated the distribution of SNP around the QTN and the ability to identify QTN depending on the N_e in simulated populations. They found that identifying QTN in populations with small N_e (i.e., 60) required three times more genotyped animals with phenotypes than in populations with large N_e (i.e., 600). However, not all simulated QTN were identified, independently of the N_e or amount of data. Distinguishing between noise and the true signal is more difficult in small populations because of longer chromosome segments and the uncertainty about the exact QTN location. Additionally, the level of noise may mask the signal, preventing associations. With sequence data, a clear GWAS resolution for small populations may be even harder to achieve due to the reasons mentioned above.

Although it is well-known that increasing the sample size for GWAS improves the resolution, the links among the number of genotyped individuals, N_e, M_e, and GWAS resolution

are missing. Additionally, understanding the appropriate sample size for variant discovery, especially with sequence data, can help to alleviate both the economic and computational costs for practical applications. Based on the limited dimensionality of the genomic information, there may be an optimal number of animals that carry all the independent chromosome segments segregating in the population, and consequently, all the genomic information available in the population (Pocrnic et al., 2016a). When animals have lots of information, GEBV are estimated with high accuracy. Knowing that GEBV can be backsolved to SNP effects raises the question on whether GWAS resolution is high when M_e animals with high accuracy GEBV are used. Therefore, we hypothesize that the ability to identify causative variants is high when the sample size for GWAS approaches M_e, and using a larger sample size may not further improve GWAS resolution. Here we used the number of eigenvalues explaining different proportions of the variance in **G** to assess the dimensionality of the genomic information and used this number as the sample size in GWAS. We used simulated populations with varying N_e, number of QTN, and amount of information on genotyped individuals. We also evaluated the impact of incorporating the pre-selected variants, from GWAS with different sample sizes based on dimensionality, to a 50k SNP chip for genomic prediction using weighted and unweighted single-step GBLUP.

MATERIALS AND METHODS

Animal Care and Use Committee approval was not needed as data were simulated

Data simulation

QMSim (Sargolzaei and Schenkel, 2009) was used to simulate a quantitative trait with 0.3, 0.9, and 0.99 heritability. Different heritabilities mimicked limited or extensive phenotypic information. The historical population was simulated for 2,000 non-overlapping generations with

an increase in size from 1,000 (generation -2,000) to 50,000 (generation -1,000), and a decrease from 50,000 (generation -999) to 20,000 (generation 0) to create LD and mutation-drift equilibrium. Random mating and no selection or migration were assumed in the historical population. Recent populations of N_e equal 20 (N_e20) and 200 (N_e200) were simulated by changing the number of breeding males from 5 to 50 but keeping the number of females at 15,000. The founders of the recent populations came from generation 0 of the historical population. Twenty generations of random mating were carried out, considering a replacement rate of 80% for sires and 30% for dams. Animals were randomly selected and culled based on age. A total of 315,005 and 315,050 animals were generated in the recent population for N_e20 and N_e200, respectively. However, only animals from generations 11−20 had phenotypic and pedigree information that was used for the current study. Of those, 75,000 animals from generations 16-20 were genotyped (N = 15,000 in each generation). The phenotype was the sum of an overall mean equal to 1.0, true breeding value (TBV), and random residual effect. The phenotypic variance was set to 1.0, whereas the additive genetic variance was 0.3, 0.9, or 0.99, all explained by the simulated QTN.

To mimic the bovine genome, we simulated 29 chromosomes with a total length of 23.19 Morgans. The overall number of SNP was 500,000, all with minor allele frequency greater than 0.05, whereas QTN numbers were 200 and 2,000 for the scenarios Q200 and Q2000, respectively. Biallelic SNP and QTN were randomly placed on each chromosome, with numbers varying from 9,000−35,000 SNP and 8−31 (Q200) or 80−320 (Q2000) QTN. The QTN effects were sampled from a gamma distribution with shape parameter 0.4 and scale parameter calculated internally for a genetic variance of 0.3, 0.9, and 0.99, depending on the scenario. A recurrent mutation rate of 2.5×10^{-5} was assumed for both SNP and QTN. A regular 50k panel was created for genomic predictions (GP) that included one every ten simulated SNP. Because different simulation

replicates would generate different QTN positions and effects, no replicate was used to obtain consistent GWAS results.

Genotype scenarios – heritability and the presence of QTN in the data

Trait heritabilities of 0.3, 0.9, and 0.99 were simulated to represent the animals with low reliability of EBV (H30), high reliability of EBV (H90), and very high reliability of EBV (H99), respectively. Therefore, higher heritabilities mean more information was added to the simulated animals without directly changing the number of records assigned to them (Pocrnic et al., 2019). Further, sequence data scenarios were created after the simulation. We assumed that QTN were in the data (withQTN), following the general assumption for sequence data; therefore, the SNP and QTN files were combined based on the corresponding maps. Descriptions for all the scenarios and combinations are in Table 5.1.

Discovery, training, and test sets

Before the GWAS analyses, all genotyped animals were separated into three non-overlapping data sets: discovery, training, and test. The test set was composed of genotyped animals from the last generation (N = 15,000), whereas the remaining genotyped animals (N = 60,000) were randomly assigned to the discovery and training sets (N = 30,000, respectively). To test the possible bias in GP by using the same data set for discovery and training, two different schemes were designed: 1) discovery = training: genotyped animals used for discovery were further used for training, and 2) discovery ≠ training: a different set of genotyped animals were used for discovery and training.

EIGx scenarios for discovery and training

Different scenarios were made based on the dimensionality of the genomic information to investigate the effect of sample size for discovery and training. The number of genotyped animals

in each discovery and training set (EIG*x*) was equivalent to the number of largest eigenvalues explaining *x* percent of the variance in **G**, where *x* assumed the values 50, 60, 70, 80, 90, 95, 98, or 99. For example, the number of largest eigenvalues explaining 50% of the variance in **G** was 530 in the N_e200 Q2000 H30 scenario (Table 5.2); thus, the size of discovery and training sets in scenario EIG50 was set to 530. In addition, one extra scenario (ALL) in which the discovery and training sets consisted of all available genotyped animals (N = 30,000) was also evaluated. The number of largest eigenvalues explaining *x* percent (50, 60, 70, 80, 90, 95, 98, 99) of the variance in **G** was computed by squaring the singular values from the matrix of genotypes centered for current allele frequencies (**M**), using all the simulated genotyped animals (N = 75,000). The singular value decomposition was done in preGSf90 (Misztal et al., 2014). For that, **G** was constructed using all simulated SNP without QTN. All genotyped animals for each discovery and training set were randomly selected beginning from the scenario explaining the least proportion of variance (EIG50). Ensuring consistent results involved keeping all the animals from a previous scenario when moving to the next one, e.g., genotyped animals in EIG60 contained the ones from EIG50. The number of genotyped animals for all scenarios used as discovery and training sets is described in Table 5.2.

Preselection of variants (TOPv scenarios)

Different numbers of variants were selected from GWAS to be included in the 50k SNP panel for GP. Each QTN scenario had a specific number of selected SNP based on the order of the p-values (TOP*v*) or statistical significance using Bonferroni corrected p-values (SIG). For Q200, *v* corresponded to 10, 50, 100, 200, and 400, whereas for Q2000 *v* assumed the values of 10, 100, 500, 1000, 2000, and 4000.

Association among number of significant QTN, sample size, and EBV reliability

In the current study, we approximated sample size based on the total proportion of variance explained by significantly identified QTN from the results of different heritability scenarios. This approximation was done by local polynomial regression (Cleveland et al., 1992) using the 'loess' and 'approx' function in the R, and the resulting sample size was represented by SS_{pol}.

The main purpose of comparing scenarios with different heritabilities was to investigate the effect of using genotyped animals with low to high reliability of EBV on GWAS performance; however, not all the genotyped animals have high reliability of EBV. Therefore, in this study, we investigated the corresponding sample size from low to high heritability scenarios at the point where the same percentage of variance was explained; we estimated the sample size using H30 as a benchmark. This helped us identify how many samples are needed for GWAS given the average reliability of breeding values for the animals in the population and the benchmark reliability. For that, derived an equation to estimate the sample size relating the total number of samples, M_e, N_e, proportion of additive genetic variance explained by significant QTN ($\%Var$), and reliability of EBV as:

$$SS_{rel} = \frac{N_s rel(\log({}_{\sigma^2_{QTN}} N_e))\lambda\sigma^2_{QTN}}{rel_t(\sigma^2_{QTN} + \ln(M_e))}$$

in which SS_{rel} is the approximated sample size for target reliability, N_s is the benchmark sample size, rel and rel_t are the benchmark and target reliabilities of EBV (heritability), λ is a constant equal to 0.4, and σ^2_{QTN} is the %var explained by identified QTN. The total proportion of genetic variance explained by the identified QTN was calculated as the sum of the genetic variance explained by each QTN. As QTN effects were given by the simulation, the percentage of genetic variance explained by individual QTN was calculated as:

$$\%Var = 2pq(\beta)^2/\sigma^2_a$$

where the p and q are the major and minor allele frequency of the QTN, β is the QTN effect, σ_a^2 is the total additive genetic variance of the model.

Models and analysis

Genome-wide associations

Efficient mixed-model association expedited (EMMAX) was performed using Gemma software (Zhou and Stephens, 2012), with the following model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}_i b_i + \mathbf{Zu} + \mathbf{e},$$

where **y** is a vector of phenotypes, μ is an overall mean, $\mathbf{X}_i$ is a vector of genotypes for i^{th} SNP, b_i is the substitution effect of the i^{th} SNP, **Z** is an incidence matrix for vector **u**, and **u** is a vector of random additive genetic effects, with $\mathbf{u} \sim N(0, \mathbf{G}\sigma_u^2)$, and **e** is a vector of residuals, with $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ and **I** an identity matrix. The **G** was computed as in Zhou (Zhou and Stephens, 2012):

$$\mathbf{G} = \frac{1}{n_s} \sum_{i=1}^{p} (x_i - 1_n \bar{x}_i)(x_i - 1_n \bar{x}_i)^T$$

where the x_i is the i^{th} SNP locus column, $\bar{x}_i$ is the marker sample mean of the i^{th} locus, n and n_s are numbers of genotyped animals and SNP.

Genomic prediction

A linear mixed model was used to compute GP:

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

where **y** is a vector of phenotypes, μ is an overall mean, **Z** is an incidence matrix for vector **u**, and **u** is a vector of random additive genetic effects, with $\mathbf{u} \sim N(0, \mathbf{H}\sigma_u^2)$ and **H** is the realized relationship matrix, and **e** is a vector of residuals, with $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. The GP was carried out with ssGBLUP and weighted ssGBLUP (WssGBLUP) (Wang et al., 2012) using the BLUPF90 family

of programs (Misztal et al., 2014). For the mixed model equations in ssGBLUP and WssGBLUP, $\mathbf{H}^{-1}$ combines both pedigree and genomic relationships (Aguilar et al., 2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix},$$

where $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix and $\mathbf{A}_{22}^{-1}$ is the inverse of the pedigree relationship matrix for the genotyped animals. The **G** was created as in VanRaden (VanRaden, 2008):

$$\mathbf{G} = \frac{\mathbf{MDM'}}{2\sum p_i(1-p_i)},$$

where **M** was defined before, p_i is the minor allele frequency of the i^{th} SNP, **D** is the diagonal matrix of SNP weights with dimensions equivalent to the number of SNP. In ssGBLUP, all SNP were assumed to have homogeneous weights, meaning that **D** was an identity matrix. To avoid singularity issues, **G** was blended with 5% of $\mathbf{A}_{22}$.

For the WssGBLUP, SNP effects were back-solved from GEBV ($\hat{\mathbf{u}}$) as described in Wang et al. (2012):

$$\hat{\mathbf{a}} = \sigma_a^2 \sigma_u^{-2} \mathbf{DM'G^{-1}}\hat{\mathbf{u}},$$

where $\hat{\mathbf{a}}$ is a vector of estimated SNP effects, σ_a^2 is the SNP variance, σ_u^{-2} is the genetic variance, **D** is the diagonal matrix of SNP weights (**I** in ssGBLUP), and **M** is the centered matrix of genotypes. After SNP effects were estimated, the variance for the i^{th} SNP was calculated using the non-linearA method (VanRaden, 2008):

$$d_i = CT^{\frac{|\hat{a}_i|}{\sigma(\hat{a})}-2},$$

where CT is a constant that determines the departure from normality when deviating from 1; $|\hat{a}_i|$ is the absolute estimated SNP effect for i^{th} SNP; and $\sigma(\hat{\mathbf{a}})$ is the standard deviation of the vector

of estimated SNP effects. This study used CT as 1.125, an empirical value based on polygenic traits in dairy cattle populations. Results from the second iteration of weights were used in this study to maximize the prediction accuracy, as suggested by Zhang et al. (2016); Lourenco et al. (2017).

Validation of genomic predictions

In each scenario, prediction accuracy was calculated as the correlation between TBV and genomic estimated breeding value (GEBV). Besides, the regression coefficient (b_1) of TBV on GEBV was used as an indicator of inflation or deflation of GEBV. When b_1 is lower than one, it is indicative of inflation and deflation otherwise. As replicates were not used in this study, standard errors (SE) were computed using the bootstrapping method (Canty, 2002).

RESULTS

Variant identification

The preliminary analysis showed similar results for both GWAS and GP when QTN were included (withQTN) or not included in the data. Therefore, results of withQTN are only described. The results of GWAS are shown in Fig. 5.1 ~ 5.4. As most of the quantitative traits are highly polygenic, only results of Q2000 with H30 and H99 are shown with two N_e scenarios (N_e20, N_e200), respectively. In addition, the GWAS results of EIG60, EIG70, and EIG80 were not included in those figures due to their insignificance. Considering a population with N_e equal to 20 and using EIG50, EIG90, EIG95, EIG98, and EIG99 as the sample size for GWAS in Q2000 was not enough to detect significant QTN when the amount of information for genotyped animals was small (Fig. 5.1). However, when the sample size increased to 30,000 (i.e., ALL), three significant QTN were identified. In contrast, using genotyped animals with high reliability increased the

ability to identify simulated QTN correctly (Fig. 5.2). EIG95 could capture three significant QTN, and as the sample size increased to EIG98, EIG99, and ALL, 17, 33, and 142 QTN were identified, respectively.

Different patterns were observed for a population with N_e equal to 200 when having contrasting EIG*x* as the sample size (Fig. 5.3 and Fig. 5.4). Although EIG50, EIG90, and EIG95 were not sufficient to capture the significant QTN in H30, a sample size of EIG98 allowed the identification of seven QTN (Fig. 5.3). Moreover, increasing the number of genotyped animals to EIG99 and ALL helped detect more QTN and improve the GWAS resolution, even though genotyped animals had low reliability. When N_e was 200, but the animals had high reliability, EIG90 is an adequate sample size to detect the QTN with the largest effect size (Fig. 5.4). For this scenario, EIG98 provided a clear resolution, similar to EIG99 and ALL. It is important to note that the number of largest eigenvalues explaining a certain proportion of the variance in **G** was different for N_e20 and N_e200 (Table 5.2). With all available genotyped animals (i.e., ALL), N_e200 had more significant QTN discovered in GWAS than N_e20. For example, in Q2000 and H30, the three significant QTN in N_e20 captured 3.9% of the additive genetic variance, whereas, in N_e200, 15 QTN were capturing 13.9%. In both Ne scenarios, using genotyped animals with high reliability helped better detect QTN. For a less polygenic trait (Q200), fewer genotyped animals were required to identify the simulated QTN than in a more polygenic trait (Q200).

Association between variance explained by identified QTN and sample size

The proportion of variance explained by identified QTN is shown in Fig. 5.5. When the sample size increased, the proportion of variance explained by the identified QTN increased regardless of heritability, N_e, and number of QTN. Comparing the proportion of variance explained according to the heritability, high heritability scenarios better identified significant QTN. For

example, using all genotyped animals in N_e20 and Q200 with H30 helped identify QTN explaining 44.9% of the variance (Fig. 5.5a). However, H90 and H99 identified QTN explaining 72.9% and 74.3% of the variance, respectively, with the same number of genotyped animals. For a more polygenic trait (Fig. 5.5b), QTN explained 3.9% and 43.1% of the variance with H30 and H99, respectively. The latter was similar to the variance explained in Q200, H30. Similar patterns were observed between N_e20 and N_e200; however, QTN identified in N_e200 explained a greater proportion of variance than N_e20 for Q200 and Q2000 (Fig. 5.5c-d). For example, the maximum proportion of variance explained by QTN in N_e20 (Fig. 5.5a-b) was 74.3% and 43.1% for Q200 and Q2000 with H99, whereas those values were 96% and 65.1% in N_e200 (Fig. 5.5c-d). One remarkable discovery with a less polygenic trait was that the use of EIG98 and EIG99 showed a similar proportion of variance explained by identified QTN as in ALL (Fig. 5.5a and 5.5c). For H99, when EIG98, EIG99, and ALL were used, the variance explained by the identified QTN was 59.9%, 68%, and 74.3%, respectively (Fig. 5.5a). In the case of Q2000, the scenarios mentioned above identified QTN explaining 11.2%, 18.9%, and 43.1% of the variance (Fig. 5.5b); therefore, the proportion of variance explained increased by almost fourfold from using EIG98 to ALL, whereas this increase was only 25% with Q200. In a larger population (N_e200) and H99 (Fig. 5.5c), EIG98, EIG99, and ALL schemes detected QTN explaining 95.2%, 95.6%, and 96% of the variance for Q200. Even for the more polygenic scheme (Q2000), EIG98, EIG99, and ALL captured QTN explaining 52.5%, 59.7%, and 65.1% (Fig. 5.5d). Similar patterns were observed for the other two heritability scenarios in Fig. 5.5c-d.

To investigate the corresponding sample size from the use of low to high heritability scenarios at the point where the same percentage of variance was explained, we used H30 as a benchmark. Table 5.3 shows the approximated sample size when the largest discovery set (ALL,

N = 30,000) was used. The results of the N_e20 Q200 scenario showed that around 2223 and 1662 genotyped animals were required for H90 and H99 to reach the same magnitude of genetic variance explained by the identified QTN in H30 (Table 5.3). In N_e20 Q2000, using 3049 and 2378 genotyped animals with H90 and H99 helped identify QTN explaining 3.9% of the variance, which was accomplished using ALL in H30. Similar patterns were observed in N_e200 scenarios. One remarkable difference between N_e20 and N_e200 was that the sample size to reach the same proportion of variance explained by QTN using ALL in H30 was equivalent to EIG90 ~ EIG98 in N_e20 but EIG80 ~ EIG90 in N_e200 when considering H90 and H99.

In the current study, in addition to the approximated sample size by local polynomial regression, we derived the equation to estimate the sample size relating the number of samples, M_e, N_e, percentage of variance explained by the identified QTN, and reliability of EBV. For example, to approximate the sample size of H90 in N_e20 Q200 scenario when %var was 44.9, it would be

$$\mathrm{SS_{rel}} = \frac{30000 \times 0.3 \times (log_{44.9}20) \times 0.4 \times 44.9}{0.9 \times (44.9 + \ln(1840))}$$

Those sample sizes approximated by the proposed equation are described in Table 5.3. Approximating the sample size in Q2000 was more accurate than in Q200 for N_e20 and N_e200. Additionally, the formula and the polynomial regression provided sample sizes that were within the same EIGx range, except for one scenario (N_e20 Q200 H99). As an example, N_e20 Q2000 resulted in 3049 (H90) and 2378 (H99) samples in $\mathrm{SS_{pol}}$ and 3007 (H90) and 2734 (H99) in $\mathrm{SS_{rel}}$; however, the sample size in both cases laid within EIG95~98. Differences between $\mathrm{SS_{pol}}$ and $\mathrm{SS_{rel}}$ across more polygenic scenarios were not large.

Genomic predictions

Beforehand, we investigate the possible bias of GP by using the same set of genotyped animals for both discovery and training. Using different groups of genotyped animals for discovery and training resulted in less inflation of GEBV than utilizing the same animals for both processes (results are not shown). Therefore, the GP analyses were done with training animals different from the discovery set.

We noticed very small standard errors for prediction accuracy (<0.005) and b_1 (<0.01) by bootstrapping. The results of prediction accuracy and inflation/deflation indicator of GEBV (b_1) are shown in Fig. 5.6 and Fig. 5.7, respectively. Those accuracies and b_1 were calculated as the average of all genotyped scenarios: 50k, TOP10, TOP50, TOP100, TOP200, TOP400, and 'SIG' for Q200 and 50k, TOP10, TOP100, TOP500, TOP1000, TOP2000, TOP4000, and 'SIG' for Q2000 scenarios. Standard deviations were less than 0.02 for all scenarios. Fig. 5.6 shows the prediction accuracy depending on the number of QTN (Q2000 or Q200), trait heritability (H30, H90, and H99), and training data scenarios (EIGx and ALL). In general, as the training set size increased, prediction accuracy also increased. Different patterns were observed between N_e20 and N_e200. In N_e200, the prediction accuracy increased consistently as the training set size increased; however, the increase in prediction accuracy was not constant in N_e20. For example, in N_e20 Q200 H30, while the training data set went up from the EIG50 to EIG95, prediction accuracy increased only by about 0.03. A similar pattern was observed for N_e20 Q2000 H30, which showed a gain of about 0.02. In general, N_e20 showed greater prediction accuracy than N_e200. For example, when the smallest sample size (EIG50) was used, the average accuracy in N_e20 was 0.82 ± 0.03 and N_e200 was 0.74 ± 0.09, a difference of about 0.08. This difference became smaller with the largest sample size (ALL), which was 0.04. Therefore, no substantial differences in prediction accuracy between the N_e20 and N_e200 were found when the largest training set was used. Overall, the

difference in prediction accuracy between EIG98 or EIG99 and ALL was smaller in N_e200 (0.88, 0.90, and 0.92 on average for EIG98, EIG99, and ALL) than in N_e20 (0.90, 0.92, and 0.96 on average for EIG98, EIG99, and ALL), indicating a sample for training with the size of EIG98 or EIG99 would suffice.

The number of QTN marginally affected the prediction accuracy for N_e20 and N_e200 scenarios. For both N_e scenarios, Q200 showed greater prediction accuracy, especially for the low heritability with the smaller training sets. For example, in N_e20 H30, the prediction accuracy from EIG50 to EIG70 was 0.81 with Q200 and 0.79 with Q2000. With larger training set sizes and higher heritability, the difference between Q200 and Q2000 decreased. Prediction accuracies were highly influenced by the heritability of the trait, particularly with a larger effective population size. For instance, with N_e200 Q200, when EIG50 was used, the prediction accuracy was 0.64, 0.81, and 0.81 for H30, H90, and H99, respectively. Even with the most extensive training set (ALL), prediction accuracy was 0.84, 0.96, and 0.98, following the previous order. Similar patterns were observed for N_e200 Q2000. Thus, using low to high-reliability genotyped animals for training could affect the prediction accuracy more in populations with large N_e. Expanding the training set from EIG95 to EIG98 in N_e200 H30 increased prediction accuracy by 5.5% and 6.82% for Q200 and Q2000, respectively. However, the maximum increase for N_e200 H90 and H99 was observed when the training set size expanded from EIG80 to EIG90 (3.8% ~ 4.61%). Unlike N_e200, great improvements in accuracy were observed when moving from EIG99 to ALL in H30 and H90 for both Q200 and Q2000.

Regression coefficients (b_1) used as indicators of inflation/deflation of GEBV are shown in Fig. 5.7. When the training set size was small, less inflation was observed in N_e200 than in N_e20. In both N_e scenarios, using a large training set alleviated the inflation, so when ALL was used for

training, all scenarios had b_1 close to 1, especially for N_e20. Interestingly, more variation between the models was observed in N_e20 than in N_e200.

Fig. 5.8 shows the prediction accuracy with 50k compared to 50k plus SIG, TOP400 (Q200), and TOP4000 (Q2000), together with the percentage of gain by adding possible causative variants. As the increase was not major across analyses combining the 50k and the top SNP, only the scenarios with the largest changes are shown in Fig. 5.8; all the other scenarios are in Additional file 2. Overall, the percentage of gain was greater in N_e200 (3.29% ~ 9.01%) than in N_e20 (0.86% ~ 1.98%). In addition, Q200 showed a higher percentage of gain than Q2000 in both N_e scenarios. Interestingly, the maximum accuracy gain was usually observed when the largest number of top SNP (TOP400 for Q200 and TOP4000 for Q2000) was added to 50k chip data, representing twice the number of simulated QTN. The only exceptions were Q200 H90 and H99 for N_e20 and N_e200, which had the highest accuracy gain with 50k plus SIG. This is probably because identifying significant QTN was easier in a less polygenic trait (Q200) with H90 and H99. In contrast, finding QTN in Q2000 or a low heritability trait was harder.

DISCUSSION

In this study, we comprehensively investigated the impact of using different sample sizes in GWAS based on the dimensionality of the genomic information, the implications of using genotyped animals having low to high reliability of EBV in GWAS, and the inclusion of preselected variants into a typical 50k SNP panel using ssGBLUP and WssGBLUP. These investigations brought insights into how different data structures can affect the performance of GWAS and GP under the ssGBLUP framework. We used the concept of limited dimensionality of the genomic information (Pocrnic et al., 2016a). Our results showed that this concept could be a

helpful indicator of the number of genotyped animals required for GWAS, depending on N_e, M_e, the number of QTN, and the reliability of EBV. Our results showed that having a sample size with the number of genotyped animals corresponding to that of EIG98 was appropriate for variant discovery, particularly in the population with large N_e. Additionally, using genotyped animals with high EBV reliability could help better identify significant QTN regardless of N_e and the genetic architecture of the traits. Incorporating selected variants obtained from GWAS to the 50k SNP chip could improve prediction accuracy when a training set with proper size was used; however, the gain could be limited in some scenarios.

GWAS – preselection of variants

The most prevalent workflow for GP with sequence data is 1) pre-selection of significant variants, 2) incorporation of selected variants to the commercial chip data (i.e., 50k), or fitting separate genomic matrices in the model (Fragomeni et al., 2019; Moghaddar et al., 2019; Lopez et al., 2021), 3) comparison of the GP performance with a benchmark SNP chip. Several studies have been conducted to improve GP using sequence data with either simulated or real data. However, conclusions about the advantage of using sequence data have not been very consistent in the literature, and they seem to be dependent on several factors such as the species, the genetic architecture of the trait, the size of data, and statistical methods (MacLeod et al., 2016; VanRaden et al., 2017; Fragomeni et al., 2019; Moghaddar et al., 2019).

Among those factors, the most critical is the size of data for discovery, training, and test sets. Specifically, the sample size for the variant discovery set is essential as it is the first step and, thus, predominantly affects the results of the entire study. Current results indicated that using a small number of genotyped animals could not identify the significant SNP or QTN. Lourenco et al. (2017) used two different numbers of genotyped animals (N = 2,000 and 25,000) for GWAS

and reported that the best resolution was observed when more genotyped animals were used. In the same line, de Las Heras-Saldana et al. (2020) outlined that using a larger dataset for GWAS allowed to better identify quantitative trait loci (QTL) regions for carcass traits in Hanwoo cattle.

As the number of genotyped animals has currently increased in many species, for instance, about 5 million U.S. Holsteins (https://queries.uscdcb.com/Genotype/cur_freq.html), and about 1 million American Angus (K. Retallick, American Angus Association, Saint Joseph, MO, personal communication) have been genotyped as of March 2022, it is important to know how many genotyped animals are effectively required to detect the significant variants. Current results showed that using at least the number of genotyped animals equivalent to EIG98 could identify the most informative QTN. Using EIG99 or all available genotyped animals little improved the ability to identify significant QTN in N_e200 for both Q200 and Q2000 scenarios. This result could be helpful for both small and large genotyped populations with large N_e. For breeding populations with fewer resources, the number of animals to genotype may be limited; therefore, accessing what would be the effective sample size could benefit cost-effective genotyping or sequencing. For large populations, our study showed that not all animals are needed for variant discovery, and a balanced data set should be constructed for discovery, training, and testing to avoid biases and maximize the power to detect the significant variants. When N_e was small and the trait was highly polygenic, the small sample size could not identify any significant QTN until it reached ALL and EIG98 for H30 and H90. With more information on genotyped animals (i.e., H99), a sample size equivalent to EIG95 helped identify a few QTN. Chicken and pigs had smaller N_e (32 ~ 48) than cattle among the livestock species (Pocrnic et al., 2016b). Therefore, using a sample size corresponding to less than ALL would not be enough to detect significant signals for those species. Gozalo-Marcilla et al. (2021) carried out large-scale GWAS for backfat thickness in pigs using around 15k to 55k

genotyped animals. They found 264 significant SNP across 8 different lines for traits with moderate to high heritability (0.30 ~ 0.58). As backfat thickness has been known for its polygenic architecture (more than 1400 QTL associated backfat thickness is reported in https://www.animalgenome.org/QTLdb), their discovery is supported by our findings in populations with small N_e and moderate heritability.

Pocrnic et al. (2016a) described the number of largest eigenvalues explaining a certain proportion of **G** as a function of N_e and genome length in Morgans, such that EIG90 ≈ N_eL, EIG95 ≈ $2N_eL$, and EIG98 ≈ $4N_eL$. Stam (1980) expressed the expected number of independent chromosome segments as $4N_eL$. In this study, N_e was 20 or 200 and L was 23, so M_e was approximated as 1,840 and 18,400, between EIG95 and EIG98 for N_e20, and EIG98 and EIG99 for N_e200 in Table 5.2. As N_e and M_e are proportional, smaller N_e denotes fewer M_e, indicating fewer blocks are existed in the genome with a strong LD between variants because of the close relationship among individuals. N_e also has been reported as a factor affecting the performance of GWAS (Baldwin-Brown et al., 2014; Lourenco et al., 2017). In the current study, we showed that when the same number of genotyped animals were used (ALL), N_e200 could better identify the significant QTN explaining more genetic variance than N_e20 for all heritability and QTN scenarios. This might be because of smaller chromosome segments and weaker LD between the QTN and SNP in N_e200 than in N_e20. Pinpointing QTN is harder in N_e20 because many SNP may capture the QTN signal. The noise in GWAS resolution is possibly due to the strong relationships between the SNP and QTN, established by a highly structured population over the generations in N_e20; therefore, identifying the true causative variant is not trivial in smaller populations. In general, N_e of farm animals such as chickens, pigs, dairy, and beef cattle is less than 200 and could range from 40 to 150 (Pocrnic et al., 2016b); thus, the current findings would be helpful information for the

future GWAS in those species. However, identifying all significant variants is not assured due to the polygenic nature of most traits in livestock animals, and most of the causal variants have a small effect. For example, even with the largest number of genotyped animals for GWAS in our study (ALL, N = 30k), identifying QTN with very small effects was not possible due to limited statistical power. Misztal et al. (2021) showed that identifying all simulated QTN was impossible even when all had the same effect, N_e was 600, and the sample size was 6000. For a population with N_e of 60, a sample size three times larger resulted in more true signals in GWAS, but not as many as with N_e of 600. The same authors argued that with smaller N_e, more data is required to overcome the noise stage and capture the actual signals.

Different heritability scenarios were compared to investigate the performance of GWAS when genotyped animals with low to high reliability of EBV were used for the variant discovery stage. Our findings highlighted that regardless of the number of QTN, N_e, M_e, and sample size, high heritability scenarios could capture more significant QTN explaining a larger portion of the variance. However, Takeda et al. (2020) observed no differences in power to detect QTL when heritability 0.2 and 0.5 were simulated but outlined that QTL detection was better with the increasing number of phenotyped progenies (N = 1,500, 4,500, 9,000). As the use of more phenotyped progeny data indicated higher reliability of EBV of the parents, which is the case of genotyped animals in the higher heritability scenarios in our study, those findings agreed with the current results. Besides, van den Berg et al. (2013) reported that the number of false positives in QTL detection decreased with increasing heritability and number of records. Thus, using genotyped animals with high EBV reliability could sufficiently detect the QTN although few animals were used.

Sample size approximation

Approximated sample sizes were obtained through polynomial regression and a formula relating N_e, M_e, and percentage of variance to be explained by significant QTN. The latter was useful to investigate the sample size given the average reliability of EBV in the set of animals available for GWAS. Overall, the comparison showed a better approximation with Q2000 scenarios but a very different approximation scale with Q200 scenarios (results now shown). One possible reason for the inaccurate approximated sample size would be an unbalanced simulation design for N_e and heritability. As we only simulated two scenarios of N_e: 20 and 200 for the smallest and largest N_e in livestock species with three heritability scenarios: 0.3, 0.9, and 0.99 representing low, high, and very high reliability of EBV, there is a large gap between N_e 20 and N_e200, and an irregular pattern of heritability scenarios. However, the proposed equation could be applied for more polygenic traits. For example, given the number of genotyped or sequenced animals available, reliability of those animals, reliability of target animals, and proportion of variance explained by identified QTN, i.e., identified SNP in real data, with N_e and M_e, we can approximate the sample size for GWAS.

Genomic prediction

In general, the accuracy of GP was improved as the size of training data increased, and combining selected variants to a 50k SNP panel could improve accuracy when the GP was performed with the proper size of training sets. It was demonstrated that increasing the number of animals in training sets improved the accuracy of GP (Daetwyler et al., 2008; Hayes et al., 2009; Boddhireddy et al., 2014). Our findings support those results, although only a tiny improvement (< 1.0%) was reported when using training sets with the number of genotyped animals equal to EIG50 to EIG70 in N_e20. Moser et al. (2009) observed no improvements in prediction accuracy when the training size was enlarged from 1,239 to 1,880 in Australian dairy cattle. Therefore,

adding a substantial number of genotyped animals to the training set is necessary to improve prediction accuracy. As the current study suggested, the training set size can be based on the number of eigenvalues explaining a certain percentage of the variance in **G**. Those improvement patterns were very similar in both N_e20 and N_e200; however, the prediction accuracies were generally smaller for the N_e200 when the same number of genotyped animals were used. Daetwyler et al. (2010) investigated the impact of the genomic structure of the population (N_e and M_e) on the accuracy of GBLUP. In their study, with the same number of individuals in the training sets, smaller N_e showed better accuracy than larger N_e regardless of the number of QTL.

Daetwyler et al. (2008) proposed the following equation for prediction accuracy: $r_{g\hat{g}G} = \sqrt{N_p h^2 / N_p h^2 + M_e}$, where N_p is a training set size, h^2 is the trait heritability, and M_e is the number of independent chromosome segments. Equation to approximate the M_e proposed by Stam (1980) was $4N_eL$, where M_e is proportional to N_e, thus current results that showed greater accuracy with smaller N_e theoretically sounded when the N_p and h^2 are equivalent. In addition, the small size of N_e means that fewer M_e to estimate, thus a smaller prediction error variance would be estimated (Pocrnic et al., 2016a).

We selected variants based on a p-value of 0.05 with a Bonferroni correction for multiple testing and the order of the significance level; however, the Bonferroni correction might generate a stringent threshold, increasing the number of false negatives. Therefore, we tested GP by combining selected variants based on sample size (TOP*v*) and significant variants. We demonstrated that when a large training set incorporated a relatively large number of variants (i.e., twice the number of simulated QTN), prediction accuracy improved by up to 9%. Several studies used selected variants from imputed sequence data to improve GP in single-breed populations. Veerkamp et al. (2016) reported that when selected variants were used for GP, accuracy decreased,

and bias increased. However, VanRaden et al. (2017) observed an improvement in accuracy by up to 5% when 16k selected variants were added to 60k chip data. In single-breed populations, an improvement in prediction accuracy using selected variants from sequence data could be limited due to long-range LD; thus, precise identification of variants is much harder than multi-breed or across-breed (Veerkamp et al., 2016).

Fragomeni et al. (2017) outlined that including causative QTN in the unweighted **G** through ssGBLUP increased accuracies by 0.04 when the number of QTN was 100 and 1000, which was similar to our results (0.02 ~ 0.06). Additionally, when those authors added weights derived from SNP effects to **G**, accuracies increased by 0.10 and 0.03 for 100 and 1000 QTN scenarios, respectively, meaning that weighting SNP was more important for the scenario with a smaller number of QTN (less polygenic). However, in the current study, WssGBLUP resulted in no improvement in accuracy, and more inflation of GEBV was observed compared to ssGBLUP. However, the inflation of GEBV was reduced when more genotyped animals were added to the training set. The major difference between ssGBLUP and WssGBLUP is that ssGBLUP assumes that all SNP explain the same amount of genetic variance, whereas WssGBLUP assigns different variances for each SNP (Wang et al., 2012). In general, weighting **G** may not increase the accuracy of GP but may improve the GWAS resolution (Wang et al., 2012).

In our study, the resolution of variant detection was improved using at least the number of genotyped animals corresponding to the number of eigenvalues explaining 98% of the variation in **G** for the N_e200 scenarios, meaning that when the number of genotyped animals for discovery is close to the approximated M_e, precise detection of significant variants is feasible. As the genomic information has limited dimensionality, it could be expressed as the number of non-redundant SNP, genotyped animals, and M_e (Misztal, 2016). Therefore, investigating the dimensionality of

the genomic information can help determine the sample size required for discovery and training. Since the performance of GWAS and GP depends on several factors such as the genetic architecture of the trait, population structure, heritability, and sample size, more research is needed with real data to validate our results.

CONCLUSIONS

Accurately identifying causative variants from sequence data depends on the effective population size and, therefore, the dimensionality of genomic information. This dimensionality can help identify the suitable sample size for GWAS and should be considered for variant selection. Assigning genotyped animals with high breeding value reliability to the discovery set helps better identify the significant QTN. As sequence data become available, preselecting variants, and adding them to regular chip data could improve prediction accuracy if the dimensionality of the genomic information is considered; however, the improvement is mostly limited.

REFERENCES

Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta, and T. Lawlor. 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. doi:10.3168/jds.2009–2730

Baldwin-Brown, J. G., A. D. Long, and K. R. Thornton. 2014. The power to detect quantitative trait loci using resequenced, experimentally evolved populations of diploid, sexual organisms. Molecular biology and evolution. 31(4):1040-1055. doi:10.1093/molbev/msu048

Berisa, T., and J. K. Pickrell. 2016. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics. 32(2):283. doi:10.1093/bioinformatics/btv546

Boddhireddy, P., M. Kelly, S. Northcutt, K. Prayaga, J. Rumph, and S. DeNise. 2014. Genomic predictions in Angus cattle: comparisons of sample size, response variables, and clustering methods for cross-validation. Journal of animal science. 92(2):485-497. doi:10.2527/jas.2013-6757

Canty, A. J. 2002. Resampling methods in R: the boot package. The Newsletter of the R Project Volume 2:3

Cleveland, W., E. Grosse, and W. Shyu. 1992. Local regression models. Chapter 8 in Statistical models in S (JM Chambers and TJ Hastie eds.), 608 p. Wadsworth & Brooks/Cole, Pacific Grove, CA

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

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PloS one. 3(10):e3395. doi:10.1371/journal.pone.0003395

de Las Heras-Saldana, S., B. I. Lopez, N. Moghaddar, W. Park, J.-e. Park, K. Y. Chung, D. Lim, S. H. Lee, D. Shin, and J. H. van der Werf. 2020. Use of gene expression and whole-genome sequence information to improve the accuracy of genomic prediction for carcass traits in Hanwoo cattle. Genetics Selection Evolution. 52(1):1-16. doi:10.1186/s12711-020-00574-2

Fragomeni, B., D. Lourenco, A. Legarra, P. VanRaden, and I. Misztal. 2019. Alternative SNP

weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants. Journal of dairy science. 102(11):10012-10019. doi:10.3168/jds.2019-16262

Fragomeni, B. O., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2017. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. Genetics Selection Evolution. 49(1):59. doi:10.1186/s12711-017-0335-0

Gozalo-Marcilla, M., J. Buntjer, M. Johnsson, L. Batista, F. Diez, C. R. Werner, C.-Y. Chen, G. Gorjanc, R. J. Mellanby, and J. M. Hickey. 2021. Genetic architecture and major genes for backfat thickness in pig lines of diverse genetic backgrounds. Genetics Selection Evolution. 53(1):1-14. doi:10.1186/s12711-021-00671-w

Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics research. 91(1):47-60. doi:10.1017/S0016672308009981

Lopez, B. I. M., N. An, K. Srikanth, S. Lee, J.-D. Oh, D.-H. Shin, W. Park, H.-H. Chai, J.-E. Park, and D. Lim. 2021. Genomic Prediction Based on SNP Functional Annotation Using Imputed Whole-Genome Sequence Data in Korean Hanwoo Cattle. Frontiers in genetics. 1523. doi:10.3389/fgene.2020.603822

Lourenco, D., B. Fragomeni, H. Bradford, I. Menezes, J. Ferraz, I. Aguilar, S. Tsuruta, and I. Misztal. 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. Journal of Animal Breeding and Genetics. 134(6):463-471. doi:10.1111/jbg.12288

MacLeod, A., C. Haley, J. Woolliams, and P. Stam. 2005. Marker densities and the mapping of ancestral junctions. Genetics Research. 85(1):69-79. doi:10.1017/s0016672305007329

MacLeod, I., P. Bowman, C. Vander Jagt, M. Haile-Mariam, K. Kemper, A. Chamberlain, C. Schrooten, B. Hayes, and M. Goddard. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC genomics. 17(1):1-21. doi:10.1186/s12864-016-2443-6

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

Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. Genetics. 202(2):401-409. doi:10.1534/genetics.115.182089

Misztal, I., I. Pocrnic, and D. Lourenco. 2021. Factors Influencing Accuracy of Genomic Selection with Sequence Information. Journal of animal science . p 20-20. doi:10.1093/jas/skab235.034

Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, and Z. Vitezica. 2014. Manual for BLUPF90 family of programs. Athens: University of Georgia

Moghaddar, N., M. Khansefid, J. H. van der Werf, S. Bolormaa, N. Duijvesteijn, S. A. Clark, A. A. Swan, H. D. Daetwyler, and I. M. MacLeod. 2019. Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. Genetics Selection Evolution. 51(1):72. doi:10.1186/s12711-019-0514-2

Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, and H. W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution. 41(1):56. doi:10.1186/1297-9686-41-56

Pérez-Enciso, M., J. C. Rincón, and A. Legarra. 2015. Sequence-vs. chip-assisted genomic selection: accurate biological information is advised. Genetics Selection Evolution.

47(1):1-14. doi:10.1186/s12711-015-0117-5

Pocrnic, I., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016a. The dimensionality of genomic information and its effect on genomic prediction. Genetics. 203(1):573-581. doi:10.1534/genetics.116.187013

Pocrnic, I., D. A. Lourenco, Y. Masuda, and I. Misztal. 2016b. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. Genetics Selection Evolution. 48(1):1-9. doi:10.1186/s12711-016-0261-6

Pocrnic, I., D. A. Lourenco, Y. Masuda, and I. Misztal. 2019. Accuracy of genomic BLUP when considering a genomic relationship matrix based on the number of the largest eigenvalues: a simulation study. Genetics Selection Evolution. 51(1):1-10. doi:10.1186/s12711-019-0516-0

Sargolzaei, M., and F. S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25(5):680-681. doi:10.1093/bioinformatics/btp045

Stam, P. 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genetics Research. 35(2):131-155. doi:10.1017/S0016672300014002.

Takeda, M., Y. Uemoto, and M. Satoh. 2020. Effect of genotyped bulls with different numbers of phenotyped progenies on quantitative trait loci detection and genomic evaluation in a simulated cattle population. Animal Science Journal. 91(1):e13432. doi:10.1111/asj.13432

van den Berg, I., S. Fritz, and D. Boichard. 2013. QTL fine mapping with Bayes C (π): a simulation study. Genetics Selection Evolution. 45(1):1-11. doi:10.1186/1297-9686-45-19

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. Journal of dairy science. 91(11):4414-4423. doi:10.3168/jds.2007-0980

VanRaden, P. M., M. E. Tooker, J. R. O'connell, J. B. Cole, and D. M. Bickhart. 2017. Selecting

sequence variants to improve genomic predictions for dairy cattle. Genetics Selection Evolution. 49(1):32. doi:10.1186/s12711-017-0307-4

Veerkamp, R. F., A. C. Bouwman, C. Schrooten, and M. P. Calus. 2016. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. Genetics Selection Evolution. 48(1):95. doi:10.1186/s12711-016-0274-1

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 2017. 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics. 101(1):5-22. doi:10.1016/j.ajhg.2017.06.005

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics Research. 94(2):73-83. doi:10.1017/S0016672312000274

Zhang, C., R. A. Kemp, P. Stothard, Z. Wang, N. Boddicker, K. Krivushin, J. Dekkers, and G. Plastow. 2018. Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. Genetics Selection Evolution. 50(1):1-13. doi:10.1186/s12711-018-0387-9

Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal. 2016. Weighting strategies for single-step genomic BLUP: an iterative approach for accurate calculation of GEBV and GWAS. Frontiers in genetics. 7:151. doi:10.3389/fgene.2016.00151

Zhou, X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. Nature genetics. 44(7):821-824. doi:10.1038/ng.2310

TABLES

Table 5.1. Description of all GWAS scenarios

Scenario description	Ne	Number of QTN	Heritability
Ne20 Q200 H30	20	200	0.3
Ne20 Q200 H90	20	200	0.9
Ne20 Q200 H99	20	200	0.99
Ne20 Q2000 H30	20	2000	0.3
Ne20 Q2000 H90	20	2000	0.9
Ne20 Q2000 H99	20	2000	0.99
Ne200 Q200 H30	200	200	0.3
Ne200 Q200 H90	200	200	0.9
Ne200 Q200 H99	200	200	0.99
Ne200 Q2000 H30	200	2000	0.3
Ne200 Q2000 H90	200	2000	0.9
Ne200 Q2000 H99	200	2000	0.99

Table 5.2. Number of genotyped animals for all scenarios in both discovery and training sets

(a) Ne20

	Ne20 Q200 H30	Ne20 Q200 H90	Ne20 Q200 H99	Ne20 Q2000 H30	Ne20 Q2000 H90	Ne20 Q2000 H99
EIG50	80	80	80	80	80	80
EIG60	130	140	140	140	140	130
EIG70	220	240	230	220	230	220
EIG80	390	410	410	400	400	400
EIG90	860	900	890	890	880	840
EIG95	1,700	1,770	1,800	1,800	1,750	1,700
EIG98	4,000	4,100	4,100	4,100	4,100	4,000
EIG99	6,900	7,100	7,100	7,100	7,000	6,900
All	30,000	30,000	30,000	30,000	30,000	30,000

(b) Ne200

	Ne200 Q200 H30	Ne200 Q200 H90	Ne200 Q200 H99	Ne200 Q2000 H30	Ne200 Q2000 H90	Ne200 Q2000 H99
EIG50	510	520	550	530	500	510
EIG60	900	910	940	920	900	900
EIG70	1,500	1,550	1,600	1,540	1,500	1500
EIG80	2,600	2,650	2,700	2,650	2,600	2,600
EIG90	5,100	5,250	5,300	5,300	5,100	5,100
EIG95	8,600	8,800	8,800	8,800	8,600	8,600
EIG98	15,000	15,200	15,200	15,200	15,000	15,000
EIG99	21,000	22,000	21,400	22,000	22,000	21,000
All	30,000	30,000	30,000	30,000	30,000	30,000

Table 5.3. Approximated sample size based on local polynomial regression and proposed equation using 'ALL' as benchmark

Scenario	Heritability	%Var[1]	SS_{pol}[2]	$EIGx_{app1}$[3]	SS_{rel}[4]	$EIGx_{app2}$[5]	Diff[6]
Ne20 Q200	0.3	44.9	30000				
	0.9	44.9	2223	EIG95~98 (1770~4100)	2698	EIG95~98 (1770~4100)	475
	0.99	44.9	1662	EIG90~95 (890~1800)	2453	EIG95~98 (1800~4100)	791
Ne20 Q2000	0.3	3.9	30000				
	0.9	3.9	3049	EIG95~98 (1750~4100)	3007	EIG95~98 (1750~4100)	-42
	0.99	3.9	2378	EIG95~98 (1700~4000)	2734	EIG95~98 (1700~4000)	356
Ne200 Q200	0.3	77.4	30000				
	0.9	77.4	5205	EIG80~90 (2650~5250)	4324	EIG80~90 (2650~5250)	-881
	0.99	77.4	3814	EIG80~90 (2700~5300)	3931	EIG80~90 (2700~5300)	117
Ne200 Q2000	0.3	13.9	30000				
	0.9	13.9	4524	EIG80~90 (2600~5100)	4719	EIG80~90 (2600~5100)	195
	0.99	13.9	3664	EIG80~90 (2600~5100)	4290	EIG80~90 (2600~5100)	626

*%Var1: percentage of variance explained by significantly identified QTN
*SS_{pol}2: Approximated sample size using local polynomial regression
*$EIGx_{app1}$3: EIGx scenario range including $Sample_{app1}$
*SS_{rel}4: Approximated sample size using proposed equation
*$EIGx_{app2}$5: EIGx scenario range including $Sample_{app2}$
*Diff6: Difference between $Sample_{app2}$ and $Sample_{app1}$
*Dotted arrows in Fig. 2 matched with each line of H90 and H99 scenario was approximated sample size using local polynomial regression and those were described above as $Sample_{app1}$.
*EIGx scenario ranges including either $Sample_{app1}$ ($EIGx_{app1}$) or $Sample_{app2}$ ($EIGx_{app2}$) are described with the difference between $Sample_{app1}$ and $Sample_{app2}$ as $Sample_{app2} - Sample_{app1}$ (Diff) in Table

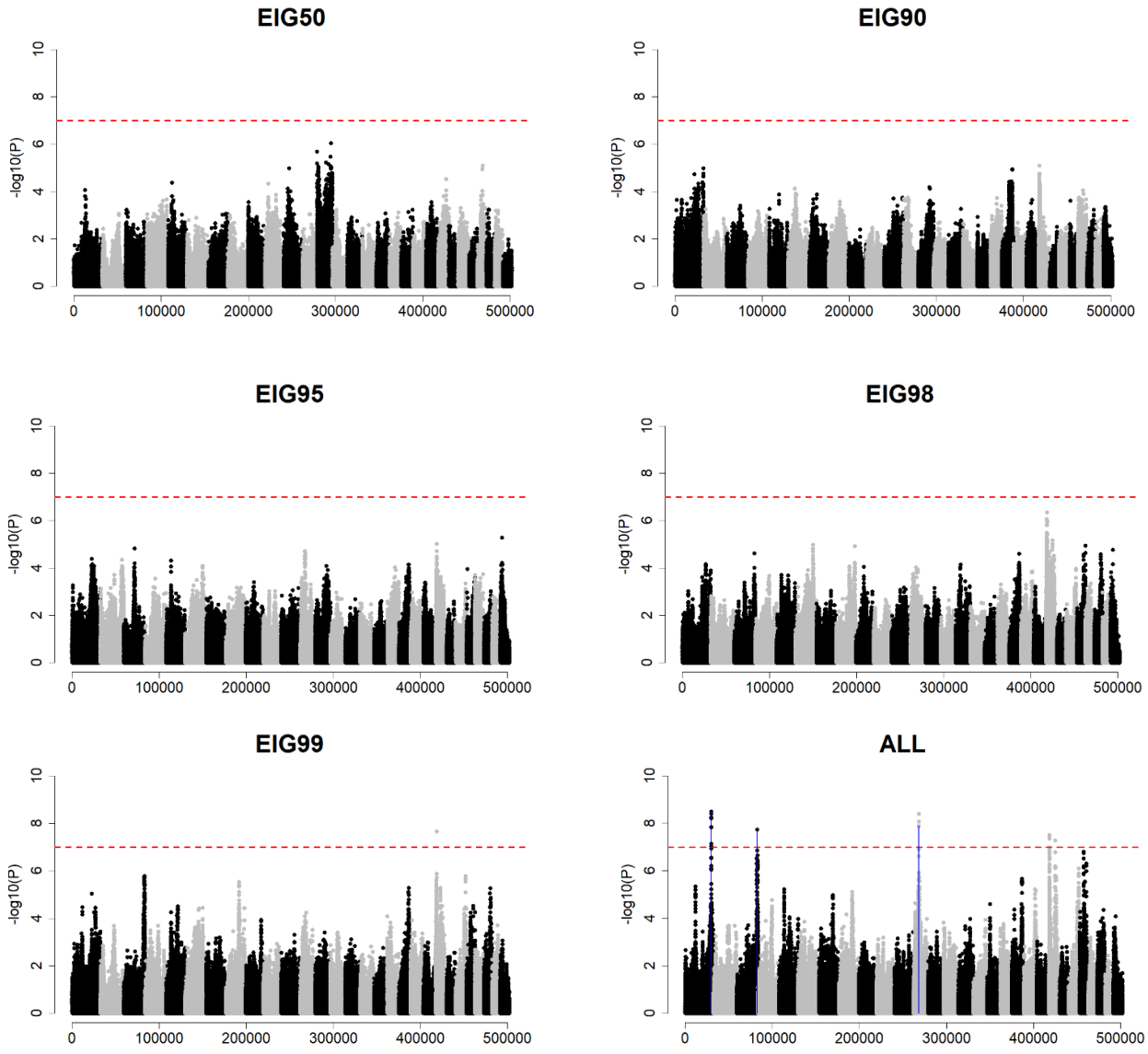


Figure 5.1. GWAS results – EIG50, EIG90, EIG95, EIG98, EIG99, All – Ne20 Q2000 H30

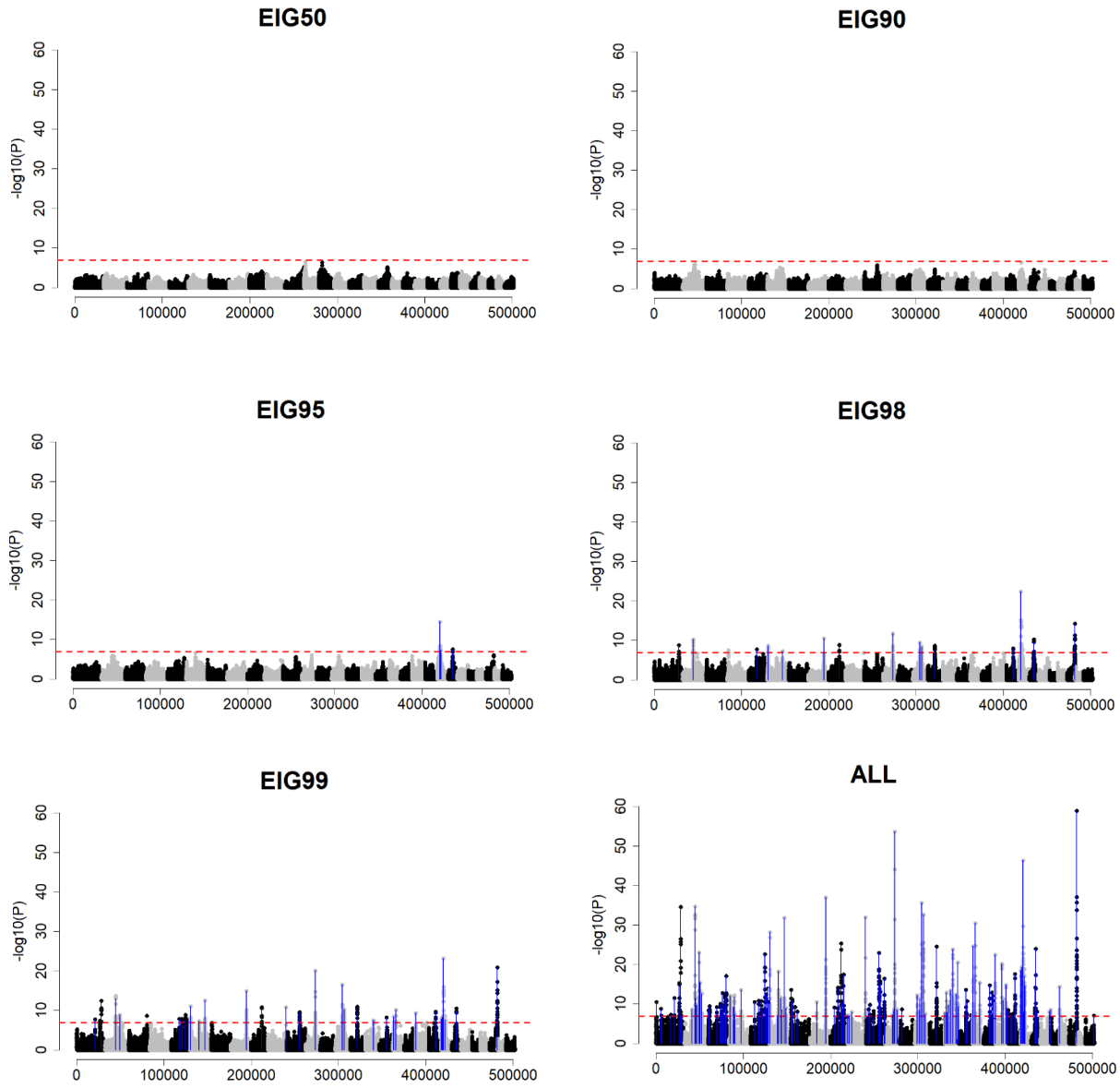


Figure 5.2. GWAS results – EIG50, EIG90, EIG95, EIG98, EIG99, All – Ne20 Q2000 H99

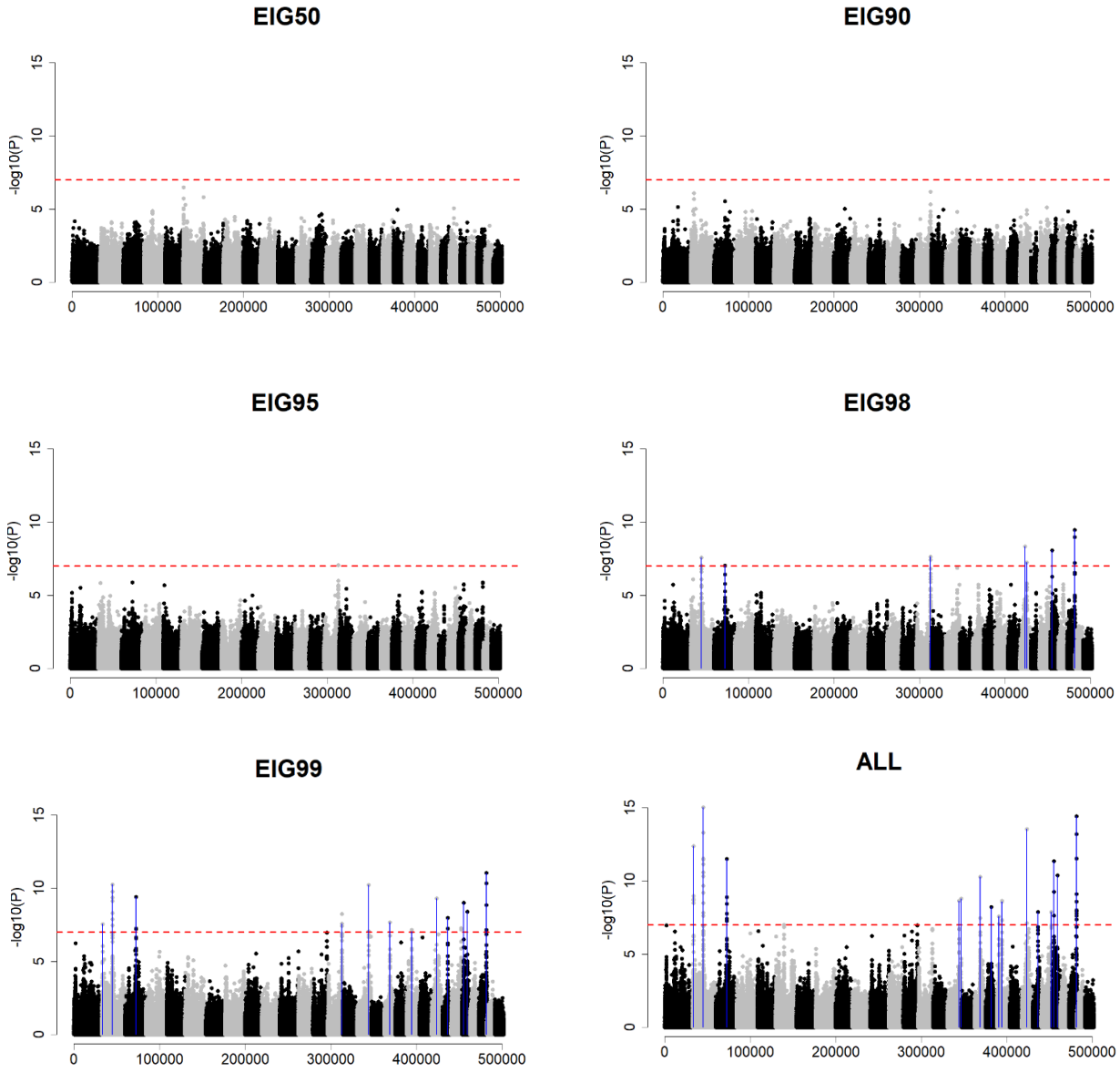


Figure 5.3. GWAS results – EIG50, EIG90, EIG95, EIG98, EIG99, All – Ne200 Q2000 H30

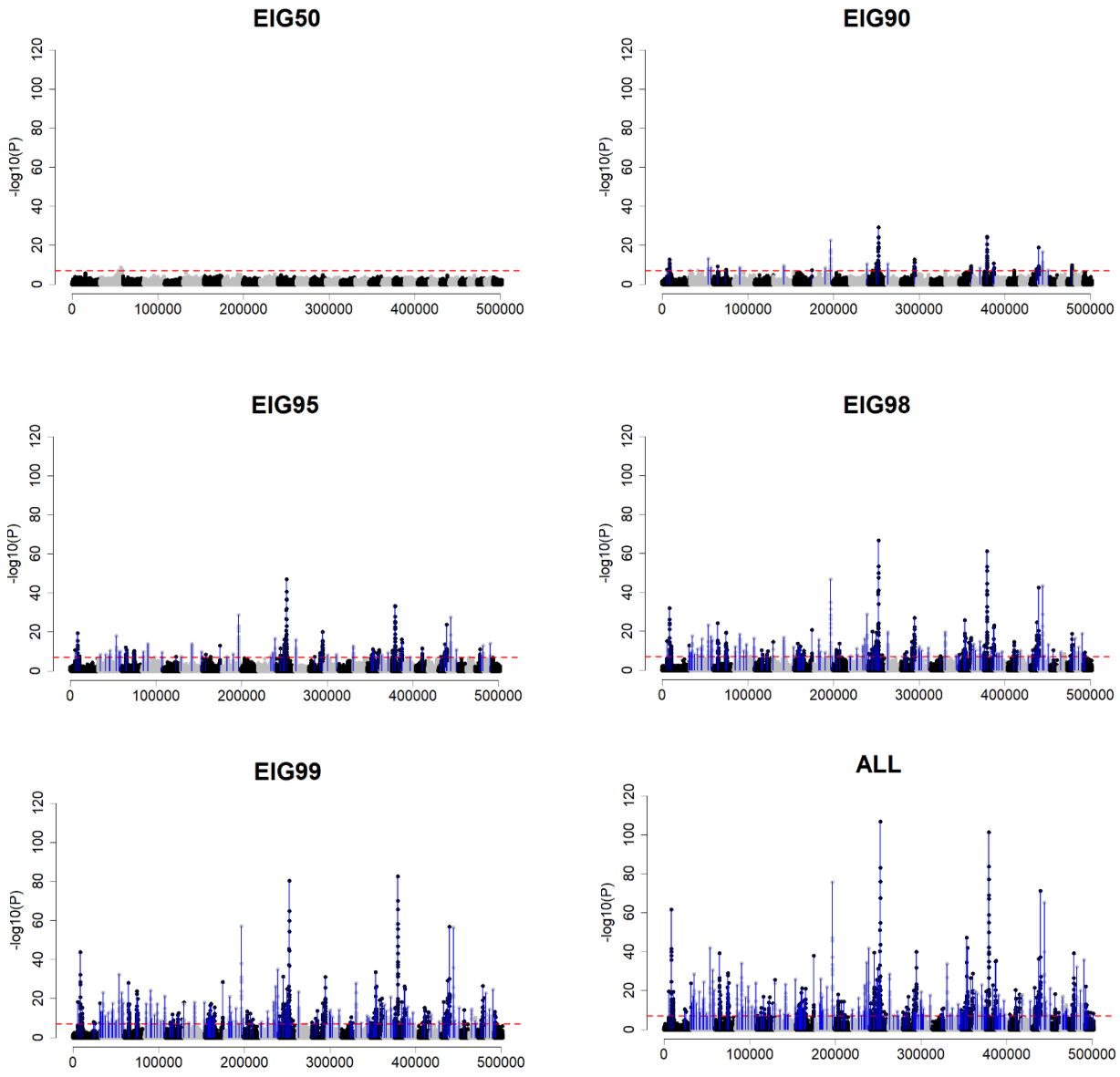
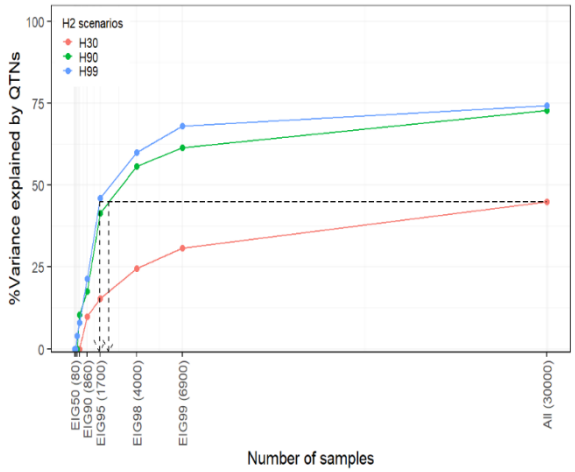
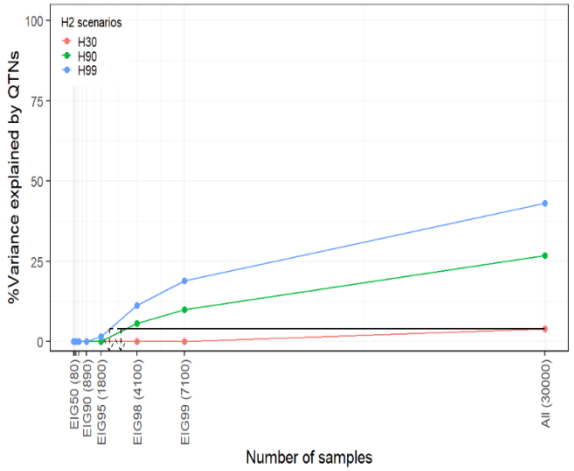


Figure 5.4. GWAS results – EIG50, EIG90, EIG95, EIG98, EIG99, All – Ne200 Q2000 H99

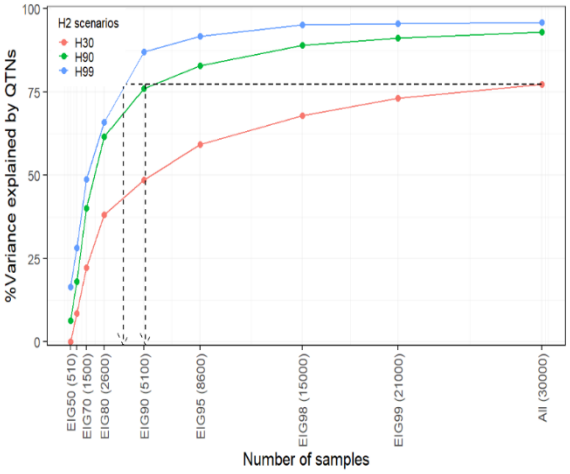
(a) Ne20 Q200



(b) Ne20 Q2000



(c) Ne200 Q200



(d) Ne200 Q2000

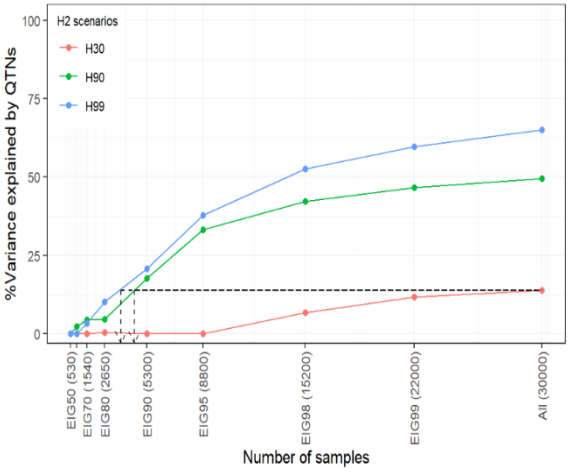


Figure 5.5. Total variance explained by significant QTN across the different sample sizes and heritabilities

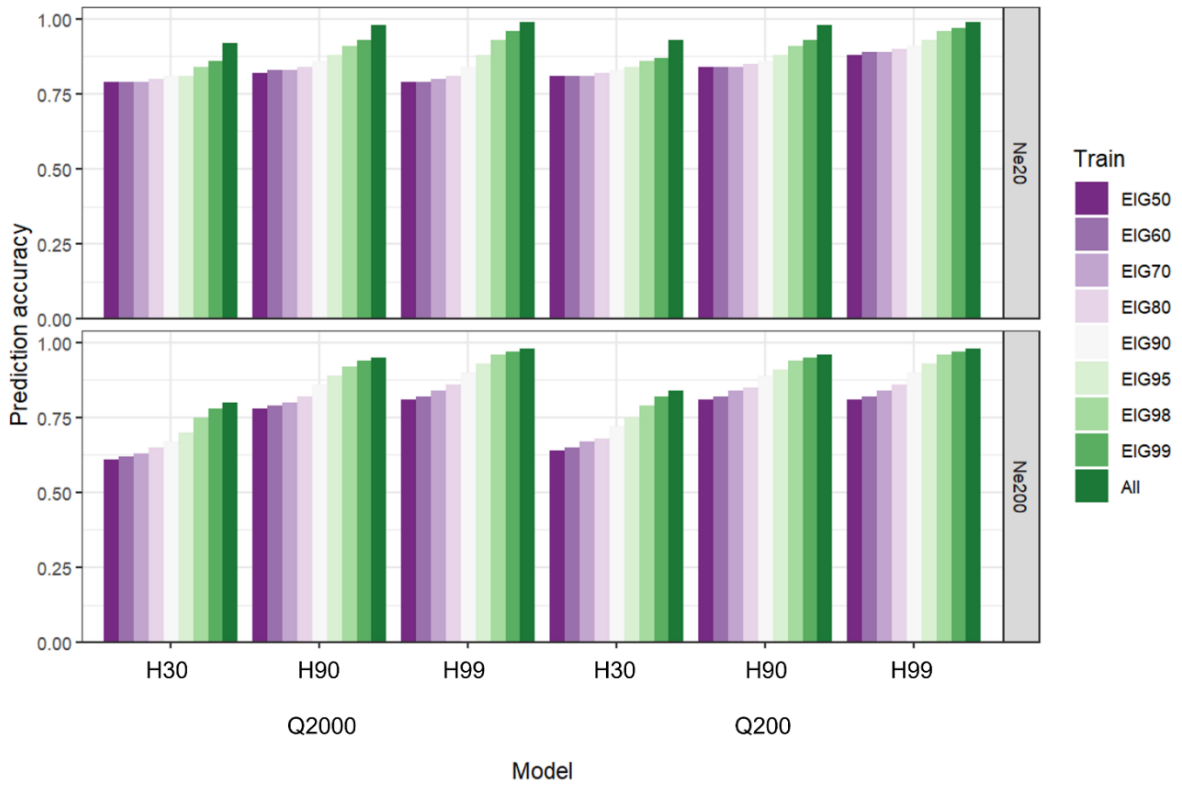


Figure 5.6. Prediction accuracy

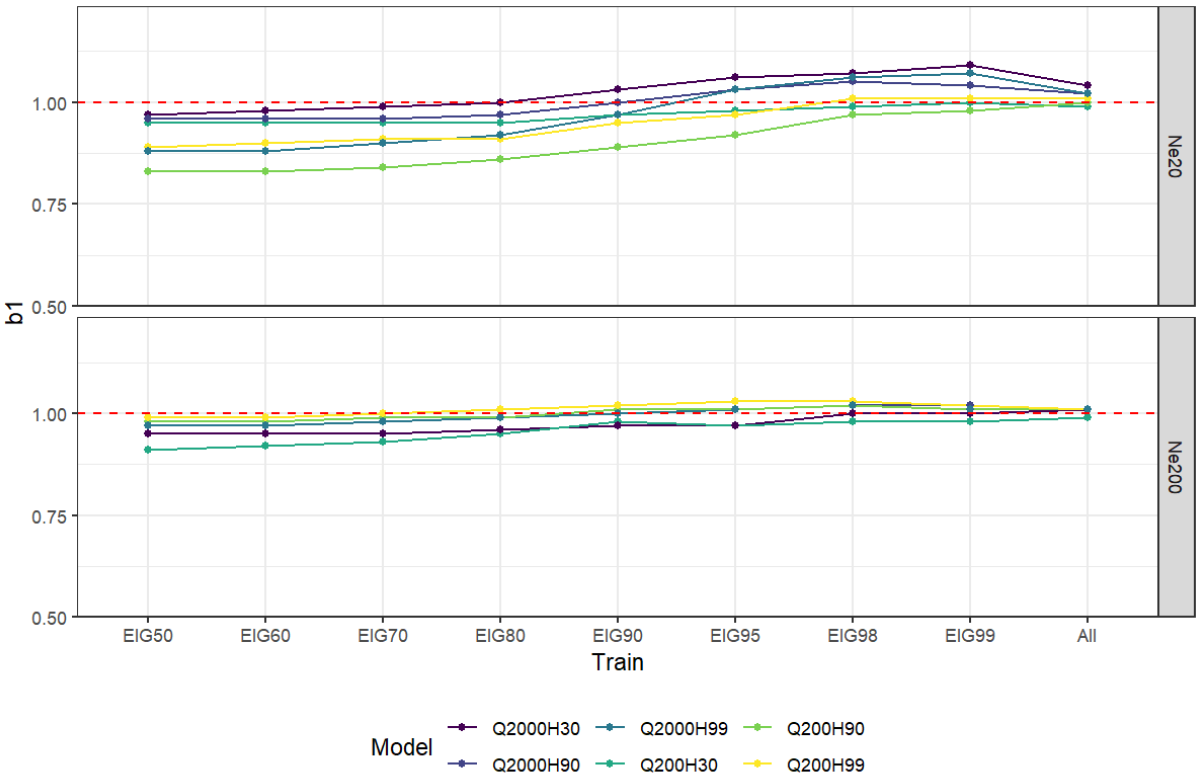
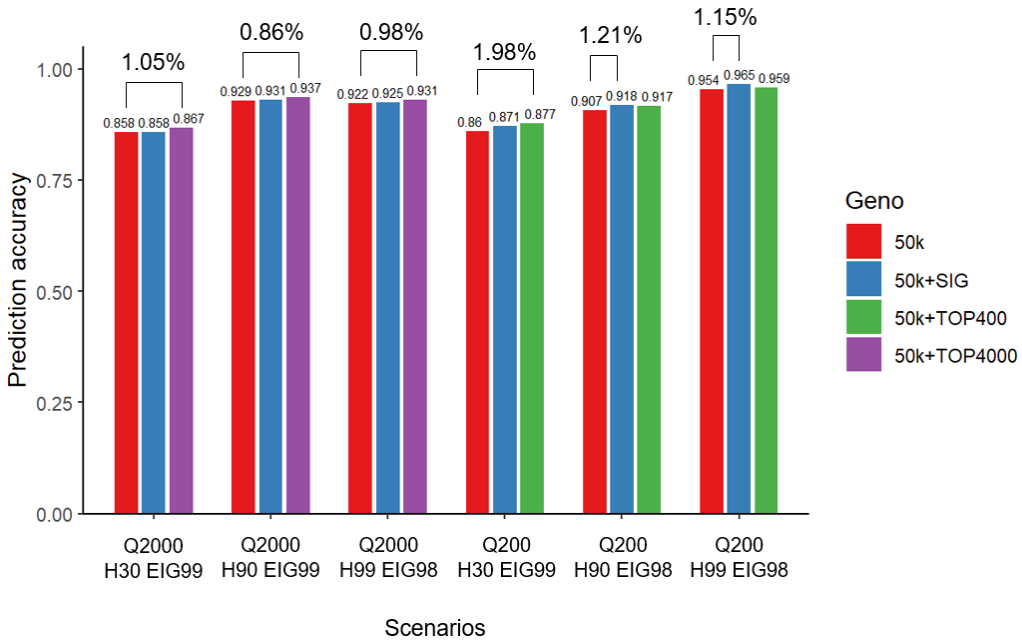


Figure 5.7. Regression coefficients (b1)

(1) Ne20



(2) Ne200

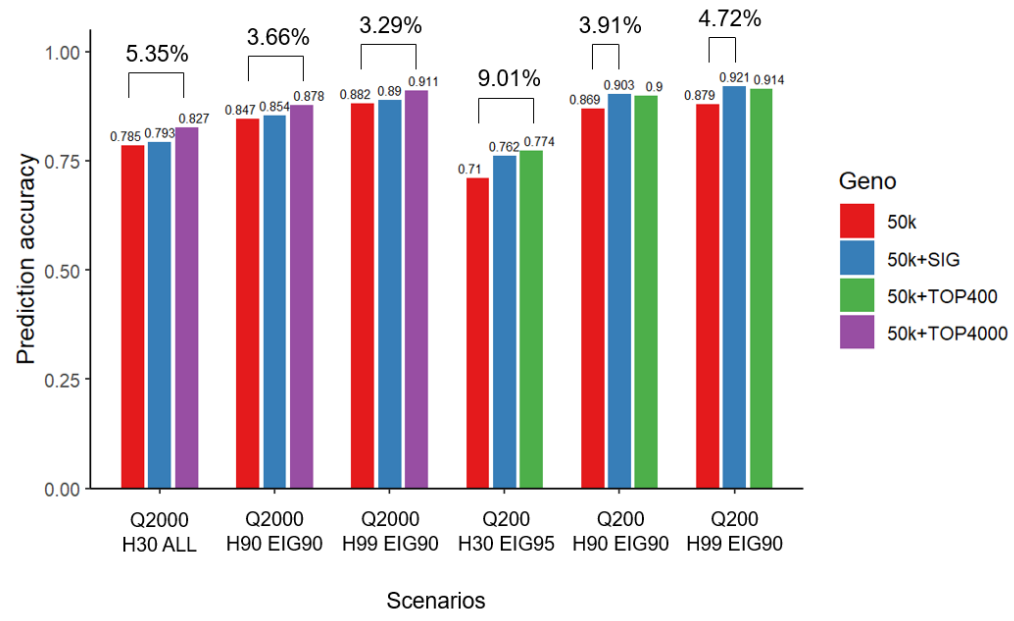


Figure 5.8. Prediction accuracy of 50k and TOPv scenario which showed a maximum gain

CHAPTER 6

CONCLUSIONS

Preselected significant variants from whole-genome sequence can help improve the accuracy of genomic predictions in large maternal and terminal pig lines, although the advantage is limited. The benefits of using different preselected variant sets depend on the genetic architecture of traits, lines, size of the genotyped population, and how those variants were selected. Multi-line genomic evaluation in pigs with the addition of unknown parent groups or metafounders to account for the genetic differences could improve the prediction accuracy, bias, and dispersion of GEBV compared to the single-line genomic evaluation although the improvement is limited. Using preselected variant sets from whole-genome sequence on genomic prediction did not outperform Chip, showing varied results depending on the lines and traits. Weighted ssGBLUP for multi-line evaluations did not show considerable improvements. Accurately identifying causative variants from sequence data depends on the effective population size and, therefore, the dimensionality of genomic information. This dimensionality of genomic information can help identify the suitable sample size for genome-wide association and should be considered for variant selection. Assigning genotyped animals with high breeding value reliability to the discovery set helps better identify the significant variants. As sequence data become available, preselecting variants, and adding them to regular chip data could improve prediction accuracy if the dimensionality of the genomic information is considered; however, the improvement is mostly limited.

APPENDIX A

INCLUSION OF SIRE BY HERD INTERACTION EFFECT IN THE GENOMIC EVALUATION FOR WEANING WEIGHT OF AMERICAN ANGUS[1]

[1] Sungbong Jang, Daniela Lourenco, and Stephen Miller. *Journal of Animal Science*. 100(3). Reprinted here with permission of the publisher.

ABSTRACT

A spurious negative genetic correlation between direct and maternal effects of weaning weight (WW) in beef cattle has historically been problematic for researchers and industry. Previous research has suggested the covariance between sires and herds may be contributing to this relationship. The objective of this study was to estimate the variance components (VC) for WW in American Angus with and without sire by herd (SxH) interaction effect when genomic information is used or not. Five subsets of approximately 100k animals for each subset were used. When genomic information was included, genotypes were added for 15,637 animals. Five replicates were performed. Four different models were tested, namely, M1: without SxH interaction effect and with covariance between direct and maternal effect (σ_{am}) ≠ 0; M2: with SxH interaction effect and σ_{am} ≠ 0; M3: without SxH interaction effect and with σ_{am} = 0; M4: with SxH interaction effect and σ_{am} = 0. VC were estimated using the restricted maximum likelihood (REML) and single-step genomic REML (ssGREML) with the average information algorithm. Breeding values were computed using single-step genomic BLUP (ssGBLUP) for the models above and one additional model, which had the covariance zeroed after the estimation of VC (M5). The ability of each model to predict future breeding values was investigated with the linear regression method. Under REML, when the SxH interaction effect was added to the model, both direct and maternal genetic variances were greatly reduced, and the negative covariance became positive (i.e., when moving from M1 to M2). Similar patterns were observed under ssGREML, but with less reduction in the direct and maternal genetic variances and still a negative covariance. Models with the SxH interaction effect (M2 and M4) had a better fit according to the Akaike Information Criteria (AIC). Breeding values from those models were more accurate and had less bias than the other three models. The rankings and breeding values of Artificial Insemination (AI)

sires (N = 1,977) greatly changed when the SxH interaction effect was fit in the model. Although the SxH interaction effect accounted for 3% to 5% of the total phenotypic variance and improved the model fit, this change in the evaluation model will cause severe reranking among animals.

INTRODUCTION

In beef cattle, the genetic covariance between the direct and maternal effects of weaning weight has shown an antagonistic effect that hinders the progress in a selection program (Meyer, 1992; Pollak et al., 1994). Several simulation studies reported this antagonistic estimate could arise from ignoring the additional variance among sires such as SxH and sire by year interaction effects (Robinson, 1996; Lee and Pollak, 1997). In Australian beef cattle, various studies reported significant SxH or sire by herd-year interaction effects for many traits accounted for approximately 5 to 10% of the phenotypic variation. Additionally, including the SxH interaction effect greatly reduced the negative covariance between direct and maternal effects on 200-day weight (Notter et al., 1992; Bradfield, 1999; Meyer and Graser, 1999). As a result, the Australian evaluation system, BREEDPLAN, began to include SxH interaction effect in its national evaluation model in 1999 (Graser et al., 1999).

The major reasons for the variation due to SxH interaction have not been completely determined, but several possible sources are reported: (1) preferential treatment, (2) non-random mating, (3) use of selected sires, which could lead to heterogeneous residual and additive genetic variance among herds, and (4) extensive use of specific sires in particular herds. Therefore, ignoring SxH interaction effect in the evaluation model could inflate the genetic variance and overestimate the estimated breeding value (EBV) (Tong et al., 1977; Meyer, 1987; Banos and Shook, 1990). When the SxH interaction effect was fit in the genetic evaluation model, the direct

and maternal variances were lower compared to the model without the SxH interaction effect (Baschnagel et al., 1999; Dodenhoff et al., 1999). Specifically, Dodenhoff et al. (1999) used data from the American Angus Association (AAA; St. Joseph, MO) and recommended the inclusion of the SxH interaction effect in routine genetic evaluations to avoid biased estimates.

The estimation of VC has been mostly computed using the pedigree relationship matrix (**A**). If the population is undergoing selection based on pedigree and phenotypes with a proper model, the VC based on those two sources of information would be unbiased (Kennedy et al., 1988). However, genomic information is now available and used for selection, so adding this source of information to VC estimation models makes sense. It is common fact in livestock populations that only a fraction of animals are genotyped, so using a genomic relationship matrix (**G**) instead of **A** could result in biased VC because the information on non-genotyped animals would not be used; therefore, the population would not be well represented (Cesarani et al., 2019). Veerkamp et al. (2011); Cesarani et al. (2019) recommended using the single-step methodology (Aguilar et al., 2010; Christensen and Lund, 2010) to estimate VC when genotyped and non-genotyped animals coexist in the pedigree. In single-step, **G** and **A** are combined into a realized relationship matrix (**H**), so the information on genotyped and non-genotyped animals can be used.

Because the estimated VC could differ with the choice of the covariance structure among animals and the presence of the SxH interaction, the EBV can also change causing animals to change rank. Changes in EBV and the ranking of animals are problematic in the commercial marketplace. However, if those changes are moving EBV in the appropriate direction, the modifications should be acceptable. Because most of the routine genetic evaluations ignore the negative covariance between additive direct and maternal effects, room for improvements could be explored if SxH interaction is deemed important. Therefore, the first objective of this study was

to investigate the impact of a random SxH interaction effect on the VC of WW in American Angus cattle in the presence or absence of genomic information. The second objective was to evaluate the prediction models in terms of accuracy, bias, and dispersion using the LR method (Legarra and Reverter, 2018). The last objective was to investigate the changes in EPD and the ranking of AI sires among different models.

MATERIALS AND METHODS

Animal care and Use Committee approval was not needed because the information was obtained from the pre-existing databases.

Data

All datasets were provided by AAA. Over 9.4 million WW phenotypes collected from 1955 to 2020 were available for almost 9.9 million animals. All WW were pre-adjusted for the age of dam and age of calf using the adjustment factors from the standard AAA national cattle evaluation. Data filtering for the VC estimation was performed to remove the following: 1) animals without WW and herd information; 2) contemporary groups with less than 50 animals; 3) animals with registration ID other than AAA and beef improvement records (BIR). After all filtering processes, 2,474,202 animals remained. Five random samples of approximately 100k animals with WW records were taken for the analysis, which mimics the current procedures and data structures for VC estimation by AAA. Each sample contained all animals in the selected herd over time. Table 1 depicts summary statistics for WW along with the number of animals, herds, sires, and SxH interactions in each replicate.

Among those animals, 180,733 were genotyped. Because of the computing limitation of ssGREML, a subset of 15,637 animals born from 1972 to 2017 was selected among 180,733

genotyped animals who had phenotypes for WW and at least one progeny as sire or dam. The animals were genotyped for 54,609 single nucleotide polymorphisms (SNP) originally present in the BovineSNP50k v2 BeadChip (Illumina Inc., San Diego, CA). Quality control of genomic data removed SNP with call rate < 0.9, minor allele frequency < 0.05, and those located on the sex chromosomes. After the quality control, 39,733 SNPs were available for animals born from 1972 to 2017. For the estimation of breeding values, a larger dataset was used which included phenotypes for 2,474,202 animals, 180,733 genotyped animals, and a four-generation pedigree including 869,583 animals in total. Because of the large number of genotyped animals, the algorithm for proven and young (APY) was used to obtain $\mathbf{G^{-1}}$ without the direct inversion of $\mathbf{G}$, as proposed by Misztal et al. (2014a). The number of core animals was set to 19,019, which has been used for routine genomic evaluations by the AAA. Among all those animals, 1,977 were AI sires under investigation for ranking and EPD changes under different models. AI sires in this data are a combination of old sires with many progenies and young sires with no progeny in production yet. These AI sires had direct progeny ranging from 0 to 6,053 with a mean of 117.02 and the number of progenies raised by daughters ranged from 0 to 19 with a mean of 0.23.

Models and Analysis

The following four different linear mixed models were used for the VC estimation.

M1: $\mathbf{y = Xb + Z_1a + Z_2m + Z_3mpe + e}$

M2: $\mathbf{y = Xb + Z_1a + Z_2m + Z_3mpe + Z_4sh + e}$

M3: $\mathbf{y = Xb + Z_1a + Z_2m + Z_3mpe + e}$; with $\sigma_{am} = 0$

M4: $\mathbf{y = Xb + Z_1a + Z_2m + Z_3mpe + Z_4sh + e}$; with $\sigma_{am} = 0$

where **y** is a vector of WW records; **b** is a vector of the fixed effects of contemporary group (CG), where CG was composed to represent animals of the same sex, born and weaned in the same herd,

in the same year and part of the same management group within that herd; **a**, **m**, and **mpe** are random vectors of additive direct genetic effect, additive maternal genetic effect, and maternal permanent environmental effect, respectively; **sh** is a random vector of SxH interaction effect as an additional uncorrelated random effect; $\mathbf{X}$, $\mathbf{Z_1}$, $\mathbf{Z_2}$, $\mathbf{Z_3}$, and $\mathbf{Z_4}$ are the incidence matrices for the effects in **b**, **a**, **m**, **mpe**, and **sh**, respectively; **e** is the vector of random residuals. Hence, variances for the random effects in models M1 and M3 were:

$$\mathrm{V}\begin{pmatrix}\mathbf{a}\\ \mathbf{m}\\ \mathbf{mpe}\\ \mathbf{e}\end{pmatrix}=\begin{pmatrix}\mathbf{A}\sigma_a^2 & \mathbf{A}\sigma_{am} & 0 & 0\\ \mathbf{A}\sigma_{am} & \mathbf{A}\sigma_m^2 & 0 & 0\\ 0 & 0 & \mathbf{I}\sigma_{mpe}^2 & 0\\ 0 & 0 & 0 & \mathbf{I}\sigma_e^2\end{pmatrix}$$

where **A** and **I** denote pedigree relationship and identity matrices; under single-step (i.e., ssGBLUP and ssGREML), the realized relationship matrix (**H**) was used instead of **A**. Models M1 and M2 considered covariance between direct and maternal effects, whereas M3 and M4 forced this covariance to zero.

Models M2 and M4 had a random SxH interaction effect, so the variance structure for the random effects was:

$$\mathrm{V}\begin{pmatrix}\mathbf{a}\\ \mathbf{m}\\ \mathbf{mpe}\\ \mathbf{sh}\\ \mathbf{e}\end{pmatrix}=\begin{pmatrix}\mathbf{A}\sigma_a^2 & \mathbf{A}\sigma_{am} & 0 & 0 & \mathbf{0}\\ \mathbf{A}\sigma_{am} & \mathbf{A}\sigma_m^2 & 0 & 0 & 0\\ 0 & 0 & \mathbf{I}\sigma_{mpe}^2 & 0 & 0\\ 0 & 0 & 0 & \mathbf{I}\sigma_{sh}^2 & 0\\ 0 & 0 & 0 & 0 & \mathbf{I}\sigma_e^2\end{pmatrix}$$

Phenotypic variance (σ_p^2) was computed based on all the variances in each model. For example, in M2 and M4:

$$\sigma_p^2=\sigma_a^2+\sigma_m^2+\sigma_{am}+\sigma_{mpe}^2+\sigma_{sh}^2+\sigma_e^2$$

Therefore, the direct and maternal heritabilities were estimated as:

$$h_a^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_m^2+\sigma_{am}+\sigma_{mpe}^2+\sigma_{sh}^2+\sigma_e^2},$$

$$h_m^2 = \frac{\sigma_m^2}{\sigma_a^2+\sigma_m^2+\sigma_{am}+\sigma_{mpe}^2+\sigma_{sh}^2+\sigma_e^2}$$

where σ_a^2, σ_m^2, σ_{am}, σ_{mpe}^2, σ_{sh}^2, and σ_e^2 are additive genetic direct variance, maternal genetic variance, the covariance between direct and maternal genetic effects, maternal permanent environment variance, SxH variance, and residual variance, respectively. The formulas for heritability had no σ_{sh}^2 for M1 and M3, and σ_{am} was zero for M3 and M4.

Two methods were used to estimate VC, which included REML and ssGREML. In REML, the assumption was $\mathbf{a} \sim N(0, \mathbf{A}\sigma_a^2)$, where $\mathbf{A}$ is the pedigree relationship matrix. Conversely, the assumption under ssGREML was $\mathbf{a} \sim N(0, \mathbf{H}\sigma_a^2)$, where $\mathbf{H}$ is the realized relationship matrix combining $\mathbf{A}$ with the genomic relationship matrix ($\mathbf{G}$). In the ssGREML algorithm, the inverse of $\mathbf{H}$ is required (Aguilar et al., 2010):

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

VC were estimated using the average information (AI) REML algorithm as implemented in AIREMLF90 (Misztal et al., 2014b), which has been modified to incorporate the YAMS package (Masuda et al., 2015) for optimized sparse matrix computations. Genomic EBV (GEBV) was estimated for all four models using ssGBLUP. One additional model was used as a benchmark, mimicking the current procedure in the AAA evaluations. This model was labeled model 5 (M5), and was similar to M1, except for the covariance between direct and maternal effects was zeroed after the VC estimation. As our objective herein was to compare genomic predictions between the models, not between methods, only ssGBLUP evaluations were carried out. Akaike Information Criteria (AIC) was used to compare models. In most cases, VC from non-genomic models are used

to obtain genomic predictions; however, in this study, GEBV were also computed using VC from genomic models. The VC used were averaged across five replicates. Changes in ranking and predictions for AI bulls were presented in the EPD scale, which was computed as one-half EBV. Ranking changes were calculated by comparing the ranking of animals in M1 to M4 against M5; the same was done for investigating EPD changes.

Validation

The LR validation method (Legarra and Reverter, 2018) was used to evaluate model performance. A total of 23,021 young genotyped animals born in 2019 were selected as validation animals and had their phenotypes removed from the evaluation, along with phenotypes for their contemporaries. The total number of records in this dataset was 2,451,181. This will be referred to as the partial data and will be represented by the subscript *p*. On the other hand, the entire data will be represented by the subscript *w* and had no phenotype truncation. Under the LR method, the accuracy of GEBV was calculated as $\widehat{acc} = \sqrt{\frac{cov(\hat{\boldsymbol{a}}_w,\ \hat{\boldsymbol{a}}_p)}{(1-\bar{F})\hat{\sigma}_a^2}}$, where a is the vector of GEBV and $\bar{F}$ is the average inbreeding coefficient for validation animals; $\hat{\sigma}_a^2$ was model-specific under REML or ssGREML. Bias was calculated as the difference between the mean of partial and whole GEBV, which is $\mu_{w,p} = \bar{\hat{a}}_p - \bar{\hat{a}}_w$, with an expected estimator of 0 if unbiased. Dispersion of GEBV was assessed as the deviation of the regression coefficient (b_1) from 1, where b_1 was obtained from the regression of $\hat{\boldsymbol{a}}_w$ on $\hat{\boldsymbol{a}}_p$: $\hat{\boldsymbol{a}}_w = b_0 + b_1\hat{\boldsymbol{a}}_p$. Under the condition of neither over nor under dispersion, the expectation of this estimator would be 1.

RESULTS AND DISCUSSION

Genetic parameter estimation

VC can be estimated considering the covariance structure among animals is given by the pedigree relationship matrix, the genomic relationship matrix, or by the realized relationship matrix. In this study, the first and third assumptions were used to examine the differences in VC when the pedigree information is combined with genomic information or not (Table 2). Under REML, M1 resulted in larger direct and maternal genetic variances compared to the other 3 models. In addition, M1 had greater negative covariance between direct and maternal genetic effects compared to M2. When SxH interaction was fit into the model (M2), both direct and maternal genetic variances were reduced by a ratio of almost 2.3 and 1.6, respectively. However, the residual variance was 16% greater in M2 compared to M1. Remarkably, the negative covariance between direct and maternal effects became positive when moving from M1 to M2. Therefore, adding the SxH interaction effect could mitigate the issue with negative covariance between direct and maternal effects.

Baschnagel et al. (1999); Dodenhoff et al. (1999) also reported larger estimates of direct and maternal genetic variance and negative covariance between those effects when the SxH interaction effect was not fit in the models. Meyer (1992) outlined that a negative estimate of covariance between direct and maternal effects increased both direct and maternal genetic variances in crosses between Hereford and Zebu cattle, but the same was not true in Angus because the covariance was positive. In the current study, M1 showed negative covariance between direct and maternal effects as well as larger estimates of direct and maternal genetic variance among all the models. Nonetheless, these estimates decreased when the SxH interaction effect was considered, and a positive covariance between direct and maternal effects was observed. Several studies with simulated data also reported biased VC without SxH interaction effect in the model

(Robinson, 1996; Lee and Pollak, 1997), supporting the hypothesis of overestimated genetic variances in the models without the SxH interaction effect.

In our study, larger additive direct genetic variances were observed in ssGREML compared to REML. In contrast, smaller estimates of maternal genetic variance, SxH variance, and residual variances were observed in ssGREML; all with smaller standard errors. The large negative covariance between direct and maternal effects was reduced when SxH was added to the model (M1 vs. M2) but was still negative.

When the covariance between direct and maternal effects was ignored in M3 and M4, most of the variances decreased, whereas the residual increased for both REML and ssGREML. One opposite pattern was observed in the comparison of M2 vs M4 under REML, which showed increased estimates of direct and maternal variances, with a decrease in residual variance. This current study's results agree with Meyer (1992) that the overestimation of both direct and maternal genetic variances is due to a negative covariance between these effects. Based on the current observations, biased direct and maternal genetic variances could be caused by ignoring the additional SxH interaction effect and allowing the negative estimation of a covariance component between direct and maternal effects. Therefore, if a negative covariance is mitigated by adding the SxH interaction effect (M1 to M2), including the covariance may give less overestimated genetic variances. AIC values were calculated for all models to determine the best model fitting the data (Table 2). As the amount of data was different for REML and ssGREML, AIC was not used for comparisons across the methods but only for the comparison of models within each method. In the results of both REML and ssGREML, M2 and M4 showed lower AIC values than models without SxH interaction effect (M1 and M3) although the differences were not very large.

Direct and maternal heritabilities, together with the proportion of the phenotypic variance explained by the SxH interaction effect, are shown in Fig. 1 for REML (a) and ssGREML (b). Overall estimates of direct heritability from ssGREML across all models were larger than the ones from REML. When the SxH interaction effect was considered under REML, direct heritabilities were reduced by a factor of 2.2 from M1 to M2, and by 1.7 from M3 to M4. The reduction was also observed under ssGREML but to a lesser extent (i.e., a factor of 1.5 and 1.25, respectively). The rationale for a larger reduction in the direct heritability when SxH interaction was added under REML is the decrease in direct variance combined with larger SxH interaction and residual variances and a larger phenotypic variance compared to ssGREML.

Overall, the estimation of VC with genomic information is affected by several factors: (1) genotyping strategy, (2) presence of selection, (3) parameters for the construction of $\mathbf{G}$, and (4) proportion of genotyped animals (Jensen, 2016; Cesarani et al., 2019; Wang et al., 2020). Because genomic selection has been applied to many livestock species, estimating unbiased VC using $\mathbf{A}$ becomes more challenging as it does not account for the impact of genomic selection (Jensen, 2016). In the AAA, the initial genotyping strategy included donor dams and proven sires because of the high costs; more recently, about half of the newly registered animals are genotyped each year, so the process is less selective. Wang et al. (2020) reported that VC estimated using $\mathbf{H}$ as the covariance structure among animals are sensitive to the genotyping strategy and proportion of genotyping. They emphasized that the strong selective genotyping and the high proportion of genotyped animals could produce overestimated variances; however, the level of overestimation observed in their study has not been confirmed.

In this study, genotyped animals were sampled that had phenotypes and at least one progeny either as a sire or dam, but animals were not filtered based on their phenotypic values.

This sampling strategy was expected to reduce the selective genotyping effect while meeting the computing limitation of ssGREML. However, as the AAA breeders practiced selective genotyping at the very early stages of genomic selection and still even less selective genotyping existed, those genotyped animals generally showed heavier adjusted WW than the non-genotyped animals (Fig. 2, t-value = 67.445 with p-value < 2.2e-16). This could be one possible reason why the direct heritability by ssGREML was larger than the estimation by REML among all the models (Fig. 1). Another possible reason could be the small proportion of genotyped animals. In the current study, the proportion of genotyping animals for each replicate is about ~8% which could produce a similar estimate or a modest overestimation in the ssGREML results (Wang et al., 2020).

Forni et al. (2011) reported similar variance component estimates between REML and ssGREML, but smaller standard errors in ssGREML as it could use more data than REML. Moreover, adding genomic information could help to solve possible issues caused by missing or incorrect pedigree information, frequent in many animal species (Banos et al., 2001). Using both genotypes and pedigree for estimating VC might be useful for populations with a high error rate in the pedigree. Cesarani et al. (2019) carried out a simulation study to compare VC using REML, GREML, and ssGREML under different genotyping strategies. Those authors reported biased VC under REML with a small dataset, but no bias under REML and ssGREML with larger datasets. The dataset used in our study was large enough to estimate VC (Table 1), so the different estimates for the direct variance under REML and ssGREML may not be due to the data size.

This is the first study that has estimated VC for WW in the presence of SxH interaction using ssGREML. Therefore, the basis for the differences between estimates under REML and ssGREML is not completely clear. Aldridge et al. (2020) claimed $\mathbf{H}$ could better separate the additive direct and permanent environmental effects. If the same theory can be applied to the

additive genetic effect and the additionally random SxH interaction effect, it could be hypothesized that the additive direct variance component estimated using **H** is more accurate than **A** because **H** reflects the realized relationships among animals rather than the expected (Legarra, 2016).

In the US dairy cattle evaluations, reduced weight for multiple daughters of a given bull in the same herd is used by adjusting for SxH interaction since 1967. As the SxH variance decreased from 14% (1967) to 10% (1997), the direct heritability increased from 25% to 30% in the same period (Van Tassell et al., 1997). Additionally, Wiggans et al. (2000) reported that SxH variance in Jersey and Brown Swiss reduced to 8% when heritability increased from 30% to 35% in November of 2000. Those findings are supported by the current results. When SxH variance was 5% in REML for both M2 and M4, direct heritability was 0.15 and 0.16, respectively. On the other hand, when SxH variance decreased to 0.03 for both M2 and M4 under ssGREML, direct heritability increased to 0.26 and 0.24, respectively (Fig. 1). Lee and Pollak (1997) scrutinized the sire x year interaction effect and conjectured that the effect might be a true effect due to the different environmental factors associated with a different year. Based on their speculation, SxH interaction might also be a true effect due to the different environmental factors related to different herds. Therefore, improving the environment in specific herds could introduce heterogeneous variance among herds, which is a possible factor to generate SxH variance.

Genetic trends and genomic prediction

Genetic trends from 1972 to 2019 for all the five models are shown in Fig. 3. The genetic trends were measured as the average EPDs by year of birth. Overall, results indicate direct genetic trends have been increasing over time. The result for the direct effect (Fig. 3a) shows M2 and M4 have lower genetic trends than M1, M3, and M5. Furthermore, M1 and M5 showed almost equivalent genetic trends and were a bit greater than M3. In Fig. 3b, opposite patterns were

observed for maternal effects, in which M2 and M4 have greater genetic trends than M1, M3, and M5. Particularly, M1 showed the lowest maternal genetic trend among all the models. Like the results of the direct genetic trend, consistent increases were observed since the 1980s; however, the slopes were not very steep after the 2010s, especially for the M1. These results suggest adding SxH interaction in the evaluation model increases maternal genetic trends and reduces the direct genetic trends, which could be overestimated without SxH. Legarra and Reverter (2017) outlined that bias was expected to increase with greater genetic gains. Genetic gain is defined as the change in the average breeding value of a population over a period, and the rate of genetic gain per year could be expressed as a genetic trend. These current results show that the models with the greatest bias for the direct effect (Table 3) have larger trends. In Fig. 3, direct genetic trends of M1, M3, and M5 are larger than M2 and M4. Also, greater bias is observed (Table 3) for those M1, M3, and M5 than M2 and M4 when both of $\mathrm{Var}_{\mathrm{REML}}$ and $\mathrm{Var}_{\mathrm{ssGREML}}$ were used.

In beef cattle and many other species, the predictive ability has been used as a tool for predicting future phenotypes (progeny performance), which is calculated as the correlation between (G)EBV and phenotypes adjusted for fixed effects (Legarra et al., 2008; Lourenco et al., 2015). However, this method was difficult to apply for complex models such as binary traits, maternal effect, and multiple random effect models. Therefore, in the current study, the LR method was used to calculate both direct and maternal prediction estimators. As the LR method was recently developed, no studies have reported its performance on models with a maternal effect, although some studies validated this method with several simulations and real datasets (Silva et al., 2019; Macedo et al., 2020; Bermann et al., 2021). The estimators of the LR method are shown in Table 3. When $\mathrm{Var}_{\mathrm{REML}}$ was used, M2 and M4 showed greater accuracy for the direct effect than the other models, as well as relatively less bias. Dispersion was almost equivalent for all the

models. Similar behavior was observed when using $\mathrm{Var}_{\mathrm{ssGREML}}$. The increase in accuracy for the direct effect when adding SxH interaction in the model (M1 vs. M2) was around 24% for $\mathrm{Var}_{\mathrm{REML}}$ and 12% for $\mathrm{Var}_{\mathrm{ssGREML}}$. Additionally, bias decreased by approximately 30% and 15% for $\mathrm{Var}_{\mathrm{REML}}$ and $\mathrm{Var}_{\mathrm{REML}}$, respectively.

The accuracy of M2 and M4 for the maternal effect was also greater than M1 and M5 for both VC scenarios, whereas M3 showed the greatest accuracy among all the models although the differences compared to M2 and M4 were not very large. The largest bias was observed in M1 for both VC scenarios. On the other hand, other models showed very similar biases when $\mathrm{Var}_{\mathrm{REML}}$ was used, but those biases increased when $\mathrm{Var}_{\mathrm{ssGREML}}$ was used, especially in M2 and M4. No large differences in dispersion were seen between the models and VC methods.

In general, lower accuracies and greater biases were observed when $\mathrm{Var}_{\mathrm{ssGREML}}$ was used. In the LR method, the dispersion estimator may indicate overdispersion of GEBV (if $b_1 < 1$) or under-dispersion of GEBV ($b_1 > 1$). The b_1 across five models did not differ either with $\mathrm{Var}_{\mathrm{REML}}$ or $\mathrm{Var}_{\mathrm{ssGREML}}$. Remarkably, M2 and M4 had the greatest accuracy under the $\mathrm{Var}_{\mathrm{REML}}$ scenario; however, those accuracies dropped about 16.8% and 12%, respectively, when $\mathrm{Var}_{\mathrm{ssGREML}}$ was used. Such a large reduction was not observed in other models. This pattern was also observed for the bias. When $\mathrm{Var}_{\mathrm{ssGREML}}$ was used for M2 and M4, the bias increased up to 21.4% and 17.2%, respectively. However, these observed increases were to a very small extent for M1, M3, and M5 (4.3~5.3%). Based on our findings, fitting SxH interaction in the model (M2 and M4) resulted in more accurate and less biased breeding values for the validation group, regardless of the choice of the covariance structure among the animals ($\mathbf{A}$ vs. $\mathbf{H}$) for estimating VC. However, it could also be speculated that the use of $\mathrm{Var}_{\mathrm{ssGREML}}$ for genomic prediction, especially with the SxH interaction effect, could decrease the accuracy and increase the bias compared to the results with

$\mathrm{Var}_{\mathrm{REML}}$ because of σ_a^2, which is part of the denominator of the accuracy formula, was larger when using genomic information, therefore, reducing the accuracy.

Accuracy, bias, and dispersion are the main features to examine the performance of genomic predictions. These three components could reflect the predictability of response to selection, correctness of model, use of inappropriate VC, and several unaccounted effects in the models (Reverter et al., 1994; Legarra and Reverter, 2018; Macedo et al., 2020). Macedo et al. (2020) applied the LR method to examine the possible bias and lower accuracy with the use of wrong heritability and unaccounted environmental effects. In that study, they concluded that if the incorrect genetic model was used for genomic evaluations, the LR method could estimate the bias when the model was not severely misspecified. The current results for the models without SxH interaction effect (M1, M3, and M5) support that discovery. These models showed a large bias for direct GEBV and some level of bias for maternal GEBV. Henderson (1975) reported that the use of an incorrect variance and covariance matrix could result in greater prediction error variance (PEV) for the solutions. Schaeffer (1984) extended that theory and concluded that the increase in PEV is directly related to the differences between true and estimated correlations. Therefore, we would argue that M2 and M4 had more appropriate variance components because of the SxH interaction effect. However, the large bias still observed in all models may be due to the effects that could not be accounted for in the models, affecting the estimation of GEBV. Wang et al. (2020) reported that the inflation of (G)EBV could reflect the bias in variance component estimation. However, the inflation of (G)EBV (i.e., dispersion) was very consistent among models and VC methods. Therefore, based on our results and reports from the literature, we could conjecture that M1, M3, and M5 used inappropriate VC (estimates without SxH effect) and did not account for the hidden trend in the data (not fitting the SxH effect). Additionally, the use of

negative covariance between direct and maternal effects might result in biased estimates, especially for the maternal GEBV (M1 vs. M5).

Wang et al. (2020) tested genomic predictions using VC estimated from $\mathbf{A}$ and $\mathbf{H}$ for commercial and simulated datasets. These results agree with the results from the current study in the sense that accuracies of GEBV were greater when using VC estimated from $\mathbf{A}$ than from $\mathbf{H}$; however, no clear explanation was provided in the previous study. One possible reason could be selective genotyping. In general, accuracy is the correlation between true breeding value (TBV) and (G)EBV or a function of (G)$EBV_{partial}$ and (G)EBV_{whole} in the LR method. Therefore, greater accuracy reflects the greater relatedness between TBV and (G)EBV or (G)$EBV_{partial}$ and (G)EBV_{whole}. If the VC used for genomic predictions were estimated with only selected genotyped animals, the relatedness between true and estimated BV would be more distant than if true VC were used. In this sense, it could be recommended to use VC from $\mathbf{A}$, especially under the selective genotyping strategy; although more precisely estimated VC are expected from $\mathbf{H}$ as it has a more accurate relationship structure among the animals.

One finding that deserves a deeper investigation is the large increase in accuracy and decrease in bias from $Var_{ssGREML}$ to Var_{REML} when the SxH interaction effect was added (M2 and M4 in Table 3). Further research is needed to understand the changes in predictions and VC when an additional random sire interaction effect is fitted in the model.

Changes in EPD and ranking of AI sires

The changes in the rank of AI sires among the models are illustrated in Fig. 4. The horizontal dotted lines were drawn to specify each change on +50, +100, -50, 0, and -100 scales. G1 to G4 represents the animals having no changes (G1), changes within the interval from -50 to +50 (G2), changes within -50 to -100 or within +50 to +100 (G3), changes more than $\pm$100 (G4).

Overall, considerable ranking changes were observed, especially for (b) M2 vs. M5 and (d) M4 vs. M5 compared to (a) M1 vs. M5 and (c) M3 vs. M5. Only a few AI sires had the same ranking among comparisons (82, 17, 81, 16 for (a) to (d), respectively). Because the ranking is an indicator of the genetic merit of the bulls in the population, even small changes could have a large impact, affecting the breeding decisions. Results of the change on direct EPDs of 1,977 AI sires are described in Fig. 5. Fig. 5a shows all the EPDs changed randomly within a very small range (from -2 to 4) regardless of the ranks of AI sires. On the contrary, Fig. 5b-d show changes that agree with the changes in the rankings of AI sires. Interestingly, the top AI sires had a greater reduction in EPDs as indicated by the greater negative values on the left-hand side of each plot (Fig. 5b-d). Additionally, a few bottom sires also had greater changes as observed on the right-hand side of the plots (Fig. 5b-d). Although similar patterns are observed in Fig. 5b-d, the range of EPD changes in Fig. 5c is smaller than that of Fig. 5b and 5d. These results imply adding the SxH interaction effect in the evaluation model could generate large changes in rank and direct EPDs on AI sires although it showed unbiased VC estimation along with a better prediction model.

The results of ranking changes of maternal EPD for AI sires among the models are in Fig. 6. The horizontal dotted lines and G1 to G4 have the same description as in Fig. 4. Similar patterns are detected in Fig. 4b and 4d and Fig. 6b and 6d, showing large ranking changes. Different from Fig. 4a, Fig. 6a also showed very large changes in rankings, implying the negative covariance between direct and maternal effects may have been the reason for such changes in the maternal effect. Fig. 7 shows changes in maternal EPDs for AI sires. Most of the AI sires had reduced maternal EPDs (Fig. 7a). Many sires had larger maternal EPDs in M2 and M4 than in M5 (Fig. 7b and Fig. 7d, respectively), in addition, the bottom sires had larger maternal EPDs in these models.

A similar pattern was observed in Fig. 7c, but a lot of sires had reduced maternal EPDs in M3 with a relatively small magnitude.

CONCLUSIONS

The inclusion of the SxH interaction effect in the model for weaning weight reduces the direct and maternal genetic variances and results in a positive covariance between direct and maternal effects when genomic information is not used. With genomics, the reduction is less, and the covariance is still negative. Using VC without genomic information may result in greater LR accuracy because of a lower additive genetic variance, with a similar level of dispersion. Adding the SxH interaction effect showed the best estimates of accuracy and bias for the direct effect but not for the maternal effect. Larger additive genetic variance with genomic information may be an artifact of selective genotyping. Fitting the SxH interaction effect in the model is recommended; however, further research is needed to investigate the improvement of prediction accuracy of maternal effects when SxH interaction is considered. Additionally, breeders should expect large changes in EPDs and ranking of animals, especially at the tails of the distributions, if this extra effect were fit into the genetic evaluation model. Before such changes are implemented in practice, more research is needed to ensure the resulting breeding values are better. The results of this study justify further investigation in this area for American Angus.

ACKNOWLEDGMENTS

We thank Dale Van Vleck for inspiring this study through his continued research in this area and for generously sharing his ideas.

REFERENCES

Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta, and T. Lawlor. 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

Aldridge, M. N., J. Vandenplas, R. Bergsma, and M. P. Calus. 2020. Variance estimates are similar using pedigree or genomic relationships with or without the use of metafounders or the algorithm for proven and young animals. Journal of animal science 98(3):skaa019

Banos, G., and G. Shook. 1990. Genotype by environment interaction and genetic correlations among parities for somatic cell count and milk yield. Journal of Dairy Science 73(9):2563-2573

Banos, G., G. Wiggans, and R. Powell. 2001. Impact of paternity errors in cow identification on genetic evaluations and international comparisons. Journal of dairy science 84(11):2523-2529

Baschnagel, M. B., J. Moll, and N. Künzi. 1999. Comparison of models to estimate maternal effects for weaning weight of Swiss Angus cattle fitting a sire× herd interaction as an additional random effect. Livestock production science 60(2-3):203-208

Bermann, M., A. Legarra, M. K. Hollifield, Y. Masuda, D. Lourenco, and I. Misztal. 2021. Validation of single-step GBLUP genomic predictions from threshold models using the linear regression method: An application in chicken mortality. Journal of Animal Breeding and Genetics 138(1):4-13

Bradfield, M. J. 1999. Genetic evaluation of cattle managed under extensive conditions in northern Australia, University of New England

Cesarani, A., I. Pocrnic, N. P. Macciotta, B. O. Fragomeni, I. Misztal, and D. A. Lourenco. 2019. Bias in heritability estimates from genomic restricted maximum likelihood methods under different genotyping strategies. Journal of Animal Breeding and Genetics 136(1):40-50

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genetics Selection Evolution 42(1):1-8

Dodenhoff, J., L. D. Van Vleck, and D. Wilson. 1999. Comparison of models to estimate genetic effects of weaning weight of Angus cattle. Journal of animal science 77(12):3176-3184

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43(1):1

Graser, H., D. Johnston, and B. Tier. 1999. Sire× herd interaction effect in BREEDPLAN. In: Proceedings of the Association for the Advancement of Animal Breeding and Genetics. p 197-198

Henderson, C. R. 1975. Comparison of alternative sire evaluation methods. Journal of Animal Science 41(3):760-770

Jensen, J. 2016. Estimation of genetic variance in the age of genomics. Wiley Online Library.

Kennedy, B., L. Schaeffer, and D. Sorensen. 1988. Genetic properties of animal models. Journal of Dairy Science 71:17-26

Lee, C., and E. Pollak. 1997. Relationship between sire× year interactions and direct-maternal genetic correlation for weaning weight of Simmental cattle. Journal of Animal Science 75(1):68-75

Legarra, A. 2016. Comparing estimates of genetic variance across different relationship models. Theoretical population biology 107:26-30

Legarra, A., and A. Reverter. 2017. Can we frame and understand cross-validation results in animal breeding. In: Proceedings of the 22nd conference association for the advancement of animal breeding and genetics. p 2-5

Legarra, A., and A. Reverter. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Genetics Selection Evolution 50(1):53

Legarra, A., C. Robert-Granié, E. Manfredi, and J.-M. Elsen. 2008. Performance of genomic selection in mice. Genetics 180(1):611-618

Lourenco, D., S. Tsuruta, B. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. Bertrand, T. Amen, L. Wang, and D. Moser. 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. Journal of animal science 93(6):2653-2662

Macedo, F., A. Reverter, and A. Legarra. 2020. Behavior of the Linear Regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models. Journal of Dairy Science 103(1):529-544

Masuda, Y., I. Aguilar, S. Tsuruta, and I. Misztal. 2015. Acceleration of sparse operations for average-information REML analyses with supernodal methods and sparse-storage refinements. Journal of animal science 93(10):4670-4674

Meyer, K. 1987. Estimates of variances due to sire× herd interactions and environmental covariances between paternal half-sibs for first lactation dairy production. Livestock Production Science 17:95-115

Meyer, K. 1992. Variance components due to direct and maternal effects for growth traits of Australian beef cattle. Livestock Production Science 31(3-4):179-204

Meyer, K., and H.-U. Graser. 1999. Estimates of parameters for scan records of Australian beef cattle treating records on males and females as different traits. In: Proceedings of the Association for the Advancement of Animal Breeding and Genetics. p 385-388

Misztal, I., A. Legarra, and I. Aguilar. 2014a. Using recursion to compute the inverse of the genomic relationship matrix. Journal of dairy science 97(6):3943-3952

Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, and Z. Vitezica. 2014b. Manual for BLUPF90 family of programs. Athens: University of Georgia

Notter, D., B. Tier, and K. Meyer. 1992. Sire× herd interactions for weaning weight in beef cattle. Journal of Animal Science 70(8):2359-2365

Pollak, E., C. Wang, B. Cunningham, L. Klei, and C. Van Tassell. 1994. Considerations on the validity of parameters used in national cattle evaluations. In: Proceedings for the Fourth Genetic Prediction Workshop. January. p 21-22

Reverter, A., B. Golden, R. Bourdon, and J. Brinks. 1994. Detection of bias in genetic predictions. Journal of animal science 72(1):34-37

Robinson, D. 1996. Models which might explain negative correlations between direct and maternal genetic effects. Livestock Production Science 45(2-3):111-122

Schaeffer, L. 1984. Sire and cow evaluation under multiple trait models. Journal of Dairy Science 67(7):1567-1580

Silva, R. M., J. P. Evenhuis, R. L. Vallejo, G. Gao, K. E. Martin, T. D. Leeds, Y. Palti, and D. A. Lourenco. 2019. Whole-genome mapping of quantitative trait loci and accuracy of genomic predictions for resistance to columnaris disease in two rainbow trout breeding populations. Genetics Selection Evolution 51(1):1-13

Tong, A., B. Kennedy, and J. Moxley. 1977. Sire by herd interactions for milk yield and composition traits. Canadian Journal of Animal Science 57(3):383-388

Van Tassell, C., G. Wiggans, P. VanRaden, and H. Norman. 1997. Changes in USDA-DHIA genetic evaluations (August 1997). AIPL. Res. Rpt 9(8-97)

Veerkamp, R., H. Mulder, R. Thompson, and M. Calus. 2011. Genomic and pedigree-based genetic parameters for scarcely recorded traits when some animals are genotyped. Journal of dairy science 94(8):4189-4197

Wang, L., L. L. Janss, P. Madsen, J. Henshall, C.-H. Huang, D. Marois, S. Alemu, A. Sørensen, and J. Jensen. 2020. Effect of genomic selection and genotyping strategy on estimation of variance components in animal models using different relationship matrices. Genetics Selection Evolution 52(1):1-14

Wiggans, G., P. VanRaden, R. Powell, and C. Van Tassell. 2000. Changes in USDA-DHIA genetic evaluations (November 2000). AIPL. Res. Rpt

TABLES

Table A.1. General statistics for all the replicates

		Replicate1	Replicate2	Replicate3	Replicate4	Replicate5
No. of animals		112,677	105,909	102,433	109,260	102,183
No. of herds		88	93	84	90	97
No. of sires		3,970	4,553	4,262	4,379	4,157
No. of S x H		5,723	6,128	5,668	6,286	5,808
WW	Min., lbs	211	193	262	246	196
	Mean., lbs	602.6	607.2	602.4	600	604
	Max., lbs	1044	1113	1032	1044	1014
	SD., lbs	95.43	103.83	96.20	90.20	90.93

Table A.2. Estimated variance component for the four investigated models using REML and ssGREML method

	Model[3]	σ_a^2	σ_m^2	σ_{mpe}^2	σ_{sh}^2	σ_e^2	σ_{am}	Cor (a,m)	σ_p^2	AIC
REML[1]	M1	1069.60 (47.12)	415.66 (43.31)	372.08 (27.13)	0	1623.96 (63.26)	-251.76 (35.35)	-0.38 (0.06)	3229.54 (94.64)	802049 (37686.16)
	M2	467.63 (40.63)	266.11 (45.67)	366.95 (27.00)	150.90 (21.16)	1889.96 (79.97)	42.34 (26.95)	0.12 (0.08)	3183.89 (94.35)	801606 (37619.05)
	M3	858.01 (16.89)	275.57 (54.57)	359.78 (27.69)	0	1730.24 (61.11)	0	0	3223.60 (98.88)	802150 (37669.09)
	M4	517.79 (20.99)	290.99 (55.71)	368.25 (27.13)	143.77 (17.75)	1865.80 (68.78)	0	0	3186.59 (95.41)	801607 (37620.00)
ssGREML[2]	M1	1185.56 (35.95)	371.13 (30.45)	341.52 (25.60)	0	1517.72 (52.46)	-263.24 (38.19)	-0.40 (0.06)	3152.70 (85.16)	829542 (37475.95)
	M2	803.34 (47.82)	255.05 (26.92)	335.62 (25.75)	100.11 (19.86)	1672.94 (72.20)	-55.09 (38.65)	-0.12 (0.08)	3111.96 (81.46)	829174 (37416.09)
	M3	928.88 (19.57)	236.99 (39.32)	326.15 (26.40)	0	1643.42 (55.07)	0	0	3135.44 (89.08)	829663 (37452.34)
	M4	736.42 (42.19)	226.54 (38.84)	333.03 (26.12)	106.86 (17.44)	1708.12 (61.31)	0	0	3110.97 (89.67)	829178 (37413.24)

*Standard deviation based on five replicates is in parenthesis

REML[1]: restricted maximum likelihood method using only pedigree and phenotype

ssGREML[2]: single-step genomic restricted maximum likelihood method using pedigree, phenotype, and genotype

Model[3]: M1, without SxH interaction effect and with covariance between direct and maternal effect (σ_{am}) $\neq 0$; M2, with SxH interaction effect and $\sigma_{am} \neq 0$; M3, without SxH interaction effect and with $\sigma_{am} = 0$; M4, with SxH interaction effect and $\sigma_{am} = 0$

Table A.3. Accuracy, bias, and dispersion using the LR method (ssGBLUP)

	Model[3]	Accuracy		Bias		Dispersion estimator (b_1)	
		Var_{REML}[1]	$Var_{ssGREML}$[2]	Var_{REML}	$Var_{ssGREML}$	Var_{REML}	$Var_{ssGREML}$
Direct	M1	0.72	0.69	-3.60	-3.80	1.00	0.99
	M2	0.95	0.79	-2.53	-3.22	1.01	1.00
	M3	0.76	0.75	-3.26	-3.41	1.00	1.00
	M4	0.92	0.81	-2.65	-3.09	1.01	1.00
	M5	0.71	0.68	-3.53	-3.71	1.00	1.00
Maternal	M1	0.59	0.62	0.55	0.58	0.97	0.97
	M2	0.65	0.67	-0.06	0.24	0.98	0.98
	M3	0.66	0.70	0.06	0.08	0.98	0.98
	M4	0.63	0.69	0.07	0.11	0.98	0.98
	M5	0.59	0.61	0.04	0.06	0.97	0.97

Var_{REML}[1]: ssGBLUP using the variance component estimated from REML
$Var_{ssGREML}$[2]: ssGBLUP using the variance component estimated from ssGREML
Model[3]: M1, without SxH interaction effect and with covariance between direct and maternal effect (σ_{am}) $\neq 0$; M2, with SxH interaction effect and $\sigma_{am} \neq 0$; M3, without SxH interaction effect and with $\sigma_{am} = 0$; M4, with SxH interaction effect and $\sigma_{am} = 0$; M5, equivalent to M1, except for the $\sigma_{am} = 0$ after variance component estimation

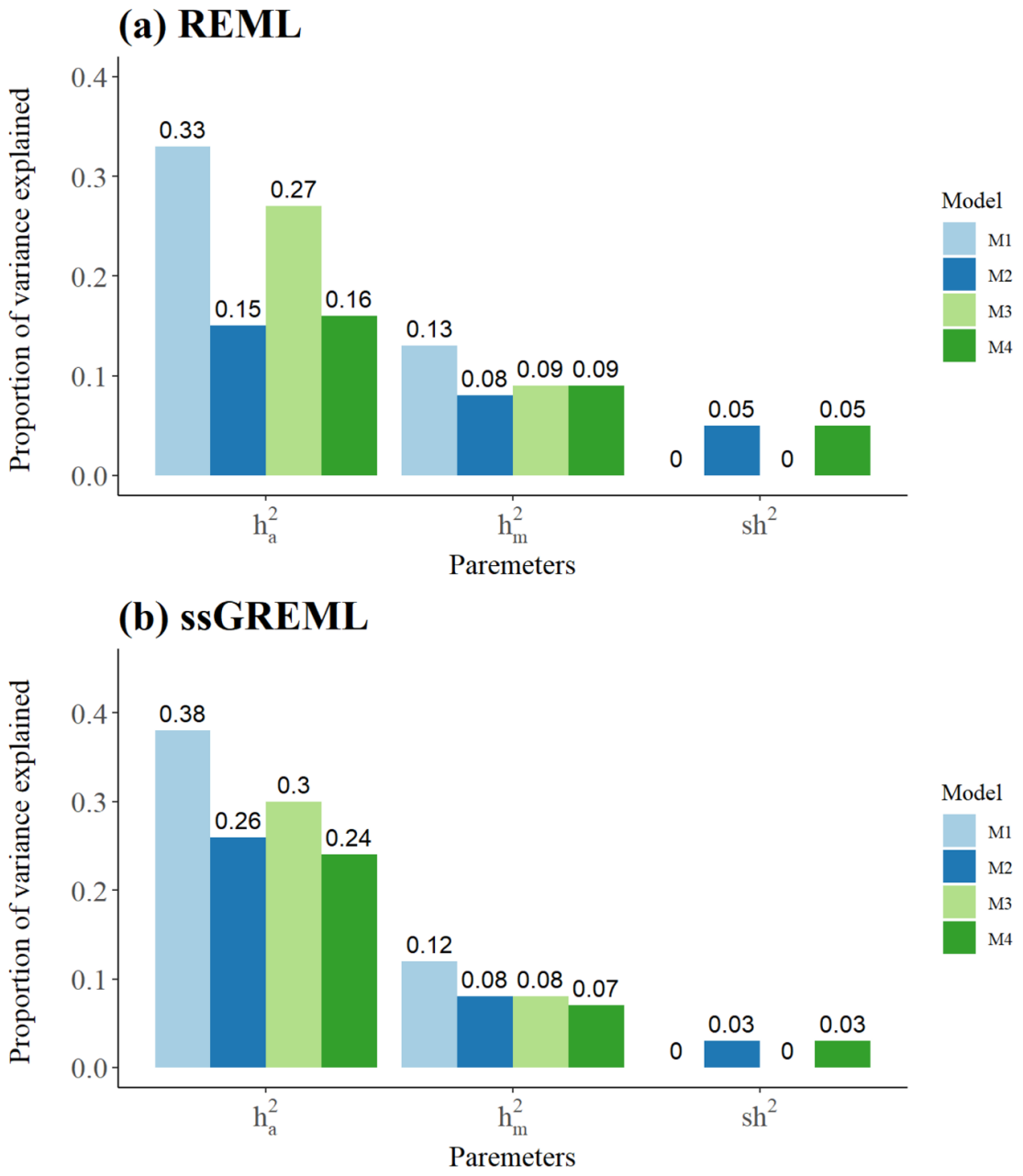


Figure A.1. Proportion of variance explained by additive direct, maternal, and sire by herd interaction effect using REML and ssGREML

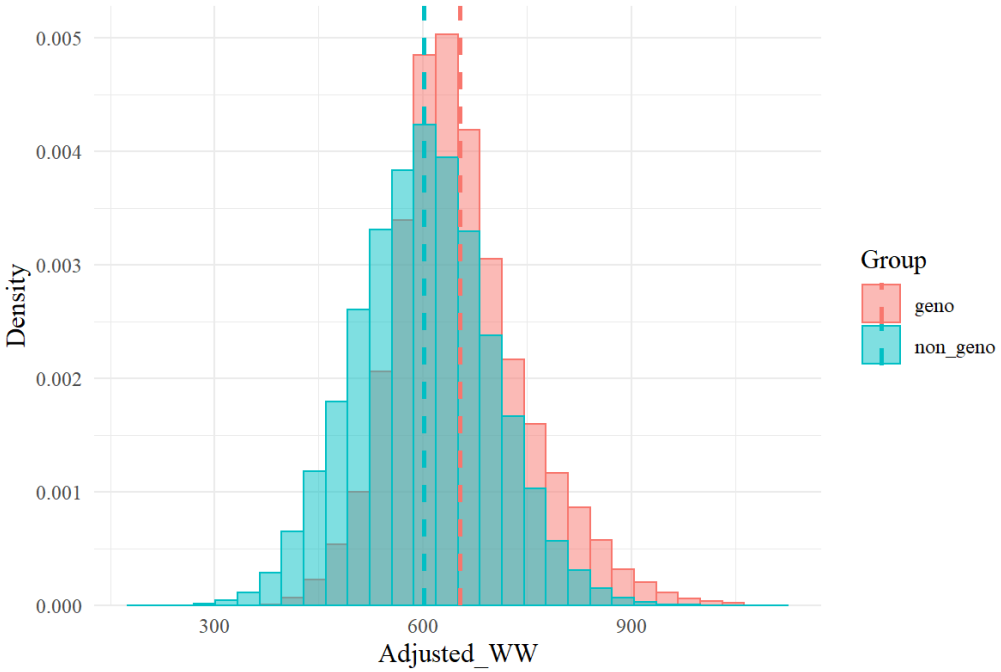


Figure A.2. Distribution of adjusted WW for genotyped and non-genotyped animals used for ssGREML. Vertical lines are indicating the average adjusted weaning weight for genotyped (geno; $\overline{X}$ = 653.30) and non-genotyped (non_geno; $\overline{X}$ = 601.43) animals

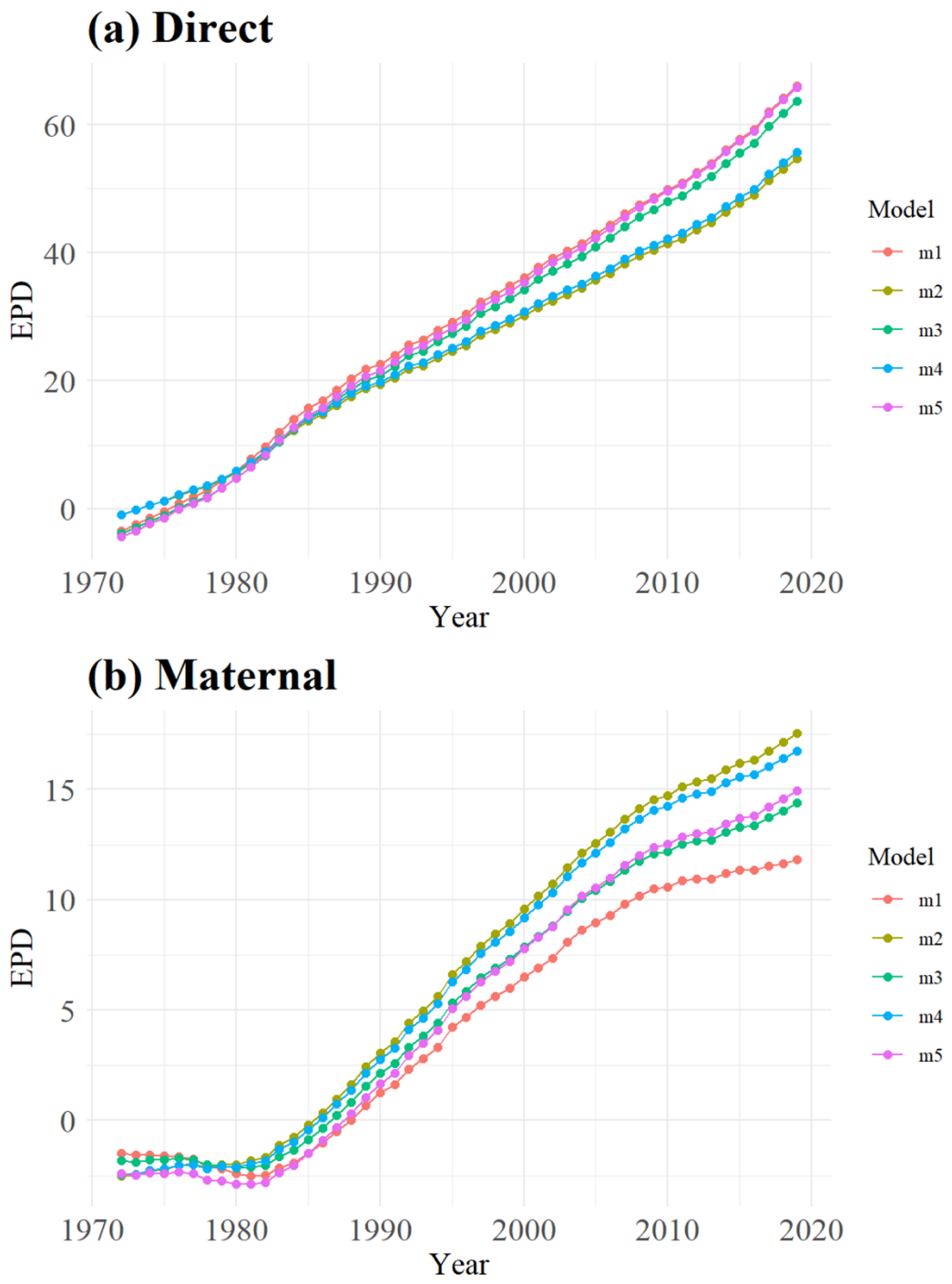


Figure A.3. Genetic trends for additive direct (a) and maternal (b) effects

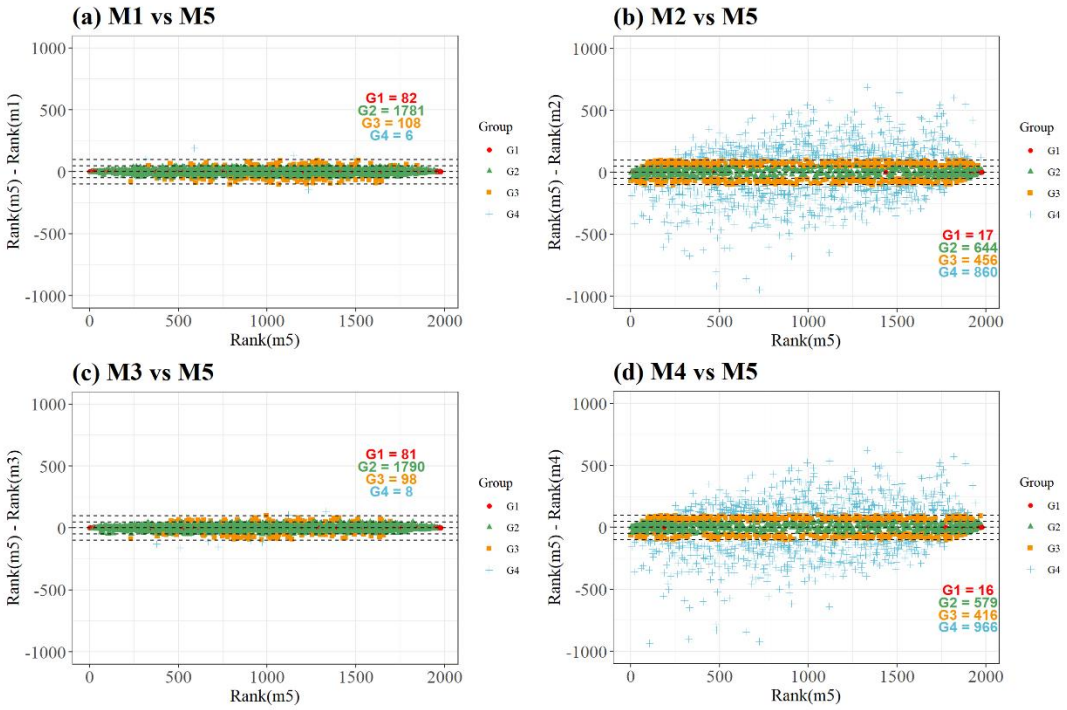


Figure A.4. Changes in the ranking of 1,977 AI sires (direct effect)

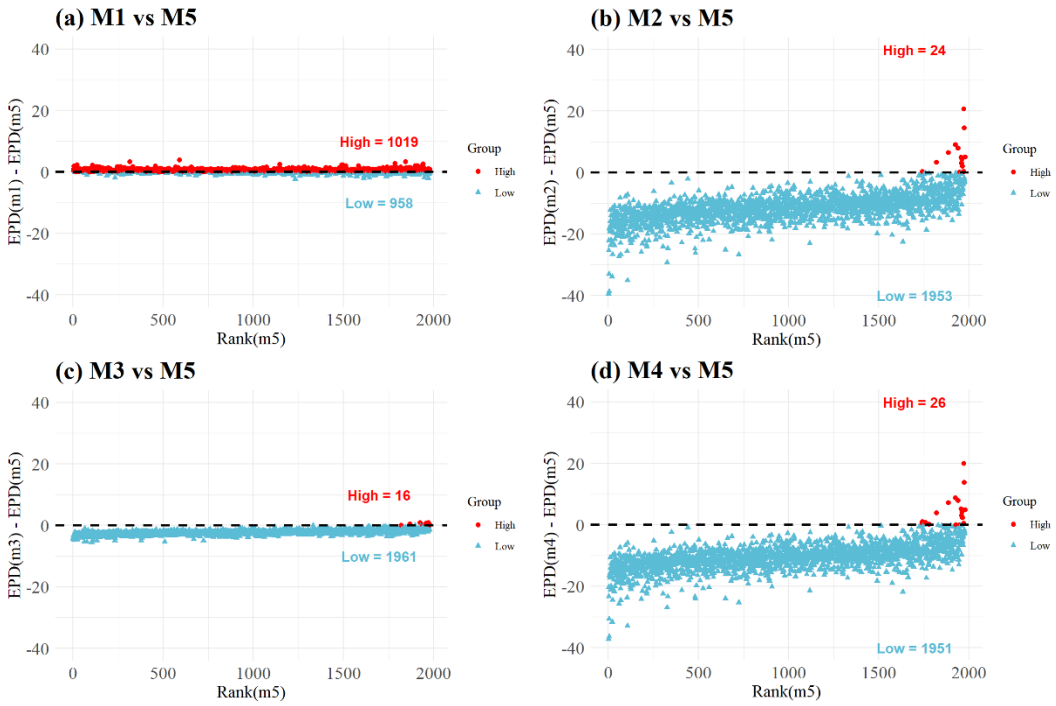


Figure A.5. Changes of EPDs for 1,977 AI sires (direct effect)

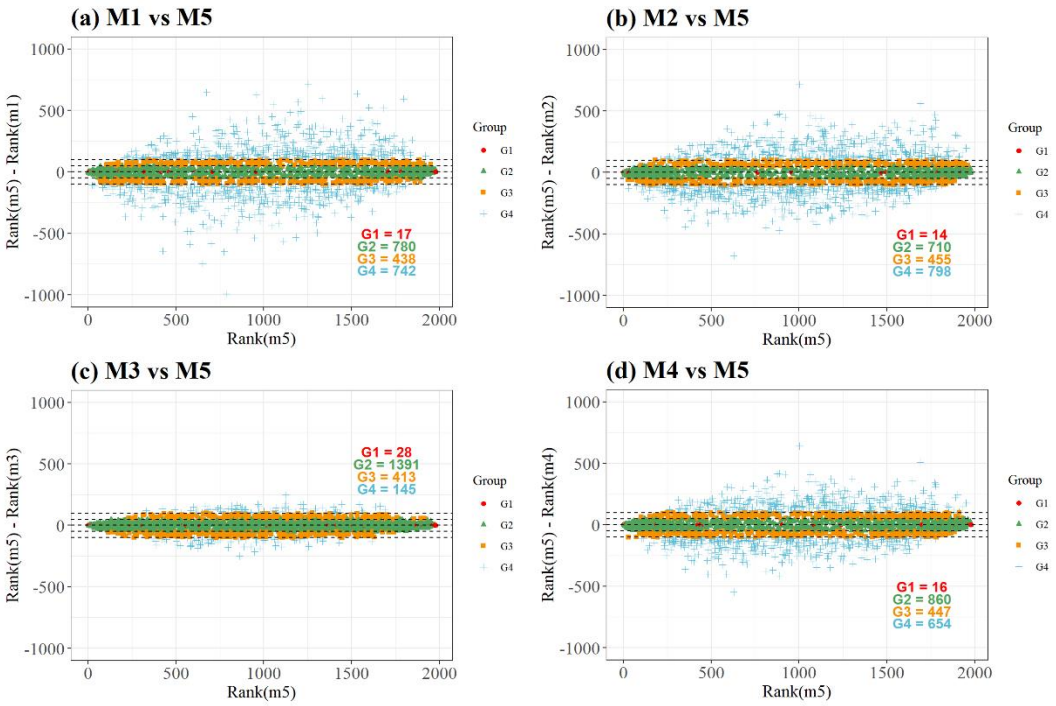


Figure A.6. Changes in the ranking of EPDs for 1,977 AI sires (maternal effect)

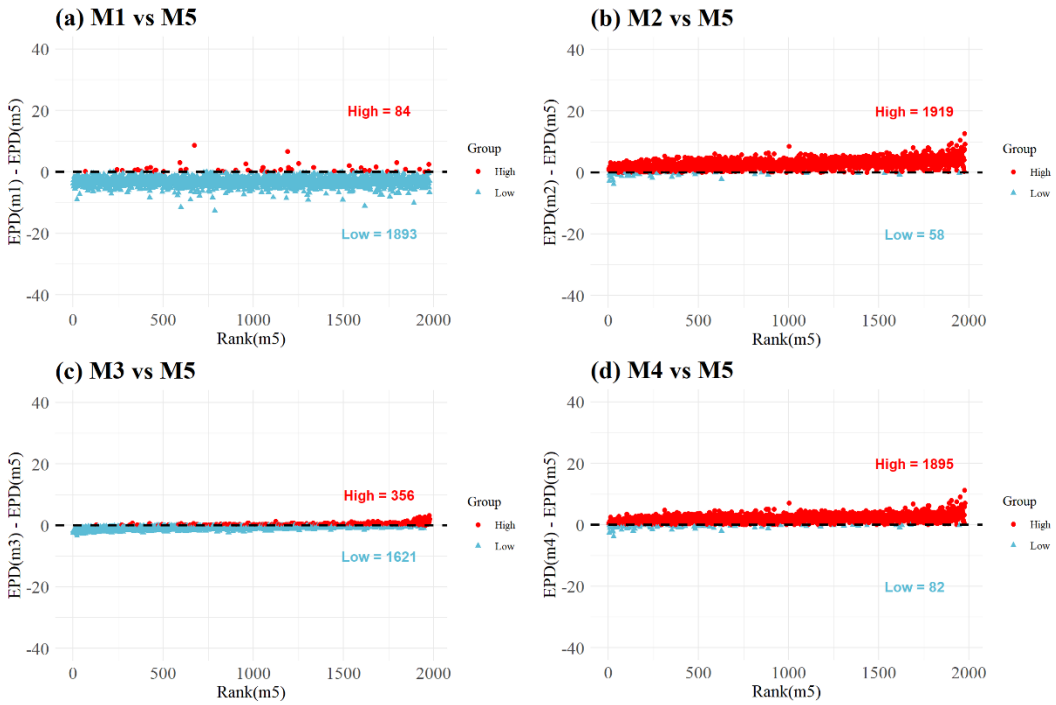


Figure A.7. Changes of EPDs for 1,977 AI sires (maternal effect)