IMPLEMENTATION OF GENOMIC SELECTION FOR CHANNEL CATFISH AND INDIRECT PREDICTIONS FOR LARGE SCALE GENOMIC EVALUATIONS

by

ANDRÉ LUIZ SECCATTO GARCIA

(Under the Direction of Daniela Lourenco)

ABSTRACT

Catfish farming is the largest segment in the US aquaculture business and among other topics, the implementation of genomic selection has been recently investigated. Using genomic information improved predictive ability by 28% for harvest weight and up to 36% for carcass traits compared to traditional evaluation. This demonstrates the benefit of genomic selection for the US catfish breeding program. Such improvements have made the use of genomic information widely adopted across many livestock and aquaculture species. With this rapid adoption, the number of genotyped animals has been steadily increasing, especially in the US dairy and beef industries. With a large number of genotyped animals, genomic evaluations may be challenging and indirect predictions (IP) can be a useful tool providing fast interim evaluations for young genotyped animals. Further, IP can be used as genomic prediction for unregistered animals not included in official evaluations. When genomic best linear unbiased prediction (GBLUP) or single-step GBLUP (ssGBLUP) are the methods of choice for genomic evaluations, IP can be obtained based on single nucleotide polymorphism (SNP) effects that are backsolved using genomically estimated breeding values (GEBV). With large number of genotyped animals, IP can be reliably obtained from (ss)GBLUP either by using direct inversion of **G** or by using the algorithm for proven and

young (APY), as long as GEBV are from a previous (ss)GBLUP evaluation. Further, in purebred beef cattle populations, a sample of at least 15,000 animals representing the whole genotyped population may also provide reliable SNP effects and IP. To make use of IP, it is important that its accuracy is comparable to the GEBV accuracy. Under (ss)GBLUP, IP accuracy can be obtained by backsolving prediction error covariance (PEC) of GEBV into PEC of SNP effects. The computational cost of PEC computations is prohibitive with large number of animals and using a subset of animals to approximate it is desirable for large scale evaluations. It is possible to reduce the number of genotyped animals in PEC computations, but accuracies may be underestimated and fine tuning is still required to scale accuracies of indirect predictions up to accuracies of GEBV.

INDEX WORDS: ssGBLUP, predictive ability, interim evaluations, accuracy

IMPLEMENTATION OF GENOMIC SELECTION FOR CHANNEL CATFISH AND INDIRECT PREDICTIONS FOR LARGE SCALE GENOMIC EVALUATIONS

by

ANDRÉ LUIZ SECCATTO GARCIA

B.S., Universidade Estadual de Maringá, Brazil, 2014

M.S., Universidade Estadual de Maringá, Brazil, 2016

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

2020

IMPLEMENTATION OF GENOMIC SELECTION FOR CHANNEL CATFISH AND INDIRECT PREDICTIONS FOR LARGE SCALE GENOMIC EVALUATIONS

by

ANDRÉ LUIZ SECCATTO GARCIA

Major Professor:	Daniela Loureco
Committee:	Ignacy Misztal
	Romdhane Rekaya
	Brian Bosworth

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
May 2020

DEDICATION

To Carlos, Natalina, Eduardo and Geovana.

ACKNOWLEDGEMENTS

I would like to thank Dr. Daniela Lourenco for accepting me as a student and for giving me so many opportunities. It has been a great learning process (as she would put it) and I am very grateful for all the teaching moments and discussions. I also would like to thank my other committee members, Dr. Ignacy Misztal, Dr. Romdhane Rekaya and Dr. Brian Bosworth for all their assistance in classes, research, writing and field experience.

I would like to thank Dr. Brian Bosworth and the USDA WARU team as well as Dr. Stephen Miller and the AGI team for the internship opportunities. Those were great experiences, very important for my professional development.

I appreciate all the informal classes from Dr. Shogo Tsuruta and Dr. Yutaka Masuda, as well as the many discussions about computers, coffee and animal breeding in general.

To not forget anyone, I want to thank all professors, postdocs, visitors, students and staff I have interacted with.

Last but not least I would like to thank my family and friends for supporting me throughout this journey and a special thank you to my girlfriend for all the support and patience.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

GENOMIC INFORMATION IN GENETIC EVALUATIONS

The publication of the human genome draft in 2001, opened the way for livestock species to have their genome sequenced as well. Later, high throughput sequencing technologies led to the development of dense single nucleotide polymorphisms (SNP) panels which generated great excitement in the animal breeding community, as the DNA information could help improving genetic gains. Although expensive at the beginning, the genotyping prices quickly declined over time making it possible to have thousands of animals genotyped, which in fact became a reality in many livestock, poultry, and aquaculture populations recently.

In animal breeding applications, SNP markers are spread across all chromosomes to cover the whole genome and should account for the linkage disequilibrium (LD) between the markers and the quantitative trait loci (QTL) affecting the traits of interest. Meuwissen et al. (2001) proposed three methods to use genomic information in genetic evaluations, each with different assumptions for the markers a priori. After that paper, many others followed, showing different methods and models to accommodate the new source of information in what became known as genomic selection (GS).

There are two main ways to incorporate the genomic information into genetic evaluations: the first focus on estimating the marker effects and the second uses the markers to obtain realized

relationships among the animals. These two approaches to genomic information led to the development of two classes of models: SNP based models (SNP-BLUP and Bayesian regression models) (Meuwissen et al., 2001) and relationship based models (GBLUP) (VanRaden, 2008). Under some assumptions these two classes of models are equivalent, which is the case of SNP-BLUP and GBLUP.

In practice, both marker effects and relationship based models use the information from genotyped animals to obtain the genomic contribution from the markers, which is later combined with the pedigree based evaluations to generate the final genomic estimated breeding value (GEBV). This method is usually called the multi-step genomic evaluation.

To accommodate all the information together and to simplify the evaluation framework, single-step versions of both classes of models were developed and the single-step genomic evaluation is becoming the method of choice in animal breeding programs. Two examples of single-step models are the single-step GBLUP (ssGBLUP) (Legarra et al., 2009; Misztal et al., 2009) and the single-step Bayesian regression (ssBR) (Fernando et al., 2014). The research presented in this dissertation focuses on the application of GBLUP and ssGBLUP models for fish and beef cattle breeding.

In GBLUP, the SNP information is used to obtain realized relationships among animals, which results in a genomic relationship matrix (**G**) that replaces the expected relationships commonly used in pedigree based models (Henderson, 1984). Therefore, in the GBLUP mixed model equations (MME), the pedigree relationship matrix (**A**) is substituted by **G**.

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{G}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'}y \\ \mathbf{W'}y \end{bmatrix} \quad (1)$$

Where **y** is the vector of observations, **β** is the vector of fixed effects and **u** is the vector of random additive genetic effects; **λ** is the ratio of residual to additive genetic variances; **X** and **W** are the

incidence matrices for **β** and **u**, respectively, and $\mathbf{G}^{-1}$ is the inverse of **G**. The initial **G** ($\mathbf{G_0}$) is often formulated as in VanRaden (2008):

$$\mathbf{G_0} = \frac{\mathbf{ZZ'}}{2\sum p_i (1-p_i)} \quad (2)$$

Where **Z** is a matrix of centered gene content and $\mathbf{p_i}$ is the minor allele frequency of SNP *i*. Ideally, allele frequencies from the base population should be used to build the **G**, but because genotypes are only available for recent generations, allele frequencies are often calculated based on current genotypes. This **G** will be singular if clones are present, if the number of markers is smaller than the number of animals or if there are some numerical dependencies. To overcome this challenge, $\mathbf{G_0}$ can be blended with the pedigree relationship matrix, making it invertible (VanRaden, 2008).

$$\mathbf{G} = \alpha\mathbf{G_0} + (1-\alpha)\mathbf{A} \quad (3)$$

Where **α** is a weight that usually assumes the value of 0.95. Once **G** is built and inverted, the MME for GBLUP can be written as in Eq. 1.

Under GBLUP, only genotyped animals are directly considered in the model, whereas the pedigree and phenotypic information from ungenotyped animals has to be incorporated later. To include all genotyped and ungenotyped animals into a single system for the genetic evaluations, Misztal et al. (2009) and Legarra et al. (2009) proposed a combined relationship matrix (**H**):

$$\mathbf{H} = \begin{bmatrix} \mathbf{A_{11}} + \mathbf{A_{12}A_{22}^{-1}(G-A_{22})A_{22}^{-1}A_{21}} & \mathbf{A_{12}A_{22}^{-1}G} \\ \mathbf{GA_{22}^{-1}A_{21}} & \mathbf{G} \end{bmatrix}, \quad (4)$$

where **A** and **G** are the pedigree and genomic relationship matrices and the subscripts **1** and **2** refer to ungenotyped and genotyped animals, respectively.

Although **H** has a complicated form, Aguilar et al. (2010) showed that it has a simple inverse:

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}-A_{22}^{-1}} \end{bmatrix}. \quad (5)$$

Once $\mathbf{H}^{-1}$ is available, it can replace the inverses of **A** or **G** in the same traditional MME, and this method is called single-step GBLUP. The MME for ssGBLUP can be written as follows:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{H}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'}y \\ \mathbf{W'}y \end{bmatrix} \quad (6)$$

Recently, ssGBLUP has become the method of choice in genomic evaluations for many species, for instance: broiler chicken (Chen et al., 2011; Lourenco et al., 2015b), layers (Yan et al., 2018), pigs (Forni et al., 2011; Lourenco et al., 2016), meat sheep (Brown et al., 2018), dairy sheep and goats (Rupp et al., 2016) and beef cattle (Lourenco et al., 2015a; Johnston et al., 2018).

GENOMIC SELECTION IN AQUACULTURE BREEDING PROGRAMS

As genomic resources were being developed for livestock species, it did not take long until researchers started investigating the possibilities for aquaculture species as well. As early as 2009, researchers started evaluating the performance of genomic evaluations and investigating strategies for efficient implementation using simulated data (Nielsen et al., 2009; Sonesson and Meuwissen, 2009).

Because of the reproductive characteristics of many aquaculture species (e.g. thousands of progeny per spawn) first, it was important to understand which animals as well as how many should be genotyped to make GS cost-effective and feasible for practical applications (Lillehammer et al., 2013; Ødegård and Meuwissen, 2014).

Early research as well as the experience from other agricultural species have contributed to the implementation of GS in many important aquaculture species in recent years. A few examples are: Atlantic salmon (*Salmo salar*) (Bangera et al., 2017; Correa et al., 2017; Sae-Lim et al., 2017), Coho salmon (*Oncorhynchus kisutch*) (Barría et al., 2018; Barría et al., 2019), rainbow trout (*Oncorhynchus mykiss*) (Vallejo et al., 2018; Yoshida et al., 2018; Silva et al., 2019),

European sea bass (*Dicentrarchus labrax*) (Palaiokostas et al., 2018a; Besson et al., 2019), tilapia (*Oreochromis niloticus*) (Yoshida et al., 2019; Joshi et al., 2020), common carp (*Cyprinus carpio*) (Palaiokostas et al., 2018b) and pacific oyster (*Crassostrea gigas*) (Gutierrez et al., 2018).

Typically, aquaculture breeding programs are based on a family structure, and genomic information is valuable because it allows for the exploration of the variation within families, making it possible to identify the best animals within the best families. This is especially useful for traits that cannot be measured on the selection candidates such as carcass traits and disease resistance (Yáñez et al., 2014).

With the genomic resources available and methods developed, more and more species will enter the genomic era and adopt genomic evaluations as a common practice in their breeding programs. In chapter two, we discuss the feasibility of implementing a genomic evaluation for the US channel catfish (*Ictalurus punctatus*) population using ssGBLUP.

LARGE SCALE GENOMIC EVALUATIONS

As genomic selection becomes a mature technology and genotyping costs keep decreasing, the number of genotyped animals is steadily increasing in some applications. One remarkable example is the US dairy industry that pioneered the field releasing its first genomic evaluation in 2009 (VanRaden, 2008; VanRaden et al., 2009) and now has over three million genotyped animals (queries.uscdcb.com/Genotype/cur_density.html). Another example is the American Angus Association with more than 750,000 genotyped animals (Steve Miller, Angus Genetics Inc., personal communication).

Such numbers demonstrate the rapid adoption of the technology by the industry. Although standard GS methods have been developed, accommodating such large number of genotyped animals into routine genetic evaluations can be challenging.

Some of the challenges have a computational nature, and one example is the inversion of **G** in GBLUP and ssGBLUP. Matrix inversion has a cubic cost with the number of genotyped animals and is not feasible for over 150,000 animals (Fragomeni et al., 2015). To solve this issue, Misztal et al. (2014) and Misztal (2016) proposed the algorithm for proven and young (APY). The APY is based on the idea that the genomic information has limited dimensionality due to small effective population size in livestock populations. The dimensionality is limited by the minimum of number of independent SNP and chromosome segments and the number of genotyped animals (Misztal, 2016). In the APY formulation, the genotyped population is divided into core (**c**) and noncore (**n**) such that the only direct inversion needed is for the core portion and the other components are obtained through recursions, dramatically reducing computing costs. In APY, **G** is represented as:

$$\mathbf{G}= \begin{bmatrix} \mathbf{G_{cc}} & \mathbf{G_{cn}} \\ \mathbf{G_{nc}} & \mathbf{G_{nn}} \end{bmatrix} \quad (7)$$

And $\mathbf{G^{-1}_{APY}}$ is calculated as follows:

$$\mathbf{G^{-1}_{APY}}= \begin{bmatrix} \mathbf{G^{-1}_{cc}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{-G^{-1}_{cc}G_{cn}} \\ \mathbf{I} \end{bmatrix} \mathbf{M^{-1}_{nn}} \begin{bmatrix} \mathbf{-G_{nc}G^{-1}_{cc}} & \mathbf{I} \end{bmatrix} \quad (8)$$

With each element of $\mathbf{M_{nn}}$ obtained for the *ith* non-core animal as:

$$m_{nn,i}= g_{ii} - \mathbf{G_{ic}G^{-1}_{cc}G_{ci}} \quad (9)$$

Pocrnic et al. (2016) showed that the number of core animals in APY can be obtained as the number of largest eigenvalues explaining 98-99% of the variance in **G**. Many studies have investigated the stability of GEBV when using APY and found that as long as the number of core

animals represents the dimensionality of the genomic information, the choice of animals is arbitrary and correlations between GEBV from ssGBLUP with and without APY are typically ≥0.99 (Fragomeni et al., 2015; Masuda et al., 2016; Bradford et al., 2017).

Another challenge is that out of all the genotyped animals, many are young and unregistered animals, therefore, do not have any phenotypes and sometimes may have incomplete pedigrees. These animals do not contribute to the evaluations of older animals and because of the amount of such incoming genotypes, they may slow down the official evaluations. Furthermore, Bradford et al. (2017) and Bradford et al. (2019) pointed out that including many animals with missing pedigrees into evaluations may decrease accuracy and increase inflation on GEBV.

These issues raise the question whether to include all the genotyped animals into one main evaluation or to find an alternative way to provide genomic predictions for young and unregistered genotyped animals without including them in routine evaluations.

Indirect predictions (IP) can be a helpful tool in this context. Because they are much faster to compute compared to official evaluations, IP can be used as interim genomic predictions allowing for weekly or even daily evaluations for young genotyped animals (Wiggans et al., 2015). Also, IP can also be used as genomic predictions for unregistered animals without having to include them in routine evaluations.

Lourenco et al. (2015a) investigated the use of IP from a ssGBLUP model for American Angus and found that using IP can be beneficial as they can be used as quick genomic predictions for young animals without running a complete evaluation.

When (ss)GBLUP is the method of choice for the genomic evaluation, SNP effects are not available by default, but can be obtained from GEBV. This is because SNP-BLUP and GBLUP are equivalent models; therefore, SNP effects can be calculated based on GEBV and the inverse

of the **G** for genotyped animals in GBLUP (VanRaden, 2008; Strandén and Garrick, 2009) and in ssGBLUP (Wang et al., 2012; Legarra et al., 2018) as follows:

$$\hat{\mathbf{a}} = \lambda \mathbf{D}\mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{u}} \quad (10)$$

Where $\hat{\mathbf{a}}$ is a vector of SNP effects; $\hat{\mathbf{u}}$ is a vector of GEBV, $\boldsymbol{\lambda}$ is the ratio of SNP to additive genetic variance, **D** is a diagonal matrix of SNP weights (**D**=**I**), and **Z** is a matrix of centered gene content. Once SNP effects are available, IP can be computed as $\mathbf{IP}=\mathbf{Z}\hat{\mathbf{a}}$, for any number of genotyped animals. In chapter three, we discuss the use of IP for large genotyped populations when (ss)GBLUP is used for genomic evaluations.

ACCURACY OF PREDICTIONS

Before the implementation of genomic selection, it is common to test the performance of different GS models regarding their ability predict future performance of animals, for the traits of interest, in a given population. This is done using different validation methods, depending on the prediction objectives, and it gives an idea about the model accuracy. Although very useful, the validation accuracy or predictive ability is a "population parameter", meaning that it does not provide a measure of accuracy for the individual breeding values.

Because the genetic gain depends on the accuracy of EBV, it is important for practical applications that (G)EBV are obtained with a measure of accuracy that reflects the standard error of the prediction. With traditional BLUP, Henderson (1984) showed that accuracies of EBV can be obtained based on the prediction error variance (PEV) by directly inverting the left hand side (LHS) matrix of BLUP MME. Once PEV is available, the accuracy for a given animal can be calculated as:

$$\mathrm{acc_i} = \sqrt{1 - \frac{\mathrm{PEV_i}}{(1+\mathrm{F_i})\sigma_u^2}} = \sqrt{1 - \frac{\mathrm{LHS^{ii}}}{(1+\mathrm{F_i})\sigma_u^2}} \quad (11)$$

where $\mathbf{F_i}$ is the inbreeding coefficient for the animal *i* and $\boldsymbol{\sigma}^2_{\mathbf{u}}$ is the additive genetic variance.

Although the method to obtain accuracy is available, when the system of equations is too big it becomes impossible to invert the LHS matrix even without genomic information, therefore accuracies are not easily available. To overcome this problem, methods to approximate accuracies have been proposed for traditional (Misztal and Wiggans, 1988; Meyer, 1989; VanRaden and Wiggans, 1991) and genomic evaluations (Misztal et al., 2013; Liu et al., 2017; Erbe et al., 2018).

Similarly, if IP are to be used as genomic predictions, it is of interest to have a measure of accuracy that is comparable to that of GEBV to be published with IP. Under the SNP-BLUP model, prediction error covariance (PEC) for SNP effects are easily available and can be used to calculate accuracy for IP (Liu et al., 2017). Because ssGBLUP is widely used for genomic evaluations, it is important to obtain SNP PEC from ssGBLUP MME to avoid running an extra SNP BLUP model to get accuracies for IP.

Under (ss)GBLUP SNP PEC can be obtained by converting PEC for genotyped animals into PEC for SNP effects (Gualdron Duarte et al., 2014; Aguilar et al., 2019). When SNP effects are backsolved from (ss)GBLUP, PEC can then be obtained as follows:

$$\text{var}(\hat{a})=\text{PEC}=\text{var}\left(\mathbf{Z'}\frac{1}{2\sum p_i(1-p_i)}\mathbf{G^{-1}}\hat{u}\right) \quad (12)$$

Then,

$$\text{var}(\hat{a})=\text{PEC}=\frac{1}{2\sum p_i(1-p_i)}\mathbf{Z'G^{-1}}(\mathbf{G}\sigma^2_u-\mathbf{LHS^{u2u2}})\mathbf{G^{-1}Z}\frac{1}{2\sum p_i(1-p_i)} \quad (13)$$

Therefore,

$$\text{var}(\hat{a})=\text{PEC}=\frac{1}{2\sum p_i(1-p_i)}\mathbf{Z'G^{-1}Z}\sigma^2_u-\mathbf{Z'G^{-1}LHS^{u2u2}G^{-1}Z}\frac{1}{2\sum p_i(1-p_i)} \quad (14)$$

Where $\mathbf{LHS^{u2u2}}$ is the inverse of the LHS matrix corresponding to genotyped animals.

Following ideas presented by Liu et al. (2017), once PEC is available, the accuracy of IP for an animal *i* can be calculated as follows:

$$ACC_{IPi} = \sqrt{1 - \frac{z_i \mathbf{var}(\hat{\mathbf{a}}) z_i'}{\sigma_u^2}} \quad (15)$$

Note that, to obtain SNP PEC, the inverse of the LHS matrix of MME is also required but may not be available for large genotyped populations. Therefore, strategies are needed to approximate SNP PEC for large datasets. The use of SNP PEC to compute accuracy of IP under a ssGBLUP model is investigated in chapter four.

REFERENCES

Aguilar, I., A. Legarra, F. Cardoso, Y. Masuda, D. Lourenco, and I. Misztal. 2019. Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. Genetics Selection Evolution 51 (1):28. doi: https://doi.org/10.1186/s12711-019-0469-3

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

Bangera, R., K. Correa, J. P. Lhorente, R. Figueroa, and J. M. Yáñez. 2017. Genomic predictions can accelerate selection for resistance against Piscirickettsia salmonis in Atlantic salmon (Salmo salar). BMC Genomics 18 (1):121. doi: https://doi.org/10.1186/s12864-017-3487-y

Barría, A., K. A. Christensen, G. Yoshida, A. Jedlicki, J. S. Leong, E. B. Rondeau, J. P. Lhorente, B. F. Koop, W. S. Davidson, and J. M. Yáñez. 2019. Whole genome linkage disequilibrium

and effective population size in a coho salmon (Oncorhynchus kisutch) breeding population using a high-density SNP array. Frontiers in genetics 10 (498) doi: https://doi.org/10.3389/fgene.2019.00498

Barría, A., K. A. Christensen, G. M. Yoshida, K. Correa, A. Jedlicki, J. P. Lhorente, W. S. Davidson, and J. M. Yáñez. 2018. Genomic predictions and genome-wide association study of resistance against Piscirickettsia salmonis in Coho Salmon (Oncorhynchus kisutch) using ddRAD sequencing. G3: Genes|Genomes|Genetics doi: https://doi.org/10.1534/g3.118.200053

Besson, M., F. Allal, B. Chatain, A. Vergnet, F. Clota, and M. Vandeputte. 2019. Combining individual phenotypes of feed intake with genomic data to improve feed efficiency in sea bass. Frontiers in genetics 10 (219) doi: https://doi.org/10.3389/fgene.2019.00219

Bradford, H. L., Y. Masuda, P. M. VanRaden, A. Legarra, and I. Misztal. 2019. Modeling missing pedigree in single-step genomic BLUP. Journal of Dairy Science 102 (3):2336-2346. doi: https://doi.org/10.3168/jds.2018-15434

Bradford, H. L., I. Pocrnić, B. O. Fragomeni, D. A. L. Lourenco, and I. Misztal. 2017. Selection of core animals in the Algorithm for Proven and Young using a simulation model. Journal of Animal Breeding and Genetics 134 (6):545-552. doi: https://doi.org/10.1111/jbg.12276

Brown, D., A. Swan, V. Boerner, L. Li, P. Gurman, A. McMillan, J. V. D. Werf, H. Chandler, B. Tier, and R. Banks. 2018. Single Step Genetic Evaluations in the Australian Sheep Industry Proceedings of the World Congress on Genetics Applied to Livestock Production No. Species - Ovine. p 460.

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. Journal of animal science 89 (9):2673-2679. doi: 10.2527/jas.2010-3555

Correa, K., R. Bangera, R. Figueroa, J. P. Lhorente, and J. M. Yáñez. 2017. The use of genomic information increases the accuracy of breeding value predictions for sea louse (Caligus rogercresseyi) resistance in Atlantic salmon (Salmo salar). Genetics Selection Evolution 49 (1):15. doi: https://doi.org/10.1186/s12711-017-0291-8

Erbe, M., C. Edel, E. C. G. Pimentel, J. dodenhoff, and K. U. Gotz. 2018. Approximation of reliability in single step models using the interbull standardized genomic reliability method. Interbull Bulletin (54)

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

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. O., D. A. L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. J. Lawlor, and I. Misztal. 2015. Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. Journal of Dairy Science 98 (6):4090-4094. doi: https://doi.org/10.3168/jds.2014-9125

Gualdron Duarte, J. L., R. J. Cantet, R. O. Bates, C. W. Ernst, N. E. Raney, and J. P. Steibel. 2014. Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. BMC Bioinformatics 15:246. doi: https://doi.org/10.1186/1471-2105-15-246

Gutierrez, A. P., O. Matika, T. P. Bean, and R. D. Houston. 2018. Genomic selection for growth traits in Pacific Oyster (Crassostrea gigas): potential of low-density marker panels for breeding value prediction. Frontiers in genetics 9 (391) doi: http://doi.org/10.3389/fgene.2018.00391

Henderson, C. R. 1984. Applications of linear models in animal breeding. University of Guelph Guelph.

Johnston, D., M. Ferdosi, N. Connors, V. Boerner, J. Cook, C. Girard, A. Swan, and B. Tier. 2018. Implementation of single-step genomic BREEDPLAN evaluations in Australian beef cattle Proceedings of the World Congress on Genetics Applied to Livestock Production No. Theory to Application 1. p 269.

Joshi, R., A. Skaarud, M. de Vera, A. T. Alvarez, and J. Ødegård. 2020. Genomic prediction for commercial traits using univariate and multivariate approaches in Nile tilapia (Oreochromis niloticus). Aquaculture 516:734641. doi: https://doi.org/10.1016/j.aquaculture.2019.734641

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. doi: http://doi.org/10.3168/jds.2009-2061

Legarra, A., D. A. Lourenco, and Z. Vitezica. 2018. Bases for genomic prediction. http://genoweb.toulouse.inra.fr/~alegarra/GSIP.pdf.

Lillehammer, M., T. H. E. Meuwissen, and A. K. Sonesson. 2013. A low-marker density implementation of genomic selection in aquaculture using within-family genomic breeding values. Genetics Selection Evolution 45 (1):39. doi: https://doi.org/10.1186/1297-9686-45-39

Liu, Z., P. VanRaden, M. H. Lidauer, M. P. Calus, H. Benhajali, H. Jorjani, and V. Ducrocq. 2017. Approximating genomic reliabilities for national genomic evaluation. Interbull Bulletin 51

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

Lourenco, D. A. L., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015b. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. Genetics Selection Evolution 47 (1):56. doi: https://doi.org/10.1186/s12711-015-0137-1

Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, C. Y. Chen, W. O. 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

Masuda, Y., I. Misztal, S. Tsuruta, A. Legarra, I. Aguilar, D. A. L. Lourenco, B. O. Fragomeni, and T. J. Lawlor. 2016. Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. Journal of Dairy Science 99 (3):1968-1974. doi: https://doi.org/10.3168/jds.2015-10540

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.

Meyer, K. 1989. Approximate accuracy of genetic evaluation under an animal model. Livestock Production Science 21 (2):87-100. doi: https://doi.org/10.1016/0301-6226(89)90041-9

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: https://doi.org/10.1534/genetics.115.182089

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

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

Misztal, I., S. Tsuruta, I. Aguilar, A. Legarra, P. M. VanRaden, and T. J. Lawlor. 2013. Methods to approximate reliabilities in single-step genomic evaluation. Journal of Dairy Science 96 (1):647-654. doi: https://doi.org/10.3168/jds.2012-5656

Misztal, I., and G. R. Wiggans. 1988. Approximation of prediction error variance in large-scale animal models. Journal of Dairy Science 71:27-32. doi: https://doi.org/10.1016/S0022-0302(88)79976-2

Nielsen, H. M., A. K. Sonesson, H. Yazdi, and T. H. E. Meuwissen. 2009. Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes. Aquaculture 289 (3):259-264. doi: https://doi.org/10.1016/j.aquaculture.2009.01.027

Ødegård, J., and T. H. Meuwissen. 2014. Identity-by-descent genomic selection using selective and sparse genotyping. Genetics Selection Evolution 46 (1):3. doi: https://doi.org/10.1186/1297-9686-46-3

Palaiokostas, C., S. Cariou, A. Bestin, J.-S. Bruant, P. Haffray, T. Morin, J. Cabon, F. Allal, M. Vandeputte, and R. D. Houston. 2018a. Genome-wide association and genomic prediction of resistance to viral nervous necrosis in European sea bass (Dicentrarchus labrax) using RAD sequencing. Genetics Selection Evolution 50 (1):30. doi: https://doi.org/10.1186/s12711-018-0401-2

Palaiokostas, C., M. Kocour, M. Prchal, and R. D. Houston. 2018b. Accuracy of genomic evaluations of juvenile growth rate in Common Carp (Cyprinus carpio) using genotyping by sequencing. Frontiers in genetics 9 (82) doi: http://doi.org/10.3389/fgene.2018.00082

Pocrnic, I., D. A. L. 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: https://doi.org/10.1534/genetics.116.187013

Rupp, R., S. Mucha, H. Larroque, J. McEwan, and J. Conington. 2016. Genomic application in sheep and goat breeding. Animal Frontiers 6 (1):39-44. doi: 10.2527/af.2016-0006

Sae-Lim, P., A. Kause, M. Lillehammer, and H. A. Mulder. 2017. Estimation of breeding values for uniformity of growth in Atlantic salmon (Salmo salar) using pedigree relationships or single-step genomic evaluation. Genetics Selection Evolution 49 (1):33. doi: https://doi.org/10.1186/s12711-017-0308-3

Silva, R. M. O., J. P. Evenhuis, R. L. Vallejo, G. Gao, K. E. Martin, T. D. Leeds, Y. Palti, and D. A. L. 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):42. doi: https://doi.org/10.1186/s12711-019-0484-4

Sonesson, A. K., and T. H. Meuwissen. 2009. Testing strategies for genomic selection in aquaculture breeding programs. Genetics Selection Evolution 41:37. doi: https://doi.org/10.1186/1297-9686-41-37

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

Vallejo, R. L., R. M. O. Silva, J. P. Evenhuis, G. Gao, S. Liu, J. E. Parsons, K. E. Martin, G. D. Wiens, D. A. L. Lourenco, T. D. Leeds, and Y. Palti. 2018. Accurate genomic predictions for BCWD resistance in rainbow trout are achieved using low-density SNP panels: Evidence that long-range LD is a major contributing factor. Journal of Animal Breeding and Genetics 135 (4):263-274. doi: https://doi.org/10.1111/jbg.12335

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

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92 (1):16-24. doi: https://doi.org/10.3168/jds.2008-1514

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

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics research 94 (2):73-83. doi: https://doi.org/10.1017/s0016672312000274

Wiggans, G. R., P. M. VanRaden, and T. A. Cooper. 2015. Technical note: Rapid calculation of genomic evaluations for new animals. Journal of Dairy Science 98 (3):2039-2042. doi: https://doi.org/10.3168/jds.2014-8868

Yan, Y., G. Wu, A. Liu, C. Sun, W. Han, G. Li, and N. Yang. 2018. Genomic prediction in a nuclear population of layers using single-step models. Poultry Science 97 (2):397-402. doi: https://doi.org/10.3382/ps/pex320

Yáñez, J. M., R. D. Houston, and S. Newman. 2014. Genetics and genomics of disease resistance in salmonid species. Frontiers in genetics 5:415. doi: https://doi.org/10.3389/fgene.2014.00415

Yoshida, G. M., R. Bangera, R. Carvalheiro, K. Correa, R. Figueroa, J. P. Lhorente, and J. M. Yáñez. 2018. Genomic Prediction Accuracy for Resistance Against Piscirickettsia salmonis in Farmed Rainbow Trout. G3: Genes|Genomes|Genetics 8 (2):719-726. doi: https://doi.org/10.1534/g3.117.300499

Yoshida, G. M., J. P. Lhorente, K. Correa, J. Soto, D. Salas, and J. M. Yáñez. 2019. Genome-Wide Association Study and Cost-Efficient Genomic Predictions for Growth and Fillet Yield in Nile Tilapia (Oreochromis niloticus). G3: Genes|Genomes|Genetics 9 (8):2597-2607. doi: https://doi.org/10.1534/g3.119.400116

CHAPTER 2

DEVELOPMENT OF GENOMIC PREDICTIONS FOR HARVEST AND CARCASS WEIGHT IN CHANNEL CATFISH[1]

[1] Garcia A.L.S, Bosworth B., Waldbieser G., Misztal I., Tsuruta S. and Lourenco D.A.L. 2018. *Genetics Selection Evolution*. 50(1):66. Reprinted here with permission of the publisher.

ABSTRACT

Catfish farming is the largest segment of US aquaculture and research is ongoing to improve production efficiency, including genetic selection programs to improve economically important traits. The objectives of this study were to investigate the use of genomic selection to improve breeding value accuracy and to identify major single nucleotide polymorphisms (SNPs) associated with harvest weight and residual carcass weight in a channel catfish population. Phenotypes were available for harvest weight (n = 27,160) and residual carcass weight (n = 6020), and 36,365 pedigree records were available. After quality control, genotypes for 54,837 SNPs were available for 2911 fish. Estimated breeding values (EBV) were obtained with traditional pedigree-based best linear unbiased prediction (BLUP) and genomic (G)EBV were estimated with single-step genomic BLUP (ssGBLUP). EBV and GEBV prediction accuracies were evaluated using different validation strategies. The ability to predict future performance was calculated as the correlation between EBV or GEBV and adjusted phenotypes. Compared to the pedigree BLUP, ssGBLUP increased predictive ability up to 28% and 36% for harvest weight and residual carcass weight, respectively; and GEBV were superior to EBV for all validation strategies tested. Breeding value inflation was assessed as the regression coefficient of adjusted phenotypes on breeding values, and the results indicated that genomic information reduced breeding value inflation. Genome-wide association studies based on windows of 20 adjacent SNPs indicated that both harvest weight and residual carcass weight have a polygenic architecture with no major SNPs (the largest SNPs explained 0.96 and 1.19% of the additive genetic variation for harvest weight and residual carcass weight respectively). Genomic evaluation improves the ability to predict future

performance relative to traditional BLUP and will allow more accurate identification of genetically superior individuals within catfish families.

INTRODUCTION

Catfish farming is the largest aquaculture segment in the US, accounting for approximately 50% of US food-fish production (Vilsack and Reilly, 2013). The US catfish industry is based on the production of channel catfish (*Ictalurus punctatus*) and the hybrid between the channel and blue catfish (*Ictalurus furcatus*). To provide a centralized source for US catfish production research, the USDA-ARS Warmwater Aquaculture Research Unit (WARU) was established in Stoneville, MS. As part of its mission to improve catfish production efficiency, the WARU has conducted a channel catfish breeding program since 2006, primarily selecting fish for increased growth and carcass yield.

Traditional evaluation using pedigree-based best linear unbiased prediction (BLUP) has been applied since the beginning of the breeding program at WARU. To investigate the potential for implementing genomic selection in the WARU catfish breeding program, animals were genotyped using a 57K single nucleotide polymorphism (SNP) array. Dense markers are used as an extra source of information to estimate breeding values (Meuwissen et al., 2001) in breeding programs for several livestock species because of the potential increase in accuracy of estimated breeding values (EBV). Another advantage of genomic selection, which is particularly important to aquaculture breeding, is the ability to exploit within-family genetic variation for animals that do not have records (Daetwyler et al., 2007).

One of the methods available for genomic evaluation is single-step genomic BLUP (ssGBLUP) (Aguilar et al., 2010). This method combines phenotypes, pedigree, and genotypes,

and potentially gives more accurate and less biased genomic EBV (GEBV) than multistep methods (Legarra et al., 2014). In ssGBLUP, the relationship matrix is a combination of pedigree and genomic relationships (Aguilar et al., 2010; Christensen and Lund, 2010); therefore, information on all animals can be used in the evaluation, regardless of genotyping status.

The accuracy of genomic evaluation depends on several factors including linkage disequilibrium (LD) between markers and quantitative trait loci (QTL), effective population size (N_e), and the relationship among individuals in training and validation data (Muir, 2007; Hayes et al., 2009). Thus, investigating the N_e and the extent of LD can give clues about how much genetic gain can be obtained by adopting genomic selection, how many animals should be genotyped, and potentially, how many SNPs should be included in the marker panel. The possibilities of using lower density SNP chips to reduce costs and promote adoption of genomic selection and searching for individual SNPs explaining major portions of variance should also be explored. If major SNPs explain a reasonable proportion of the genetic variance observed for a trait, selection based on a limited number of SNPs can be performed.

The first objective of this study was to investigate the feasibility of implementing genomic evaluation in US channel catfish by using ssGBLUP. The second objective was to determine the presence of potential regions in the genome that contain SNPs with major effects on harvest weight and residual carcass weight (i.e. carcass weight adjusted for harvest weight).

MATERIALS AND METHODS

DATA

Data from the USDA-ARS Warmwater Aquaculture Research Unit (WARU) were available for this study. Harvest weight and carcass weight (i.e., the weight of a fish with intact

skin, but removed head and viscera) were recorded from 2008 to 2015, with a total of 27,160 and 6020 records, respectively, and pedigree information was available for 36,365 fish. Among those, 27,883 had either phenotypes/genotypes or were related to phenotyped/genotyped fish.

This population constitutes the Delta Select strain that was developed based on 10 to 13 egg-masses collected from eight commercial catfish farms in the spring of 2006 (total = 97 egg masses). Each egg-mass was assumed to be a single full-sib family and families were assumed to be unrelated to each other. Each egg-mass was hatched in a separate hatching tank, fry were reared in separate full-sib family tanks until the fingerling stage when ~ 50 fish per family were tagged with passive integrated transponders (PIT tags) and stocked communally in earthen ponds where they were grown until the fall of 2007. At harvest, gender and weight of all fish were recorded, and an average of seven males and six females were randomly selected from each full-sib family and kept as broodfish. In addition to these fish, mature fish were obtained from two additional farms (40 males and 39 females from one farm, and 20 males and 59 females from the other farm). The broodfish from the base population were allowed to mate at random until 2 and 3 years old, and offspring represent the 2008 and 2009 year-class. Parentage was determined by genotyping fish for 16 microsatellites (Waldbieser and Bosworth, 2013). In total, 181 and 198 families were produced in 2008 and 2009, respectively. The families were reared separately until tagging (about 280 days old). Approximately 30 fish per family were tagged and reared communally in earthen pounds. Harvest weight was recorded when the animals were about 16 months old and a month later, approximately seven fish per family were processed for carcass weight recording.

Variance components and EBV were estimated and broodfish were selected using an index, which was the average standardized EBV for harvest weight and residual carcass weight. This approach was used to equalize selection emphasis on each trait. The fish selected from the 2008

and 2009 year-class (first generation of selection) were spawned in ponds in 2011 and 2012 as 2-, 3- and 4-year old fish. Performances of the 2011 and 2012 year-class progeny reflect effects of one generation of selection. Progeny from the 2011 and 2012 year-classes were evaluated and selected on the same index, spawned in ponds in 2014 and 2015 as 2-, 3- and 4-year old. Progeny from the 2014 and 2015 year-class were evaluated as described previously, and their performance reflects effects of the second generation of selection. Approximately 10% of the harvested fish from each year-class were kept as broodfish and no more than 10% of selected broodfish were from a single full-sib family to limit inbreeding. From 110 to 198 full-sib families were evaluated for each year-class and 954 and 752 full-sib families were evaluated for harvest weight and residual carcass weight, respectively.

Broodfish were stocked in March of each spawning year into 0.04 to 0.1 ha earthen ponds at a rate of 800 to 1000 kg per ha and stockings were designed to prevent mating among full-sibs. Male to female ratios in brood ponds ranged from 1:1 to 1:2. In early April, weighted plastic 'spawning-cans' were placed in ponds to provide spawning sites, and cans were inspected for the presence of egg-masses two or three times a week. Egg-masses were collected from ponds and transported to the hatchery. Fry were reared in separate full-sib tanks until the fingerling stage at which point, they were tagged and stocked communally in earthen ponds and fed daily. Appropriate commercial catfish diets were provided, and proper water quality was maintained throughout the study.

Genomic DNA from 49 founders of the Delta Select strain (described above) was sequenced with 2x150 bp reads on the NextSeq 500 platform (Illumina Inc., San Diego, CA) to obtain approximately 5X genome coverage per individual (25 to 40 million read pairs per individual). Paired sequences were aligned to the reference genome (Liu et al., 2016) using BWA-

MEM (Li, 2013) and variants were identified using the Genome Analysis ToolKit (DePristo et al., 2011). The GATK best practices workflow was used to identify SNPs and indels in individuals (HaplotypeCaller) and then jointly across the population (GenotypeGVCFs). The analysis produced more than 15 million raw variants (SNPs plus indels) and more than 12 million raw SNPs. Filtering for strand bias, map quality, and depth of coverage ($\leq$ mean + 2 standard deviations) reduced the number of high-quality putative SNPs to 7,445,905. Further filtration to identify SNPs that were positioned at least 50 bp from another SNP or indel and with a minor allele frequency higher than 0.05 reduced the number of candidate SNPs to 1,661,221.

An Axiom custom screening array (ThermoFisher Scientific, Waltham, MA) was produced using 660,000 SNPs, and 162 channel catfish were genotyped to validate the selected SNPs. Six doubled haploid (homozygous) catfish were also included to identify false heterozygosity at loci within genomic repeats. A total of 489,390 loci were called as polymorphic, high resolution loci on the array, and 340,737 loci were uniquely located on the catfish genome assembly. After the removal of 17,635 loci that demonstrated heterozygosity in the doubled haploids, 323,102 converted SNPs were available. A custom python script (Guangtu Gao, personal communication) was used to select SNPs that were evenly distributed across each of the 29 chromosomes. A new custom Axiom genotyping array was produced, which contained 57,354 SNPs with an average distance between markers of 13.3 kb. The final genotype data included 2911 animals, each genotyped at 54,837 SNPs after quality control. The SNPs excluded in the quality control had a minor allele frequency lower than 0.05, were monomorphic or had a call rate lower than 90%. Genotyped animals were excluded if the call rate was lower than 90% (i.e., 10% of the genotypes were missing). Among the animals that passed the quality control, 2826 had records on harvest

weight and 969 on carcass weight. The distribution of genotypes and phenotypes based on year-class is in Table 2.1.

MODEL AND ANALYSIS

Single-trait animal models were used for harvest weight and residual carcass weight. For harvest weight, the model was:

$$y_w = \mathbf{X}b + \mathbf{Z}u + \mathbf{W}p + e \quad (1)$$

where $\mathbf{y_w}$ is a vector of harvest weight; $\mathbf{b}$ is a vector of fixed effect of year-sex-pond interaction, and age (ranging from 391 to 620 days) as a linear covariable nested within sex; $\mathbf{u}$ is a vector of additive direct genetic effect; $\mathbf{p}$ is a vector of common environmental effect, which accounts for the fact that full-sibs from the same spawn were raised in the same tank until they reach an age and weight suitable for tagging (average tagging weight of 119.3 g and average tagging age of 271 days); $\mathbf{e}$ is the vector of residuals; $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{W}$ are incidence matrices for the effects contained in $\mathbf{b}$, $\mathbf{u}$, and $\mathbf{p}$, respectively.

For residual carcass weight, the model was:

$$y_c = \mathbf{X_1}b_1 + \mathbf{X_2}b_2 + \mathbf{Z}u + \mathbf{W}p + e \quad (2)$$

where $\mathbf{y_c}$ is a vector of carcass weight; $\mathbf{b_1}$ is a vector of linear covariables for body weight nested within year-sex interaction; $\mathbf{b_2}$ is a vector of fixed effect of year-sex-pond interaction; $\mathbf{u}$, $\mathbf{p}$, and $\mathbf{e}$ are described above; $\mathbf{X_1}$ and $\mathbf{X_2}$ are incidence matrices for the effects contained in $\mathbf{b_1}$ and $\mathbf{b_2}$. The term residual carcass weight arose from the fact that adjusting carcass weight to a common body weight allows identification of fish that have a higher proportion of whole weight as saleable carcass. The idea is similar to the residual feed intake which is widely used in livestock breeding.

Traditional BLUP and ssGBLUP analyses were performed using the BLUPF90 family of programs (Misztal et al., 2016). In the mixed model equations for ssGBLUP, the inverse of the

pedigree relationship matrix ($\mathbf{A}^{-1}$) is replaced by $\mathbf{H}^{-1}$ (Aguilar et al., 2010), the realized relationship matrix that combines pedigree and genomic relationships:

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix} \quad (3)$$

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

$$\mathbf{G} = \frac{\mathbf{MDM'}}{2\sum p_j(1-p_j)} \quad (4)$$

where $\mathbf{M}$ is a matrix of genotypes centered by twice the current allele frequencies ($\mathbf{p}$); j is the j^{th} locus; $\mathbf{D}$ is a diagonal matrix of SNP weights with a dimension equal to the number of SNPs. All SNPs were assumed to have homogeneous weights in ssGBLUP, meaning that $\mathbf{D}$ was an identity matrix ($\mathbf{I}$). To avoid singularity problems, $\mathbf{G}$ was blended with 5% of $\mathbf{A}_{22}$.

VALIDATION

The main interest in fish breeding is to better predict genetic merit of a fish as broodstock; however, the data collected so far during this first development of genomic predictions for catfish in the US do not allow a comparison between mid-parent GEBV and progeny performance, but this comparison will soon be possible. In our study, most of the genotyped animals with phenotypes were from the same year-class (i.e., 2015), precluding the use of validation on progeny performance and also forward prediction (i.e., future performance on individual fish). Therefore, to compare predictive ability of traditional and genomic evaluations, we conducted validations using several strategies to split fish into training and validation datasets.

Strategies 1 and 2 were used for both harvest weight and residual carcass weight. Strategy 1 was a random *k-fold* cross-validation, where the dataset was randomly split into k folds, predicting one fold based on k-1 folds. Genotyped animals with phenotypes were randomly split

into 5 or 10 mutually exclusive groups (k = 5 or k = 10, respectively). In each round of cross-validation, phenotypes from one group (i.e., validation group) were removed from the dataset and the remaining folds (i.e., training group) were used to predict the future performance for animals in the validation group. This k-folds cross-validation was replicated five times and results are presented as the mean and standard error for the five replicates. In the validation strategy 2, genotyped full-sibs were split into two groups with one group used for training and the other group used for validation, and all phenotypes of the validation group were removed from the evaluation. This scenario is most important when phenotypes are measured on sibs of the selection candidates.

Validation strategies 3 and 4 were conducted for residual carcass weight only to evaluate the importance of collecting genotypes on fish that will be slaughtered for phenotype recording. Carcass weight requires the slaughtering of many animals and thus their removal from the pool of selection candidates and is also considerably more expensive to measure than harvest weight. Harvest weight is quickly and inexpensively measured on all selection candidates and therefore, evaluating scenarios 3 and 4 for harvest weight provided no realistic benefit. Strategy 3 was similar to strategy 2 except that we assumed that only half of the full-sibs in the training population had phenotypes. This third validation strategy would be especially important for carcass traits to reduce the number of genotyped animals that are slaughtered to collect phenotypes. The validation group remained the same as in scenario 2.

In strategy 4, training animals had genotypes, but no phenotypes and the validation group remained the same. The ssGBLUP method uses all available information in the evaluation, meaning that phenotypes for 5051 ungenotyped, slaughtered fish were included. In this way, genotyped animals could benefit from phenotypes of ungenotyped animals if both groups are related through the pedigree relationship matrix although no genotyped animals had phenotypes

for carcass weight. This scenario was proposed because the cost of genotyping fish can be as high as the value of a fish itself. If genotyped fish have to be slaughtered for phenotype recording and they are removed as selection candidates, the cost of implementation of genomic selection would likely increase.

Trait heritabilities with the full data were 0.27 and 0.34 for harvest weight and residual carcass weight, respectively. As we changed the data structure by creating different training datasets for each validation strategy, we also estimated updated variance components to evaluate how changing the animals used in the training set analysis (which also changed the subsequent variance components) impacted predictive ability and inflation of (G)EBV. Reverter et al. (1994) pointed out that breeding value inflation or deflation can be introduced if variance components do not reflect the actual data.

Ability to predict performance was used to compare traditional and genomic models. It was calculated as the correlation between (G)EBV for validation animals and phenotypes adjusted for fixed effects ($\mathbf{y}^*$), as described in (1) and (2), which were estimated based on the full data:

$$\text{predictive ability}=\text{cor}[(\text{G})\text{EBV}, y^*] \quad (5)$$

In addition, the regression coefficient ($\mathbf{b_1}$) of adjusted phenotypes on (G)EBV was used as a measure of inflation of breeding values.

$$y^*=b_0+b_1\times(\text{G})\text{EBV} + e \quad (6)$$

A regression coefficient lower than one indicates (G)EBV inflation, whereas a value higher than one indicates deflation.

GENOME-WIDE ASSOCIATION

A genome-wide association study (GWAS) was performed to identify possible regions of the genome containing SNPs with major effects on harvest weight or residual carcass weight.

Weighted ssGBLUP (WssGBLUP; Wang et al. (2012)) implemented in postGSf90 from the BLUPF90 family of programs (Misztal et al., 2016) was used for the GWAS. In the first implementation of WssGBLUP, Wang et al. (2012) suggested that SNP weights should be calculated as $d_j=\hat{u}_j^2 2p_j\left(1-p_j\right)$, following the formula for genetic variance due to an additive locus (Falconer and Mackay, 1996). However, Lourenco et al. (2017) showed that this method did not reach convergence under a more polygenic scenario because of extreme weights. Therefore, the SNP weights used in this study were described by VanRaden (2008) as non-linearA weights:

$$d_j=CT^{\frac{|\hat{u}_j|}{sd(\hat{u})}-2} \quad (7)$$

where **CT** is a constant that determines the departure from normality; $|\hat{\mathbf{u}}_\mathbf{j}|$ is the absolute estimated SNP effect for marker *j*, and $\mathbf{sd(\hat{u})}$ is the standard deviation of the vector of estimated SNP effects. Non-linearA weights had good convergence properties and avoided extreme values (Breno Fragomeni personal communication). This is because the maximum change in weights is limited by the minimum between 5 and the exponent of **CT**. In our study, **CT** received a value of 1.125 based on Legarra et al. (2018) and VanRaden (2008). Although these values were empirically derived based on dairy cattle populations, they resulted from tests over several traits with a more polygenic architecture.

The WssGBLUP is an iterative process. Wang et al. (2012) and Zhang et al. (2016) suggested that two iterations of weights were sufficient to maximize genomic accuracy and to correctly identify major SNPs in WssGBLUP. Based on the non-linearA weights, the number of iterations to reach convergence may vary from 5 to 10 (Breno Fragomeni personal communication) Therefore, we chose five iterations and checked the stability of predictive ability and regression coefficients of adjusted phenotypes on GEBV. Predictive ability and inflation can be used as indicators for convergence when computing SNP weights in WssGBLUP (Wang et al., 2012).

After investigating which iteration had the highest predictive ability, based on reduced data, WssGBLUP was applied to the full data for harvest weight and residual carcass weight, and Manhattan plots were drawn for that iteration.

Manhattan plots were drawn based on the proportion of additive genetic variance explained by windows of 20 adjacent SNPs. The concept of SNP windows is rather abstract and tries to approximate haplotype blocks; therefore, it assumes that windows may be inherited together, which may not always be the case for all assumed windows.

LINKAGE DISEQUILIBRIUM AND EFFECTIVE POPULATION SIZE

We used the first medium density SNP array (55K SNP) developed for channel catfish in this study. However, we also examined linkage disequilibrium (LD) to determine the feasibility of using a lower cost, reduced SNP panel for genomic selection in this population.

In our study, LD was calculated with preGSf90 using the following equation:

$$r^2=\frac{D^2}{P_A P_a P_B P_b} \quad (8)$$

where $D= P_{AB} - P_A P_B$; $\mathbf{P_{AB}}$ is the frequency of the genotype AB; $\mathbf{P_A}$, $\mathbf{P_a}$, $\mathbf{P_B}$ and $\mathbf{P_b}$ are the allele frequencies. The LD was calculated as the average of adjacent SNPs within chromosomes and across the genome.

A curve that fits the LD decay with distance between markers for each chromosome was calculated by fitting the equation proposed by Sved (1971):

$$E[r_t^2]= \frac{1}{1+4Ne_t d_{ij}} \quad (9)$$

Where $\mathbf{d_{ij}}$ is the distance between markers i and j in Morgan and $\mathbf{Ne_t}$ is the effective population size for the chromosome t, calculated as proposed by Saura et al. (2015):

$$Ne_t=(4d_t)^{-1}\left[\left(r_t^2-N^{-1}\right)^{-1}-\alpha\right] \quad (10)$$

With $\mathbf{d_t}$ as the average chromosome length in Morgan; $\mathbf{r_t^2}$ is the average LD at chromosome *t;* $\mathbf{N^{-1}}$ is the adjustment term for sample size (number of genotyped animals); and $\boldsymbol{\alpha}$ is a fixed parameter that is assumed to be 1 if mutation is not considered and 2 if it is considered; we considered $\boldsymbol{\alpha=2}$.

Besides chromosome-based N_e, we also calculated N_e based on the rate of inbreeding by generation using the of formula Falconer and Mackay (1996):

$$Ne_F = \frac{1}{2\Delta F} \quad (11)$$

where

$$\Delta F = \frac{F_n - F_{n-1}}{1 - F_{n-1}} \quad (12)$$

with $\mathbf{F_n}$ as the inbreeding coefficient in the $\mathbf{n^{th}}$ generation.

RESULTS AND DISCUSSION

PREDICTIVE ABILITY AND INFLATION

Table 2.2 shows the predictive ability for both traits under different validation strategies. In all validations, using genomic information through ssGBLUP improved the ability to predict future fish performance relative to traditional BLUP.

In general, cross-validation scenarios using either k = 5 or k = 10-fold scenarios had very similar predictive ability. In addition, updating the variance components for different training datasets did not affect predictive ability, as expected (Reverter et al., 1994). Including genomic information increased predictive ability by 28% (for both 5 and 10-fold) for harvest weight, and by 29% and 33% (5 and 10-fold, respectively) for residual carcass weight relative to traditional BLUP.

Validation strategy 2 (splitting full sibs into training and validation sets) resulted in overall

predictive abilities for traditional BLUP and ssGBLUP that were greater compared to k-fold cross-validations. This was likely due to closer relationships between animals in training and validation groups (Tsai et al., 2016) in strategy 2. The ssGBLUP outperformed BLUP by 23% for harvest weight and by 36% for residual carcass weight in strategy 2. Genomic information may have more impact on traits that cannot be measured on the selection candidates (Meuwissen et al., 2016), such as carcass and disease resistance traits. For instance, in our study the greatest increase in predictive ability was for residual carcass weight.

Validation strategy 3, where only a portion of the full-sibs in the training set had phenotypes, had a predictive ability slightly higher than strategy 4 (no phenotypes on genotyped animals), but lower than those for validation on full-sibs with genotypes and phenotypes (strategy 2) and k-folds cross-validation (strategy 1). The gain in predictive ability of GEBV over EBV in strategy 3 was 22% for residual carcass weight. The drop in predictive ability for residual carcass weight for strategy 3 relative to strategies 1 and 2 was caused by the reduction in the number of phenotypes available to estimate breeding values.

Validation strategy 4 represented the situation where genotyped fish had no phenotypes in the dataset, which would eliminate the need to process genotyped fish. Predictive ability for residual carcass weight EBV decreased from 0.24 to 0.22, and of GEBV from 0.31 to 0.24. These results suggest that having genotypes for fish that are slaughtered for carcass weight recording is important and translates into the greatest benefit from genomic selection. Having phenotypes for genotyped individuals is important not only in aquaculture genomics, but in general livestock genomics. In a simulation study, Pszczola et al. (2012) showed that the highest accuracies from genomic evaluation were obtained when animals from both reference (phenotyped) and evaluated (non phenotyped) populations were genotyped. Furthermore, Lourenco et al. (2015a) showed only

one point increase in predictive ability in the genomic evaluation for calving ease in American Angus and related that to the small number of genotyped animals with records on difficult calving.

Although predictive ability decreased considerably when carcass records for genotyped fish were removed, ssGBLUP still outperformed traditional BLUP by about 9%. The improved performance of ssGBLUP in this situation is due to the fact that the **H** matrix connects genotyped animals without phenotypes to ungenotyped animals with phenotypes, if they are connected through the pedigree.

Overall, the use of genomic information improved the calculation of relationships among animals and allowed for a better estimation of Mendelian sampling, promoting an increase in predictive ability and allowing the use of within-family variation. Without genomic information, young full-sib fish (i.e., without phenotype or progeny) would have the same EBV for a trait, which equals to parent average (Daetwyler et al., 2010).

Lourenco et al. (2015b) showed that when an animal is genotyped but has no phenotype and progeny, the GEBV is composed of:

$$GEBV = w_1PA + w_2GP - w_3PP \quad (13)$$

where **PA** is the parent average EBV for the animal, **GP** is the portion of prediction due to the genomic information, coming from **G**, and **PP** is pedigree prediction that comes from $\mathbf{A}_{22}$; weights **w1** to **w3** sum to 1. Quaas (1988) described that the breeding value of an animal is the average of EBV from parents (**PA**) plus a random term that considers the uncertainty about which 50% of the genes were passed to progeny (i.e., Mendelian sampling):

$$EBV = 0.5EBV_S + 0.5EBV_D + \varphi \quad (14)$$

where $\mathbf{EBV_S}$ is EBV from sire; $\mathbf{EBV_D}$ is EBV form dam and $\boldsymbol{\varphi}$ is the Mendelian sampling term. If the first portion of the formula corresponds to **PA**, $\boldsymbol{\varphi}$ can be partially estimated by the genomic information present in **GP**, as shown in Eq. (13), because genomic data helps to estimate part of the uncertainty about which alleles and the proportion of alleles shared among individuals. Therefore, genotyped full-sibs that are selection candidates (i.e., young) have unique GEBV (not just **PA**) and the best candidates can be identified within families. Figure 2.1 shows the distribution of GEBV for a family of 34 full-sibs that had no phenotypes for residual carcass weight but were genotyped. Without genomic information, all 34 full-sibs had only **PA**, which is equal to 4.64 g. After including genomic information for all full-sibs, we observed a distribution ranging from 1.24 to 7.65. Use of GEBV would allow selection of fish within a family based on individual genetic merit for carcass weight, avoiding random selection of fish within a family based on BLUP $\mathbf{EBV_S}$, which could result in selecting fish with in fact low genetic merit.

The ability to identify selection candidates within a family that have higher genetic merit is a key benefit for a trait such as carcass weight in fish, which is not measured on selection candidates, and for quite large full-sib family sizes. Studies on other fish species such as Atlantic salmon (Odegard et al., 2014; Tsai et al., 2015; Bangera et al., 2017; Correa et al., 2017) and rainbow trout (Vallejo et al., 2017; Yoshida et al., 2018) have demonstrated increases in predictive ability or accuracy of GEBV compared to EBV, confirming the benefits of genomic selection for aquaculture species.

Tables 2.3 and 2.4 present EBV and GEBV inflation ($\mathbf{b_1}$) for harvest weight and residual carcass weight. In all validation scenarios, GEBV were less inflated or deflated compared to EBV, meaning that GEBV were closer in scale to the adjusted phenotypes. Updating variance components for each training dataset was beneficial for estimating inflation for both EBV and

GEBV. The benefit comes from the fact that the variance components used to predict (G)EBV reflect the true state of the population after removing phenotypes for validation animals and therefore, less inflation is expected. Wiggans et al. (2011) suggested that one way to reduce inflation of genomic evaluations of US cows would be to reduce heritability; this would be in line with a reduced additive genetic variation in recent generations. In our study, when variance components were re-estimated, the regression coefficients became closer to 1 and were the most beneficial for the cross-validation scenario for harvest weight, in which $\mathbf{b_1 = 1}$ for GEBV, meaning that GEBV and adjusted phenotypes had similar dispersion.

GENOME-WIDE ASSOCIATION

Manhattan plots from the GWAS for harvest weight and residual carcass weight are shown in Figures. 2.2 and 2.3, respectively. The plots were drawn for the first iteration of WssGBLUP, because it had the greatest predictive ability and least inflation. In the first iteration, GEBV were computed assuming that all SNPs had the same weight. The GEBV were then back-solved to SNP effects and new weights were calculated and plotted as percentage of variance explained. Although predictive ability had to be computed based on the reduced dataset, the Manhattan plots were drawn based on the full dataset. The proportion of additive genetic variance explained by windows of 20 adjacent SNPs was up to 0.96% for harvest weight and up to 1.19% for residual carcass weight, which indicates that both traits are extremely polygenic. A single window explaining close to 1% of the additive genetic variation for harvest weight was located on chromosome 19, whereas, for residual carcass weight the top windows were located on chromosomes 13 and 21.

In an experimental population of less than 600 genotyped progeny of F1 males (channel x blue catfish) and channel catfish females, Li et al. (2018) found a significant association between SNPs on chromosome 5 and body weight. These SNPs explained from 3.69 to 6.72% of the

phenotypic variance for body weight. In a rainbow trout population from the National Center for Cool and Cold Water Aquaculture, Gonzalez-Pena et al. (2016) found windows of 20 SNPs that explained more than 1% of the additive genetic variance for body weight at 10 and 13 months on chromosome 5, for fillet weight and yield on chromosome 9, and for carcass weight on chromosomes 9, 17, and 27. In our study, the windows that explained the top variance did not overlap with windows already described in the literature for the same species or trait.

The fact that top windows do not overlap even in populations from the same species has been described in the literature. Silva et al. (2018) found very few overlapping genomic windows that explained more than 1% of the additive genetic variance for columnaris disease in two different rainbow trout populations. Fragomeni et al. (2014) showed that, in a selected commercial broiler chicken population, the location of the windows with the largest effect was not consistent across different generations.

With a polygenic architecture and windows of SNPs explaining small proportions of the additive genetic variance, genomic selection for harvest weight and residual carcass weight in this catfish population is preferred over marker-assisted selection (MAS). Using MAS with such an architecture would not provide successful results given that only a small proportion of variance can be explained by individual SNPs.

Under a polygenic architecture, the use of Bayesian alphabet (e.g., BayesA, BayesB) and GBLUP-based methods that allow SNPs to explain a different proportion of variance (i.e., different SNP weightings; (Daetwyler et al., 2010; Zhang et al., 2016)) may not help to increase the predictive ability or accuracy of GEBV. In fact, we observed that predictive ability for harvest weight and residual carcass weight did not change over the iterations of WssGBLUP when using non-linearA weights (results not shown). In addition, inflation slightly increased from iterations 1

to 3, reaching a plateau in later iterations (results not shown). When the best results for predictive ability and inflation are obtained in the first iteration of WssGBLUP, we can assume that using different weights is not beneficial, and, in this case, GEBV obtained from WssGBLUP are the same as in ssGBLUP. In a simulation study using linear SNP weights (i.e., $\mathbf{d_j=\hat{u}_j^2 2p_j\left(1-p_j\right)}$), Lourenco et al. (2017) found that for more polygenic traits, decreases in accuracy or increases in inflation/deflation for WssGBLUP could be caused by the shrinkage of SNP weights for SNPs with smaller effects.

Although Manhattan plots were drawn based on the first iteration of WssGBLUP, the percentage of variance explained by SNPs did not change considerably over iterations. In fact, there was no change from iterations 2 to 5 for harvest weight and 3 to 5 for residual carcass weight. This possibly shows that non-linearA weights are not overestimated and they converge at some point. This convergence occurs because the formula contains a maximum limit for SNP weight. In an attempt to use the linear weights, we observed a constant increase in the proportion of variance explained (results not shown). This increase is due to the fact SNP weights keep changing over iterations without a limit for maximum change.

LINKAGE DISEQUILIBRIUM AND EFFECTIVE POPULATION SIZE

The overall whole-genome LD was 0.22 and ranged from a low value of 0.12 (chromosome 29) to a high value of 0.37 (chromosome 17). The LD was moderate even at long distances as shown in the LD decay plots in Figure. 2.4. There was a large, conservative LD block, which did not decay even at long distances (20 Mb) on chromosome 17, and a more in-depth investigation is needed to understand what might have caused this LD pattern.

The effective population size calculated based on LD and that based on inbreeding did not differ much, i.e. 27 and 28, respectively. Compared to livestock species, $\mathbf{N_e}$ in catfish is relatively

small. Pocrnic et al. (2016b) showed that $\mathbf{N_e}$ for broiler chicken, swine, Angus cattle, Jersey, and Holstein cattle were 44, 32, 113, 101, and 149, respectively. In studies based on simulated populations, Pocrnic et al. (Pocrnic et al., 2016a) and Muir (Muir, 2007) associated $\mathbf{N_e}$ with the dimensionality of the genomic information and showed higher accuracy of genomic predictions for smaller $\mathbf{N_e}$. When $\mathbf{N_e}$ is small, there are fewer and longer LD blocks, which can be well estimated even when the number of genotyped animals is less than 5000 (Lourenco et al., 2017). In this way, the small $\mathbf{N_e}$ in this catfish population may have contributed to the great increase in predictive ability even when only 2911 fish were genotyped (i.e., 8% of the population).

Considering the small effective population size and the long-range LD in this population, it might be possible to reduce the number of markers needed for genomic selection. Other studies have demonstrated similar accuracies when comparing low- and high-density SNP panels in salmonid species (Odegard et al., 2014; Tsai et al., 2016; Bangera et al., 2017; Yoshida et al., 2018). Recently, Vallejo et al. (2018) reported gains of accuracy (relative to traditional BLUP) of 88% for a 35K SNP panel and 42% with a greatly reduced 200 SNP panel with ssGBLUP for bacterial cold water disease resistance in rainbow trout. The authors related the efficiency of the reduced SNP panel to the strong long-range LD in that rainbow trout population.

Reducing the density of markers in the panel would likely reduce genotyping costs and improve the cost efficiency of genomic selection in fish. More studies are necessary to investigate the overall cost and benefit of different SNP panel densities on implementation of genomic selection in this catfish population.

CONCLUSIONS

Genomic information is beneficial for channel catfish breeding because it provides a

greater ability to predict future performance and reduces inflation of breeding values. For carcass traits, it is important to record carcass weight phenotypes on genotyped fish to obtain the largest advantage of genomic selection. Genomic information also allows the estimation of Mendelian sampling, enabling the identification of genetically superior individuals within families, which is not possible with pedigree information only. Genome-wide association suggests that harvest weight and residual carcass weight have a polygenic architecture, indicating that using many SNPs in a genome-wide selection approach would be superior to using fewer SNPs in a marker-assisted selection type of approach.

REFERENCES

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

Bangera, R., K. Correa, J. P. Lhorente, R. Figueroa, and J. M. Yáñez. 2017. Genomic predictions can accelerate selection for resistance against Piscirickettsia salmonis in Atlantic salmon (Salmo salar). BMC Genomics 18 (1):121. doi: https://doi.org/10.1186/s12864-017-3487-y

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

Correa, K., R. Bangera, R. Figueroa, J. P. Lhorente, and J. M. Yáñez. 2017. The use of genomic information increases the accuracy of breeding value predictions for sea louse (Caligus

rogercresseyi) resistance in Atlantic salmon (Salmo salar). Genetics Selection Evolution 49 (1):15. doi: https://doi.org/10.1186/s12711-017-0291-8

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: https://doi.org/10.1534/genetics.110.116855

Daetwyler, H. D., B. Villanueva, P. Bijma, and J. A. Woolliams. 2007. Inbreeding in genome-wide selection. Journal of Animal Breeding and Genetics 124 (6):369-376. doi: https://doi.org/10.1111/j.1439-0388.2007.00693.x

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. Del Angel, M. A. Rivas, and M. Hanna. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 43 (5):491. doi: https://doi.org/10.1038/ng.806

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. ed. Longman, Essex, England :.

Fragomeni, B. d. O., I. Misztal, D. L. Lourenco, I. Aguilar, R. Okimoto, and W. M. Muir. 2014. Changes in variance explained by top SNP windows over generations for three traits in broiler chicken. Frontiers in genetics 5 (332) doi: https://doi.org/10.3389/fgene.2014.00332

Gonzalez-Pena, D., G. Gao, M. Baranski, T. Moen, B. M. Cleveland, P. B. Kenney, R. L. Vallejo, Y. Palti, and T. D. Leeds. 2016. Genome-Wide Association Study for Identifying Loci that Affect Fillet Yield, Carcass, and Body Weight Traits in Rainbow Trout (Oncorhynchus mykiss). Frontiers in genetics 7 (203) doi: https://doi.org/10.3389/fgene.2016.00203

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

Legarra, A., O. F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. Livestock Science 166:54-65. doi: https://doi.org/10.1016/j.livsci.2014.04.029

Legarra, A., D. A. Lourenco, and Z. Vitezica. 2018. Bases for genomic prediction. http://genoweb.toulouse.inra.fr/~alegarra/GSIP.pdf 2018).

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997

Li, N., T. Zhou, X. Geng, Y. Jin, X. Wang, S. Liu, X. Xu, D. Gao, Q. Li, and Z. Liu. 2018. Identification of novel genes significantly affecting growth in catfish through GWAS analysis. Molecular Genetics and Genomics 293 (3):587-599. doi: https://doi.org/10.1007/s00438-017-1406-1

Liu, Z., S. Liu, J. Yao, L. Bao, J. Zhang, Y. Li, C. Jiang, L. Sun, R. Wang, Y. Zhang, T. Zhou, Q. Zeng, Q. Fu, S. Gao, N. Li, S. Koren, Y. Jiang, A. Zimin, P. Xu, A. M. Phillippy, X. Geng, L. Song, F. Sun, C. Li, X. Wang, A. Chen, Y. Jin, Z. Yuan, Y. Yang, S. Tan, E. Peatman, J. Lu, Z. Qin, R. Dunham, Z. Li, T. Sonstegard, J. Feng, R. G. Danzmann, S. Schroeder, B. Scheffler, M. V. Duke, L. Ballard, H. Kucuktas, L. Kaltenboeck, H. Liu, J. Armbruster, Y. Xie, M. L. Kirby, Y. Tian, M. E. Flanagan, W. Mu, and G. C. Waldbieser. 2016. The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. Nature Communications 7:11757. doi: https://doi.org/10.1038/ncomms11757

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

Lourenco, D. A. L., B. O. Fragomeni, H. L. Bradford, I. R. Menezes, J. B. S. 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: https://doi.org/10.1111/jbg.12288

Lourenco, D. A. L., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015b. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. Genetics Selection Evolution 47 (1):56. doi: https://doi.org/10.1186/s12711-015-0137-1

Meuwissen, T., B. Hayes, and M. Goddard. 2016. Genomic selection: A paradigm shift in animal breeding. Animal Frontiers 6 (1):6-14. doi: https://doi.org/10.2527/af.2016-0002

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., S. Tsuruta, D. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. 2016. Manual for BLUPF90 family of programs.

Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding and Genetics 124 (6):342-355. doi: https://doi.org/10.1111/j.1439-0388.2007.00700.x

Odegard, J., T. Moen, N. Santi, S. A. Korsvoll, S. Kjoglum, and T. H. Meuwissen. 2014. Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). Frontiers in genetics 5:402. doi: https://doi.org/10.3389/fgene.2014.00402

Pocrnic, I., D. A. L. 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: https://doi.org/10.1534/genetics.116.187013

Pocrnic, I., D. A. L. 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 48:82. doi: https://doi.org/10.1186/s12711-016-0261-6

Pszczola, M., T. Strabel, J. A. M. van Arendonk, and M. P. L. Calus. 2012. The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection. Journal of Dairy Science 95 (9):5412-5421. doi: https://doi.org/10.3168/jds.2012-5550

Quaas, R. L. 1988. Additive Genetic Model with Groups and Relationships. Journal of Dairy Science 71 (5):1338-1345. doi: https://doi.org/10.3168/jds.S0022-0302(88)79691-5

Reverter, A., B. Golden, R. Bourdon, and J. Brinks. 1994. Method R variance components procedure: application on the simple breeding value model. Journal of animal science 72 (9):2247-2253. doi: https://doi.org/10.2527/1994.7292247x

Saura, M., A. Tenesa, J. A. Woolliams, A. Fernández, and B. Villanueva. 2015. Evaluation of the linkage-disequilibrium method for the estimation of effective population size when generations overlap: an empirical case. BMC Genomics 16 (1):922. doi: https://doi.org/10.1186/s12864-015-2167-z

Silva, R. M. O., E. J.P, R. Vallejo, G. Gao, K. E. Martin, I. Misztal, T. D. Leeds, D. A. Lourenco, and Y. Palti. 2018. GWAS for Detecting QTL Associated with Columnaris Disease in Two Rainbow Trout Breeding Populations. In: Plant & Animal Genome, San Diego

Sved, J. A. 1971. Linkage Disequilibrium and Homozygosity of chromosome segments in finite populations.pdf. Theoretical popularion biology 2:125- 141.

Tsai, H.-Y., A. Hamilton, A. E. Tinch, D. R. Guy, J. E. Bron, J. B. Taggart, K. Gharbi, M. Stear, O. Matika, R. Pong-Wong, S. C. Bishop, and R. D. Houston. 2016. Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations. Genetics Selection Evolution 48 (1):47. doi: https://doi.org/10.1186/s12711-016-0226-9

Tsai, H.-Y., A. Hamilton, A. E. Tinch, D. R. Guy, K. Gharbi, M. J. Stear, O. Matika, S. C. Bishop, and R. D. Houston. 2015. Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. BMC Genomics 16 (1):969. doi: https://doi.org/10.1186/s12864-015-2117-9

Vallejo, R. L., T. D. Leeds, G. Gao, J. E. Parsons, K. E. Martin, J. P. Evenhuis, B. O. Fragomeni, G. D. Wiens, and Y. Palti. 2017. Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree-based model in rainbow trout aquaculture. Genetics Selection Evolution 49 (1):17. doi: https://doi.org/10.1186/s12711-017-0293-6

Vallejo, R. L., R. M. O. Silva, J. P. Evenhuis, G. Gao, S. Liu, J. E. Parsons, K. E. Martin, G. D. Wiens, D. A. L. Lourenco, T. D. Leeds, and Y. Palti. 2018. Accurate genomic predictions for BCWD resistance in rainbow trout are achieved using low-density SNP panels: Evidence that long-range LD is a major contributing factor. Journal of Animal Breeding and Genetics doi: https://doi.org/10.1111/jbg.12335

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

Vilsack, T., and J. T. Reilly. 2013. Census of aquaculture 2013. In: U. S. D. o. Agriculture (ed.) No. 3. p 1-98. USDA, National Agricultural Statistics Service.

Waldbieser, G. C., and B. G. Bosworth. 2013. A standardized microsatellite marker panel for parentage and kinship analyses in channel catfish, Ictalurus punctatus. Animal Genetics 44 (4):476-479. doi: https://doi.org/10.1111/age.12017

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics research 94 (2):73-83. doi: https://doi.org/10.1017/s0016672312000274

Wiggans, G. R., T. A. Cooper, P. M. VanRaden, and J. B. Cole. 2011. Technical note: Adjustment of traditional cow evaluations to improve accuracy of genomic predictions. Journal of Dairy Science 94 (12):6188-6193. doi: https://doi.org/10.3168/jds.2011-4481

Yoshida, G. M., R. Bangera, R. Carvalheiro, K. Correa, R. Figueroa, J. P. Lhorente, and J. M. Yáñez. 2018. Genomic Prediction Accuracy for Resistance Against Piscirickettsia salmonis in Farmed Rainbow Trout. G3: Genes|Genomes|Genetics 8 (2):719-726. doi: https://doi.org/10.1534/g3.117.300499

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: https://doi.org/10.3389/fgene.2016.00151

TABLES

Table 2.1 Distribution of phenotypes and genotypes by year-class.

Year-class	Full-sib families	Harvest weight	Carcass weight	Genotyped animals
Before 2006	-	-	-	70
2006	-	-	-	2
2008	181	4762	829	78
2009	198	5686	1352	44
2011	180	1982	-	38
2012	110	4484	924	133
2014	113	4141	955	189
2015	172	6105	1960	2357
Total	954	27,160	6020	2911

Table 2.2 Predictive ability for harvest weight and residual carcass weight under BLUP and ssGBLUP for all validation scenarios.

Validation strategy	Validation scenarios[a]	Harvest weight		Residual carcass weight	
		BLUP	ssGBLUP	BLUP	ssGBLUP
1	5-fold cross-validation[b]	0.29 $^{0.001}$	0.37 $^{0.001}$	0.24 $^{0.002}$	0.31 $^{0.002}$
1	10-fold cross-validation[b]	0.29 $^{0.0003}$	0.37 $^{0.0004}$	0.24 $^{0.002}$	0.32 $^{0.002}$
2	Full sib validation	0.31	0.38	0.25	0.34
3	Half of the full sibs with phenotypes	-	-	0.23	0.28
4	No phenotypes for all genotyped animals	-	-	0.22	0.24

[a]Updating variance components or not produced exactly the same predictive ability for all scenarios. [b]Average and standard error across five replicates.

Table 2.3 Regression coefficients of adjusted phenotypes on EBV or GEBV for harvest weight.

Validation strategy	Validation scenario	Same variance components		Updated variance components	
		BLUP	ssGBLUP	BLUP	ssGBLUP
1	5-fold cross-validation[a]	0.87 $^{0.002}$	0.92 $^{0.002}$	0.97 $^{0.002}$	1.00 $^{0.002}$
1	10-fold cross-validation	0.87 $^{0.001}$	0.92 $^{0.001}$	0.96 $^{0.001}$	1.00 $^{0.001}$
2	Full sib validation	0.94	0.98	1.05	1.04

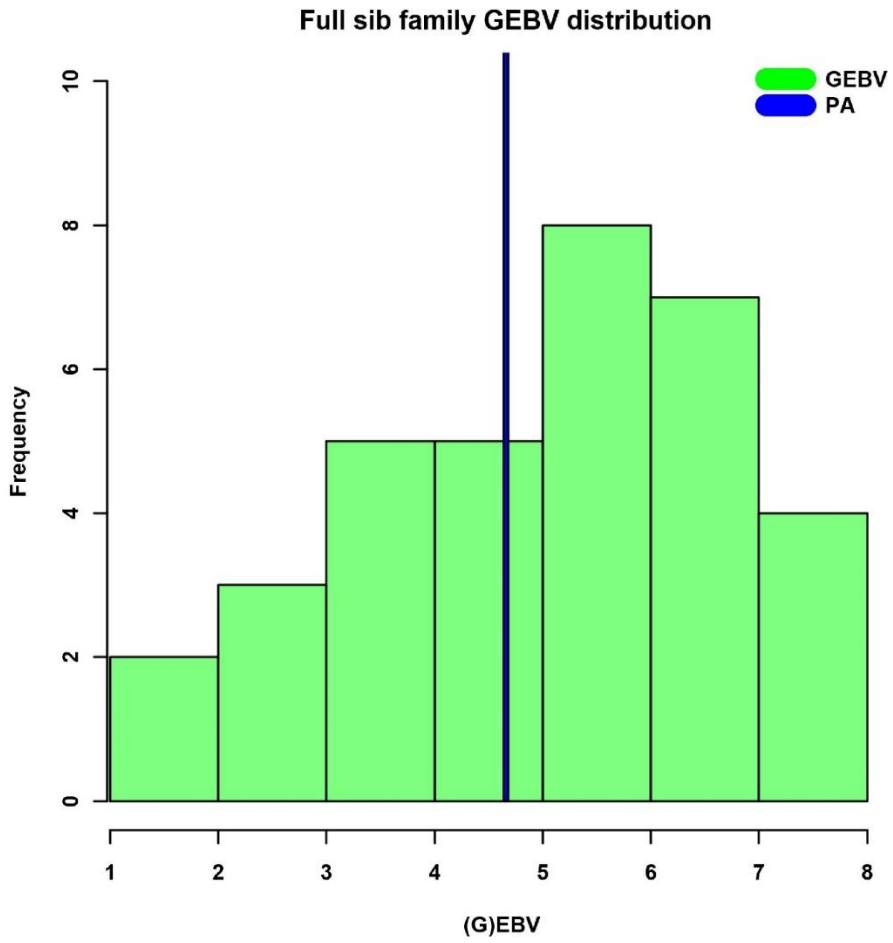
[a]Average and standard error across five replicates.

Table 2.4 Regression coefficients of adjusted phenotypes on EBV or GEBV for residual carcass weight.

Validation strategy	Validation scenario	Same variance components		Updated variance components	
		BLUP	ssGBLUP	BLUP	ssGBLUP
1	5-fold cross-validation[a]	0.80 $^{0.008}$	0.91 $^{0.007}$	0.89 $^{0.03}$	0.94 $^{0.007}$
1	10-fold cross-validation[a]	0.80 $^{0.008}$	0.92 $^{0.005}$	0.82 $^{0.008}$	0.95 $^{0.005}$
2	Full sib validation	0.83	1.08	0.85	1.10
3	Half of the full sibs with phenotypes	0.75	0.95	0.77	0.98
4	No phenotypes for all genotyped animals	0.76	0.87	0.79	0.90

[a]Average and standard error across five replicates.

FIGURES



3.1 Distribution of genomic EBV for residual carcass weight (g) in a family of 34 young genotyped full-sibs.

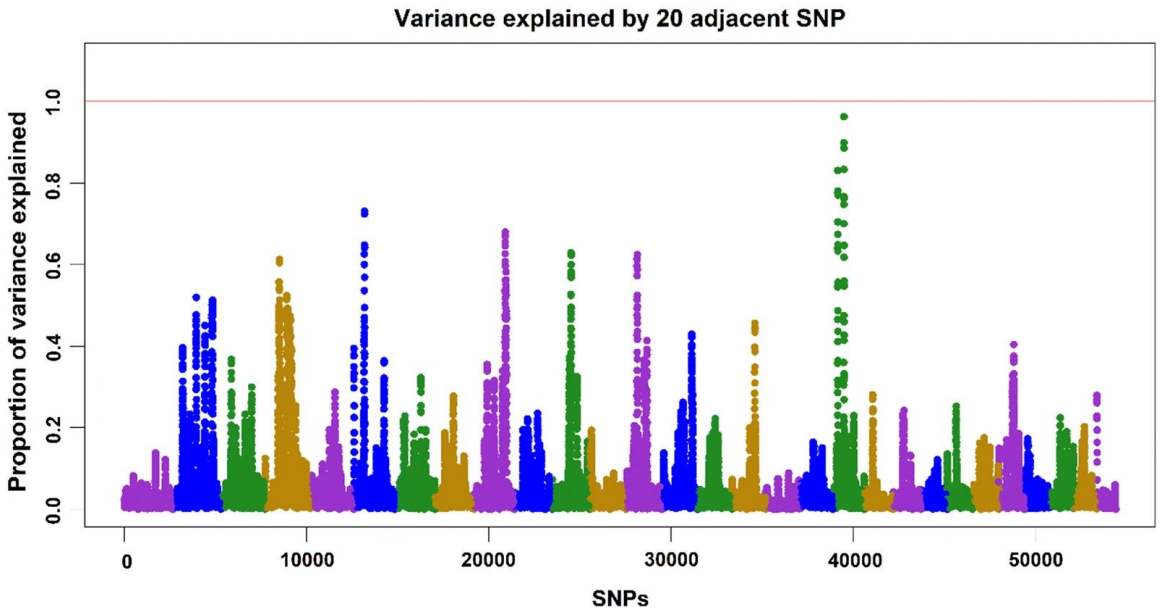


Figure 2.2 Manhattan plot for harvest weight in the 1st iteration of WssGBLUP, with the proportion of additive genetic variance explained by windows of 20 adjacent SNPs.

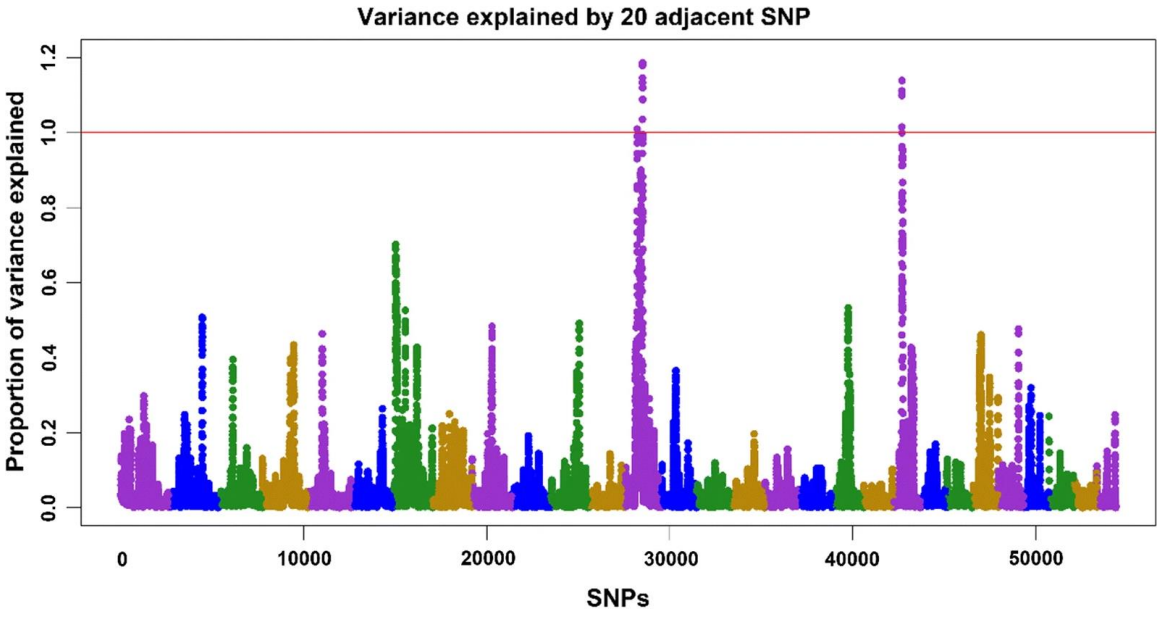


Figure 2.3 Manhattan plot for residual carcass weight in the 1st iteration of WssGBLUP, with the proportion of additive genetic variance explained by windows of 20 adjacent SNPs.

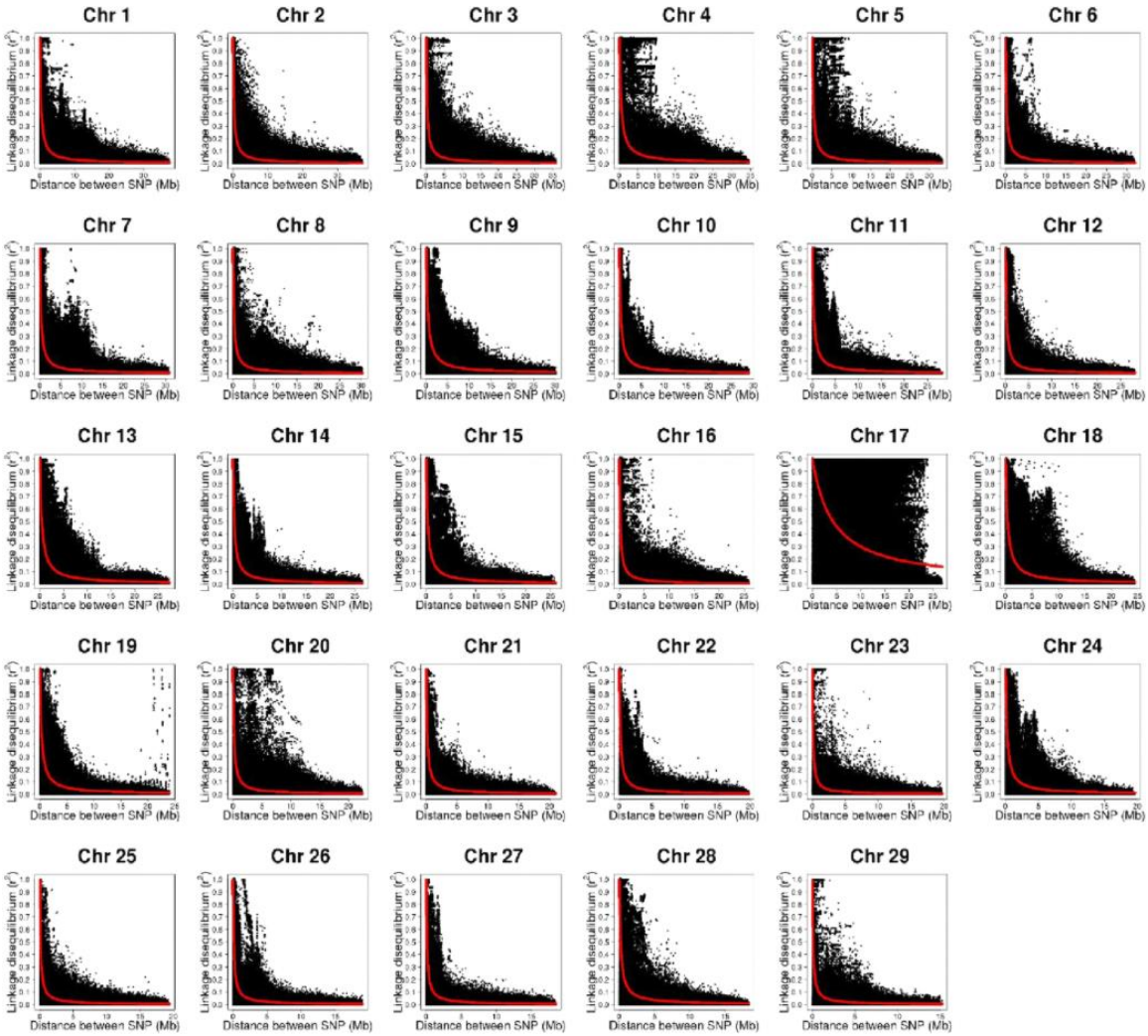


Figure 2.4 LD decay plots for 29 chromosomes.

CHAPTER 3

INDIRECT PREDICTIONS WITH A LARGE NUMBER OF GENOTYPED ANIMALS USING THE ALGORITHM FOR PROVEN AND YOUNG [1]

[1]Garcia A.L.S., Masuda Y., Tsuruta S., Miller S., Misztal I., and Lourenco D.A.L. Submitted to *Journal of Animal Science*, 02/11/2020.

ABSTRACT

Obtaining single nucleotide polymorphism (SNP) effects from genomic BLUP (GBLUP) and single-step GBLUP (ssGBLUP) may be of interest because they can be used to calculate indirect predictions (IP), which can be useful as interim evaluations for young genotyped animals, or as genomic predictions animals not included in official evaluations. When a large number of genotyped animals is available, there is a question about the number of animals needed to reliably calculate SNP effects and IP. The objectives of this study were to evaluate the quality of IP with increasing number of genotyped animals and to investigate how many animals are needed to reliably calculate such predictions. The data were provided by the American Angus Association and had genotypes and phenotypes for birth and weaning weight, and post-weaning gain. Genotyped animals were divided in three year-classes: born up to 2013 (n= 114, 937), 2014 (n= 183,847), and 2015 (n= 280,506). The number of animals with phenotypes was > 3.8 million. A three-trait model was fit using the algorithm for proven and young (APY) with 19,021 animals as core, under two definitions: first, core animals were the same for all year-classes with animals born up to 2013 (core 2013); and second, core changed for different year-classes with animals born up to 2014 (core 2014) and 2015 (core 2015). While GBLUP used only phenotypes of genotyped animals, ssGBLUP used all phenotypes available. The SNP effects were calculated based on genomically estimated breeding values (GEBV) from all or only core animals. Correlations between GEBV from GBLUP and IP, when SNP effects were backsolved with core 2013, were ≥0.99 for animals in 2013 but as low as 0.07 for animals in 2014 and 2015. Under ssGBLUP, those correlations were ≥0.99 for animals in 2013, 2014, and 2015. Predictive ability when GEBV were computed by ssGBLUP and SNP effects were backsolved based on only core animals was as high

as based on all animals. When the number of animals for computing SNP effects varied, correlations between GEBV and IP from ssGBLUP were ≥0.76, ≥0.90, and ≥0.98 with 2K, 5K and 15K animals. If GEBV are computed based on GBLUP and the SNP effects based on the proper number of core animals, IP is adequate for the current generation but there is a considerable drop in accuracy for the next generation. Such IP based on a large number of phenotypes from non-genotyped animals (ssGBLUP) has persistent accuracy in further generations.

INTRODUCTION

The availability of genomic resources in the form of dense single nucleotide polymorphisms (SNP) panels has allowed for the implementation of genomic selection in many livestock species. Once the deoxyribonucleic acid (DNA) markers are available, methods such as SNP-best linear unbiased prediction (SNP-BLUP), genomic BLUP (GBLUP) and single-step genomic BLUP (ssGBLUP) can be used to obtain genomic predictions (Meuwissen et al., 2001; Aguilar et al., 2010; Christensen and Lund, 2010).

As genomic selection becomes popular and genotyping costs decrease, the number of animals being genotyped steadily increases. Examples of this are the U.S. dairy industry with more than three million genotyped animals (queries.uscdcb.com/Genotype/cur_density.html) and the American Angus Association with more than 750,000 genotyped animals (Steve Miller, Angus Genetics Inc., personal communication). When GBLUP and ssGBLUP are used for such large genomic datasets, the computing cost becomes a problem because inverting the genomic relationship matrix (**G**) has a cubic cost with the number of genotyped animals, which is not feasible for over 150,000 animals (Fragomeni et al., 2015). To solve this problem, Misztal et al. (2014a) proposed the algorithm for proven and young (APY). In the APY formulation, the

genotyped population is divided into core and noncore such that the only direct inversion needed is for the core portion and the other components are obtained through recursions, dramatically reducing computing costs.

Even with appropriate tools, the addition of newly genotyped animals will increase computing time on routine evaluations which can increase the time between collecting a DNA sample and obtaining the actual predictions (Wiggans et al., 2015). This timing is important since most of the genotypes come from young animals and producers rely on genomic predictions to make the decision of whether to keep an animal or not. Being able to quickly decide which animals to keep and which ones to cull will potentially decrease rearing costs at the farm level (Nicolazzi et al., 2018). Genomic predictions are also important for producers outside the seedstock market raising unregistered animals, as they might help them to make better management decisions. Such predictions on commercial non-registered Angus females are available now and are marketed as GeneMax Advantage.

One common issue in the genomic era is that often animals are genotyped before phenotypes are collected, and sometimes pedigree information is missing; therefore, those animals may not contribute information to the official evaluations and in fact, their inclusion in the evaluations might even decrease accuracy and increase inflation of genomically estimated breeding values (GEBV) because of their incomplete pedigrees (Bradford et al., 2017; Bradford et al., 2019). If SNP effects are available, Indirect Predictions (IP) can be used as interim evaluation providing quick genomic predictions for newly genotyped and also non-registered animals, without affecting routine evaluations (Lourenco et al., 2015).

Because SNP-BLUP and GBLUP are equivalent models, SNP effects can be calculated based on GEBV and the inverse of $\mathbf{G}$ ($\mathbf{G}^{-1}$) for genotyped animals in GBLUP (VanRaden, 2008;

Strandén and Garrick, 2009) and in ssGBLUP (Wang et al., 2012). As the process of backsolving GEBV into SNP effects involves $\mathbf{G}^{-1}$, using all genotyped animals to compute SNP effects might be prohibitive and tools such as APY (Misztal et al., 2014a) can help to surpass this limitation. Lourenco et al. (2018) investigated IP from ssGBLUP using almost 81,000 genotyped animals from the American Angus Association data, and their results show that accurate IP can be obtained from ssGBLUP with $\mathbf{G}^{-1}$ calculated using APY or only a set of core animals. Although their study shows the feasibility of obtaining IP from ssGBLUP with APY, the number of genotyped animals used was small compared to the current database, and the impact of adding new genotypes was not investigated. Therefore, the purposes of this study were to: 1) test the stability of IP and check if the core group should be updated when large numbers of genotyped animals are added to the database; 2) investigate the choice of core animals to calculate SNP effects that are used for IP, i.e., whether all animals or only core should be used; 3) assess the ideal number of genotyped animals needed, when backsolving GEBV into SNP effects, to obtain reliable IP.

MATERIALS AND METHODS

DATA AND MODEL

The dataset used in the study was provided by the American Angus Association. Phenotypes were available for birth weight (BW; N= 7,574,765), weaning weight (WW; N= 8,302,222), and post-weaning gain (PWG; N= 4,145,166), and the pedigree included 9,145,109 animals, from which 280,506 animals born up to 2015 were genotyped for 39,774 markers after quality control.

The following three-trait model was used:

$$y_t = \mathbf{X}b + \mathbf{W_1}u + \mathbf{W_2}m + \mathbf{W_3}p + e \quad (1)$$

Where **t** refers to each trait, BW, WW, and PWG; **y** and **b** are the vectors of phenotypes and fixed effect of contemporary group; **u**, **m**, and **p** are the vectors of random effects of additive direct, maternal, and maternal permanent environmental effects; **e** is the vector of residuals. The **X**, $\mathbf{W_1}$, $\mathbf{W_2}$, and $\mathbf{W_3}$ are incidence matrices for the effects in **b**, **u**, **m**, and **p** respectively. All random effects were present for WW, but only **u**, **m**, and **e** for BW, and **u** and **e** for PWG.

ANALYSES

Genomic BLUP provides a simple framework to test the quality of IP because when the SNP effects are derived from GBLUP, one should be able to obtain IP and GEBV on the same scale. Whereas, in ssGBLUP, a mean has to be added to IP to consider the fact **G** is tuned to match **A** (Lourenco et al., 2018). Genomic analyses were performed using GBLUP and ssGBLUP models, although the process of obtaining IP on the same scale as GEBV under ssGBLUP is still under investigation. As a similar scale is obtained by adding a constant that reflects the average GEBV in the population used to compute SNP effects (Legarra et al., 2018; Lourenco et al., 2018), correlations investigated in the present study are not affected.

In ssGBLUP, the inverse of the relationship matrix combining pedigree and genomic relationships ($\mathbf{H^{-1}}$) was constructed as in Aguilar et al. (2010):

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1} - A_{22}^{-1}} \end{bmatrix} \quad (2)$$

where $\mathbf{G^{-1}}$ is the inverse of the genomic relationship matrix and $\mathbf{A_{22}^{-1}}$ is the inverse pedigree relationship matrix for genotyped animals.

In both models, the initial genomic relationship matrix ($\mathbf{G_0}$) was constructed following VanRaden (2008):

$$\mathbf{G_0} = \frac{\mathbf{ZZ'}}{2\sum p_i (1-p_i)} \quad (3)$$

where $\mathbf{Z}$ is a matrix of centered gene content and $\mathbf{p_i}$ is the minor allele frequency of SNP *i*. Allele frequencies were calculated based on current genotypes. To avoid singularity problems, in GBLUP $\mathbf{G} = 0.99\mathbf{G_0} + 0.01\mathbf{A_{22}}$, whereas in ssGBLUP $\mathbf{G} = 0.95\mathbf{G_0} + 0.05\mathbf{A_{22}}$. The impact of other blending proportions was also investigated under ssGBLUP.

Given the number of genotyped animals used in the present study, the direct inversion of $\mathbf{G}$ is not feasible; therefore, APY was used to compute the inverse of $\mathbf{G}$ ($\mathbf{G^{-1}_{APY}}$) as proposed by Misztal et al. (2014a) and Misztal (2016). In APY, the genotyped animals are partitioned as core (c) and noncore (n):

$$\mathbf{G} = \begin{bmatrix} \mathbf{G_{cc}} & \mathbf{G_{cn}} \\ \mathbf{G_{nc}} & \mathbf{G_{nn}} \end{bmatrix} \quad (4)$$

And $\mathbf{G^{-1}_{APY}}$ is calculated as follows:

$$\mathbf{G^{-1}_{APY}} = \begin{bmatrix} \mathbf{G^{-1}_{cc}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{-G^{-1}_{cc}G_{cn}} \\ \mathbf{I} \end{bmatrix} \mathbf{M^{-1}_{nn}} \begin{bmatrix} \mathbf{-G_{nc}G^{-1}_{cc}} & \mathbf{I} \end{bmatrix} \quad (5)$$

With each element of $\mathbf{M_{nn}}$ obtained for the *ith* non-core animal as:

$$m_{nn,i} = g_{ii} - \mathbf{G_{ic}G^{-1}_{cc}G_{ci}} \quad (6)$$

Using the APY formula, the only direct inversion needed is the part of $\mathbf{G}$ containing relationships among core animals, whereas the other components are obtained through recursions.

Pocrnic et al. (2016) showed that the number of core animals can be obtained as the number of eigenvalues explaining 98% – 99% of the variance of $\mathbf{G_0}$. Because the eigenvalue decomposition of $\mathbf{G_0}$ is computationally more expensive than the equivalent singular value decomposition of $\mathbf{Z}$, the latter was used, and eigenvalues were obtained as the square of singular values. The number of core animals corresponding to 99% of the variance was 19,021 animals. This was used in this study and corresponds to the number of core animals used in routine evaluations by the American Angus Association.

SNP EFFECTS, INDIRECT PREDICTIONS AND VALIDATION

To evaluate the impact of increasing the number of genotyped animals in the calculation of SNP effects and IP, the genotyped animals were divided into three year-classes: animals born up to 2013 (N= 114,937), 2014 (N= 183,847), and 2015 (all animals; N= 280,506). The number of records included in GBLUP and ssGBLUP for each year-class, as well as heritability of the traits are presented in Table 3.1.

While the number of core animals remained the same in all analyses (19,021), two core definitions for APY were tested:

1) Core 2013: core animals were randomly sampled from animals born up to 2013 and remained the same across all year-classes;
2) Core 2014 and core 2015: core animals were randomly sampled from animals born up to 2014 and 2015 for year-classes 2014 and 2015, respectively;

After the core groups were defined, GEBV were calculated using (ss)GBLUP with APY for each year-class dataset. Then, SNP effects were backsolved using either $\mathbf{G}^{-1}_{\mathbf{APY}}$ or $\mathbf{G}^{-1}$ only for core animals ($\mathbf{G}^{-1}_{\mathbf{core}}$), using the formula derived by Wang et al. (2012):

$$\hat{a}_{Full} = \lambda \mathbf{DZ'G}^{-1}_{\mathbf{APY}}\hat{u} \quad (7)$$

$$\hat{a}_{core} = \lambda \mathbf{DZ}'_{\mathbf{core}}\mathbf{G}^{-1}_{\mathbf{core}}\hat{u}_{core} \quad (8)$$

Where $\hat{\mathbf{a}}$ is a vector of SNP effects; $\hat{\mathbf{u}}$ is a vector of GEBV for all genotyped animals; $\hat{\mathbf{u}}_{\mathbf{core}}$ is a vector with GEBV for core animals; $\boldsymbol{\lambda}$ is the ratio of SNP to additive genetic variance, $\mathbf{D}$ is a diagonal matrix of SNP weights ($\mathbf{D}$=$\mathbf{I}$ in our case), and $\mathbf{Z}$ ($\mathbf{Z}_{core}$) is a matrix of centered gene content for all (core) genotyped animals; $\mathbf{G}^{-1}_{\mathbf{APY}}$ and $\mathbf{G}^{-1}_{\mathbf{core}}$ are genomic relationship matrices for all genotyped animals (computed using APY) and for core animals only, respectively.

Once SNP effects were available, IP were calculated as follows:

$$IP_{Full}=\mathbf{Z}\hat{a}_{Full} \quad (9)$$

$$IP_{core}=\mathbf{Z}\hat{a}_{core} \quad (10)$$

Where **Z** is the centered gene content matrix for all genotyped animals within each year class.

In the GBLUP context, $\hat{\mathbf{u}}|\hat{\mathbf{a}}=\mathbf{Z}\hat{\mathbf{a}}$, therefore, to evaluate the quality of the IP and how good they are in retrieving GEBV given that SNP effects are known, the correlation between IP **(***i.e.*, IP_{Full} and IP_{core}) and GEBV was calculated for each year-class and core definition.

Typically, IP are calculated for young genotyped animals not included in the evaluations used to compute GEBV and SNP effects, thus we also performed a validation study using genotyped animals born in 2016 (N= 54,997), as validation animals. Such animals had genotypes and phenotypes for all traits; however, their data was not included in previous analyses. Using SNP effects previously calculated from (ss)GBLUP models with year-class 2015 data and both core definitions, IP were calculated for validation animals. Further, genotypes for validation animals were included in evaluations with reduced data and GEBV were obtained. Predictive ability was calculated as the correlation between adjusted phenotypes (based on traditional BLUP with full data) and IP or GEBV for validation animals.

NUMBER OF ANIMALS TO COMPUTE IP

To investigate the minimum number of animals needed to compute SNP effects, whereas keeping correlations between IP and GEBV >0.99, we randomly assigned genotyped animals into subsets with size varying from 500 to 40,000 (i.e., 500, 1K, 2K, 3K, 4K, 5K, 10K, 15K, 20K, 30K, and 40K) from the whole population (280,506). Once the subsets were created, SNP effects were calculated as:

$$\hat{a}_{subset}= \lambda \mathbf{D}\mathbf{Z}'_{\mathbf{subset}}\mathbf{G}^{-1}_{\mathbf{subset}}\hat{u}_{subset} \quad (11)$$

Where $\mathbf{G}^{-1}_{\mathbf{subset}}$ is the direct $\mathbf{G}^{-1}$ being computed for each subset of genotyped animals and $\hat{\mathbf{u}}_{\mathbf{subset}}$ is a vector of GEBV for the subset animals; GEBV were calculated using ssGBLUP with APY based on all genotyped animals and core 2013. Indirect predictions were then calculated for all genotyped animals as $\mathbf{IP}_{\mathbf{subset}}=\mathbf{Z}_{\mathbf{subset}}\hat{\mathbf{a}}_{\mathbf{subset}}$, and correlations between IP and GEBV are shown for each subset.

All the analyses were performed using software from the BLUPF90 family of programs (Misztal et al., 2014b) and in-house bash and R (R core team, 2019) scripts.

RESULTS AND DISCUSSION

GEBV

The correlation between GEBV across core definitions using all genotyped animals (year-class 2015), were ≥0.99 for all traits, which indicates that changes in GEBV coming from APY computations were minimal with different core definitions. Previous studies with simulated and real datasets have investigated changes in GEBV when using APY and found that as long as the number of core animals reflects the dimensionality of the genomic information (i.e., number of eigenvalues explaining at least 98% of the variance of **G**), the choice of core animals is arbitrary (Fragomeni et al., 2015; Masuda et al., 2016; Bradford et al., 2017).

INDIRECT PREDICTIONS WITH $\mathbf{G}^{-1}_{\mathbf{APY}}$ AND $\mathbf{G}^{-1}_{\mathbf{CORE}}$

When $\mathbf{G}^{-1}_{\mathbf{APY}}$ was used, the correlations between IP_{Full} and GEBV were ≥0.96 for all traits and scenarios for ssGBLUP and GBLUP models (Tables 3.2 and 3.3).

With the number of core animals constant and the addition of new genotyped animals (i.e., different year-classes) the number of noncore animals in $\mathbf{G}^{-1}_{\mathbf{APY}}$ is increased. Our results show that as long as the number of core animals represents the dimensionality of genomic information, APY

delivers robust IP under both models, regardless the addition of a large number of genotyped animals and the core definition.

For the computation of SNP effects based on $\mathbf{G}^{-1}_{\mathbf{core}}$, the results differed by model. While correlations between IP_{core} and GEBV were ≥0.99 for ssGBLUP regardless of core definition (Table 3.2), under GBLUP, there was a dramatic decrease in correlations when core 2013 was used (Table 3.3). Correlations decreased from 0.99 to 0.64 for BW, from 0.99 to 0.12 for WW, and from 0.99 to 0.07 for PWG. When core animals were chosen from the recent population (i.e., core 2014 and core 2015), correlations were restored to 0.99 (Table 3.3). Although in both cases the GEBV were computed using APY with all genotyped animals, SNP effects and IP were computed based on $\mathbf{G}^{-1}$ that contained only relationships for core genotyped animals. In this case, the backsolving process uses only a portion of the equations. The core based on 2013 represented a population with 114,937 genotyped animals, whereas the 2015 core was a random sample based on all 280,506 animals. Using the core 2013 to compute IP for all animals up to 2015 may not reflect the current state of the population, under GBLUP. On the other hand, the fact that ssGBLUP uses much more data than GBLUP may have contributed to a more robust estimation of GEBV, and therefore, SNP effects and IP.

Pocrnic et al. (2019) investigated the accuracy of GBLUP in terms of the number of eigenvalues of the genomic relationship matrix. They found that with little phenotypic information, eliminating 90% of the smallest eigenvalues did not reduce accuracy. With large amounts of phenotypic information, considering more eigenvalues increased accuracy. Additionally, 10% of the largest eigenvalues explained 90% of the variation in **G**. Using only *n* largest eigenvalues or *n* core animals in the APY algorithm resulted in similar accuracy. They claimed that the largest eigenvalues represent many chromosome segments, and a small amount of data is adequate to

estimate a few eigenvalues, which may explain a large portion of the genetic variation in **G**. Therefore, intermediate accuracy can be achieved with small amounts of phenotypic data, but further increases in accuracy require much larger amounts of data to estimate the remaining eigenvalues. The clusters of chromosome segments considered with small data sets may be different in future generations, leading to low persistence of predictions. On the other hand, when data is sufficiently large to estimate nearly all eigenvalues and indirectly, chromosome segments, the persistence is likely to be better. Similar accuracy with the same number of eigenvalues or core animals suggest that *n* animals contain information on almost the same chromosome segments as the *n* largest eigenvalues.

The decrease in correlation between GEBV and IP with $\mathbf{G}^{-1}_{\mathbf{APY}}$ and $\mathbf{G}^{-1}_{\mathbf{core}}$ was also reflected by the correlations between SNP effects. With core 2013, the correlations decreased when moving from 2013 to 2015 year-class in both models, but the decrease was much smaller under ssGBLUP compared to GBLUP (Table 3.4); for instance, for PWG, correlations decrease from 0.92 to 0.88 (0.04 points) in ssGBLUP, but from 0.95 to 0.73 (0.22) in GBLUP. For core 2014 and core 2015 scenarios, the SNP effect correlations were very similar between the two models, and although still showing a small decrease for different year-classes, this decrease was much smaller compared to the core 2013 scenario, especially in GBLUP (Table 3.4). This behavior shows that with core 2013, SNP effects are not as well estimated under GBLUP as more genotyped animals are added.

Even though a decrease in correlations between IP_{core} and GEBV using core 2013 and 2014 under GBLUP were observed for all traits, birth weight seemed to be more persistent (Table 3.3). This could be because of heritability and selection intensity. Birth weight has almost double the heritability compared to WW and PWG (Table 3.1). With higher heritability, more eigenvalues

of smaller effect are accounted for and their information contributes to higher accuracy (Pocrnic et al., 2019) or in our study, higher persistence.

In a study with layer chickens, Wolc et al. (2011) showed that traits with higher heritability had more persistent accuracy across generations as opposed to lowly heritable traits.

Regarding selection, in a simulation study with a population under selection, zeroing the first eigenvalues of **G** and using the reconstructed matrix for genomic evaluations decreased selection response by almost 40%, indicating strong effect of selection on persistence, especially if the dataset is limited (Yvette Stein, University of Georgia, Athens GA, personal communication). Figure 3.1 shows genetic trends standardized by additive genetic standard deviation for all traits. Although there is genetic improvement for all traits, selection pressure on BW is different compared to WW and PWG. Low BW is desirable to avoid calving problems; however, BW is positively correlated with WW and PWG, therefore, selecting for increased WW and PWG while decreasing BW requires extra selection pressure on the latter. In this way, persistence of predictions for WW and PWG is expected to be different from BW given lower heritabilities and different selection pressure.

IMPACT OF BLENDING AND TUNING

In ssGBLUP, **G** has to be blended and tuned to make it invertible and compatible with the pedigree relationships in **A** (VanRaden, 2008; Vitezica et al., 2011). If these steps are not considered, IP will be affected with changes in blending parameters. Preliminary analyses using different blending strategies (1% $\mathbf{A_{22}}$, 5% $\mathbf{A_{22}}$, and 10%$\mathbf{A_{22}}$) showed that the highest the blending percentage with $\mathbf{A_{22}}$, the lowest the correlation between IP and GEBV. Additionally, the more animals used, the bigger the impact of blending (IP_{Full} vs IP_{core}) (Table 3.5). Table 3.2 shows that

correlations between IP_{core} and GEBV are slightly higher compared to IP_{Full} which was likely due to the impact of blending.

Lourenco et al. (2018) investigated the impact of not accounting for tuning on IP and showed that under GBLUP $\mathbf{E(u)=0}$ and $\hat{\mathbf{u}}|\hat{\mathbf{a}}=\mathbf{Z}\hat{\mathbf{a}}$, but in ssGBLUP this assumption does not hold because genotyping is more recent compared to the entire pedigree, which creates a difference between genetic bases from pedigree and genomic data. The authors recommended adding the average GEBV to IP such that $\hat{\mathbf{u}}|\hat{\mathbf{a}}=\hat{\boldsymbol{\mu}}+\mathbf{Z}\hat{\mathbf{a}}$, which makes the two predictions comparable. More recently, Legarra et al. (2018) derived formulas taking blending and tuning parameters into account when computing SNP effects from ssGBLUP:

$$\hat{a}= b\alpha\lambda\mathbf{DZ'G^{-1}}\hat{u} \quad (12)$$

where $\boldsymbol{\alpha}$ and $\mathbf{b}$ are the blending and tuning parameters, with $\mathbf{b}$ as in Vitezica et al. (2011).

VALIDATION

Our validation study represents a more realistic scenario of IP in which young genotyped animals are predicted based only on their genotypes without being part of the routine evaluations. The same patterns of the previous results were observed in our validation study. When $\mathbf{G^{-1}_{APY}}$ was used to calculate GEBV, SNP effects, and subsequently IP, the correlations with adjusted phenotypes were highest regardless of core definition and method (Table 3.6). On the other hand, when $\mathbf{G^{-1}_{core}}$ was used to compute SNP effects and IP, GBLUP and ssGBLUP behaved differently with a fixed set of core animals (core 2013). Under GBLUP, predictive ability decreased from 0.42 to 0.30 for BW, from 0.36 to 0.06 for WW and from 0.30 to 0.04 for PWG, whereas under ssGBLUP predictive ability remained stable at 0.44 for BW, 0.38 for WW and 0.31 for PWG. Therefore, the behavior of validation was similar to the correlation between GEBV and IP with

core 2013. When an updated set of core animals was used (core 2014 and core 2015), predictive ability from GBLUP was restored to the same levels of ssGBLUP (Table 3.6).

Another interesting aspect of our validation was that when $\mathbf{G}^{-1}_{\mathbf{APY}}$ was used, predictive ability for ssGBLUP and GBLUP were very similar (Table 3.6). Once there is enough information available to estimate most of the chromosome segments, accuracies are similar regardless of the model (Pocrnic et al., 2019). Karaman et al. (2016) investigated accuracies of genomic prediction using different models and concluded that when the reference population was big enough, different models (i.e., GBLUP, BayesB, and BayesC) "converged" to the same accuracy.

The results from Lourenco et al. (2018) and from our current study showed that the algorithm for proven and young can be used to calculate SNP effects from ssGBLUP and GBLUP to obtain reliable IP with large genotyped populations. Furthermore, with current implementation of APY in the BLUPF90 family of programs (Misztal et al., 2014b), SNP effects and indirect predictions can be obtained using a large number of genotyped animals without constraints in computing time and memory usage. Additionally, the use of a subset of core animals to compute IP is also a viable option when ssGBLUP is the model of choice for official evaluations.

NUMBER OF ANIMALS USED TO COMPUTE IP

Even when all genotyped animals can be used to backsolve SNP effects from GBLUP or ssGBLUP using tools such as APY, assuming that a representative set of genotyped animals with GEBV from previous evaluation is available, we investigated the minimum number of animals needed to obtain reliable estimates of SNP effects and indirect predictions. Using from 500 to 40,000 animals, the results are presented in Figures 3.2, 3.3, and 3.4 for birth weight, weaning weight, and post-weaning gain, respectively. Indirect predictions were calculated for all genotyped

animals, and correlations between GEBV and IP are presented as a function of number of animals used.

The results followed an exponential trend, showing that, for beef cattle populations, more than 5,000 animals are needed for a reliable estimation of SNP effects and IP. Once the number of animals reached 10,000, correlations were ≥0.97 for all traits, and when 15,000 or more animals were used correlations were ≥0.98 for all traits, reaching a plateau at what seems to be a minimum number of animals needed. Interestingly, this optimal number of animals to reach correlations ≥ 0.98 is close to the number of eigenvalues explaining 98% of the variance of **G** (Figures 3.2-3.4); therefore, the theory of limited dimensionality of genomic information (Misztal, 2016) seems to play a role in the amount of information needed for the estimation of SNP effects.

These results agree with Lourenco et al. (2015) who investigated reference populations with 2K, 8K, and 33K animals to calculate indirect predictions from ssGBLUP. The authors suggested the use of approximately 33K animals to obtain reliable predictions. In our study, we examined a wider range of animals, which allowed us to obtain a clearer view on how many animals are needed to obtain stable indirect predictions. Building up on the results showed by Lourenco et al. (2015), when the number of animals used to calculate SNP effects is large enough and their GEBV is available from previous official evaluations (Wiggans et al., 2015), it is possible to obtain reliable indirect predictions from ssGBLUP and GBLUP. Assuming that the ideal number of animals to compute SNP effects depends on the dimensionality of genomic information, this number will possibly vary by species as shown in Pocrnic et al. (2016). In their study, the number of eigenvalues explaining 98% of the variance of **G** was 14K for Holsteins, 11.5K for Jerseys, 10.6K for Angus, and 4.1K for pigs and chickens. Therefore, using a smaller subset of animals can

be possibly safe if the number of animals represents the dimensionality of the genomic information and if such subset is a fair representation of the genotyped population.

As pointed out by Wiggans et al. (2015), indirect predictions are much faster to compute compared to the official evaluations and they allow for weekly or even daily evaluations, shortening the interval between the DNA sampling and genomic prediction. Additionally, they can be used as genomic predictions for non-registered animals without having to include them into official evaluations, because their inclusion could potentially lead to problems due to lack of phenotypes and missing pedigrees. In these scenarios, indirect predictions may become a useful tool to provide quick and reliable genomic predictions for young and non-registered genotyped animals.

CONCLUSIONS

With increasing numbers of genotyped animals, using all available genotypes and GEBV from previous official evaluations to compute SNP effects is a practical approach to ensure that indirect predictions are stable and reliable. The algorithm for proven and young is a feasible option to calculate SNP effects from GBLUP and ssGBLUP when the number of genotyped animals is large. Under GBLUP, if a subset of animals is used to compute SNP effects, the number and the choice of animals has a considerable impact on the quality of indirect predictions. In purebred beef cattle populations, a sample of at least 15,000 animals representing the whole genotyped population should provide reliable SNP effects and indirect predictions; however, using information on all genotyped animals from the previous official evaluation is the usual procedure. In large datasets, ssGBLUP is less sensitive to the distance between the core and the more recent genotyped population, providing more persistent genomic predictions.

REFERENCES

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

Bradford, H. L., Y. Masuda, P. M. VanRaden, A. Legarra, and I. Misztal. 2019. Modeling missing pedigree in single-step genomic BLUP. Journal of Dairy Science 102 (3):2336-2346. doi: https://doi.org/10.3168/jds.2018-15434

Bradford, H. L., I. Pocrnić, B. O. Fragomeni, D. A. L. Lourenco, and I. Misztal. 2017. Selection of core animals in the Algorithm for Proven and Young using a simulation model. Journal of Animal Breeding and Genetics 134 (6):545-552. doi: https://doi.org/10.1111/jbg.12276

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

Fragomeni, B. O., D. A. L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. J. Lawlor, and I. Misztal. 2015. Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. Journal of Dairy Science 98 (6):4090-4094. doi: https://doi.org/10.3168/jds.2014-9125

Karaman, E., H. Cheng, M. Z. Firat, D. J. Garrick, and R. L. Fernando. 2016. An Upper Bound for Accuracy of Prediction Using GBLUP. PLOS ONE 11 (8):e0161054. doi: https://doi.org/10.1371/journal.pone.0161054

Legarra, A., D. A. Lourenco, and Z. Vitezica. 2018. Bases for genomic prediction. http://genoweb.toulouse.inra.fr/~alegarra/GSIP.pdf.

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

Lourenco, D. A. L., A. Legarra, S. Tsuruta, D. Moser, S. Miller, and I. Misztal. 2018. Tuning indirect predictions based on SNP effects from single-step GBLUP. Interbull Bulletin (53)

Masuda, Y., I. Misztal, S. Tsuruta, A. Legarra, I. Aguilar, D. A. L. Lourenco, B. O. Fragomeni, and T. J. Lawlor. 2016. Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. Journal of Dairy Science 99 (3):1968-1974. doi: https://doi.org/10.3168/jds.2015-10540

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. 2016. Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size. Genetics 202 (2):401-409. doi: https://doi.org/10.1534/genetics.115.182089

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: https://doi.org/10.3168/jds.2013-7752

Misztal, I., S. Tsuruta, D. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. 2014b. Manual for BLUPF90 family of programs.

Nicolazzi, E. L., J. W. Durr, and G. R. Wiggans. 2018. Genomics in the US dairy industry: Current and future challenges. Interbull bulletin 53

Pocrnic, I., D. A. L. Lourenco, Y. Masuda, and I. Misztal. 2016. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. Genetics Selection Evolution 48:82. doi: https://doi.org/10.1186/s12711-016-0261-6

Pocrnic, I., D. A. L. 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):75. doi: https://doi.org/10.1186/s12711-019-0516-0

R core team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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

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

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. doi: https://doi.org/10.1017/S001667231100022X

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics research 94 (2):73-83. doi: https://doi.org/10.1017/s0016672312000274

Wiggans, G. R., P. M. VanRaden, and T. A. Cooper. 2015. Technical note: Rapid calculation of genomic evaluations for new animals. Journal of Dairy Science 98 (3):2039-2042. doi: https://doi.org/10.3168/jds.2014-8868

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, and J. C. Dekkers. 2011. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genetics Selection Evolution 43 (1):23. doi: https://doi.org/10.1186/1297-9686-43-23

TABLES

Table 3.1: Number of phenotypic records included in ssGBLUP and GBLUP in each year class.

Trait	h^2	ssGBLUP			GBLUP		
		2013	2014	2015	2013	2014	2015
BW	0.42	6,944,152	7,250,456	7,574,765	73,850	120,389	188,241
WW	0.20	7,659,259	7,972,273	8,302,222	75,428	122,838	191,792
PWG	0.24	3,835,752	3,985,075	4,145,166	56,254	91,422	140,975

Table 3.2: Correlations between IP and GEBV calculated based on ssGBLUP model with $\mathbf{G^{-1}_{APY}}$ (IP_{Full}) and $\mathbf{G^{-1}_{core}}$ **(**IP_{core}**)** for all year classes and core definitions.

Core definition	Year class	BW		WW		PWG	
		IP_{Full}	IP_{core}	IP_{Full}	IP_{core}	IP_{Full}	IP_{core}
2013	2013	0.98	0.99	0.99	1.00	0.99	1.00
	2014	0.97	0.99	0.99	1.00	0.99	1.00
	2015	0.96	0.99	0.99	1.00	0.99	1.00
2013[1]	2013	0.97	0.99	0.99	1.00	0.99	1.00
2014	2014	0.96	0.99	0.99	1.00	0.99	1.00
2015	2015	0.98	0.99	0.99	1.00	0.99	1.00

1- Results from year-class 2013 are the same.

Table 3.3: Correlations between IP and GEBV calculated based on GBLUP model with $\mathbf{G^{-1}_{APY}}$ (IP_{Full}) and $\mathbf{G^{-1}_{core}}$ **(**IP_{core}**)** for all year classes and core definitions.

Core definition	Year class	BW		WW		PWG	
		IP_{Full}	IP_{core}	IP_{Full}	IP_{core}	IP_{Full}	IP_{core}
2013	2013	0.99	0.99	0.99	0.99	0.99	0.99
	2014	0.98	0.82	0.99	0.34	0.99	0.31
	2015	0.97	0.64	0.99	0.12	0.99	0.07
2013[1]	2013	0.99	0.99	0.99	0.99	0.99	0.99
2014	2014	0.98	0.99	0.99	0.99	0.99	0.99
2015	2015	0.97	0.99	0.99	0.99	0.99	0.99

1- Results from year-class 2013 are the same.

Table 3.4: Correlations between SNP effects calculated based on $\mathbf{G}^{-1}_{APY}$ and $\mathbf{G}^{-1}_{core}$ in different year-classes within the same core definition.

Core definition	Year class	BW		WW		PWG	
		ssGBLUP	GBLUP	ssGBLUP	GBLUP	ssGBLUP	GBLUP
2013	2013	0.86	0.88	0.92	0.92	0.92	0.95
	2014	0.82	0.83	0.90	0.85	0.90	0.86
	2015	0.78	0.78	0.87	0.75	0.88	0.73
2013[1]	2013	0.86	0.88	0.92	0.92	0.92	0.95
2014	2014	0.82	0.84	0.89	0.90	0.90	0.93
2015	2015	0.78	0.79	0.86	0.88	0.88	0.91

1- Results from year-class 2013 are the same.

Table 3.5: Correlation between IP and GEBV with different blending strategies in ssGBLUP.

Blending [*]	BW		WW		PWG	
	IP_{Full}	IP_{core}	IP_{Full}	IP_{core}	IP_{Full}	IP_{core}
1% A_{22}	0.96	0.99	0.98	0.99	0.99	1.00
5% A_{22}	0.94	0.98	0.97	0.99	0.98	0.99
10% A_{22}	0.92	0.97	0.95	0.98	0.96	0.98

[*] Year class 2015 and core 2015 definition.

Table 3.6: Predictive ability for validation animals born in 2016 for ssGBLUP and GBLUP models.

Core Definition	Model	BW			WW			PWG		
		IP_{Full}	IP_{core}	GEBV	IP_{Full}	IP_{core}	GEBV	IP_{Full}	IP_{core}	GEBV
2013	ssGBLUP	0.43	0.44	0.44	0.38	0.38	0.38	0.31	0.31	0.32
	GBLUP	0.42	0.30	0.43	0.36	0.06	0.37	0.30	0.04	0.30
2015	ssGBLUP	0.43	0.44	0.45	0.38	0.38	0.38	0.31	0.32	0.32
	GBLUP	0.42	0.43	0.43	0.36	0.37	0.37	0.30	0.30	0.30

FIGURES

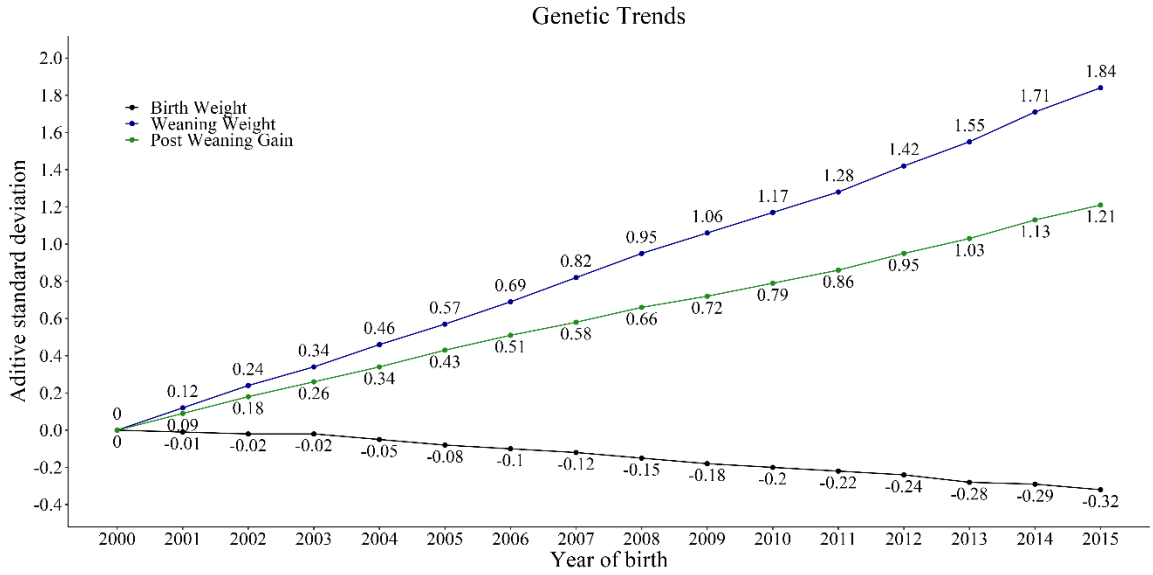


Figure 3.1: Genetic trend for all traits. Genetic trends are presented as additive genetic standard deviations and genetic base is adjusted to 2000.

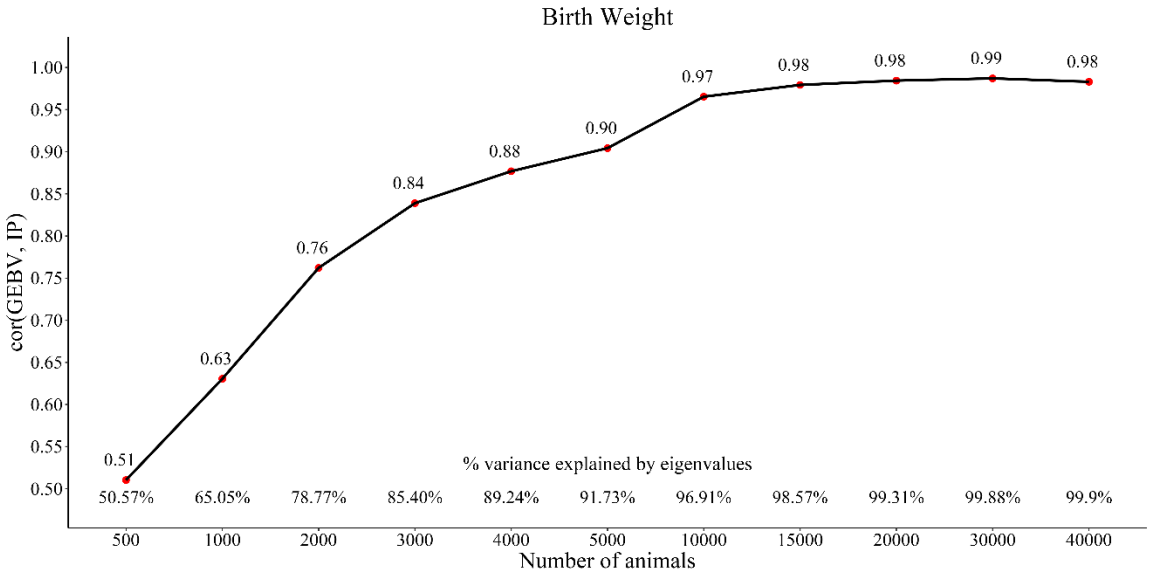


Figure 3.2: Correlations between GEBV and indirect predictions for birth weight with increasing number of genotyped animals used to calculate SNP effects.

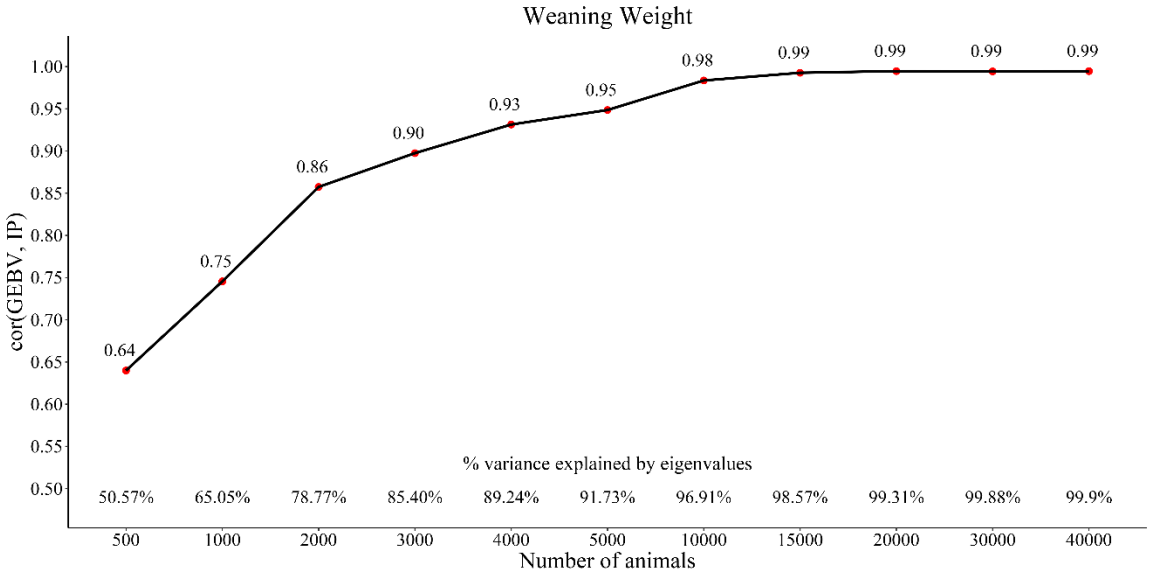


Figure 3.3: Correlations between GEBV and indirect predictions for weaning weight with increasing number of genotyped animals used to calculate SNP effects.

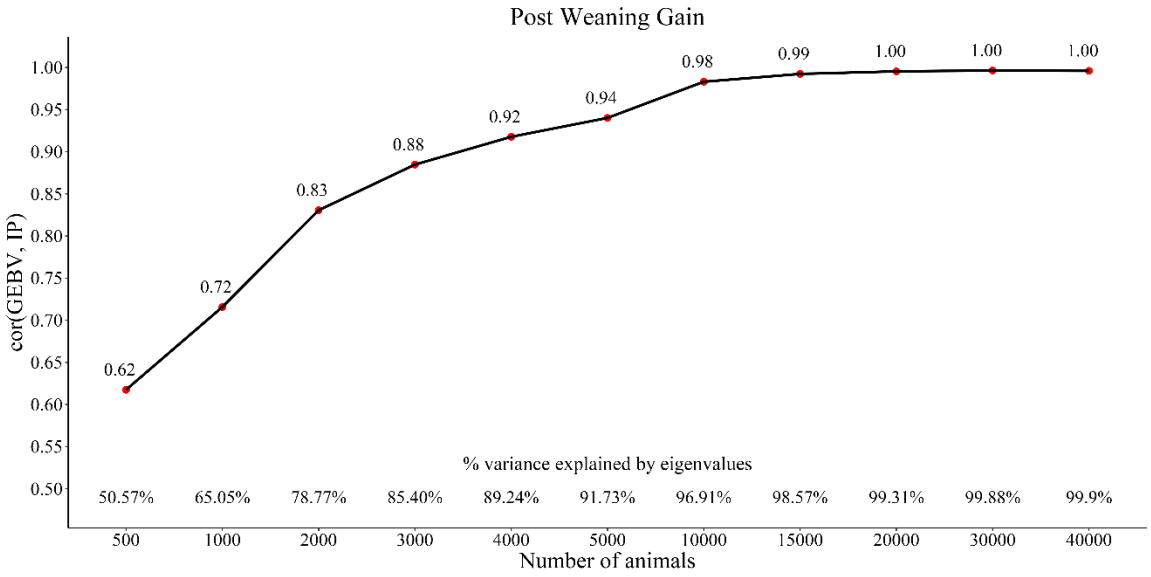


Figure 3.4: Correlations between GEBV and indirect predictions for post-weaning gain with increasing number of genotyped animals used to calculate SNP effects.

CHAPTER 4

GENOMIC ACCURACY FOR INDIRECT PREDICTIONS BASED ON SNP EFFECTS FROM SINGLE-STEP GBLUP[1]

[1]Garcia A.L.S., Aguilar I., Legarra A., Miller S., Tsuruta S., Misztal I., Lourenco D.A.L. To be submitted to *Genetics Selection Evolution*.

ABSTRACT

When single-step GBLUP (ssGBLUP) is the method of choice for genomic evaluations, SNP effects can be backsolved from GEBV and indirect prediction (IP) can be calculated as the sum of the SNP effects weighted by the gene content. Indirect predictions can be useful when the number of genotyped animals is large, when genotyped animals are not included in the official evaluations, and when interim evaluations are needed to reduce the time between DNA collection and management decisions at the farm level. Having IP is beneficial if their accuracy is comparable to GEBV accuracy. Our first objective was to implement formulas to compute accuracy of IP by backsolving prediction error covariance (PEC) of GEBV into PEC of SNP effects, and to investigate the feasibility of this method. The second objective was to investigate the number of genotyped animals needed to obtain robust IP accuracy in large genotyped populations. An application was done in a beef cattle population with up to 60,000 genotyped animals. Using SNP effects from ssGBLUP evaluation, correlations between GEBV and IP were ≥0.99. When all genotyped animals were used for PEC computations, correlations between GEBV accuracy and IP accuracy were ≥0.99. Additionally, IP accuracies were compatible with GEBV accuracies either with direct inversion of the genomic relationship matrix (**G)** or using the algorithm for proven and young (APY) to obtain the inverse of **G**. As the number of genotyped animals included in PEC computations decreased up to 15,000, correlations were still ≥0.96, but IP accuracies were biased downwards. Indirect prediction accuracy can be successfully obtained by computing SNP PEC from ssGBLUP equations using direct or APY **G** inverse. It is possible to reduce the number of genotyped animals in PEC computations, but accuracies may be underestimated. Further research

is needed to approximate SNP PEC from ssGBLUP when the inverse of the left hand side of the mixed model equations if prohibitive because of a large number of genotyped animals.

INTRODUCTION

One of the ways to deal with the ever-increasing number of genotyped animals in single-step GBLUP (ssGBLUP) evaluations may be to use only genotyped animals with complete information in the official evaluation and compute indirect predictions (IP) for the remaining young genotyped animals. Additionally, IP can be a useful tool to provide fast, interim evaluations for registered animals and also a sort of prediction for animals not included in official evaluations. Such predictions help to decrease the timing between collecting a DNA sample and getting predictions on young animals, allowing farmers to make faster management decisions which could reduce raising costs by culling animals earlier (Wiggans et al., 2015; Nicolazzi et al., 2018). When genomic BLUP (GBLUP) or ssGBLUP is the method of choice for genomic evaluations, SNP effects are not readily available but can be easily backsolved from genomically estimated breeding values (GEBV) using formulas showed by VanRaden (2008) and Wang et al. (2012). Once SNP effects are calculated, IP can be obtained for young animals as the sum of the SNP effects weighted by the gene content.

Typically in animal breeding programs, not only a prediction is needed (EBV, GEBV, IP) but also a measure of accuracy for such predictions, to help in the selection decisions. Henderson (1984) showed that accuracies of EBV can be obtained based on the prediction error variance (PEV) by directly inverting the coefficient matrix of BLUP mixed model equations (MME). Although a good measure of accuracy of EBV, when the system of equations is too big it becomes impossible to invert the coefficient matrix to obtain PEV even with modern computers. To overcome this

problem, approximations have been proposed and implemented for pedigree based evaluations (Misztal and Wiggans, 1988) and when genomic information is included (Misztal et al., 2013; Liu et al., 2017; Erbe et al., 2018; Pocrnic et al., 2019). Similarly, it is of interest to have a measure of accuracy that is comparable to that of GEBV to be published along with IP to help producers make decisions.

Using a SNP-BLUP model, Liu et al. (2017) showed how to calculate accuracies for IP or direct genomic value (DGV) based on the prediction error covariance (PEC) of SNP effects and explained that the cost of obtaining such reliabilities is smaller because the size of the LHS matrix depends mainly on the number of SNP markers rather than the number of genotyped animals. Since SNP-BLUP and GBLUP are equivalent models, it is also possible to obtain SNP PEC for SNP effects calculated using (ss)GBLUP, although the computational cost increases with the number of genotyped animals. Derivations to obtain SNP PEC under ssGBLUP model were described by Gualdron Duarte et al. (2014) and Aguilar et al. (2019).

Pocrnic et al. (2019) investigated the accuracy of genomic selection under a GBLUP model using the algorithm for proven and young (APY) and showed that only a small number of eigenvalues from the genomic relationship matrix (GRM) was enough to account for a large portion of the genetic variation. Because the dimensionality of the genomic information is limited (Pocrnic et al., 2016a; Pocrnic et al., 2016b), it is possible to reduce the number of animals needed to calculate SNP effects and IP (Lourenco et al., 2018; Garcia et al., 2020). Likewise, the limited dimensionality could also allow for a reduction in the number of animals needed to obtain SNP PEC and accuracies for IP under (ss)GBLUP.

The objectives of this study were to: 1) implement formulas to compute accuracy of IP by backsolving prediction error covariance (PEC) of GEBV into PEC of SNP effects, and to

investigate the feasibility of this method; 2) investigate the number of genotyped animals needed to obtain robust IP accuracy in large genotyped populations.

MATERIALS AND METHODS

DATA AND MODEL

Data for the study were provided by the American Angus Association and included 230,639 animals in the pedigree and 38,000 post-weaning gain (PWG) phenotypes. Genotypes for 39,774 markers, after quality control, were available for 60,000 animals born up to 2018. To mimic a real situation where animals being indirectly predicted only have genotypes available, genotyped animals born in 2018 (N= 5,467) were considered as validation and had their phenotypes and pedigree omitted from all the analyses. Their genotypes were also omitted in a reduced dataset (N= 54,533) to calculate SNP PEC.

Single-step GBLUP was used with the model $\mathbf{y=cg+u+e}$, where **y** is a vector of post-weaning gain phenotypes and **cg** is a vector of fixed contemporary group effects; **u** is the vector of random additive genetic effect and **e** is the vector of random residuals. In ssGBLUP the inverse of the relationship matrix combining pedigree and genomic information ($\mathbf{H^{-1}}$) was constructed as in Aguilar et al. (2010):

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

Where $\mathbf{G^{-1}}$ is the inverse of the genomic relationship matrix and $\mathbf{A_{22}^{-1}}$ is the inverse pedigree relationship matrix for genotyped animals. The initial genomic relationship matrix was constructed as in VanRaden (2008):

$$\mathbf{G_0} = \frac{\mathbf{ZZ'}}{2\sum p_i (1-p_i)} \quad (2)$$

Where **Z** is a matrix of centered gene content and $\mathbf{p_i}$ is the minor allele frequency of SNP *i*. Allele frequencies were calculated based on current genotypes. Often in ssGBLUP, **G** is constructed as:

$$\mathbf{G}=(1-\alpha)(\mathbf{11}'a+b\mathbf{G_0})+\alpha\mathbf{A_{22}} \quad (3)$$

Where α=0.05 and refers to blending (VanRaden, 2008), and a and b are tuning parameters calculated as in Vitezica et al. (2011):

$$a = \frac{1}{n^2}(\sum_i \sum_j \mathbf{A}_{22\ i,j} - \sum_i \sum_j \mathbf{G}_{i,j}) \text{ and } b=1-\frac{1}{2}a \quad (4)$$

After tuning and blending steps, **G** is invertible and compatible with the pedigree relationships.

For large-scale genomic evaluations, it becomes infeasible to directly invert **G** and to overcome this limitation, the algorithm for proven and young (APY) was proposed by Misztal et al. (2014a) and Misztal (2016). In APY, the genotyped animals are divided into core (c) and non-core (n):

$$\mathbf{G}= \begin{bmatrix} \mathbf{G_{cc}} & \mathbf{G_{cn}} \\ \mathbf{G_{nc}} & \mathbf{G_{nn}} \end{bmatrix} \quad (5)$$

And $\mathbf{G^{-1}_{APY}}$ is calculated as follows:

$$\mathbf{G^{-1}_{APY}}= \begin{bmatrix} \mathbf{G^{-1}_{cc}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{-G^{-1}_{cc}G_{cn}} \\ \mathbf{I} \end{bmatrix} \mathbf{M^{-1}_{nn}} \begin{bmatrix} \mathbf{-G_{nc}G^{-1}_{cc}} & \mathbf{I} \end{bmatrix} \quad (6)$$

With elements of $\mathbf{M_{nn}}$ obtained for the *ith* non-core animal as:

$$m_{nn,i}= g_{ii} - \mathbf{G_{ic}G^{-1}_{cc}G_{ci}} \quad (7)$$

The number of core animals for APY can be obtained as the number of eigenvalues explaining 98-99% of the variance in **G**, which can be found by the eigenvalue decomposition of **G** or the singular value decomposition of **Z** (Pocrnic et al., 2016b). For our study, the number of eigenvalues explaining 99% of the variance was 15,000 and core animals were randomly selected from the genotyped animals in the reduced dataset.

Once $\mathbf{H^{-1}}$ is built, the ssGBLUP MME are:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W+H^{-1}}\lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'}y \\ \mathbf{W'}y \end{bmatrix} \quad (8)$$

Where $\mathbf{X}$ and $\mathbf{W}$ are incidence matrices for fixed effects and the animal effect; λ is the variance ratio $\frac{\sigma_e^2}{\sigma_u^2}$, and $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are the estimates of fixed effects and GEBV respectively.

BENCHMARK GEBV AND ACCURACY

A ssGBLUP evaluation using the complete data, i.e., 60K genotyped animals with pedigree and phenotypes up to 2017, was run to obtain benchmark GEBV accuracy (ACC_{GEBV}) for validation animals. The ACC_{GEBV} was calculated based on PEV from the inverse of the LHS of MME (6) as follows:

$$acc_i = \sqrt{1 - \frac{PEV_i}{\sigma_u^2}} = \sqrt{1 - \frac{LHS^{ii}}{\sigma_u^2}} \quad (9)$$

INDIRECT PREDICTIONS AND ACCURACY

Before calculating IP, SNP effects from ssGBLUP were obtained as described in Wang et al. (2012), using POSTGSF90 (Misztal et al., 2014b). Recently, Legarra et al. (2018) showed that under ssGBLUP, blending and tuning parameters need to be taken into account when backsolving GEBV into SNP effects:

$$\hat{a}|\hat{u} = (1-\alpha)b\mathbf{Z'}\frac{1}{2\sum p_i(1-p_i)}\mathbf{G^{-1}}\hat{u} \quad (10)$$

Where, $\boldsymbol{\alpha}$ and $\mathbf{b}$ are the blending and tuning parameters as described above and $\hat{\mathbf{u}}$ if a vector of GEBV from previous ssGBLUP evaluation. Once SNP effects are available, IP can be calculated as $\mathbf{IP=Z_{validation}}\hat{\mathbf{a}}$.

Liu et al. (2017) showed how to compute accuracies for direct genomic values (DGV; same as IP in our study), from a SNP-BLUP model using SNP PEC as follows:

$$acc_i^{IP}=\sqrt{1-\frac{z_i\mathbf{C}^{gg}z_i'}{\sigma_u^2}} \quad (11)$$

Where $\mathbf{C}^{gg}$ is the portion of the inverse of the LHS of the SNP-BLUP MME corresponding to marker effects (SNP PEC matrix) and $\mathbf{z}_i$ is the row vector from the $\mathbf{Z}$ matrix, that contains the genotypes for animal *i*. Since SNP-BLUP and GBLUP are equivalent models, one should be able to extend this idea using the same backsolving process that is used to obtain SNP effects, to obtain SNP PEC from (ss)GBLUP. Gualdron Duarte et al. (2014) and Aguilar et al. (2019), in an attempt to obtain formulas for the computation of p-values in GBLUP and ssGBLUP, respectively, showed that PEC of SNP effects can be calculated as follows:

$$var(\hat{a})=PEC=var\left((1-\alpha)b\mathbf{Z}'\frac{1}{2\sum p_i(1-p_i)}\mathbf{G}^{-1}\hat{u}\right) \quad (12)$$

Then,

$$var(\hat{a})=PEC=\frac{1}{2\sum p_i(1-p_i)}(1-\alpha)b\mathbf{Z}'\mathbf{G}^{-1}\left(\mathbf{G}\sigma_u^2-\mathbf{C}^{u2u2}\right)\mathbf{G}^{-1}\mathbf{Z}(1-\alpha)b\frac{1}{2\sum p_i(1-p_i)} \quad (13)$$

Therefore,

$$var(\hat{a})=PEC=\frac{1}{2\sum p_i(1-p_i)}(1-\alpha)b\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}\sigma_u^2-\mathbf{Z}'\mathbf{G}^{-1}\mathbf{C}^{u2u2}\mathbf{G}^{-1}\mathbf{Z}(1-\alpha)b\frac{1}{2\sum p_i(1-p_i)} \quad (14)$$

Note that **α** and **b** are blending and tuning parameters, accounted for in PEC computations, and $\mathbf{C}^{u2u2}$ is the inverse of the LHS of MME (8) corresponding to genotyped animals.

Once SNP PEC is available, accuracy for IP (ACC_{IP}) for an animal *i* can be calculated as:

$$ACC_{IPi}=\sqrt{1-\frac{(1-\alpha)b\,z_i\mathbf{var}(\hat{a})z_i'}{\sigma_u^2}} \quad (15)$$

While accuracy of IP can be easily obtained with small datasets, for large scale evaluations, obtaining $\mathbf{C}^{u2u2}$ becomes impractical as the number of genotyped animals increase. To overcome this limitation, the dimensionality of genomic information was exploited by using the APY algorithm to compute $\mathbf{G}^{-1}$. Lourenco et al. (2018) and Garcia et al. (2020) showed that correlations

between IP obtained based on SNP effects from all genotyped animals or only core animals from APY under ssGBLUP were >0.98; with a reduced computing cost when using only core animals. In an effort to reduce computations for SNP PEC, additional scenarios were tested with reduced number of genotyped animals. The scenarios were as follows:

1) *direct*: All genotyped animals (54,533) and phenotypes with direct $\mathbf{G}^{-1}$

2) *apy*: All genotyped animals (54,533) and phenotypes with APY $\mathbf{G}^{-1}$

3) 5*0k-2k*: All phenotypes and decreasing the number of genotyped animals from 50K to 2K

4) *core*: Genotypes for core animals only (15K) and all phenotypes

5) *hacc*: Genotypes for high accuracy animals only (15K) and all phenotypes

6) *core_prog*: Genotypes and phenotypes for core animals plus their progeny phenotypes

7) *hacc_prog*: Genotypes and phenotypes for high accuracy animals plus their progeny phenotypes

The first two scenarios (*direct* and *apy*) used all animals in the reduced data and reflect an extreme case when all animals in the evaluation are used to calculate SNP PEC and ACC_{IP} and serve as a test to compare the impact of direct or APY inversion of **G** in PEC computations. The other scenarios represent a situation when only a subset of the animals is used. In scenario five (*hacc*), 15,000 animals with the highest accuracy based on the benchmark ($GEBV_{ACC}$) were selected. In all scenarios, the pedigree for animals with phenotypes and/or genotypes was traced 3 generations back. The number of animals with genotypes, phenotypes, and pedigree for each scenario is shown in Table 4.1. Once SNP PEC were available, ACC_{IP} was calculated as in equation (12) for validation animals in each scenario. Regardless of the number of animals used to obtain PEC in each scenario, GEBV used to backsolve SNP effects were always obtained from the first scenario including phenotypes and genotypes in the reduced dataset. This is to mimic the real situation where GEBV are available from an official evaluation.

To check the quality of the IP and ACC_{IP}, correlation between GEBV and IP as well as the correlation between ACC_{GEBV} and ACC_{IP} were calculated for validation animals. Further, a regression model was fitted as $\mathbf{ACC_{GEBV}=b0+b1\times ACC_{IP}}$, to investigate the presence of scale differences and dispersion in ACC_{IP} calculation. All the analyses were performed using the BLUPF90 family of programs (Misztal et al., 2014b) after modifications to compute PEC of SNP accounting for blending and tuning.

RESULTS AND DISCUSSION

IP AND ACCURACY OF IP

The correlations between GEBV and IP were ≥0.99 when 10K or more genotyped animals were used to backsolve SNP effects. Previous studies have shown that IP can be safely obtained when using the APY algorithm or by using a subset of the genotyped animals, as long as the GEBV and genotypes used to backsolve SNP effects come from previous ssGBLUP evaluations (Lourenco et al., 2015; Lourenco et al., 2018; Garcia et al., 2020).

The quality of the IP accuracies was evaluated based on correlations and the regression of ACC_{GEBV} on ACC_{IP} and results for all scenarios are presented in Table 4.2. Correlations were ≥0.89 across all scenarios and ≥0.99 when 20k or more genotyped animals were used to calculate PEC for SNP effects. Our results show that as long as the number of genotyped animals used to calculate PEC represent the dimensionality of the genomic information (98-99% of the variance in **G**), correlations between ACC_{GEBV} and ACC_{IP} were ≥0.96.

Using high accuracy animals resulted in slightly lower correlations (*hacc*= 0.97 and *hacc_prog*= 0.96) which indicates that randomly selecting the animals from the whole genotyped population would be a better strategy for PEC computations. This would allow a better

representation of genotyped animals in all generations, although differences are not great. When phenotypes and pedigree information were available only from own and progeny records (*core_prog* and *hacc_prog*), correlations did not drop dramatically although the dispersion increased.

Even when correlations between accuracies are high, we need to make sure ACC_{IP} is unbiased and in the same scale as $GEBV_{ACC}$. This will assure that IP can be used as interim evaluations or permanent replacements for GEBV when the number of genotyped animals becomes extremely large to use all young animals in the evaluation. For all the scenarios, the coefficient of the regression (**b1**) of ACC_{GEBV} on ACC_{IP} was used to evaluate dispersion and the intercept (**b0**) was used to check the scale. If there is no dispersion, **b1**=1, and deviations from one indicate either under or overestimation of ACC_{IP}. Regression coefficient and intercept for each scenario are presented in Table 4.2. No bias or scale differences were found when all genotyped animals in the reduced data were used to calculate SNP PEC in scenarios *direct* and *apy*, and for scenarios *50k* and *40k*, **b1**≥0.92 and **b0**≤0.08. Using APY did not result in any differences in accuracy calculations and the results were basically identical to using direct inversion of **G** matrix.

As the number of genotyped animals decreased, ACC_{IP} were underestimated and the difference in scale between ACC_{IP} and ACC_{GEBV} increased. For instance, **b1** was as low as 0.42 and **b0** as high as 0.56 for the *2k* scenario. Typically, when **b1** is smaller than one, the conclusion is that the predictions are overestimated; however, this is true when **b0** is close to 0. When 30k or less genotyped animals were used to compute PEC, the intercept was not zero and despite **b1**<1, ACC_{IP} were underestimated rather than overestimated. As a matter of illustration, plots of ACC_{GEBV} versus ACC_{IP} are shown for two scenarios. Figure 4.1 shows the *direct* scenario where there was no dispersion; and Figure 4.2 shows the *core* scenario, where ACC_{IP} were clearly

underestimated. This underestimation can be easily seen in the descriptive statistics in Table 4.3. For instance, the average ACC_{GEBV} was 0.73 but the average ACC_{IP} was as low as 0.57 for the *core_prog* scenario and even lower (0.41) for the *2k* scenario.

While with a smaller subset of genotyped animals (*50k* and *40k*), we were able to successfully approximate SNP PEC and obtain good ACC_{IP}, as the number of genotyped animals decreased, ACC_{IP} deteriorated. As the number of genotyped animals decrease, the contributions due to the $\mathbf{G^{-1}-A_{22}^{-1}}$ block of MME are reduced and the approximation of PEC becomes poor, resulting in underestimated IP accuracies.

Even with the number of animals in the pedigree and with records remaining constant in most of the scenarios (Table 4.1), the changes in ACC_{IP} are a function of the number of genotyped animals used to compute SNP PEC. Further, using only own and progeny records, did not result in increased dispersion compared to using complete data and pedigree information (*core* vs *core_prog* and *hacc* vs *hacc_prog* scenarios in Table 4.2). It is worth noticing that the number of records and animals in the pedigree was nearly halved comparing *core* and *core_prog* scenarios.

This indicates that including enough genotyped animals with own phenotypes, and the addition of their phenotyped progeny may be enough to account for the contributions due to phenotypes and pedigrees as well as $\mathbf{G^{-1}-A_{22}^{-1}}$ and obtain reasonable SNP PEC for IP accuracy.

Using SNP PEC from a SNP-BLUP model, Erbe et al. (2018) found that the reference population composition affected the quality of the final GEBV accuracies approximation from the Interbull Standardized Genomic Reliability Model (ISGRM), and pointed out that under ssGBLUP, the definition of such reference population is not as clear as in the multi-step procedure, which would require further investigation to define which animals should be included in PEC computations from ssGBLUP.

The inversion of the LHS to obtain SNP PEC from a ssGBLUP model is the most demanding step in the process of calculating accuracies for IP, therefore, reducing the overall size of the MME to be inverted, and specially reducing the number of genotyped animals is of interest for routine applications. Compared to the approach presented by Liu et al. (2017) for an SNP-BLUP model, obtaining SNP PEC from ssGBLUP may be difficult because it depends on the number of animals rather than the number of markers included in the system of equations, therefore reducing the number of genotyped animals for PEC computations is critical.

With 40 to 50K genotyped animals it was possible to obtain ACC_{IP} without severe dispersion. Additionally, our results suggest that using as few as 15K genotyped animals can yield correlations between ACC_{IP} and ACC_{GEBV} that are as high as 0.98. Although it is important to note that with smaller number of animals, even with blending and tuning parameters considered, there was still a scaling issue and ACC_{IP} were underestimated. To be able to use smaller subsets of animals in PEC computations, fine tuning of formulas will be needed to overcome this issue.

More research is needed to investigate whether SNP PEC computed from a smaller subset of genotyped animals can be used to approximate ACC_{IP} based on a number of genotyped animals that is larger than that included in our study. Such tests may become hard to accomplish because obtaining ACC_{GEBV} based on PEV as a benchmark is not feasible for large datasets.

With the formulas and implementation presented in our study it is feasible to obtain SNP PEC from ssGBLUP and it is a more straightforward approach than using a SNP-BLUP model, as it does not require an extra run to compute SNP PEC.

The SNP PEC accounts for the genomic contributions from ssGBLUP MME and a combination of our approach with existing PEV approximations may be useful to obtain GEBV accuracies for large scale evaluations.

CONCLUSIONS

Indirect prediction accuracy can be successfully obtained by computing SNP PEC from single-step MME using direct inversion of **G** or by the APY algorithm, with the same formulas. With at least 40K genotyped animals included in PEC calculations, robust indirect predictions accuracies can be obtained without dispersion. To reduce computational costs of inverting the LHS even further, PEC can be approximated by using a smaller subset of the genotyped animals. This yields high correlations but a fine tuning is still required to scale accuracies of indirect predictions up to accuracies of GEBV. Further studies are needed to investigate fine tuning of PEC approximation for large scale genomic data.

REFERENCES

Aguilar, I., A. Legarra, F. Cardoso, Y. Masuda, D. Lourenco, and I. Misztal. 2019. Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. Genetics Selection Evolution 51 (1):28. doi: https://doi.org/10.1186/s12711-019-0469-3

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

Erbe, M., C. Edel, E. C. G. Pimentel, J. dodenhoff, and K. U. Gotz. 2018. Approximation of reliability in single step models using the interbull standardized genomic reliability method. Interbull Bulletin (54)

Garcia, A. L. S., Y. Masuda, S. Tsuruta, S. Miller, I. Misztal, and D. Lourenco. 2020. Indirect predictions with a large number of genotyped animals using the algorithm for proven and young Journal of animal science (Under review)

Gualdron Duarte, J. L., R. J. Cantet, R. O. Bates, C. W. Ernst, N. E. Raney, and J. P. Steibel. 2014. Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. BMC Bioinformatics 15:246. doi: https://doi.org/10.1186/1471-2105-15-246

Henderson, C. R. 1984. Applications of linear models in animal breeding. University of Guelph Guelph.

Legarra, A., D. A. Lourenco, and Z. Vitezica. 2018. Bases for genomic prediction. http://genoweb.toulouse.inra.fr/~alegarra/GSIP.pdf.

Liu, Z., P. VanRaden, M. H. Lidauer, M. P. Calus, H. Benhajali, H. Jorjani, and V. Ducrocq. 2017. Approximating genomic reliabilities for national genomic evaluation. Interbull Bulletin 51

Lourenco, D. A. L., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. Genetics Selection Evolution 47 (1):56. doi: https://doi.org/10.1186/s12711-015-0137-1

Lourenco, D. A. L., A. Legarra, S. Tsuruta, D. Moser, S. Miller, and I. Misztal. 2018. Tuning indirect predictions based on SNP effects from single-step GBLUP. Interbull Bulletin (53)

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: https://doi.org/10.1534/genetics.115.182089

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: https://doi.org/10.3168/jds.2013-7752

Misztal, I., S. Tsuruta, I. Aguilar, A. Legarra, P. M. VanRaden, and T. J. Lawlor. 2013. Methods to approximate reliabilities in single-step genomic evaluation. Journal of Dairy Science 96 (1):647-654. doi: https://doi.org/10.3168/jds.2012-5656

Misztal, I., S. Tsuruta, D. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. 2014b. Manual for BLUPF90 family of programs.

Misztal, I., and G. R. Wiggans. 1988. Approximation of prediction error variance in large-scale animal models. Journal of Dairy Science 71:27-32. doi: https://doi.org/10.1016/S0022-0302(88)79976-2

Nicolazzi, E. L., J. W. Durr, and G. R. Wiggans. 2018. Genomics in the US dairy industry: Current and future challenges. Interbull bulletin 53

Pocrnic, I., D. A. L. 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: https://doi.org/10.1534/genetics.116.187013

Pocrnic, I., D. A. L. 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 48:82. doi: https://doi.org/10.1186/s12711-016-0261-6

Pocrnic, I., D. A. L. 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):75. doi: https://doi.org/10.1186/s12711-019-0516-0

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

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. doi: https://doi.org/10.1017/S001667231100022X

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics research 94 (2):73-83. doi: https://doi.org/10.1017/s0016672312000274

Wiggans, G. R., P. M. VanRaden, and T. A. Cooper. 2015. Technical note: Rapid calculation of genomic evaluations for new animals. Journal of Dairy Science 98 (3):2039-2042. doi: https://doi.org/10.3168/jds.2014-8868

TABLES

Table 4.1: Number of animals with genotypes, phenotypes and pedigree information in each scenario.

Scenario	Genotypes	Phenotypes	Pedigree
direct	54,533	38,000	230,639
apy	54,533	38,000	230,639
50k	50,000	38,000	230,639
40k	40,000	38,000	230,639
30k	30,000	38,000	230,639
20k	20,000	38,000	230,639
10k	10,000	38,000	230,639
5k	5,000	38,000	230,639
2k	2,000	38,000	230,639
core	15,000	38,000	230,639
hacc	15,000	38,000	230,639
core_prog	15,000	22,625	101,837
hacc_prog	15,000	32,673	106,051

Table 4.2: Accuracy correlations and regression coefficients

Scenario	Correlation	b0	b1
direct	>0.99	-0.01	1.00
apy	>0.99	-0.01	1.01
50k	>0.99	0.02	0.98
40k	>0.99	0.08	0.92
30k	0.99	0.16	0.84
20k	0.99	0.25	0.74
10k	0.97	0.37	0.62
5k	0.94	0.47	0.53
2k	0.89	0.56	0.42
core	0.98	0.31	0.69
hacc	0.97	0.35	0.62
core_prog	0.97	0.34	0.68
hacc_prog	0.96	0.37	0.60

Table 4.3: Descriptive statistics for ACC_{GEBV} and ACC_{IP} for all scenarios

Scenario	Average	Min	Max	SD
GEBV	0.73	0.27	0.82	0.03
direct	0.73	0.28	0.82	0.03
apy	0.74	0.28	0.82	0.03
50k	0.73	0.26	0.82	0.03
40k	0.71	0.21	0.80	0.03
30k	0.68	0.10	0.79	0.04
20k	0.64	0.00	0.76	0.04
10k	0.57	0.00	0.71	0.05
5k	0.50	0.00	0.67	0.05
2k	0.41	0.00	0.62	0.06
core	0.61	0.00	0.74	0.04
hacc	0.62	0.00	0.76	0.05
core_prog	0.57	0.00	0.70	0.04
hacc_prog	0.61	0.00	0.75	0.05

FIGURES

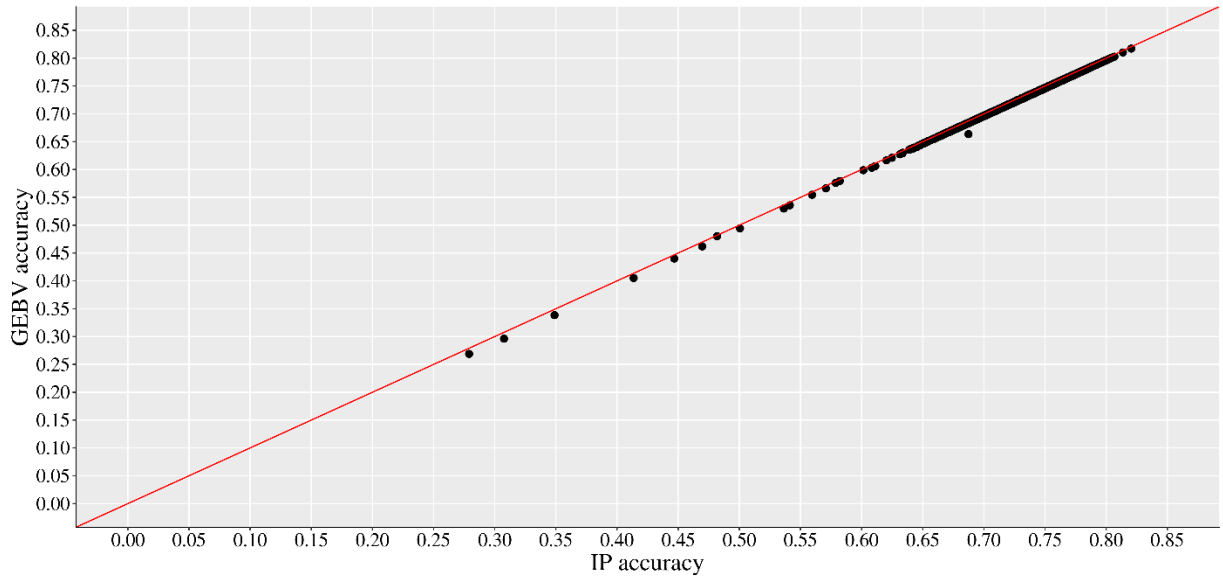


Figure 4.1: Accuracies for GEBV and IP from direct scenario

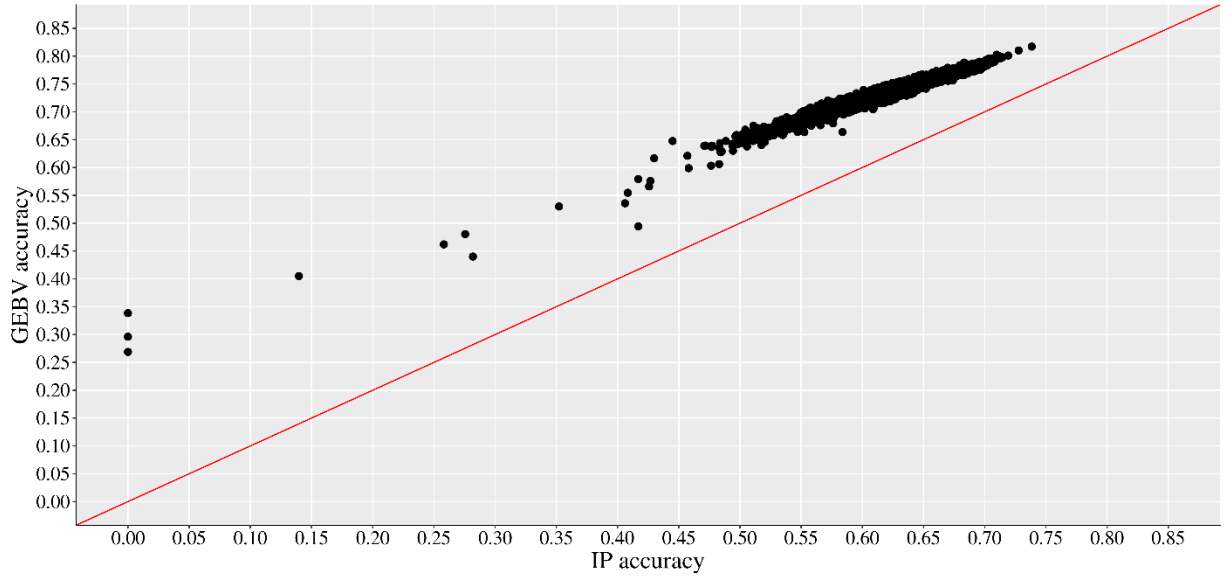


Figure 4.2: Accuracies for GEBV and IP from core scenario

CHAPTER 5

CONCLUSIONS

Using genomic information is feasible and beneficial for the US channel catfish breeding program because it provides greater ability to predict future performance and it reduces inflation of breeding values. At the same time, phenotype recording is essential to obtain the maximum advantage of genomic selection especially for carcass traits.

For larger genotyped populations with many young genotyped animals, indirect predictions are a robust tool for prediction when SNP effects are backsolved using GEBV from previous (ss)GBLUP evaluation. In purebred beef cattle populations, computing cost can be further reduced by using a sample of at least 15,000 animals representing the whole genotyped population to obtain SNP effects, as long as their GEBV comes from the previous (ss)GBLUP evaluation.

When indirect predictions from ssGBLUP are used as interim evaluations or to provide genomic predictions for unregistered animals, their accuracy is available by computing SNP PEC from MME either with direct inversion of **G** or by using the APY algorithm. With at least 40K genotyped animals included in PEC calculations, robust indirect predictions accuracies can be obtained without dispersion. To reduce computational costs of inverting the LHS even further, PEC can be approximated by using a smaller subset of the genotyped animals. This yields high correlations but a fine tuning is still required to scale accuracies of indirect predictions up to accuracies of GEBV. Further studies are needed to investigate fine tuning of PEC approximation for large scale genomic data.