IMPROVING GENOMIC PREDICTION METHODS AND ESTIMATION OF GENETIC PARAMETERS FOR LARGE POPULATIONS

by

MARY KATE HOLLIFIELD

(Under the Direction of Ignacy Misztal)

ABSTRACT

Improving the methods for genomic prediction and genetic parameters will increase accuracy and decrease the generation interval, resulting in increased genetic gain. The objective of this dissertation was to introduce methods to improve the accuracy, efficiency, and understanding of genomic prediction and genetic parameter estimation in large populations. Simulated datasets and datasets from dairy cattle and pig populations were used to test and analyze the methods. The differences in bias, accuracy, and computing time using ssGBLUP were negligible when blending the genomic relationship matrix with different proportions of the identity or pedigree relationship matrix for genotyped animals. However, a new algorithm was introduced that reduced the bottleneck in computing time from approximately 2 hours to less than one second. The number of independent chromosomes in a population is assumed to be $4N_eL$, where N_e is the effective population size and L is the genome length in Morgans (M). A segment effect model was compared with the true accuracy to test if all genetic variation could be explained in $4N_eL$ segments. Segment accuracies maximized at $4N_eL$ but were not as high as the true accuracy, suggesting a more biologically reasonable definition for segments is needed. Using genomic

information in heritability estimation in large populations is computationally expensive. Method R using genomics was compared with AI-REML with genomics and reduced computing time from 9.5 to 1.6 hours. However, the heritability estimates were not as precise and had large standard errors compared to the AI-REML estimates. Improvement in high-throughput phenotyping methods is also needed to incorporate this information into genetic evaluations and increase genetic gain. Behavior traits were recorded using digital phenotyping in a pig population. The data quality was analyzed, and genetic parameters of the behavior traits and relating behavior traits to production traits were estimated. The behavior traits analyzed had heritabilities ranging from 0.19 to 0.57 and had low to moderate genetic correlations with production traits. As the amount of phenotypic and genomic information is increasing rapidly, methods must be improved continuously to utilize the information and incorporate it into genetic evaluations.

INDEX WORDS: big data, digital phenotyping, heritability, independent chromosome segments, prediction accuracy, single-step GBLUP

IMPROVING GENOMIC PREDICTION METHODS AND ESTIMATION OF GENETIC PARAMETERS FOR LARGE POPULATIONS

by

MARY KATE HOLLIFIELD

B.S., North Carolina State University, 2019

M.S., University of Georgia, 2021

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

2024

IMPROVING GENOMIC PREDICTION METHODS AND ESTIMATION OF GENETIC PARAMETERS FOR LARGE POPULATIONS

by

MARY KATE HOLLIFIELD

Major Professor:	Ignacy Misztal
Committee:	Daniela Lourenco Romdhane Rekaya

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2024

DEDICATION

To my grandfather, Cecil Lee Hollifield, who encouraged me to strive for greatness.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisor, Dr. Ignacy Misztal. Thank you for believing in me, especially when the results were not as expected. We always found and solved the errors and "created a story" to fit the unknown. I will forever hear your voice when writing or analyzing results. Thank you for making me a better scientist.

Dr. Daniela Lourenco, thank you for all that you taught me. Both academically and personally. Thank you for the constant reassurance and for showing me how to be a great leader. Thank you for helping me find the pieces to my projects when they have been lost. I couldn't have finished without your continuous support.

Dr. Romdhane Rekaya, thank you for teaching me how to think as a geneticist, statistician, data scientist, biologist, etc. combined. The discussions with you throughout my graduate school journey will be with me for life. Thank you for bringing us up when we feel defeated by the class material. Thank you for challenging us and then reminding us that we will be valued in the future for the education we have.

Dr. Jorge Hidalgo, you have been such an inspiration to me during my entire graduate school journey. Your dedication to research and teaching, as well as your dedication to family and leadership in the lab, is incomparable. Your kindness and support are highly appreciated.

Dr. Ching-Yi Chen, thank you for taking me under your wing for a summer internship at PIC. I gained so much valuable experience and knowledge that prepared me for the next step of my journey.

To the postdocs, students, and visitors in the lab, thank you for being there for me when I needed you. Thank you for sharing your wealth of knowledge on research and life. You have shaped me into the scientist and person that I am today. Thank you for sharing your cultures, your friendship, and your support. I give a special thank you to those who were my "cubical neighbors" during my time in graduate school and were my day-to-day support system: Yvette Steyn, Matias Bermann, Jennifer Richter, Joe Tabet, and Fernando Bussiman. I could not have finished without you all.

To my non-academic family, thank you to Covenant Grove Farms for providing me a paradise of a home for myself and my many animals. Thank you for being the most welcoming, blissful escape from graduate school.

Lastly, to my family. Thank you for believing in me, for trusting me, and, most of all, for supporting me. You have inspired me to be the best that I can be. Thank you for allowing me to achieve my goals. I could not have survived without your encouragement. I hope I have made you proud and can be an inspiration to the next generation. ❤

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Genetic evaluations are used in populations to quantify the genetic performance of individuals to enhance production or health. Data and technological advancement are rapidly increasing, and some standard genomic prediction and genetic parameter estimation methods have become insufficient. Predictions and estimations must be unbiased, accurate, and efficient to maintain positive genetic gain and capture correct biological features. The quantity of phenotypic and genotypic information and the rate of new data-collecting technology are exponentially increasing; therefore, the advancement of genetic evaluation methodology must continue to be able to utilize and understand the wealth of information.

The objective of this dissertation was to introduce and analyze methods to improve genomic prediction and genetic parameter estimation in large populations. A literature review on the foundations of estimation and prediction methods, challenges of large datasets, and applications of precision agriculture is present in Chapter 2. Chapter 3 compares accuracy, bias, and computing time for blending the genomic relationship matrix with different proportions of the identity matrix or the pedigree relationship matrix for genotyped animals. Chapter 4 explains the characteristics and statistical properties of independent chromosomes and the function of chromosome segments in genomic prediction methods. A twist on an old heritability estimation method, Method R, is introduced in Chapter 5 with the inclusion of genomic information. This

chapter presents the reduction in computing time for heritability estimation with genomics using method R compared to AI-REML. Chapter 6 evaluates data collected by digital phenotyping, a data cleaning protocol, including behavior traits in a genomic evaluation and their relationship with production traits, and the application and implications of digital phenotyping. The overall conclusions of this dissertation are presented in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

GENETIC AND GENOMIC PREDICTIONS

The fundamentals of the rate of genetic change are represented in the "Breeder's Equation" (Lush, 2013):

$$\Delta G = \frac{r_{BV,\widehat{BV}} i \sigma_a^2}{L}. \quad [1]$$

Where ΔG is the rate of genetic change per unit time, $r_{BV,\widehat{BV}}$ is the accuracy of selection, i is selection intensity, σ_a^2 is the additive genetic variance, and L is the generation interval. The numerator must be increased, and the denominator must be decreased for faster genetic change. Estimated breeding values (EBV) must be calculated to determine the genetic trends. Genomic EBV (GEBV) can be further decomposed into $GEBV = w_1 PA + w_2 YD + w_3 PC + w_4 DGV - w_5 PP$, where PA is parent average, YD is yield deviation, PC is progeny contribution, DGV is direct genomic value, and PP is pedigree prediction based on the subset of animals that are genotyped (Wiggans and VanRaden, 1991). (G)EBV can be predicted from best linear unbiased predictor (BLUP) evaluations (Henderson, 1949; Henderson, 1950).

Mixed model equations allow simultaneous estimation of fixed and random effects. The standard BLUP MME can be expressed as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}, \quad [2]$$

where **X** is an incidence matrix relating fixed effects to animals, **Z** is an incidence matrix relating random effects to animals, λ is the variance ratio ($\frac{\sigma_e^2}{\sigma_a^2}$; residual variance divided by additive genetic variance), **I** is the identity matrix, $\widehat{\boldsymbol{\beta}}$ is a vector of estimated fixed effects, $\widehat{\boldsymbol{u}}$ is a vector of predicted random effects, and **y** is a vector of phenotypes. Henderson (1973) was the first to introduce pedigree relationship information into the MME. When animals are assumed unrelated, **I** is used; however, when relationships are included, **I** can be replaced with the inverse of **A** (pedigree relationship matrix), **G** (genomic relationship matrix), or **H** (combination of pedigree and genomic relationship matrix). The animal model was established when the inverse of **A** could be built directly, eliminating the computing expense of inverting the matrix (Henderson, 1976; Quaas, 1976). The solutions of the additive genetic random effect are described as (G)EBV. Assuming the animal model, the expectations and variances for equation 2 are assumed to be:

$$E\begin{bmatrix}\mathbf{y}\\ \boldsymbol{u}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\mathbf{X\beta}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix} \quad [3]$$

$$Var\begin{bmatrix}\mathbf{y}\\ \boldsymbol{u}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\mathbf{V} & \mathbf{ZG} & \mathbf{R}\\ \mathbf{GZ'} & \mathbf{G} & \mathbf{0}\\ \mathbf{R} & \mathbf{0} & \mathbf{R}\end{bmatrix}, \quad [4]$$

where $\mathbf{V} = \mathbf{ZGZ'} + \mathbf{R}$, $\mathbf{G} = \mathbf{A}\sigma_u^2$ (**A** can be replaced with **I**, **G**, or **H** depending on the model), and $\mathbf{R} = \mathbf{I}\sigma_e^2$.

The human genome project paved the way for the future of genomics (Sachidanandam et al., 2001). Sequencing technology advanced, and single nucleotide polymorphism (SNP) chips became widely used in humans, animals, and plants. Various SNP densities can be used depending on the purpose, such as identifying new mutations (high density), capturing relationships and SNP related to quantitative trait loci (QTL; medium density), or parentage verification (low density). Genotypes can also be imputed up or down to fit a desired density. Meuwissen et al. (2001)

proposed using markers to predict genetic value and increase the rate of genetic gain in populations. Marker-assisted selection (MAS) was introduced, where animals were selected based on a desired marker profile for a specific trait. Few markers were used due to the high cost of genotyping at this time. MAS worked well for traits of a few genes but not for the production traits that are quantitative in nature. The need for technology and models to detect and analyze enough SNP to explain the variation in quantitative traits became apparent. As Fisher (1919) demonstrated with the infinitesimal model, the variation in a quantitative trait is due to an infinitely large number of genes that all have a very small impact on the phenotype.

A simple marker model can be used to estimate the effect of a marker in complete or incomplete linkage disequilibrium (LD) with a QTL and is considered an additive model:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z}\boldsymbol{a} + \boldsymbol{e} \quad [5]$$

Where $\mathbf{Z}$ relates markers with animals, and $\boldsymbol{a}$ is a vector of allele effects. Lande and Thompson (1990) suggested using the markers that are associated only with traits of interest. However, associations can differ across populations, which will cause bias, reducing the power of QTL detection (Legarra et al., 2018). Meuwissen et al. (2001) suggested to assume all markers are QTL and use all available markers to reduce bias and estimate a whole genetic effect based on markers. Using all markers becomes an issue when the number of markers is greater than the number of genotyped animals, creating more parameters to be estimated than the data available. This can be accommodated by treating the SNP effects as random, which allows all effects to be estimated together.

SNP-BLUP is used for estimating marker effects and is expressed as:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix} \quad [6]$$

with $Var(\boldsymbol{a}) = \mathbf{D} = \mathbf{I}\sigma_a^2$ and $Var(\boldsymbol{e}) = \mathbf{R} = \mathbf{I}\sigma_e^2$, where $\widehat{\boldsymbol{a}}$ is a vector of marker effects and the other variables are defined above. SNP-BLUP assumes the same variance for each SNP. Other

methods exist that do not have this assumption. BayesA assumes heterogeneous variances per SNP by providing information on the marker variances *a priori* (Meuwissen et al., 2001; Gianola et al., 2009). BayesB is similar to BayesA, except it assumes a subset of SNP has no effect (Meuwissen et al., 2001; Gianola et al., 2009). BayesC assumes homogenous variance for all SNP except for a subset of SNP with no effect (Habier et al., 2011). These SNP-based methods quantify allelic similarities between individuals.

VanRaden (2008) proposed two methods for explaining genomic relationships between individuals using standardized covariances. The first method proposes that individuals are a sum over marker effects: $\boldsymbol{u} = \mathbf{Z}\boldsymbol{a}$. Whereas the (co)variance matrix can be rearranged as:

$$Var(\boldsymbol{u}) = \mathbf{Z}Var(\boldsymbol{a})\mathbf{Z}' = \mathbf{ZDZ}' = \mathbf{ZZ}'\sigma_a^2. \quad [7]$$

To standardize, the variance must be divided by the variance of the breeding values (σ_u^2) for the set of animals, and assuming the population is in Hardy-Weinberg and Linkage Equilibrium, $\sigma_u^2 = 2\sum_{i=1}^{nsnp} p_i q_i \sigma_a^2$. Therefore, the genomic relationship matrix is:

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum_{i=1}^{nsnp} p_i q_i}. \quad [8]$$

The second method is used less frequently and includes unique weights per marker: $\mathbf{G} = \mathbf{ZD}_w\mathbf{Z}'$. When **I** is replaced by **G** in eq. 2, the model is genomic BLUP (GBLUP).

The main limitation of GBLUP is the exclusion of animals without genotypes. Methods are available to combine information from genotyped and non-genotyped animals in multiple steps (VanRaden et al., 2009); however, this is time-consuming and introduces biases. Therefore, single-step GBLUP (ssGBLUP) was proposed where a relationship matrix exists for genotyped and non-genotyped animals (Legarra et al., 2009; Misztal et al., 2009; Aguilar et al., 2010; Christensen and Lund, 2010). The basis of ssGBLUP is to consider **A** as *a priori* relationship information and **G** as the observed relationship information (Legarra et al., 2009). Therefore, the non-genotyped animals

($\mathbf{u_1}$) could obtain observed relationship information via the genotyped animals ($\mathbf{u_2}$) based on the joint distribution of their breeding values:

$$p(\boldsymbol{u}_1, \boldsymbol{u}_2) = \boldsymbol{p}(\boldsymbol{u}_2)\boldsymbol{p}(\boldsymbol{u}_1|\boldsymbol{u}_2). \quad [9]$$

Where the joint relationships can be shown in matrix form as:

$$\mathbf{H} = \begin{pmatrix} var(\boldsymbol{u}_1) & cov(\boldsymbol{u}_1, \boldsymbol{u}_2) \\ cov(\boldsymbol{u}_1, \boldsymbol{u}_2) & var(\boldsymbol{u}_2) \end{pmatrix}. \quad [10]$$

After rearranging and simplification:

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} [\mathbf{G} - \mathbf{A}_{22}] [\mathbf{I} \quad \mathbf{I}] \begin{bmatrix} \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \quad [11]$$

When $\mathbf{I}$ is replaced by $\mathbf{H}$ in eq. 2, the model is called ssGBLUP. The inverse of $\mathbf{H}$ is needed to solve the MME. Directly building $\mathbf{H^{-1}}$ is possible and more straightforward than building $\mathbf{H}$ (Aguilar et al., 2010; Christensen and Lund, 2010):

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix}. \quad [12]$$

Today, ssGBLUP is the most popular methodology used in genetic evaluations when genotyped and non-genotyped animals are included (Bermann et al., 2022).

ESTIMATION OF GENETIC PARAMETERS

The two most used methods for estimating variance parameters are restricted maximum likelihood (REML) and Gibbs sampling. REML was developed by Patterson and Thompson (1971). For the model in eq. 2, REML estimates are given by maximizing the log of the likelihood function and deriving with respect to $\mathbf{G}$ and $\mathbf{R}$. Where the likelihood function is:

$$l(\mathbf{y}|\boldsymbol{u}|\boldsymbol{\beta}, \mathbf{R}, \mathbf{G}) = l(\mathbf{y}|\boldsymbol{u}|\boldsymbol{\beta}, \mathbf{R}) l(\boldsymbol{u}|\mathbf{G}). \quad [13]$$

After expanding, removing terms that are not functions of the parameters of interest, multiplying by -2, and taking the log, the log-likelihood function becomes (Verbyla, 1990):

$$-2\mathrm{L} \propto \ln|\mathbf{R}| + \ln|\mathbf{G}| + \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - 2\mathbf{y}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}'\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{u} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{u} + \boldsymbol{u}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{u} + \boldsymbol{u}'\mathbf{G}^{-1}\boldsymbol{u}. \quad [14]$$

After deriving with respect to $\mathbf{G}$ and $\mathbf{R}$, the estimates of the variance parameters are found. REML is widely used because of its resistance to selection bias, and the variance estimates are always located within the parameter space. Its disadvantages, however, involve its high cost for large or complex models. Inverting the left-hand side of the MME every iteration becomes essentially impossible with large genomic models.

The Gibbs sampler is a Monte Carlo Markov Chain (MCMC) sampling algorithm and is a special case of the Metropolis-Hastings algorithm (Geman and Geman, 1984). Gibbs sampling is based on sampling from a conditional distribution when marginalizing by integrating over a joint distribution is difficult. The full conditional distributions of the model parameters of interest ($\mathbf{G}$ and $\mathbf{R}$) are needed to implement Gibbs sampling. By sampling from the conditional distributions, the joint distribution can be approximated. These can be derived by constructing the joint posterior distribution and taking only the terms that are a function of the parameter of interest. The joint posterior distribution is the product of the likelihood function and the prior distributions, and for a multivariate mixed linear model, is represented as:

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \mathbf{R}, \mathbf{G}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{u}, \mathbf{R})p(\boldsymbol{u}|\mathbf{A}, \mathbf{G})p(\mathbf{R}|\nu_e, \mathbf{S}_e)p(\mathbf{G}|\nu_u, \mathbf{S}_u) \quad [15]$$

The full conditional distribution of the parameter of interest, $\mathbf{R}$, in eq. 2 is as follows:

$$p(\mathbf{R}|\boldsymbol{\beta}, \boldsymbol{u}, \mathbf{G}, \mathbf{y}) \sim IW[(n + \nu_e), (\mathbf{V}_e + \mathbf{S}_e) \quad [16]$$

which is a scaled inverted Wishart distribution with $(n + \nu_e)$ degrees of belief and scaling matrix $(\mathbf{V}_e + \mathbf{S}_e)$, where $\mathbf{V}_e = \begin{bmatrix} \boldsymbol{e}'_1\boldsymbol{e}_1 & \boldsymbol{e}'_1\boldsymbol{e}_2 \\ \boldsymbol{e}'_2\boldsymbol{e}_1 & \boldsymbol{e}'_2\boldsymbol{e}_2 \end{bmatrix}$, $\boldsymbol{e}_i = \boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i - \mathbf{Z}_i\boldsymbol{u}_i$, and $\mathbf{S}_e$ is a matrix of flat prior ("guesses") for the residual variances. The full conditional distribution of the parameter of interest, $\mathbf{G}$, in eq. 2 is as follows:

$$p(\mathbf{G}|\boldsymbol{\beta}, u, \mathbf{R}, \mathbf{y}) \sim IW[(q + \nu_u), (\mathbf{V}_u + \mathbf{S}_u) \quad [16]$$

which is a scaled inverted Wishart distribution with $(q + \nu_u)$ degrees of belief and scaling matrix $(\mathbf{V}_u + \mathbf{S}_u)$, where $\mathbf{V}_u = \begin{bmatrix} \boldsymbol{u}_1'\mathbf{A}^{-1}\boldsymbol{u}_1 & \boldsymbol{u}_1'\mathbf{A}^{-1}\boldsymbol{u}_2 \\ \boldsymbol{u}'_2\mathbf{A}^{-1}\boldsymbol{u}_1 & \boldsymbol{u}_2'\mathbf{A}^{-1}\boldsymbol{u}_2 \end{bmatrix}$, and $\mathbf{S}_u$ is a matrix of flat prior ("guesses") for the genetic variances. The convergence of the Gibbs sampler is slow, it requires a large number of rounds, and can have high correlations between successive samples. However, due to its ability to manage complex traits or models, it is widely used.

CHALLENGES OF LARGE DATASETS

Generally, populations of the same species share a large proportion of their genetic material; thus, phenotypic or genotypic information can be redundant. Sequencing every animal in a population will give unnecessary overlapping information. Memory and computing costs can be reduced by using SNP instead of whole genome sequence. Additionally, high throughput phenotyping is becoming more widespread. In some cases, data are collected every second or less or in a continuous video accumulation. At a point, the information related to the unique individual will be captured, and additional information will be excessive. Therefore, methods to reduce redundant information genotypically and phenotypically can be adapted.

As the number of genotyped animals increases, the inversion of $\mathbf{G}$ becomes unfeasible due to its dense properties. As livestock populations have small effective population sizes, they are highly related and $\mathbf{G}$ is not full rank. Pocrnic et al. (2016b) found that the dimensionality of genomic information can be explained by the number of largest eigenvalues that explain 99% of the variation. They tested several livestock species and found 5,570 in broiler chickens, 6,083 in pigs, 14,555 in Angus cattle, 16,645 in Jersey cattle, and 19,379 in Holstein cattle. The limited dimensionality insinuates that many haplotype blocks and chromosome segments exist in the populations. The redundancy of genomic information can be quantified by the number of

independent chromosome segments (M_e), corresponding to the number of largest eigenvalues that explain 99% of the variation (Pocrnic et al., 2016a). Stam (1980) postulates that $M_e = 4N_eL$, where N_e is the effective population size, and L is the genome length in Morgans. Thus, for populations with smaller N_e, the smaller the dimensionality of $\mathbf{G}$.

Therefore, Misztal et al. (2014) proposed a sparse representation of $\mathbf{G^{-1}}$ created by recursion on a "core" subset of genotyped animals with the size of the rank of $\mathbf{G}$. Initially, the core subset were animals with high accuracy or "proven" and the noncore subset would be animals with lower accuracy, typically younger animals; thus, the method was named the Algorithm for Proven and Young (APY). However, Fragomeni et al. (2015) showed that random animals can be used as core and give an accurate inverse. $\mathbf{G}$ can be partitioned as:

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{G}_{cn} \\ \mathbf{G}_{nc} & \mathbf{G}_{nn} \end{bmatrix}, \quad [17]$$

where c and n represent blocks for core and noncore animals, respectively. Then, $\mathbf{G}_{APY}^{-1}$ can be obtained directly as (Misztal, 2016):

$$\mathbf{G}_{\mathrm{APY}}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} [-\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1} \quad \mathbf{I}], \quad [18]$$

where $\mathbf{M}_{nn} = diag(\mathbf{G}_{nn} - \mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn})$. As the number of genotyped animals increases, the computing and memory costs increase almost linearly (Fragomeni et al., 2015). APY with ssGBLUP is now widely used in genomic evaluations.

Additionally, high-throughput phenotyping is becoming widely used in livestock phenotyping, creating big data challenges (Koltes et al., 2019). A surplus of information comes with this technology and is widely under-utilized. A challenge with big phenotypic data is that they are often not clean data (Morota et al., 2018). With the rapid development of high-throughput technologies and the influx of data, verification and validation of data quality are essential. The power of automated phenotyping can lead to faster decision-making, animal health alerts, and

overall faster genetic gain (Berckmans and Guarino, 2017; Wolfert et al., 2017). Mistakes in the automation process could be detrimental to the population in the short and long term. For example, a glitch in an automated animal health system could cause illnesses or deaths that were preventable, or selecting phenotypes that are not a biological representation of the true trait could lead to unwanted population changes. Therefore, removing noise and eliminating errors are essential for utilizing high-throughput data.

PRECISION AGRICULTURE

Precision agriculture is becoming more prevalent and can improve plant and animal production by capturing larger quantities of phenotypic data, increasing accuracy and, thus, speeding up the rate of genetic gain (Brito et al., 2020). The majority of precision agriculture is seen digitally, such as picture or video information; however, many other avenues of precision agriculture exist to capture information. For example, accelerometers to quantify activity and behavior (Andriamandroso et al., 2016), microphones to identify stress (Moura et al., 2008; Exadaktylos et al., 2014), or sensors to measure respiratory rate (Strutzke et al., 2019). Digital precision agriculture has been utilized in real-time monitoring for welfare traits (Berckmans, 2014), infrared thermography for sickness or skin damage (Harris-Bridge et al., 2018), imaging to detect body condition scores (Azzaro et al., 2011), or video to compare behavior to production traits (Obermier et al., 2023).

Two main uses of precision livestock are to capture novel traits and to gain more measurements of existing traits with higher accuracy. As manual labor becomes more costly over time, technology becomes more inexpensive. For example, automatic milking systems have proven to substantially save in labor costs (Mathijs, 2004; Bijl et al., 2007). Often, the more elite

animals have phenotypes recorded, and the commercial-level animals do not. In turn, if the commercial-level animals had phenotypes recorded, this would increase the accuracy of the breeding candidates and improve genetic gain (Pérez-Enciso and Steibel, 2021). Digital phenotyping could provide a method to implement data recording on the commercial level with reduced or no manual labor costs.

Additionally, digital phenotyping could provide a way to analyze the animal's characteristics more precisely than the human eye. For example, with categorically measured traits (such as body condition score), human technicians may score the animals in whole or half number increments from one to ten, while a digital data extracting algorithm could give a score with more detailed increments with greater accuracy. Overall, with precision agriculture, more information with higher accuracy will be implemented into genetic evaluations and the decision-making of animal production and well-being.

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 score1. Journal of Dairy Science 93(2):743-752. doi: https://doi.org/10.3168/jds.2009-2730

Andriamandroso, A., J. Bindelle, B. Mercatoris, and F. Lebeau. 2016. A review on the use of sensors to monitor cattle jaw movements and behavior when grazing. Biotechnologie, Agronomie, Société et Environnement 20

Azzaro, G., M. Caccamo, J. D. Ferguson, S. Battiato, G. M. Farinella, G. C. Guarnera, G. Puglisi, R. Petriglieri, and G. Licitra. 2011. Objective estimation of body condition score by modeling cow body shape from digital images. Journal of dairy science 94(4):2126-2137.

Berckmans, D. 2014. Precision livestock farming technologies for welfare management in intensive livestock systems. Rev. Sci. Tech 33(1):189-196.

Berckmans, D., and M. Guarino. 2017. From the Editors: Precision livestock farming for the global livestock sector No. 7. p 4-5. Oxford University Press.

Bermann, M., A. Cesarani, I. Misztal, and D. Lourenco. 2022. Past, present, and future developments in single-step genomic models. Italian Journal of Animal Science 21(1):673-685. doi: 10.1080/1828051X.2022.2053366

Bijl, R., S. Kooistra, and H. Hogeveen. 2007. The profitability of automatic milking on Dutch dairy farms. Journal of Dairy Science 90(1):239-248.

Brito, L. F., H. R. Oliveira, B. R. McConn, A. P. Schinckel, A. Arrazola, J. N. Marchant-Forde, and J. S. Johnson. 2020. Large-scale phenotyping of livestock welfare in commercial production systems: A new frontier in animal breeding. Frontiers in genetics 11:793.

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

Exadaktylos, V., M. Silva, and D. Berckmans. 2014. Chapter Automatic Identification and Interpretation of Animal Sounds, Application to Livestock Production Optimisation.

Fisher, R. A. 1919. XV.—The correlation between relatives on the supposition of Mendelian inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh 52(2):399-433.

Fragomeni, B., D. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. 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.

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence (6):721-741.

Gianola, D., G. de Los Campos, W. G. Hill, E. Manfredi, and R. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. Genetics 183(1):347-363.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12(1):186. doi: 10.1186/1471-2105-12-186

Harris-Bridge, G., L. Young, I. Handel, M. Farish, C. Mason, M. A. Mitchell, and M. J. Haskell. 2018. The use of infrared thermography for detecting digital dermatitis in dairy cattle: What is the best measure of temperature and foot location to use? The Veterinary Journal 237:26-33.

Henderson, C. R. 1949. Estimation of changes in herd environment. J. Dairy Sci 32(8):706-706.

Henderson, C. R. 1950. Estimation of genetic parameters.

Henderson, C. R. 1973. Sire evaluation and genetic trends. Journal of Animal Science 1973(Symposium):10-41.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics:69-83.

Koltes, J. E., J. B. Cole, R. Clemmens, R. N. Dilger, L. M. Kramer, J. K. Lunney, M. E. McCue, S. D. McKay, R. G. Mateescu, and B. M. Murdoch. 2019. A vision for development and utilization of high-throughput phenotyping and big data analytics in livestock. Frontiers in genetics 10:1197.

Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124(3):743-756.

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

Legarra, A., D. A. Lourenco, and Z. G. Vitezica. 2018. Bases for genomic prediction. Short course 1(1):1-141.

Lush, J. L. 2013. Animal breeding plans. Read Books Ltd.

Mathijs, E. 2004. Socio-economic aspects of automatic milking, Automatic Milking—A Better Understanding. Academic Publishers, Wageningen. p. 46-55.

Meuwissen, T. H., B. J. Hayes, and M. 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.

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

Morota, G., R. V. Ventura, F. F. Silva, M. Koyama, and S. C. Fernando. 2018. BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Machine learning and data mining advance predictive big data analysis in precision animal agriculture1. Journal of Animal Science 96(4):1540-1550. doi: 10.1093/jas/sky014

Moura, D., W. Silva, I. Naas, Y. Tolón, K. Lima, and M. Vale. 2008. Real time computer stress monitoring of piglets using vocalization analysis. Computers and Electronics in Agriculture 64(1):11-18.

Obermier, D., M. Trenahile-Grannemann, T. Schmidt, T. Rathje, and B. Mote. 2023. Utilizing NU track to Access the Activity Levels in Pigs with Varying Degrees of Genetic Potential for Growth and Feed Intake. Animals 13(10):1581.

Patterson, H. D., and R. Thompson. 1971. Recovery of Inter-Block Information when Block Sizes are Unequal. Biometrika 58(3):545-554. doi: 10.2307/2334389

Pérez-Enciso, M., and J. P. Steibel. 2021. Phenomes: the current frontier in animal breeding. Genetics Selection Evolution 53(1):22. doi: 10.1186/s12711-021-00618-1

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

Pocrnic, I., D. A. 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(1):82. doi: 10.1186/s12711-016-0261-6

Quaas, R. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics:949-953.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P.-Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, D. Altshuler, S. N. P. M. W. G. The International, L. Cold Spring Harbor, I. National Center for Biotechnology, C. The Sanger, L. Washington University in St, and M. I. T. C. f. G. R. Whitehead. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409(6822):928-933. doi: 10.1038/35057149

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

Strutzke, S., D. Fiske, G. Hoffmann, C. Ammon, W. Heuwieser, and T. Amon. 2019. Development of a noninvasive respiration rate sensor for cattle. Journal of dairy science 102(1):690-695.

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

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

Verbyla, A. P. 1990. A CONDITIONAL DERIVATION OF RESIDUAL MAXIMUM LIKELIHOOD. Australian Journal of Statistics 32(2):227-230. doi: https://doi.org/10.1111/j.1467-842X.1990.tb01015.x

Wiggans, G., and P. VanRaden. 1991. Method and effect of adjustment for heterogeneous variance. Journal of Dairy Science 74(12):4350-4357.

Wolfert, S., L. Ge, C. Verdouw, and M.-J. Bogaardt. 2017. Big Data in Smart Farming – A review. Agricultural Systems 153:69-80. doi: https://doi.org/10.1016/j.agsy.2017.01.023

CHAPTER 3

IMPACT OF BLENDING THE GENOMIC RELATIONSHIP MATRIX WITH DIFFERENT LEVELS OF PEDIGREE RELATAIONSHIPS OR THE IDENTITY MATRIX ON GENETIC EVALUATIONS[1]

[1] Hollifield M. K., M. Bermann, D. Lourenco, I. Misztal. 2022. *JDS Communications*. 3(5):343-347. Reprinted here with permission of the publisher.

SUMMARY

For single-step GBLUP, the genomic relationship matrix (**G**) is blended with a positive definite matrix for inversion to solve the mixed model equations. Conventionally, **G** is blended with a proportion of the numerator relationship matrix for genotyped animals ($\mathbf{A}_{22}$); however, blending with $\mathbf{A}_{22}$ can take excess time, while blending with an identity matrix (**I**) may take less time and guarantees the non-singularity of **G**. The purpose of this study was to compare differences in the reliability and inflation of GEBV, convergence rate, and elapsed wall-clock time when blending **G** with different proportions of $\mathbf{A}_{22}$ or **I** and introduce a more efficient blending algorithm. Using a U.S. Holstein population of 9.7 million animals in the pedigree and 569,404 genotypes, negligible differences in performance were observed in blending with $0.001\mathbf{I}$ and $0.05\mathbf{A}_{22}$. The optimized blending algorithm reduced the computing time from approximately two hours to five minutes for $\mathbf{A}_{22}$ and less than one second for **I**.

ABSTRACT

Evaluations using single-step GBLUP require blending the genomic relationship matrix (**G**) with a positive definite matrix to ensure non-singularity for solving the mixed model equations (MME). Many organizations blend **G** with a proportion of the numerator relationship matrix for genotyped animals ($\mathbf{A}_{22}$) to improve stability and possibly add a residual polygenic effect. However, when nearly all the polygenic variance is explained by **G**, blending with $\mathbf{A}_{22}$ may cause inflation and add excess computing time; thus, blending with an identity matrix (**I**) multiplied by a small value may be a better solution. The objective of this study was to evaluate changes in

reliability and inflation of GEBV, convergence rate, elapsed wall-clock time for blending **G** with different levels of $\mathbf{A}_{22}$ or **I**, and develop a more time-efficient blending method. A U.S. Holstein cattle dataset was used with 9.7 million animals in the pedigree, 569,404 animals with genotypes, and 10.1 million stature phenotypes. Blending **G** by adding a small value to the diagonal elements had comparable performance to $\mathbf{A}_{22}$ with fewer rounds to convergence required to solve the system of equations. Reliability and inflation of GEBV ranged from 0.63 to 0.68 and 0.86 to 0.89 for all blending scenarios tested. The current blending default in the BLUPF90 software is to replace **G** with $(1 - \beta)\mathbf{G} + \beta\mathbf{A}_{22}$, where β equals 0.05. In this study, β values of 0.30, 0.20, 0.05, 0.01, 0.005, and 0.001 were evaluated with $\mathbf{A}_{22}$ and **I**. Negligible differences in elapsed computing time between the blending types and levels were observed. Subsequently, the current blending algorithm used in the BLUPF90 family of programs was optimized, reducing the blending time from approximately 2 hours to five minutes for $\mathbf{A}_{22}$ and less than one second for **I**. The new time difference between blending with $\mathbf{A}_{22}$ or **I** is negligible and not computationally critical. The results indicate that blending **G** with $\mathbf{A}_{22}$ does not have clear advantages over blending with a small proportion of **I**.

MAIN BODY

Single-step genomic best linear unbiased predictor (ssGBLUP) allows estimating the breeding values jointly for genotyped and non-genotyped animals (Aguilar et al., 2010, Christensen and Lund, 2010). For solving the mixed model equations (MME), the main difference between ssGBLUP and the pedigree-based best linear unbiased predictor (PBLUP) is in the covariance matrix for the breeding values. In PBLUP, the inverse of the numerator relationship matrix $(\mathbf{A}^{-1})$ is used, whereas in ssGBLUP is replaced by $\mathbf{H}^{-1}$. The matrix $\mathbf{H}^{-1}$ is composed of

$\mathbf{A}^{-1}$, the inverse of the genomic relationship matrix $(\mathbf{G}^{-1})$, and the inverse of the numerator relationship matrix for genotyped animals $(\mathbf{A}_{22}^{-1})$ as follows:

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

Calculating such matrices can be referred to as the "genomic setup," which typically involves computing $\mathbf{G}$, adding ("blending") it to a small fraction of a positive definite matrix; usually an identity matrix ($\mathbf{I}$) or $\mathbf{A}_{22}$, to guarantee its non-singularity (VanRaden, 2008), tuning to make it compatible with $\mathbf{A}_{22}$ (Vitezica et al., 2011), and inverting it. The current default in BLUPF90 is blending implemented first and then tuning. McWhorter et al. (2021) found predictions were unbiased, accurate, and neither over- nor under-dispersed with either order. When solving the MME, $\mathbf{A}^{-1}$ is calculated following Henderson's rules (Henderson, 1976, Quaas, 1976, 1988). For small evaluations, $\mathbf{A}_{22}$ is calculated with Colleau (2002) and inverted, whereas for large-scale evaluations, a product $\mathbf{A}_{22}^{-1}\mathbf{q}$ is calculated as proposed by Masuda et al. (2016). Typically, $\mathbf{G}$ is constructed using the first method of VanRaden (2008):

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

where $\mathbf{Z} = \mathbf{M} - \mathbf{P}$, p_j is the allele frequency of the second allele at locus j, calculated based on observed allele frequencies, $\mathbf{M}$ is a genotypic matrix relating marker alleles to individuals with the number of rows equal to the number of animals and the number of columns equal to the number of SNP, and $\mathbf{P}$ is a matrix containing $2p_j$.

The blending step is frequently done as $(1-\beta)\mathbf{G} + \beta\mathbf{A}_{22}$, where β reflects the proportion of residual polygenic variance not accounted for by $\mathbf{G}$ (Habier et al., 2007, Liu et al., 2016, Mäntysaari et al., 2017). Blending in this way creates a non-singular $\mathbf{G}$ and is equivalent to fitting a residual polygenic effect (RPG) in the model (Liu et al., 2014). However, when nearly all QTL

were identified, using $\mathbf{A}_{22}$ for blending reduced accuracy while using a fraction of **I** for blending did not (Fragomeni et al., 2017). The results from the study of Himmelbauer et al. (2021) showed that blending with β greater than 0.001 introduced biases for bulls with many genotyped progeny. It is standard for β to be equal to 0.05; although, depending on the population parameters and quality of genomic information, values of β can vary from 0.01 to 0.50 (Lourenco et al., 2020). Larger values of β are used when the markers do not adequately explain the additive genetic variance or to reduce the effect of the genomic information (Meyer et al., 2018).

When the number of genotyped animals is large, $\mathbf{G}^{-1}$ cannot be computed. In such a case, $\mathbf{G}^{-1}$ can be replaced by $\mathbf{G}_{\mathrm{APY}}^{-1}$, which is calculated with the Algorithm of Proven and Young (APY; (Misztal et al., 2014, Misztal, 2016). Let **G** be partitioned as

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{G}_{cn} \\ \mathbf{G}_{nc} & \mathbf{G}_{nn} \end{bmatrix}, \tag{3}$$

where the subscripts *c* and *n* denote the blocks for core (proven) and non-core (young) animals, respectively. Using APY, the inverse of **G** ($\mathbf{G}_{APY}^{-1}$) can be obtained directly as:

$$\mathbf{G}_{APY}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} [-\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1} \quad \mathbf{I}], \tag{4}$$

where $\mathbf{M}_{nn} = diag(\mathbf{G}_{nn} - \mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn})$ is a diagonal matrix.

APY reduces computational costs by utilizing concepts of effective population size and the limited dimensionality of the genomic relationship matrix (Fragomeni et al., 2015, Masuda et al., 2016, Pocrnic et al., 2016). Previous studies have successfully tested APY in various species (Ostersen et al., 2016, Mäntysaari et al., 2017, Gonzalez-Peña et al., 2019, Nilforooshan and Lee, 2019). For Holsteins, 15,000 eigenvalues corresponded to 98% of the variation and realized accuracies peaked using this number of randomly chosen core animals (Pocrnic et al., 2016). In addition to APY, other options are available for single-step with large genomic data, such as

ssGTBLUP (Mäntysaari et al., 2017), single-step Bayesian Regression (Fernando et al., 2014), single-step SNP-BLUP (Liu et al., 2014, Taskinen et al., 2017), and using reduced-dimension singular value decomposition of the genotype matrix (Ødegård et al., 2018). Although APY makes it possible to obtain a sparse representation of $\mathbf{G}^{-1}$ ($\mathbf{G}^{-1}_{\text{APY}}$) that well-depicts the population structure without inverting the entire $\mathbf{G}$ directly, blending is still needed because of the inverse of $\mathbf{G}$ for core animals. The computing time of the genomic setup may be a limiting factor to timely accomplish large-scale evaluations when the number of genotyped animals exceeds one million. Therefore, this study aimed to compare the efficiency of blending $\mathbf{G}$ with $\mathbf{I}$ versus $\mathbf{A}_{22}$, analyze the reliability and inflation of the resulting GEBV, and develop an improved blending algorithm.

The current implementation of APY in preGSf90 follows the methods in Masuda et al. (2016), and at the time of the development, the number of genotyped animals was small; therefore, the algorithm was efficient. Now that the number of genotyped U.S. Holsteins is nearing 5 million (Council of Dairy Cattle Breeding, 2022), a more efficient blending method is required for feasible routine evaluations. Because of the structure of $\mathbf{G}^{-1}_{APY}$ and for memory efficiency, only $\mathbf{G}_{cc}$ and $\mathbf{G}_{cn}$ are stored as dense matrices, whereas for $\mathbf{G}_{nn}$, only its diagonals ($\mathbf{g}_n$) are stored. In the blending proposed by Masuda et al. (2016), all the columns of $\mathbf{A}_{22}$ are calculated, but only the elements corresponding to $\mathbf{G}_{cc}$, $\mathbf{G}_{cn}$, and $\mathbf{g}_n$ are added to these matrices. The rest of the elements are used for calculating the average of the diagonal and off-diagonal elements of $\mathbf{A}_{22}$, which are required for the tuning of $\mathbf{G}$. Since the only elements of $\mathbf{A}_{22}$ needed are those corresponding to $\mathbf{G}_{cc}$, $\mathbf{G}_{cn}$, and $\mathbf{g}_n$, we propose an optimized algorithm that computes only the rows of $\mathbf{A}_{22}$ corresponding to the core animals instead of calculating all the columns of $\mathbf{A}_{22}$. The elements of $\mathbf{A}_{22}$ corresponding to $\mathbf{g}_n$ are the inbreeding coefficients added to the value of one for the non-core animals. They are

calculated before computing any row of $\mathbf{A}_{22}$ because the method for calculating these rows requires the inbreeding coefficients (Colleau, 2002).

The primary purpose of blending **G** with $\mathbf{A}_{22}$ is to make **G** nonsingular, which is attainable with **I**. Blending with $\mathbf{A}_{22}$ may cause unwanted bias, and blending with **I** should require less computing time. To evaluate this, we compared the reliability and inflation of GEBV, number of rounds required for convergence, and elapsed wall-clock time for blending when blending **G** with various proportions (0.30, 0.20, 0.05, 0.01, 0.005, 0.001) of $\mathbf{A}_{22}$ and **I**. A U.S. Holstein dataset provided by The Council on Dairy Cattle Breeding (Bowie, MD) was used in this study. Stature phenotypes were available from 10,067,745 animals. The pedigree file included 9,730,943 animals, from which 569,404 animals had 60,671 SNP markers after quality control. SNP with minor allele frequency lower than 0.05, call rates lower than 0.9, or a difference greater than 0.15 between expected and observed frequency of heterozygous were removed during quality control. Animals with call rates lower than 0.9 or parent-progeny Mendelian conflicts were removed during quality control. Of the genotyped animals included after quality control, 21,127 were sires, 59,723 were dams, and 32,855 had stature phenotypes. For APY, 15,000 genotyped animals were randomly chosen as core.

A partial dataset was created for validation by removing phenotypes from daughters of bulls that have at least 50 daughters with records in 2014. Genomic EBV (GEBV) were calculated for the whole ($GEBV_w$) and partial ($GEBV_p$) datasets using the BLUP90IOD2OMP1 software (version 3.119; (Tsuruta et al., 2001, Tsuruta and Misztal, 2008). Estimates of daughter yield deviations (DYD) for validation bulls were obtained using the whole dataset with the method of Liu et al. (2004) and the algorithm of Mrode and Swanson (2004). The regression coefficient (b_1) and the coefficient of determination (r^2) between DYD and $GEBV_p$_x were used to measure the

inflation and reliability of predictions for validation bulls, respectively, where x denotes the blending combination tested (0.30, 0.20, 0.05, 0.01, 0.005, or 0.001 multiplied by $\mathbf{A}_{22}$ or **I**).

The results for the validation are shown in Figure 3.1. None of the b_1 nor r^2 values between DYD and $GEBV_p$_x differ by more than 0.05, indicating consistent outputs of the models. The lowest r^2 values were seen with the blending coefficient of 0.30 for both matrices tested. To compare the GEBV of the various blending combinations to the current blending default in the BLUPF90 programs ($0.05\mathbf{A}_{22}$), $GEBV_w$_$0.05\mathbf{A}_{22}$ was regressed on $GEBV_w$_x, and the correlation coefficients (r) and b_1 of the two were calculated for the genotyped animals and are shown in Figure 3.2. When comparing with the same blending proportion, there were negligible differences between b_1 and r for **I** and $\mathbf{A}_{22}$. For $GEBV_w$_$0.05\mathbf{A}_{22}$ regressed on $GEBV_w$ of 0.30, 0.20, 0.01, 0.005, or 0.001 $\mathbf{A}_{22}$ blending combination, r and b_1 ranged from 0.99 to 0.98 and 0.99 to 0.97, respectively, indicating very little inflation and strong association. For $GEBV_w$_$0.05\mathbf{A}_{22}$ regressed on $GEBV_w$ of each **I** blending combination, r was 0.99 and b_1 ranged from 0.98 to 0.97. The differences observed here are negligible and suggest no differences in reliability or inflation of GEBV when blending **G** with a small value multiplied by **I** compared with $\mathbf{A}_{22}$.

The number of rounds until convergence using preconditioner conjugate gradient with iteration on data (Tsuruta et al., 2001) were compared for each blending combination to quantify the computational efforts and are shown in Table 3.1. The termination criterion was 10^{-12} with a convergence criterion of $C = \|\mathbf{b} - \mathbf{Ax}\|^2/\|\mathbf{b}\|^2$, where the mixed model equations are $\mathbf{Ax} = \mathbf{b}$. For every blending combination, the convergence patterns were steady, and there was no indication of divergence. The $0.001\mathbf{A}_{22}$ and $0.001\mathbf{I}$ blending combinations had the most iterative rounds to convergence (599 and 596, respectively). The blending scenario of $0.30\mathbf{I}$ took 251 rounds to converge, which was the fewest observed (Table 3.1). The fewer rounds necessary for convergence

suggest a more well-conditioned system of equations. However, the results do not indicate a concerningly high number of rounds or a diverging system for any blending combinations tested.

One would assume blending with **I** would drastically reduce computing time since **I** is sparse, and the direct creation of $\mathbf{A}_{22}$ would be avoided. Using the algorithm by Masuda et al. (2016), the elapsed wall-clock time for creating and blending $0.05\mathbf{A}_{22}$ and $0.05\mathbf{I}$ with $0.95\mathbf{G}$ were both around 1 hour and 54 minutes, with no notable difference (Table 3.1). This lack of difference in computing time between blending with the two matrices can be attributed to the algorithm in Masuda et al. (2016); as mentioned above, all columns of $\mathbf{A}_{22}$ were calculated for the tuning of **G**, regardless of the matrix used for blending. In contrast, only the rows of $\mathbf{A}_{22}$ relating to core animals were calculated in the optimized algorithm, which reduced computing time. With the optimized blending algorithm, blending $0.95\mathbf{G}$ with $0.05\mathbf{A}_{22}$ took 5 minutes, and with $0.05\mathbf{I}$, it took less than one second. Blending **G** for large-scale evaluations is efficient with the optimized algorithm using either $\mathbf{A}_{22}$ or **I**. Although blending with **I** is remarkably faster than with $\mathbf{A}_{22}$ using the new algorithm, an elapsed computing time of approximately five minutes is not critical. Additionally, the peak memory use was 78.57GB and did not differ between models.

The new algorithm can include virtually any number of genotyped animals in the genomic set-up for ssGBLUP with APY. Using $0.001\mathbf{I}$ for blending is enough for inverting and has no negative consequences on reliability and inflation. Moreover, the weight applied to the blending matrix should be determined by the portion of the genetic variance the markers explain. The results may differ depending on the dataset and values of β used. For each blending proportion tested, **I** had fewer convergence rounds than $\mathbf{A}_{22}$, and the least was with $0.30\mathbf{I}$. The differences in reliability and inflation of GEBV when blending **G** with various proportions of $\mathbf{A}_{22}$ and **I** were negligible, and the computing time is no longer a limiting factor with the new algorithm. Therefore, the

decision of which matrix to use to ensure the non-singularity of **G** is trivial for the implementation of ssGBLUP.

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 score1. Journal of Dairy Science 93(2):743-752.

Christensen, O. F. and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genetics Selection Evolution 42(1):2.

Colleau, J.-J. 2002. An indirect approach to the extensive calculation of relationship coefficients. Genetics Selection Evolution 34(4):1-13.

Council of Dairy Cattle Breeding. 2022. Counts of Genotyped Animals by Country Code.

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.

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.

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

Gonzalez-Peña, D., N. Vukasinovic, J. J. Brooker, C. A. Przybyla, and S. K. DeNise. 2019. Genomic evaluation for calf wellness traits in Holstein cattle. Journal of Dairy Science 102(3):2319-2329.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177(4):2389-2397.

Henderson, C. R. 1976. A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. Biometrics 32(1):69-83.

Himmelbauer, J., Schwarzenbacher, H., Fuerst, C. 2021. Implementation of single-step evaluations for fitness traits in the German and Austrian Fleckvieh and Brown Swiss populations. Interbull Bulletin 56, 82-89.

Liu, Z., M. E. Goddard, B. J. Hayes, F. Reinhardt, and R. Reents. 2016. Technical note: Equivalent genomic models with a residual polygenic effect. Journal of Dairy Science 99(3):2016-2025.

Liu, Z., M. E. Goddard, F. Reinhardt, and R. Reents. 2014. A single-step genomic model with direct estimation of marker effects. Journal of Dairy Science 97(9):5833-5850.

Liu, Z., F. Reinhardt, A. Bünger, and R. Reents. 2004. Derivation and Calculation of Approximate Reliabilities and Daughter Yield-Deviations of a Random Regression Test-Day Model for Genetic Evaluation of Dairy Cattle. Journal of Dairy Science 87(6):1896-1907.

Lourenco, D., A. Legarra, S. Tsuruta, Y. Masuda, I. Aguilar, and I. Misztal. 2020. Single-Step Genomic Evaluations from Theory to Practice: Using SNP Chips and Sequence Data in BLUPF90. Genes 11(7):790.

Mäntysaari, E. A., R. D. Evans, and I. Strandén. 2017. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals1. Journal of Animal Science 95(11):4728-4737.

Masuda, Y., I. Misztal, S. Tsuruta, A. Legarra, I. Aguilar, D. Lourenco, B. Fragomeni, and T. 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.

McWhorter, T. M., A. Garcia, M. Bermann, A. Legarra, I. Aguilar, I. Misztal, and D. Lourenco. 2021. 36 Effect of Blending and Tuning Relationship Matrices in Single-step Genomic BLUP. Journal of Animal Science 99(Supplement_3):19-20.

Meyer, K., B. Tier, and A. Swan. 2018. Estimates of genetic trend for single-step genomic evaluations. Genetics Selection Evolution 50(1):39.

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.

Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. J Dairy Sci 97(6):3943-3952.

Mrode, R. A. and G. J. T. Swanson. 2004. Calculating cow and daughter yield deviations and partitioning of genetic evaluations under a random regression model. Livestock Production Science 86(1):253-260.

Nilforooshan, M. A. and M. Lee. 2019. The quality of the algorithm for proven and young with various sets of core animals in a multibreed sheep population1. Journal of Animal Science 97(3):1090-1100.

Ødegård, J., U. Indahl, I. Strandén, and T. H. E. Meuwissen. 2018. Large-scale genomic prediction using singular value decomposition of the genotype matrix. Genetics Selection Evolution 50(1):6.

Ostersen, T., O. F. Christensen, P. Madsen, and M. Henryon. 2016. Sparse single-step method for genomic evaluation in pigs. Genetics Selection Evolution 48(1):48.

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(1):82.

Quaas, R. L. 1976. Computing the Diagonal Elements and Inverse of a Large Numerator Relationship Matrix. Biometrics 32(4):949-953.

Quaas, R. L. 1988. Additive Genetic Model with Groups and Relationships. Journal of Dairy Science 71:91-98.

Taskinen, M., E. A. Mäntysaari, and I. Strandén. 2017. Single-step SNP-BLUP with on-the-fly imputed genotypes and residual polygenic effects. Genetics Selection Evolution 49(1):36.

Tsuruta, S. and I. Misztal. 2008. Technical note: Computing options for genetic evaluation with a large number of genetic markers. Journal of Animal Science 86(7):1514-1518.

Tsuruta, S., I. Misztal, and I. Strandén. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications1. Journal of Animal Science 79(5):1166-1172.

VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91(11):4414-4423.

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.

TABLES

Table 3.1. The elapsed wall-clock time for blending in minutes and the number of rounds to reach the convergence criterion of 10^{-12} for obtaining the solutions of the system of equations for each blending scenario with the new and old algorithm. The blending scenarios are the matrices added to $(1 - \beta^1)\mathbf{G}^2$ to obtain an invertible **G**.

Blending Scenario	Algorithm	Elapsed wall-clock time for blending (min)	Rounds to Convergence[3]
$0.30\mathbf{A}_{22}$[4]	new	2.4	333
	old	54.6	333
$0.20\mathbf{A}_{22}$	new	2.9	349
	old	66.0	346
$0.05\mathbf{A}_{22}$	new	5.2	416
	old	113.6	418
$0.01\mathbf{A}_{22}$	new	2.5	487
	old	61.0	487
$0.005\mathbf{A}_{22}$	new	2.3	536
	old	98.7	534
$0.001\mathbf{A}_{22}$	new	3.9	599
	old	98.0	594
$0.30\mathbf{I}$[5]	new	2.6	251
	old	62.2	251
$0.20\mathbf{I}$	new	3.0	257
	old	61.9	257
$0.05\mathbf{I}$	new	<0.1	362
	old	113.5	362
$0.01\mathbf{I}$	new	<0.1	470
	old	111.3	468
$0.005\mathbf{I}$	new	<0.1	530
	old	77.2	532
$0.001\mathbf{I}$	new	<0.1	596
	old	76.0	598

[1]The proportion of residual polygenic variance not accounted for by **G**

[2]Genomic relationship matrix

[3]The average elapsed wall-clock time per iteration round was 7.3 seconds for all scenarios

[4]Numerator relationship matrix for genotyped animals

[5]Identity matrix

FIGURES

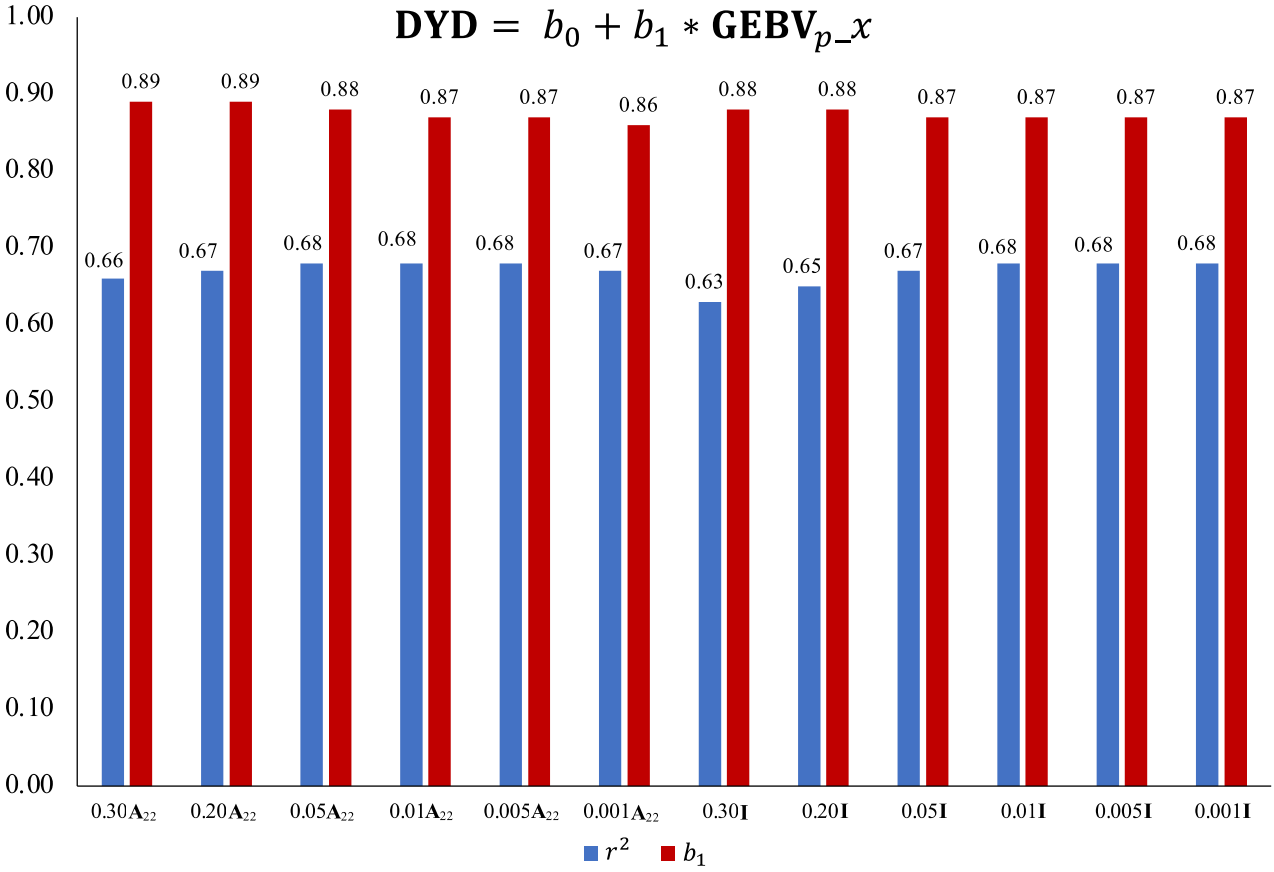

Figure 3.1. Regression coefficient (b_1) and coefficient of determination (r^2) of DYD on GEBV calculated with a partial dataset ($GEBV_p$) with all $GEBV_p\_x$ blending scenarios, where x denotes the blending combination tested, as shown on the x-axis.

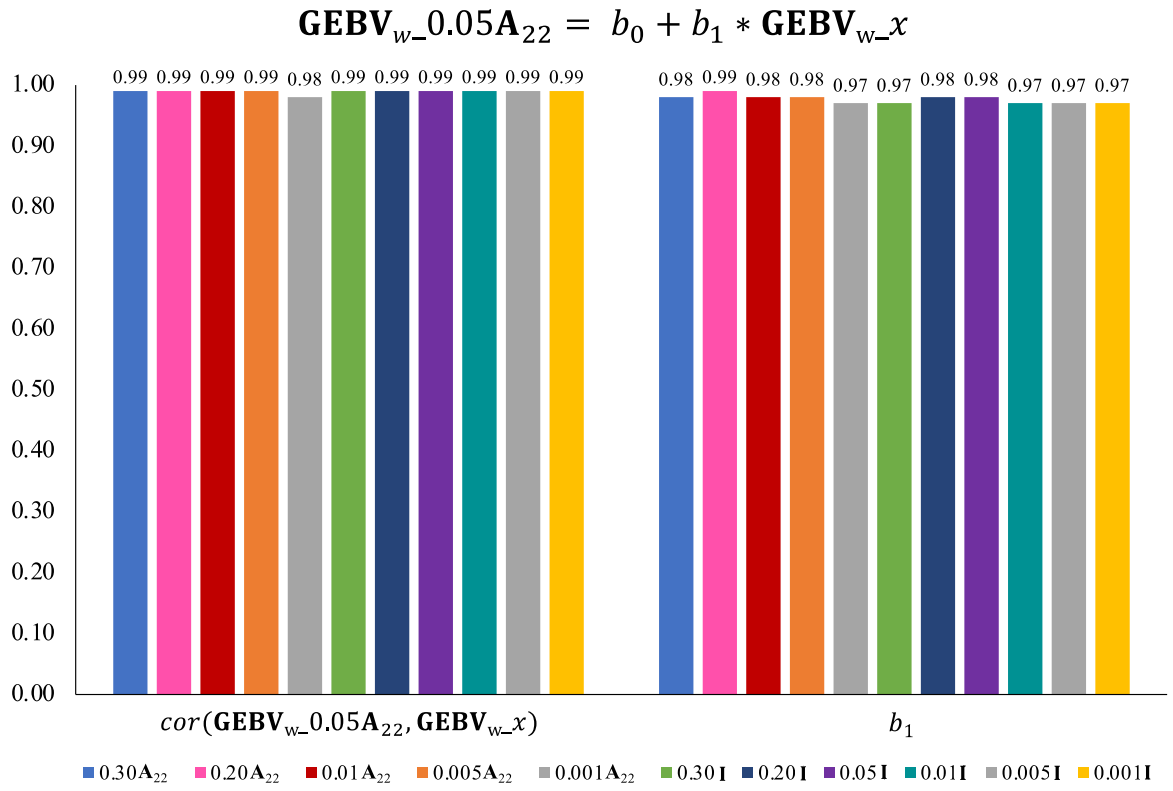

Figure 3.2. Regression coefficient (b_1) and correlation coefficient (r) of $\mathrm{GEBV}_{\mathrm{w}}$ using $0.05\mathbf{A}_{22}$ blending on $\mathrm{GEBV}_{\mathrm{w}}\_\mathrm{x}$, where x is the blending combination tested as shown in the legend, and w refers to the whole dataset.

CHAPTER 4

EXPLORING THE STATISTICAL NATURE OF INDEPENDENT CHROMOSOME SEGMENTS[2]

[2] Hollifield, M K., M. Bermann, D. Lourenco, and I. Misztal. 2023. *Livestock Science*. 270 (105207). Reprinted here with permission of the publisher.

ABSTRACT

Previous research in genomics developed the concept of independent chromosome segments and suggested that genomic prediction works on capturing the segment effects rather than on LD to QTL. The number of independent chromosome segments (M_e) was posited as $4N_eL$, a function of effective population size (N_e) and genome length (L) in Morgans (M). The objective of this study was to determine whether the physical segments are approximately consecutive haplotype blocks of length ¼ M, with the number of haplotype blocks for each physical location equal to N_e. In a simulated population, the number of animals randomly selected as reference animals is represented as N_a. For all animals, the genome was split into equal-sized segments represented as N_s. For each specific location, segments of non-reference animals were assigned the most similar segment of one reference animal. Genomic analyses estimated the value of the segment effects, and breeding values were a sum of all segment effects for a specific animal. Accuracies of segment effects were calculated by correlating the true breeding values (TBV) and the breeding values based on segments. Segment effect accuracies were compared with the true accuracy calculated by the correlation of TBV and genomic estimated breeding values (GEBV) computed using GBLUP. Accuracies were maximized at $N_a=N_e$, $N_s=4L$, but they were not as high as in GBLUP. Accuracies may be smaller using the statistical concept of segments due to approximations based on computing limitations, as the origin of each segment and the recombination sites were unknown in the simulation. Therefore, random animals served as reference animals, and each non-reference animal received the most similar segment of a reference animal instead of a linear combination of such segments. Genomic selection acts partially on ¼ M

long chromosome segments, and using the statistical definition of segments moderately explains the accuracy. More complex simulations are needed to investigate the issue thoroughly.

INTRODUCTION

The genome is inherited in large segments from each parent and is a mosaic of ancestral segments broken in each generation (Pääbo, 2003). In any population, an arbitrary past generation can be designated as the "founder" generation, and each chromosome in subsequent generations is a composite of segments that originated from this generation (MacLeod, 2005). Alleles are strongly associated with each other on the portions of segments inherited unbroken from the founder population. Thus, the potential for selection on segments exists if enough information is present to explain the independent chromosome segments (Pocrnic et al., 2019). According to Stam (1980), the number of independent chromosome segments (M_e) is based on the number of chromosome junctions in a finite population and is equal to $4N_eL$, where N_e is the effective population size, and L is genome length in Morgans (M). Goddard (2009) estimated M_e as $2N_eL/\log(4N_eL)$ due to the variation in relationships in the population, and Hayes et al. (2009) used $2N_eL$, an approximation between Stam (1980) and Goddard (2009).

Genomic selection works by estimating independent chromosome segment effects (Goddard, 2009). Experimentally, M_e has been related to the number of the largest eigenvalues that explain 98−99% of the genetic variation in the genomic relationship matrix (**G**), as the dimensionality of genomic information is approximately $4N_eL$ (Pocrnic et al., 2016a). Misztal et al. (2014a) proposed the Algorithm of Proven and Young (APY), an efficient computation of the inverse of **G** by using recursion on a small subset of animals, which serve as holders of linear combinations of independent chromosome segments. Initially, the core subset of animals used for

APY were those with high accuracy or "proven" with many records. However, further studies have shown that predictions are accurate when M_e random animals are used as the core subset for APY (Fragomeni et al., 2015; Pocrnic et al., 2016a; Bradford et al., 2017). The success of the APY algorithm affirms the validity of the segment model.

While the concept of the independent chromosome segments seems valid, the physical nature of the segments is not known. However, it can be posited as one of two concepts. In the first concept, the genome is composed of approximately ¼ M consecutive blocks, and each block holds N_e different segments (or haplotypes). Each segment is associated with an additive value, and a breeding value is a sum of the additive values of the segments. In the second concept, each founder animal contributes approximately 1M blocks to progeny, and after several generations, approximately 4NeL segments are formed; they can be of different sizes per population. Like before, each segment is associated with an additive value, and the value of a segment is a weighted sum of its components. MacLeod et al. (2005) looked at transition points between consecutive segments, suggesting that the segments form a continuous block but not necessarily a fixed size across the genome; about 12 SNP per one chromosome segment was required to detect 90% of transitions.

The objective of this study was to investigate the statistical nature of independent chromosome segments using the first concept since the second one is difficult to implement as it is uncertain how to explicitly identify independent segments after many generations. We tested if all additive genetic information can be composed in 4NeL independent chromosome segments, if equal-sized blocks well represent physical aspects of chromosome segments, and if the accuracy of segment effects is close to the true accuracy.

MATERIALS AND METHODS

DATA SIMULATION

Data were simulated using QMSim software and replicated five times (Sargolzaei and Schenkel, 2009). A historical population undergoing drift and mutation was generated and consisted of 1,000 generations, gradually increasing from 1,000 to 100,000 individuals with random mating, no selection, and no migration. From the last generation of the historical population, five males and 2,000 females were randomly sampled to create a recent population with an effective population size of 20. Ten non-overlapping generations of the recent population were created by random mating, no selection, equal sex ratio, and a litter size of one for a total population size of 20,005, with 8,966 males and 11,039 females. One polygenic trait was simulated for each animal with a phenotypic variance of 1.0 and heritability of 0.3. The last three generations were genotyped, and each genome contained ten, one M long chromosomes, for a total genome length of ten M. Each chromosome had 5,000 evenly allocated biallelic SNP markers, totaling 50,000 SNP per genome. Randomly distributed among the chromosomes were 1,660 QTL affecting the trait that were biallelic and had equal frequencies in the first generation with allelic effects sampled from a gamma distribution with a shape parameter of 0.4. The mutation rate per locus per generation was assumed to be 2.5×10^{-5} for QTL and markers with recurrence.

MODELS AND COMPUTATIONS

A group of animals was selected to generate conceptual independent chromosome segments. Hereinafter, these animals will be referred to as “reference” animals. The reference animals were selected from the genotyped animals in the last three generations. Each animal’s genome, composed of 50,000 SNP, was split into discrete segments. To test the effects of the

theoretical $4N_eL$ independent chromosome segments, N_e (20) reference animals and 4L (40) segments per animal were analyzed. We conducted two main analyses: *i)* reference animals were selected randomly or by the least related animals in the population, and *ii)* various numbers of segments and reference animals were compared.

The least related reference animals were those with the maximum difference in SNP codes compared to all other animals and were determined by $max(\sum\sum|q_{rs} - q_{ts}|)$ where r is the animal with the maximum sum of SNP codes (q) compared to the rest of the genotyped population (t), and s is the allele. A random number generator determined the animals chosen as the random reference animals. All reference and non-reference animals were genotyped, and only the genotyped animals were used in the analyses. To investigate the behavior around $4N_eL$ segment effects, the number of segments per animal (N_s) varied while the number of reference animals (N_a) remained fixed to N_e (20). Similarly, N_s was fixed to 4L (40) as N_a varied. The actual effective population size and genome length in the simulation did not change; only the number of artificially created segments changed to estimate the accuracy of segment effects. Each segment of each non-reference animal was assigned to a reference animal segment that was the most similar. Only the effects of the reference animal's chromosome segments were tested to evaluate independent chromosome segment effects. The chromosome segments of each animal were denoted as c_{ij}, where c is the j^{th} segment of the i^{th} animal. Altogether, there were N_aN_s segment effects total, and the ij^{th} segment effect was numbered as $(i-1)N_a + j$. The similarities per non-reference segment were calculated by $min(\sum|c_{kj} - c_{lj}|)$, where k is the reference animal with the minimum sum of the SNP codes in segment j for non-reference animal l. The effects of the hypothetical independent chromosome segments were estimated using in-house software written in Fortran with the model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z_s}\boldsymbol{s} + \boldsymbol{e}; \quad (1)$$

where $\mathbf{y}$ is a vector of phenotypes, $\mathbf{Z_s}$ is an incidence matrix relating the assigned reference animal segments to all animals, $\boldsymbol{s}$ is a random vector of estimated segment effects, and $\boldsymbol{e}$ is a random vector of errors. We assumed the variances were var($\boldsymbol{s}$) = 0.3$\mathbf{I}$ and var($\boldsymbol{e}$) = 0.7$\mathbf{I}$.

Accuracies were computed to investigate the theoretical number of $4N_eL$ independent chromosome segments containing all genetic information in a population. Accuracies were calculated for animals in the last generation by $cor(\mathbf{TBV}, \mathbf{Z_s}\hat{\mathbf{s}})$, where $\mathbf{Z_s}\hat{\mathbf{s}}$ is a vector of breeding values based on segments, and **TBV** is a vector of the true breeding values outputted by QMSim. Accuracies of segment effects were compared with the true accuracy (i.e., $cor(\mathbf{TBV}, \mathbf{GEBV})$). GEBV were computed in the BLUPF90 software suite (Misztal et al., 2014b) using the GBLUP model:

$$\mathbf{y} = \mathbf{1}\mu + \boldsymbol{u} + \boldsymbol{e}, \quad (2)$$

in which $\boldsymbol{u}$ is a random vector of breeding values, and $\mathbf{y}$ and $\boldsymbol{e}$ are defined as in the previous model. The variances were assumed to be var($\boldsymbol{u}$) = 0.3$\mathbf{G}$ and var($\boldsymbol{e}$) = 0.7$\mathbf{I}$. The coefficients used for the variances were the same as the variance parameters used for the simulation. A standard genomic relationship matrix was constructed for $\mathbf{G}$ using the first method of (VanRaden, 2008):

$$\mathbf{G} = \frac{\mathbf{W}\mathbf{W}'}{2\sum p_j(1-p_j)} \quad (3)$$

where $\mathbf{W}$ is a centered matrix of gene content and p_j is the allele frequency of marker j.

RESULTS AND DISCUSSION

Accuracies of estimated chromosome segment effects with random and the least related animals as reference animals are shown in Figures 4.1 and 4.2. In Figure 4.1, N_a is fixed to N_e (20), and in Figure 4.2, N_s is fixed to 4L (40). It was hypothesized that accuracy is maximized if enough

information is available to explain $4N_eL$ independent chromosome segments, which aligns with the theory that there is no new genetic information if all $4N_eL$ independent chromosome segments are explained (Pocrnic et al., 2019).

The purple lines in Figures 4.1 and 4.2 indicate where $4N_eL$ chromosome segment effects are estimated and are close to where accuracy is maximized and begins to plateau or decrease. The maximum accuracy of segment effects was 0.69 for both random and least related reference animals, and the accuracy of GBLUP was 0.90. The results suggest that 4L segments of N_e reference animals contain all additive information in the population. However, the accuracy of GBLUP is greater than that of the chromosome segment effects, suggesting that the chromosome segments do not account for all genetic variation as they were calculated. It was presumed that if the reference animals were the most unrelated to each other, accuracy would be greater since more of the variation in the population would be captured and more ancestral segments would be accounted for than randomly chosen reference animals. Similar to the initial proposition of APY, the core was defined as the "proven" animals with many records and progeny, hence the animals with the most information and greatest genetic impact on the population (Misztal et al., 2014a). Core animals have since been redefined, and random animals are appropriate to use as core and produce the same accuracy as the previous definition (Fragomeni et al., 2015).

To better represent the mosaic nature of segments, segments of non-reference animals in this study could have been composed of inherited segments of true founder animals. In this way, the contributing segments from the true founders and recombination sites can be traced throughout the successive generations. Breeding values would be a weighted sum of the combination of founder segments. The chromosome segments are held in haplotype blocks and result from recombination. In each generation, the blocks become more mosaic but retain the unbroken

independent chromosome segments. In Meuwissen et al. (2001), haplotypes were defined as 1 cM regions in the genome and using BLUP, the accuracy of these haplotype effects was 0.73. However, only a few major genes were simulated to explain the additive genetic variance of the trait, unlike the polygenic traits of interest in commercial livestock populations.

Pocrnic et al. (2016a) showed that approximately N_eL largest eigenvalues explain most variation in the **G** matrix, and $4N_eL$ largest eigenvalues were the maximum that could explain more variation. Likewise, in Figures 4.1 and 4.2, the accuracies did not increase beyond $4N_eL$ segments. Segment accuracies did not reach the true accuracy (0.9) for $4N_eL$ segments for both random and least related reference animals, suggesting the construction of the segments in this study did not align with biological properties. Stam (1980) derived the pdf of segment length as $f(x) = \frac{8N_e}{(1+4N_e x)^3}$ with the mean equal to $1/(4N_e)$. The segments in this study remained equal in length for every segment effect estimation, inferring that the largest segments were potentially not captured. Ferdosi et al. (2016) compared the optimal haplotype length for three different methods of building relationship matrices and for three traits. The three methods for creating haplotype relationship matrices were distinct windows (non-overlapping), sliding windows (overlapping), and total minimum similarity (counting the number of identical SNP in contiguous segments). For all three methods, using haplotypes of one-SNP was proved to be the same as the **G** matrix of VanRaden (2008), and using haplotypes of length greater than one segment improved the accuracy and peaked at an optimum length. The optimum length for all three methods ranged from 7-40 SNP for scrotal circumference, 7-11 for age at puberty, and 2-11 for weight at first corpus luteum. It is expected that haplotype length differs by trait due to the assumption that groups of SNP formed into haplotypes are in LD with QTL, which affects the trait of interest (Calus et al., 2008). Similarly, our results presented greater accuracies for segment length greater than one SNP but peaked or

plateaued as the size increased. Assuming 50k equidistant SNP and a genome length of 30 M, the ¼ M blocks would each be about 400 SNP long. Thus, in our model, the segments were much longer than haplotypes.

Suppose segment size is directly proportional to the amount of variance explained. In that case, a small number of the largest segments may explain the majority of the variance, while small segments or individual SNP explain the remaining variance, analogous to the distribution of eigenvalues. Pocrnic et al. (2016b) found that realized accuracies peaked with APY when the number of core animals was equal to the number of eigenvalues that explained 99% of the variation in **G**. The optimal number of core animals is about 14,000 for Holstein and Angus cattle, 12,000 for Jersey cattle, and 6,000 for broiler chickens and pigs. Considering N_e and L for each respective species, the optimal number of core animals corresponded to approximately $4N_eL$, aligning with previous literature on independent chromosome segments and redundancy in the **G** matrix (Stam, 1980; Daetwyler et al., 2008; Goddard, 2009; Hayes et al., 2009).

It is expected that if nearly all variation in **G** can be explained by the independent chromosome segments, then breeding values of n animals are linear functions of these segments. As aforementioned, DNA is inherited from ancestors as contiguous blocks formed by recombination throughout the evolution of the population (Pääbo, 2003). Biologically, in violation of present concepts of physical chromosome segments, this study used constant segment size, and recombination locations were not considered. Edriss et al. (2013) found a slight increase in accuracy when fitting haplotypes based on genealogical trees compared to fitting individual SNP, suggesting a better prediction when accounting for recombination. Other studies have shown that the accuracy of genomic predictions increase as the genetic relationships are better represented by the markers (Habier et al., 2007; Habier et al., 2010). MacLeod (2005) provides a comprehensive

description of chromosome segments and their role in inbred populations; the population used in this study is not inbred. The results presented indicate some agreement with expectation, but it is not complete; perhaps accounting for recombination will better follow the physical properties of segments. Add something here about following segments down generations and using the 4NeL most occurring segments in the population.

CONCLUSION

Approximately $4N_eL$ chromosome segments contain almost all the genetic information that can be explained in a population. It is important to notice that M_e is a statistical concept, and how those M_e segments are physically organized in haplotype blocks depends on linkage disequilibrium. Haplotype blocks become a mosaic of founder animals as generations proceed; however, further investigation is needed on the definition and characteristics of physical chromosome segments. If each segment has an additive value and can be ranked, genomic predictions can potentially be based on weighted averages of the segments.

REFERENCES

Bradford, H., I. Pocrnić, B. Fragomeni, D. 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.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics 178(1):553-561. doi: 10.1534/genetics.107.080838

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLOS ONE 3(10):e3395. doi: 10.1371/journal.pone.0003395

Edriss, V., R. L. Fernando, G. Su, M. S. Lund, and B. Guldbrandtsen. 2013. The effect of using genealogy-based haplotypes for genomic prediction. Genetics Selection Evolution 45(1):5. doi: 10.1186/1297-9686-45-5

Ferdosi, M. H., J. Henshall, and B. Tier. 2016. Study of the optimum haplotype length to build genomic relationship matrices. Genetics Selection Evolution 48(1):75. doi: 10.1186/s12711-016-0253-6

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

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136(2):245-257.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177(4):2389-2397. doi: 10.1534/genetics.107.081190

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

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

MacLeod, A. K. 2005. Detecting ancestral junctions in inbred populations, University of Edinburgh.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157(4):1819-1829. doi: 10.1093/genetics/157.4.1819

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

Misztal, I., S. Tsuruta, D. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. G. Vitezica. 2014b. Manual for BLUPF90 family of programs. http://nce.ads.uga.edu/wiki/doku.php?id=application_programs.

Pääbo, S. 2003. The mosaic that is our genome. Nature 421(6921):409-412. doi: 10.1038/nature01400

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: 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(1):82. doi: 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: 10.1186/s12711-019-0516-0

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

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

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

FIGURES

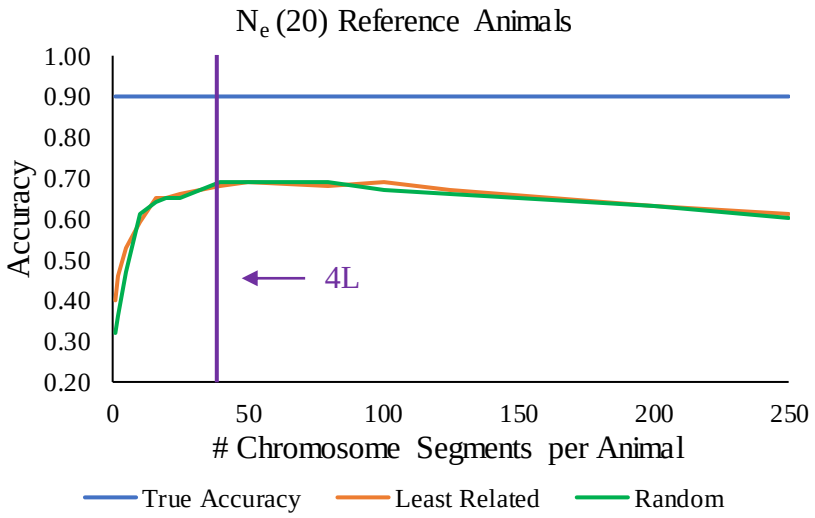

Figure 4.1. Accuracies of chromosome segment effects with random and least related reference animals. The number of reference animals is fixed to N_e (20). Accuracies of chromosome segment effects were calculated by correlating the TBV and the estimated segment effect. The true accuracy was calculated by correlating the TBV and GEBV. The purple line represents $4N_eL$ independent chromosome segments.

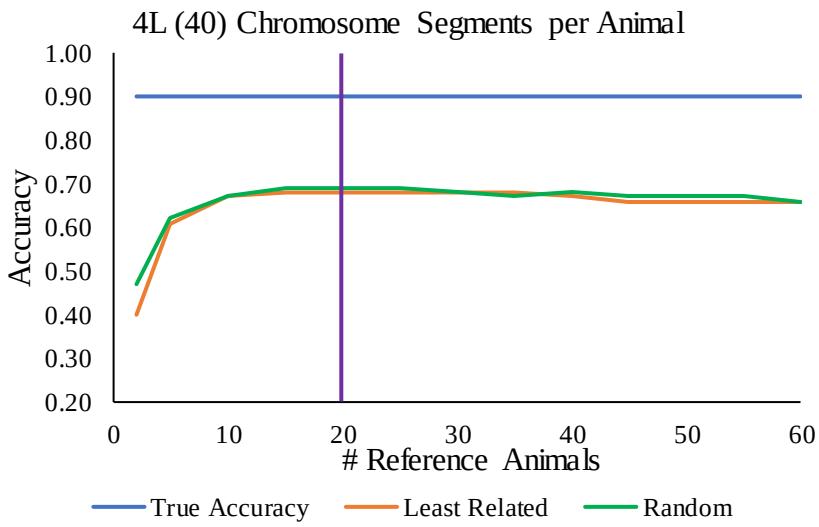

Figure 4.2. The number of chromosome segments per animal is fixed to 4L (40). The methods and components are the same as in Figure 4.1.

CHAPTER 5

ESTIMATION OF HERITABILITY WITH GENOMIC INFORMATION BY METHOD R[3]

[3] Hollifield, M. K., D. Lourenco, I. Misztal. 2024. *Journal of Animal Breeding and Genetics*. 00:1-9. Reprinted here with permission of the publisher.

ABSTRACT

Estimating heritabilities with large genomic models by established methods such as restricted maximum likelihood (REML) or Bayesian via Gibbs sampling is computationally expensive. Alternatively, heritability can be estimated indirectly by method R and by maximum predictivity, referred to as MaxPred here, at a much lower computing cost. By method R, the heritability used for predictions with whole and partial data is considered the best estimate when the predictions based on partial data are unbiased relative to those with the complete data. By MaxPred, the heritability estimate is the one that maximizes predictivity. This study aimed to compare heritability estimation with genomic information using average information REML (AI-REML), method R, and MaxPred. A simulated population was generated with ten generations of 5,000 animals each and an effective population size of 80. Each animal had one record for a trait with a heritability of 0.3, a phenotypic variance of 10.0, and was genotyped at 50k SNP. In method R, the heritability estimate is found when the expectation of a regression coefficient is equal to one. The regression is the EBV of selection candidates calculated with the whole dataset regressed on the EBV of selection candidates calculated from a partial dataset. In this study, we used the GBLUP framework, and therefore, GEBV were calculated. The partial dataset was created by removing the last generation of phenotypes. Predictivity was defined as the correlation between the adjusted phenotypes of the selection candidates and their GEBV calculated from the partial data. We estimated the heritability for populations that included between three and ten generations. In every scenario, predictivity increased as more data was used and was the highest at the simulated

heritability. However, the predictivity for all data subsets and all heritabilities compared did not differ more than 0.01, suggesting MaxPred is not the best indication for heritability estimation. For the whole dataset, the heritability was estimated as 0.30 ± 0.01, 0.26 ± 0.01, and 0.30 ± 0.04 for AI-REML without genomics, AI-REML with genomics, and method R with genomics, respectively. Heritability estimation with genomics by method R reduced timing by 83%, implying a reduction in computing time from 9.5 hours to 1.6 hours, on average, compared to AI-REML with genomics. Method R has the potential to estimate heritabilities with large genomic information at a low cost when many generations of animals are present; however, the standard error can be high when only a few iterations are used.

INTRODUCTION

Heritabilities change over time due to selection, and the variance components used in the genetic evaluations need to be updated periodically. The magnitude of changes depends on the selection intensity and available genetic variation (Falconer, 1996; Walsh and Lynch, 2018). Additionally, some changes may be due to resource allocation constraints (Rauw, 2008) that are difficult to model. With the decrease in genotyping costs and, thus, the increase in the number of genotyped animals, traditional methods for variance component estimation (VCE) are becoming too costly or impossible (Misztal et al., 2021). Under genomic selection, the selection intensity is greater, possibly causing more rapid changes in genetic parameters compared to non-genomic selection. For example, Hidalgo et al. (2020) found large changes in heritability and genetic correlation for some traits in a swine population a few years after genomic selection was implemented. Therefore, with the rapid change in genetic parameters due to selection and the increase in data size, a feasible method to estimate heritabilities is needed.

The most popular methods for estimating variance components are REML (Patterson and Thompson, 1971) and Bayesian via Gibbs sampling (Gianola and Fernando, 1986). They have good theoretical properties (Gianola et al., 2009) and, with sparse matrix implementation (Misztal, 2008), could support large models. REML, by Monte Carlo sampling, could support models with a large number of parameters (Matilainen et al., 2013). However, with genomic information, mixed model equations are no longer sparse as the genomic component creates a dense subblock (Misztal et al., 2021). For smaller datasets, computations were much faster compared with larger datasets and were facilitated by using a sparse matrix package that recognizes dense blocks in a matrix and processes them using parallel computing (Masuda et al., 2015). Since the inverse of the genomic relationship matrix can be approximated by a sparse matrix for populations with a small effective population size using the algorithm for proven and young (Misztal et al., 2014; $\mathbf{G}_{\mathrm{APY}}^{-1}$), Junqueira et al. (2022) attempted REML using $\mathbf{G}_{\mathrm{APY}}^{-1}$ in single-step GBLUP. However, no savings in computing resources were observed as the inverse of the pedigree relationship matrix for genotyped animals was dense since sparse representations such as APY do not work for this matrix.

One way to reduce computing costs when doing VCE in large, genotyped populations is to use a subset of individuals with genotypes. However, this can introduce bias due to data preselection if the animals are not randomly chosen for genotyping or phenotyping (Bussiman et al., 2023; Patry and Ducrocq, 2011). Consequently, this bias propagates to GEBV. Therefore, using the entire dataset can help avoid biases. Before genomics, one method used for heritability estimation of complete populations was method R (Reverter et al., 1994b). Method R depends on the property of mixed models, where the expectation of the regression coefficient of breeding values obtained with the complete data on breeding values obtained with the partial data equals 1 when the correct variance ratio is used. Because estimates of mixed model equations do not depend

on absolute variances but on variance ratios, method R is suitable for estimating heritabilities but not the actual variances. While models supported by method R were initially restricted to single-trait with one additive effect, formulas exist for models with correlated effects (Druet et al., 2001).

Method R was used successfully to estimate additive and dominance variance ratios, which require large data sets for small standard errors (Misztal, 1997). Estimation of variance ratios without genomic information was successful with the complete dairy, beef, and pig populations and for traits of low to high heritability (Culbertson et al., 1998; Gengler et al., 1998; Misztal et al., 1998). Because method R requires only an initial variance ratio value and the calculation of breeding values for partial and complete data sets, computations are manageable for any data size where the genetic evaluation is possible, including for the largest models. Likewise, method R should be applicable to genomic models. Some limitations of method R include its inability to estimate variances of the random effects separately from the other random effects, estimate genetic correlations between traits, and its lack of theoretical properties (Reverter et al., 1994a; Misztal, 1997). Therefore, method R would not be suitable for models with multiple correlated traits or multiple random effects. However, Druet et al. (2001) showed that estimating genetic covariances between direct and maternal effects is possible. Another method applicable to estimating heritabilities with large genomic datasets would be selecting variance components that maximize the predictivity, hereinafter denoted as MaxPred, where predictivity is the correlation between genomic estimated breeding values and adjusted phenotypes (Legarra et al., 2008). Finding such components would require a grid search and would be applicable to any data size. This study aimed to analyze heritability estimation by method R and MaxPred and compare the estimates to those obtained by average information REML (AI-REML) in the presence of genomic information.

MATERIALS AND METHODS

DATA SIMULATION

QMSim software (Sargolzaei and Schenkel, 2009) was used for data simulation. A historical population of 1,000 generations with a constant size of 50,000 animals was generated with random mating, no selection, and no migration. A bottleneck effect was simulated by reducing the population to 20,000 animals over 1,800 generations. From the last generation of the historical population, 20 unrelated males and 5,000 unrelated females were randomly selected as the breeding individuals for the first recent generation to create an effective population size equal to 80. A single phenotype was simulated for each animal for a trait with a QTL heritability of 0.3, no polygenic effect, an overall heritability of 0.3, and a phenotypic variance of 10. Ten generations were simulated with an effective population size constant at 80, litter size equal to one, equal probability of male or female progeny, random mating, replacement ratio of 0.9 for sires and 0.4 for dams, and the selection and culling decisions were based on high and low GEBV, respectively. GEBV were calculated externally using blup90iod3 (Misztal et al., 2014b) with a GBLUP model:

$$\mathbf{y} = \boldsymbol{u} + \boldsymbol{e}, \qquad [1]$$

where $\mathbf{y}$ is a vector of phenotypes, $\boldsymbol{u}$ is a random vector of breeding values, and $\boldsymbol{e}$ is a random vector of residuals. The blup90iod3 program uses an iteration on data method with a convergence criterion of 10^{-12}. The covariance matrices were assumed to be:

$$\text{Var}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G}\sigma_a^2 & 0\\ 0 & \mathbf{I}\sigma_e^2\end{bmatrix}, \qquad [2]$$

where σ_a^2 and σ_e^2 are variances for the additive genetic and residual effects, respectively, $\mathbf{I}$ is the identity matrix, and $\mathbf{G}$ is the genomic relationship matrix created using the first method of (VanRaden, 2008):

$$\mathbf{G} = \frac{\mathbf{WW}'}{2\sum p_j(1-p_j)}, \quad [3]$$

where $\mathbf{W} = \mathbf{M} - \mathbf{P}$, p_j is the allele frequency of marker j, calculated using observed allele frequencies, $\mathbf{M}$ relates individuals to marker genotypes, and $\mathbf{P}$ is a matrix containing columns of $2p_j$. To avoid singularity of $\mathbf{G}$, $\mathbf{G}$ was replaced with $\mathbf{G}^*$, where $\mathbf{G}^* = 0.95\mathbf{G} + 0.05\mathbf{I}$. For GEBV calculations for selection and culling purposes, σ_a^2 and σ_e^2 were assumed to be 3.0 and 7.0, respectively, as initialized in the simulation parameters.

All 50,020 animals in the population were genotyped, and the genome consisted of 29 chromosomes, each of length 1 Morgan, 52,896 biallelic evenly spaced markers, and 1,247 biallelic randomly spaced QTL. The allelic effects were sampled from a gamma distribution with shape parameter 0.4, and the mutation rate was recurrent per generation and assumed to be 2.5 x 10^{-5} per locus for QTL and markers. The simulation was replicated twelve times, the two replicates that deviated the most from the average phenotype and average variance components were discarded, and the remaining ten replicates were analyzed separately. Results are shown as an average of the ten replicates with their respective standard deviations.

MODELS AND COMPUTATIONS

To compare the heritability estimation by method R, variance components were estimated using the AI-REML algorithm implemented in blupf90+ (Misztal et al., 2014b) with and without genomic information. Three to ten consecutive generations beginning with the first generation were analyzed (i.e., 1-3, 1-4, …, 1-10), as Hollifield et al. (2021) showed insufficient accuracies with intervals of less than three generations. For VCE by method R and AI-REML with genomic information (AIGREML), a GBLUP-based model was assumed, and Henderson's mixed model equations were used to predict GEBV:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}, \quad [4]$$

where $\mathbf{y}$ and the covariance matrices are as mentioned above, the inverse of $\mathbf{G}$ is as constructed in equation 3, $\hat{\beta}$ is a scalar of the mean, $\hat{\boldsymbol{u}}$ is a random vector of additive genetic effects, λ is the variance ratio defined as $\frac{\sigma_e^2}{\sigma_a^2}$, and $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices relating elements of $\mathbf{y}$ to elements of $\hat{\beta}$ and $\hat{\boldsymbol{u}}$, respectively (Henderson, 1949; Henderson, 1950). Both $\mathbf{G}$ and $\mathbf{G}^{-1}$ were reconstructed for each data subset using observed allele frequencies. For AI-REML without genomic information (AIREML), a similar model was used, but $\mathbf{G}$ is replaced by $\mathbf{A}$, where $\mathbf{A}$ is the numerator relationship matrix of the genotyped animals.

Whole and partial datasets were created for validation purposes (i.e., to obtain predictivity and statistics from method R), and their solutions are denoted as $\hat{\boldsymbol{u}}_w$ and $\hat{\boldsymbol{u}}_p$, respectively. For every subset of data analyzed, the partial dataset was created by removing the most recent generation of data, and the validation animals were those that had their phenotypes removed in the partial dataset.

METHOD R

Reverter et al. (1994b) propose a technique to estimate heritability by regressing the EBV of validation animals obtained using the whole dataset on EBV of validation animals obtained with a dataset where the phenotypes of the validation animals are removed. With the correct heritability, the expectation of the regression is 1. Previous studies have compared other heritability estimation methods; however, genomics were not included (Misztal et al., 1997; Misztal, 1997; Druet et al., 2001; Duangjinda et al., 2001). In this study, the regression of breeding values obtained with the whole on breeding values obtained with the partial data was calculated as (adapted from Reverter et al., 1994a):

$$b_{w,p} = \frac{\widehat{\boldsymbol{u}}'_w \widehat{\boldsymbol{u}}_p}{\widehat{\boldsymbol{u}}'_p \widehat{\boldsymbol{u}}_p}. \quad [5]$$

Note that eq. [5] does not contain the relationship matrix as in Reverter et al. (1994a) $\left(b_{w,p} = \frac{\widehat{\boldsymbol{u}}'_w \mathbf{A}^{-1} \widehat{\boldsymbol{u}}_p}{\widehat{\boldsymbol{u}}'_p \mathbf{A}^{-1} \widehat{\boldsymbol{u}}_p} \right)$. This is because relationship information is included in the calculation of $\widehat{\boldsymbol{u}}_w$ and $\widehat{\boldsymbol{u}}_p$, and is not needed for the regression. Leaving the relationship matrices out does not change the statistics. This was also observed by Legarra & Reverter (2018) who developed the LR validation method based on Method R. Additionally, unnecessarily including $\mathbf{A^{-1}}$ or $\mathbf{G^{-1}}$ matrix in the numerator and denominator products to obtain $b_{w,p}$ increases computing time. Predictions $\widehat{\boldsymbol{u}}_w$ and $\widehat{\boldsymbol{u}}_p$ were calculated using blup90iod3 (Misztal et al., 2014b).

For method R, the variance ratios compared had numerators of 1.5, 2.0, 2.5, 3.0, 3.5, and 4.0. The same heritability was used to calculate $\widehat{\boldsymbol{u}}_w$ and $\widehat{\boldsymbol{u}}_p$, and the same $\mathbf{G}$ matrix as AIGREML was used. To keep the phenotypic variance constant for each variance ratio assessed, the variance ratio was calculated as $\lambda = \frac{\sigma_e^2 + (\sigma_a^2 - x)}{x}$, where σ_e^2 and σ_a^2 are the additive genetic and residual variances estimated by AIGREML by blupf90+ and x represents one of the values as mentioned above. Thus, the denominator of the variance ratio is one of the six values mentioned above, and the numerator is adjusted so that all variance ratio values maintain the same phenotypic variance as estimated by AIGREML. Using prior estimates from AIGREML to fix the phenotypic variance in the λ equation could create bias. Additional research should be conducted to compare method R with differing phenotypic variances. When two values of $b_{w,p}$ were found closest to and marginally above and below 1.0, a line was fit between these two values, and the estimated variance ratio was obtained when $b_{w,p}$ equals 1.0. The variance ratios were on the x-axis, and $b_{w,p}$

values were on the y-axis. Then, the x value where y is equal to 1 was found and used as the variance ratio estimate for method R.

We compared six different heritabilities to analyze the behavior of method R, where heritability was calculated as $h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$; however, six heritabilities are not required. One only needs to find where $b_{w,p}$ is marginally above and below 1. An algorithm could be implemented to iterate until a convergence criterion is met, potentially giving a more precise estimate. Method R without genomic information was not analyzed in this study. The σ_e^2 and σ_a^2 values used as inputs in the formula for λ could be obtained from AIREML without genomics or from previous estimates. In this study, variances estimated by AIGREML were used because they were obtainable with the data size. Pedigree-based estimates could provide an initial baseline for traits where a reasonable heritability estimate is unknown. The heritability estimates by method R will not be affected by the initial values used; however, more iterations may be needed to find values where $b_{w,p}$ is marginally above and below 1.

MAXIMUM PREDICTIVITY (MAXPRED)

Adjusted phenotypes ($\mathbf{y}_c$) were used for predictivity and were calculated using the solution for the mean from the model [4] with the whole data as $\mathbf{y}_c = \mathbf{y} - \mathbf{1}\hat{\mu}$. Predictivity was calculated for all generation intervals using heritabilities equal to 0.15, 0.20, 0.25, 0.30, 0.35, and 0.40. For validation, the predictivity was calculated for the validation animals by $r = cor(\mathbf{y}_c, \hat{\boldsymbol{u}}_p)$. It was hypothesized that predictivity would be maximized with the best-fitting variance ratio, and the heritability could be estimated by finding the values that maximized predictivity.

RESULTS AND DISCUSSION

HERITABILITY ESTIMATION

Heritability estimates by AIREML, AIGREML, and method R were compared using intervals with three to ten generations of data. Figure 5.1 shows heritability estimates using the three methods for intervals with the base held constant at the first generation and increasing by one generation. The heritability estimates on average were 0.30 ± 0.01, 0.26 ± 0.01, and 0.30 ± 0.05 for AIREML, AIGREML, and method R, respectively. The AIREML and AIGREML estimates had smaller standard errors consistent with varying amounts of data. In contrast, the method R estimates had larger standard errors and were more variable with varying amounts of data, which aligns with the results in Duangjinda et al. (2001). Reverter et al. (1994a) compare subsequent evaluations to determine the heritability estimate; however, other studies create a partial dataset by removing a random percentage of animals from the whole dataset (Misztal, 1997; Duangjinda et al., 2001). The standard error could be decreased if the heritability estimation was repeated with various partial datasets, reducing the sampling variance (Misztal et al., 1997).

The lower heritability estimates from AIGREML compared to AIREML are consistent with recent studies comparing heritability with and without genomic information in populations undergoing genomic selection (Hidalgo et al., 2020; Richter et al., 2022). However, the variances in this study were not scaled against the same base, which could add bias to the estimations. In older studies where genomic preselection was not yet present, similar heritabilities but lower standard errors were estimated using genomic information in a pig population (Forni et al., 2011) and in a dairy cattle population (Veerkamp et al., 2011). As genomic selection was simulated in the study, the heritability estimates with AIREML may be biased as genomics were not considered, and therefore, the information used for selection decisions was not included. However, the

heritability parameter in the simulation was set to 0.3, and all data subsets begin with generation 1.

As the properties of method R are based on finding unbiased breeding values, we analyzed $b_{w,p}$ of GEBV calculated with heritabilities estimated by AIGREML by regressing $\hat{\boldsymbol{u}}_w$ on $\hat{\boldsymbol{u}}_p$ for all generation intervals, as shown in Figure 5.2. The GEBV by AIGREML were slightly under-dispersed for all subsets of data. When more data was used, the distribution among replicates was smaller than for subsets with less data. This aligns with the results shown in Figure 5.1 and suggests that the heritability estimates by AIGREML were underestimated, and thus the GEBV are under-dispersed. As stated in Misztal (1997), a $b_{w,p}$ greater than one denotes an estimated additive variance that is too small and, thus, an estimated heritability that is lower than the correct value.

Elapsed wall-clock times in minutes for AIGREML and method R for various dataset sizes are shown in Figure 5.3. For each subset of data, the method R process included calculating and saving $\mathbf{G}^{-1}$ once per subset of data and calculating $\hat{\boldsymbol{u}}_w$ and $\hat{\boldsymbol{u}}_p$ using iteration on data for the six variance ratios compared. The times presented in Figure 5.3 include the average computing time for all factors in the method R process per replicate. Method R was faster than AIGREML for all amounts of data, and elapsed time increased as the amount of data increased. On average, for ten generations of data (50,020 genotypes and phenotypes), method R took 1.6 hours, and AIGREML took 9.5 hours. It is important to note that Method R, as implemented in this study, has a stopping criterion instead of a convergence criterion, which is based on grid search. However, an iterative algorithm could be implemented and possibly increase the resolution of the method R heritability estimates but a possible increase in computing time. Alternatively, convergence criteria could be implemented with an iterative algorithm instead of fitting a line to solve for the variance ratio

estimates. This way, more precise estimates could be possible, and research should be conducted for further testing.

Ultimately, the time savings by method R are due to its utilization of efficient methods to solve for breeding values as implemented in blup90iod3 (Misztal et al., 2014) and the avoidance of costly variance component estimation methods. Method R can have a linear cost with the number of genotyped or phenotyped animals, and it is feasible whenever a genetic evaluation is possible. Methods based on REML have quadratic memory requirements and have close to cubic costs with the number of genotyped animals. Methods based on Gibbs sampling have a quadratic cost but also a long computing time due to the high number of samples needed. Hence, method R can be applied to the large datasets when the mixed model equations are not explicitly stored. In contrast, methods based on storing mixed model equations are inherently limited. Another option is to use SNPBLUP as an alternative to GBLUP, whereas efficient SNPBLUP-based software exists (Lee and van der Werf, 2016; Lidauer et al., 2011).

MAXPRED

Predictivities are shown in Figure 5.4. The values ranged from 0.23 to 0.39 and increased as the number of generations of data increased in the interval. The values for all subsets of data increased, peaked, and decreased. However, the curves are relatively flat, and there were minimal differences between the predictivity for all heritabilities compared. The curves were flatter when all data was used, and a more substantial peak was present when fewer generations of data were used. In agreement with previous studies, predictivity is dependent on the amount of data used and the correctness of the model (Legarra et al., 2008). Specifically, in a sire model with all sires having the same number of daughters, the predictivity does not largely depend on heritability. A small

numerical example using a sire model is presented in Appendix A to show the lack of major change in predictivity as the variance ratio substantially changes.

VALIDATION

TBV and GEBV calculated with heritability estimates from AIGREML and method R were compared. Correlations, regression coefficients (b_1), and intercepts (b_0) were calculated on linear regressions of combinations of TBV, GEBV by AIGREML, and GEBV by method R. The correlation, b_1, and b_0 of the regressions using intervals of generations 1-10 and 1-3 are shown in Table 5.1. The whole dataset and a subset with the first three generations of data were used to compare the performance of method R with large and small datasets. Correlations were 1.0 between GEBV by AIGREML and GEBV by method R using all generations and for the first three generations. As the simulated population was under genomic selection, it is expected that the GEBV using variance components estimated with genomic information will be more similar compared to EBV calculated from non-genomic models. Correlations and b_1 for TBV regressed on GEBV by AIGREML and TBV regressed on GEBV by method R were closer to 1.0 when all generations were used and deviated further from 1.0 when only the first three generations were used. This is as expected, as accuracies tend to be higher with more data. The intercept, however, was closer to the expected value of 0.0 when the first three generations were used compared to all generations. Because selection was simulated in this population, the phenotypes are changing directionally with each generation; therefore, bias is expected to increase with more generations of data. All validation comparisons between GEBV by AIGREML and TBV and GEBV by method R and TBV are identical or overlap when accounting for the standard error. This suggests that the GEBV calculated with heritabilities from either AIGREML or method R have negligible

differences, and it is promising that method R can be used as a fast alternative to REML or Bayesian methods for heritability estimation in single-trait GBLUP models.

CONCLUSIONS

The main advantage of heritability estimation with genomics using method R is its ability to process any size model for which the genetic evaluation is possible. Although its heritability estimates do not have the theoretical properties of REML, method R is a quick way to compute the estimates. Additional VCE should be conducted prior to implementing the variance ratio obtained by method R into official evaluations. Further research on method R with genomics is needed for multiple trait models, complex models, and real data.

REFERENCES

Bussiman, F., C.-Y. Chen, J. Holl, M. Bermann, A. Legarra, I. Misztal, and D. Lourenco. 2023. Boundaries for genotype, phenotype, and pedigree truncation in genomic evaluations in pigs. Journal of Animal Science 101doi: 10.1093/jas/skad273

Culbertson, M. S., J. W. Mabry, I. Misztal, N. Gengler, J. K. Bertrand, and L. Varona. 1998. Estimation of dominance variance in purebred Yorkshire swine. Journal of Animal Science 76(2):448-451. doi: 10.2527/1998.762448x

Druet, T., I. Misztal, M. Duangjinda, A. Reverter, and N. Gengler. 2001. Estimation of genetic covariances with method R. Journal of Animal Science 79(3):605-615. doi: 10.2527/2001.793605x

Duangjinda, M., I. Misztal, J. K. Bertrand, and S. Tsuruta. 2001. The empirical bias of estimates by restricted maximum likelihood, Bayesian method, and method ℜ under selection for additive, maternal, and dominance models. Journal of Animal Science 79(12):2991-2996. doi: 10.2527/2001.79122991x

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.* **43**:1. doi:10.1186/1297-9686-43-1

Falconer, D. S. 1996. Introduction to quantitative genetics. Pearson Education India.

Gengler, N., I. Misztal, J. K. Bertrand, and M. S. Culbertson. 1998. Estimation of the dominance variance for postweaning gain in the U.S. Limousin population1. Journal of Animal Science 76(10):2515-2520. doi: 10.2527/1998.76102515x

Gianola, D., and R. L. Fernando. 1986. Bayesian Methods in Animal Breeding Theory. Journal of Animal Science 63:217-244.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando. 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics 183(1):347-363. doi: 10.1534/genetics.109.103952

Henderson, C. L. 1949. Estimation of Changes in Herd Environment. Journal of Dairy Science 32(8):706. doi: https://doi.org/10.3168/jds.S0022-0302(49)92104-9

Henderson, C. L. 1950. Estimation of Genetic Parameters. The Annals of Mathematical Statistics 21(2):309.

Hidalgo, J., S. Tsuruta, D. Lourenco, Y. Masuda, Y. Huang, K. A. Gray, and I. Misztal. 2020. Changes in genetic parameters for fitness and growth traits in pigs under genomic selection. Journal of Animal Science 98(2)doi: 10.1093/jas/skaa032

Hollifield, M. K., D. Lourenco, M. Bermann, J. T. Howard, and I. Misztal. 2021. Determining the stability of accuracy of genomic estimated breeding values in future generations in commercial pig populations. Journal of Animal Science 99(4)doi: 10.1093/jas/skab085

Junqueira, V. S., D. Lourenco, Y. Masuda, F. F. Cardoso, P. S. Lopes, F. F. e. Silva, and I. Misztal. 2022. Is single-step genomic REML with the algorithm for proven and young more computationally efficient when less generations of data are present? Journal of Animal Science 100(5)doi: 10.1093/jas/skac082

Lee, S. H., and J. H. J. van der Werf. 2016. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics 32(9):1420-1422. doi: 10.1093/bioinformatics/btw012

Legarra, A. s., C. l. Robert-Granié, E. Manfredi, and J.-M. Elsen. 2008. Performance of Genomic Selection in Mice. Genetics 180(1):611-618. doi: 10.1534/genetics.108.088575

Legarra, A., Reverter, A. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genet Sel Evol* **50**: 53. https://doi.org/10.1186/s12711-018-0426-6

Lidauer, M., K. Matilainen, E. Mantysaari, and I. Stranden. 2011. : solving large mixed model equations manual. Jokioinen: MTT

Masuda, Y., I. Aguilar, S. Tsuruta, and I. Misztal. 2015. Technical note: Acceleration of sparse operations for average-information REML analyses with supernodal methods and sparse-storage refinements1,2. Journal of Animal Science 93(10):4670-4674. doi: 10.2527/jas.2015-9395

Matilainen, K., E. A. Mäntysaari, M. H. Lidauer, I. Strandén, and R. Thompson. 2013. Employing a Monte Carlo Algorithm in Newton-Type Methods for Restricted Maximum Likelihood

Estimation of Genetic Parameters. PLOS ONE 8(12):e80821. doi: 10.1371/journal.pone.0080821

Misztal, I., T. J. Lawlor, and R. L. Fernando. 1997. Dominance Models with Method R for Stature of Holsteins. Journal of Dairy Science 80(5):975-978. doi: https://doi.org/10.3168/jds.S0022-0302(97)76022-3

Misztal, I. 1997. Estimation of Variance Components with Large-Scale Dominance Models. Journal of Dairy Science 80(5):965-974. doi: https://doi.org/10.3168/jds.S0022-0302(97)76021-1

Misztal, I. 2008. Reliable computing in estimation of variance components. Journal of Animal Breeding and Genetics, 125: 363-370. https://doi.org/10.1111/j.1439-0388.2008.00774.x

Misztal, I., I. Aguilar, D. Lourenco, L. Ma, J. P. Steibel, and M. Toro. 2021. Emerging issues in genomic selection. Journal of Animal Science 99(6)doi: 10.1093/jas/skab092

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

Misztal, I., S. Tsuruta, D. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. G. Vitezica. 2014b. Manual for BLUPF90 family of programs. http://nce.ads.uga.edu/wiki/doku.php?id=application_programs.

Misztal, I., L. Varona, M. S. Culbertson, J. K. Bertrand, J. W. Mabry, T. J. Lawlor, C. P. V. Tassel, and N. Gengler. 1998. Studies on the value of incorporating the effect of dominance in genetic evaluations of dairy cattle, beef cattle and swine. Biotechnologie, Agronomie, Société et Environnement 2:227-233.

Patry, C., and V. Ducrocq. 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. Journal of Dairy Science 94(2):1011-1020. doi: https://doi.org/10.3168/jds.2010-3804

Patterson, H. D., and R. Thompson. 1971. Recovery of Inter-Block Information when Block Sizes are Unequal. Biometrika 58(3):545-554. doi: 10.2307/2334389

Rauw, W. M. 2008. Resource allocation theory applied to farm animal production. CABI Publishers.

Reverter, A., B. L. Golden, R. M. Bourdon, and J. S. Brinks. 1994a. Method R variance components procedure: application on the simple breeding value model1,2,3. Journal of Animal Science 72(9):2247-2253. doi: 10.2527/1994.7292247x

Reverter, A., B. L. Golden, R. M. Bourdon, and J. S. Brinks. 1994b. Technical note: detection of bias in genetic predictions2. Journal of Animal Science 72(1):34-37. doi: 10.2527/1994.72134x

Richter, J., J. Hidalgo, V. Breen, R. Hawken, I. Misztal, and D. Lourenco. 2022. 590. Changes in genetic parameters for traits under genomic selection in poultry, Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP). p. 2445.

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

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

Veerkamp, R. F., H. A. Mulder, R. Thompson, and M. P. Calus. 2011. Genomic and pedigree-based genetic parameters for scarcely recorded traits when some animals are genotyped. *J. Dairy Sci.* **94**:4189–4197. doi:10.3168/jds.2011-4223

Walsh, B., and M. Lynch. 2018. Evolution and selection of quantitative traits. Oxford University Press.

APPENDIX 5.1

Numerical example of predictivity as heritability changes using a sire model:

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{s} + \boldsymbol{e}$$

$$\lambda = \frac{\sigma_e^2}{\sigma_s^2} = \frac{(4 - h^2)}{h^2}$$

Assuming that: $\lambda_1 = 7; \lambda_2 = 0.5; \lambda_3 = 50$

$$\mathbf{y} = [120 \quad 176 \quad 154 \quad 138 \quad 162]'$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0.5 \\ 0 & 1 & 0.5 & 0 \\ 0 & 0.5 & 1 & 0 \\ 0.5 & 0 & 0 & 1 \end{bmatrix}$$

For $\lambda_1 = 7$,

$$\lambda_1 \mathbf{A}^{-1} = \begin{bmatrix} 9.33 & 0 & 0 & -4.67 \\ 0 & 9.33 & -4.67 & 0 \\ 0 & -4.67 & 9.33 & 0 \\ -4.67 & 0 & 0 & 9.33 \end{bmatrix}$$

$$\begin{bmatrix} 5 & 2 & 1 & 1 & 1 \\ 2 & 11.33 & 0 & 0 & -4.67 \\ 1 & 0 & 10.33 & -4.67 & 0 \\ 1 & 0 & -4.67 & 10.33 & 0 \\ 1 & -4.67 & 0 & 0 & 10.33 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{s_1} \\ \widehat{s_2} \\ \widehat{s_3} \\ \widehat{s_4} \end{bmatrix} = \begin{bmatrix} 750 \\ 282 \\ 176 \\ 154 \\ 138 \end{bmatrix}$$

$$\begin{bmatrix} \hat{\mu} \\ \widehat{s_1} \\ \widehat{s_2} \\ \widehat{s_3} \\ \widehat{s_4} \end{bmatrix} = \begin{bmatrix} 150.54 \\ -2.68 \\ 3.29 \\ 1.82 \\ -2.42 \end{bmatrix}$$

$$\mathbf{y}^* = (\mathbf{y} - \mathbf{X}\widehat{\mathbf{B}})'\mathbf{Z}$$

$$\mathbf{y}^* = [-19.07 \quad 25.46 \quad 3.46 \quad -12.53]'$$

$$cor(\mathbf{y}^*, \hat{\boldsymbol{s}}) = 0.96$$

λ	$cor(\mathbf{y}^*, \hat{\boldsymbol{s}})$
0.5	0.96
7	0.96
50	0.95

TABLES

Table 5.1 Correlation, regression coefficient (b_1), and intercept (b_0) of TBV regressed on GEBV by AIGREML, GEBV by method R, and GEBV by AIGREML on GEBV by method R. Results using data from generations 1-10 and generations 1-3 are shown.

y, x	Correlation		b_1		b_0	
	1-10	1-3	1-10	1-3	1-10	1-3
TBV, AIGREML	0.95 ± 0.00	0.73 ± 0.01	1.01 ± 0.00	1.04 ± 0.02	3.93 ± 0.19	0.41 ± 0.09
TBV, Method R	0.95 ± 0.00	0.73 ± 0.01	1.01 ± 0.01	1.01 ± 0.03	3.93 ± 0.19	0.41 ± 0.09
AIGREML, Method R	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.97 ± 0.02	0.00 ± 0.00	0.00 ± 0.00

FIGURES

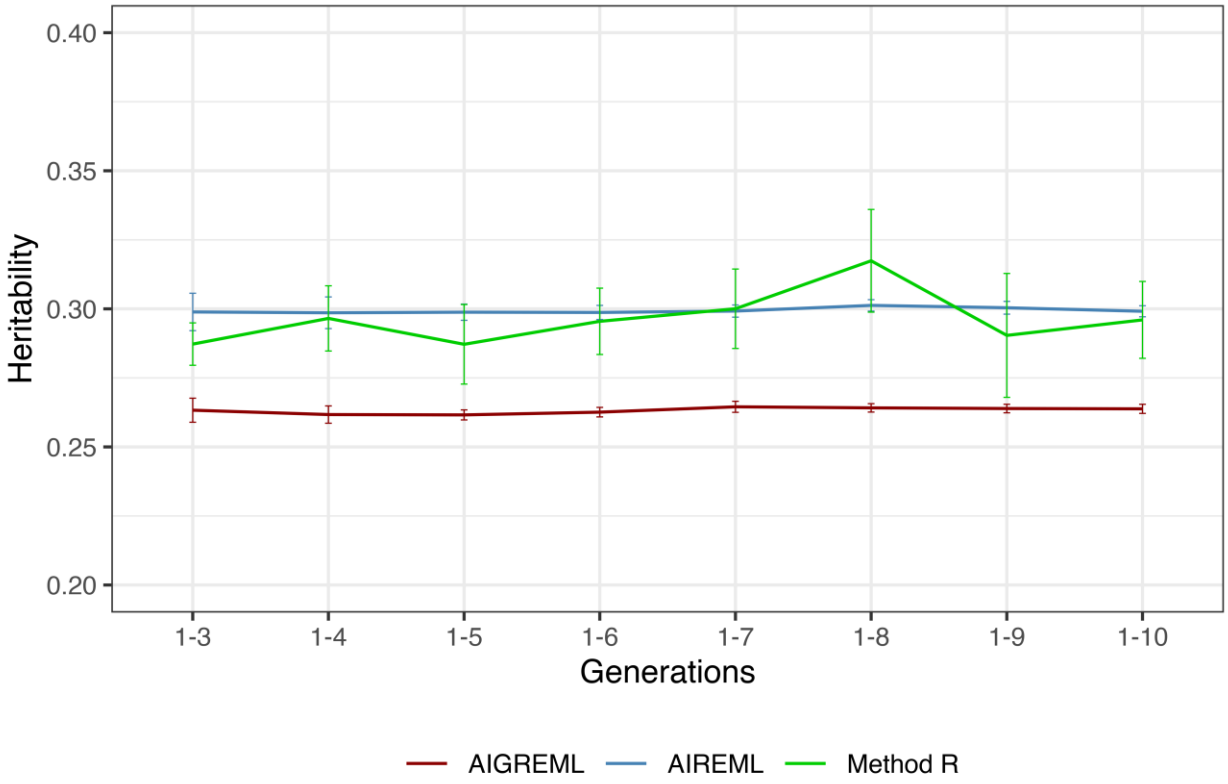

Figure 5.1 Heritability estimates using AIGREML, AIREML, and method R with genomics for generation intervals with the base generation constant at generation 1.

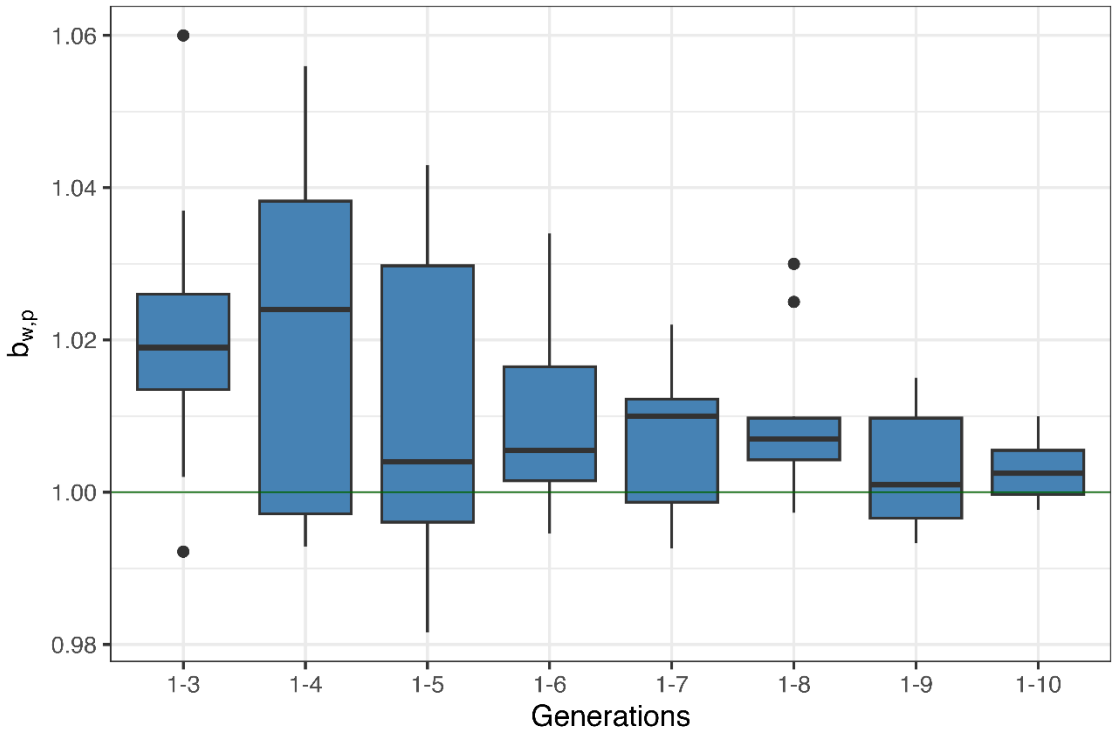

Figure 5.2 The regression coefficient ($b_{w,p}$) of GEBV from AIGREML with the whole dataset regressed on GEBV from AIGREML with the partial dataset for various subsets of data over 10 replicates.

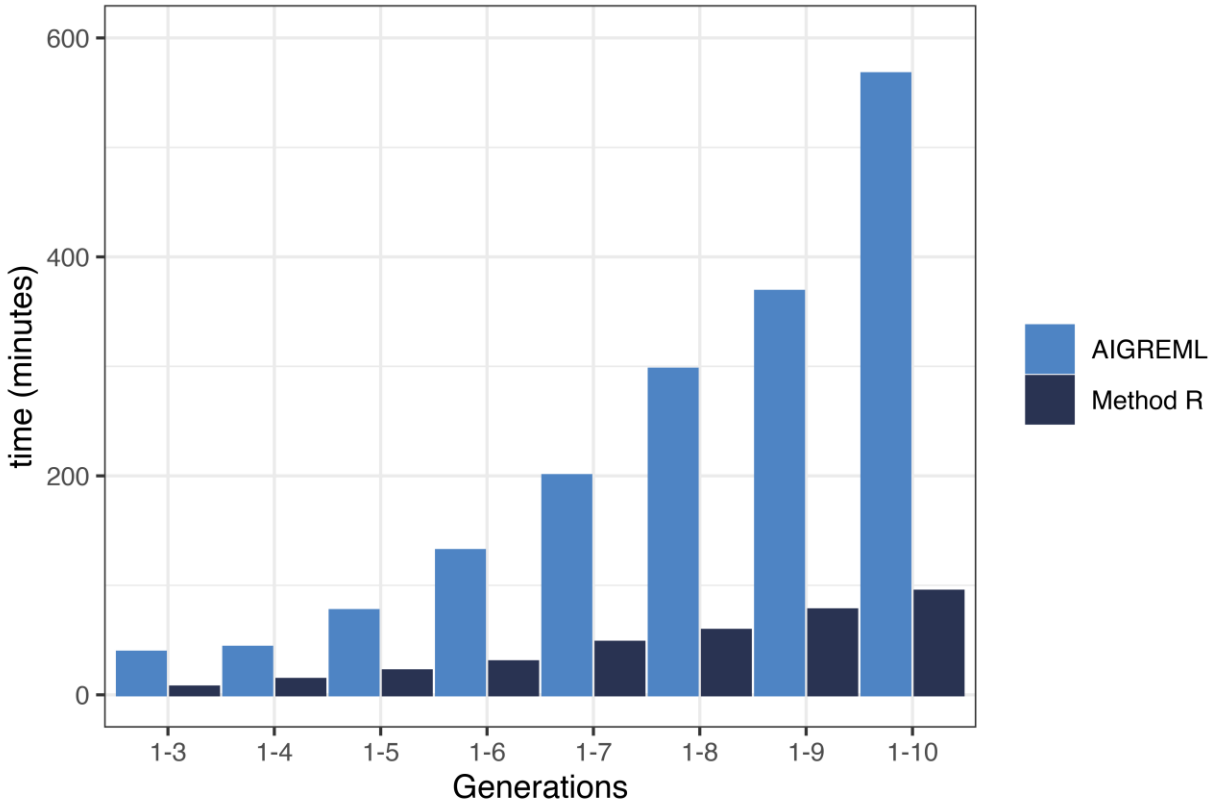

Figure 5.3 Elapsed wall-clock time in minutes for heritability estimation using AIGREML and method R for various dataset sizes.

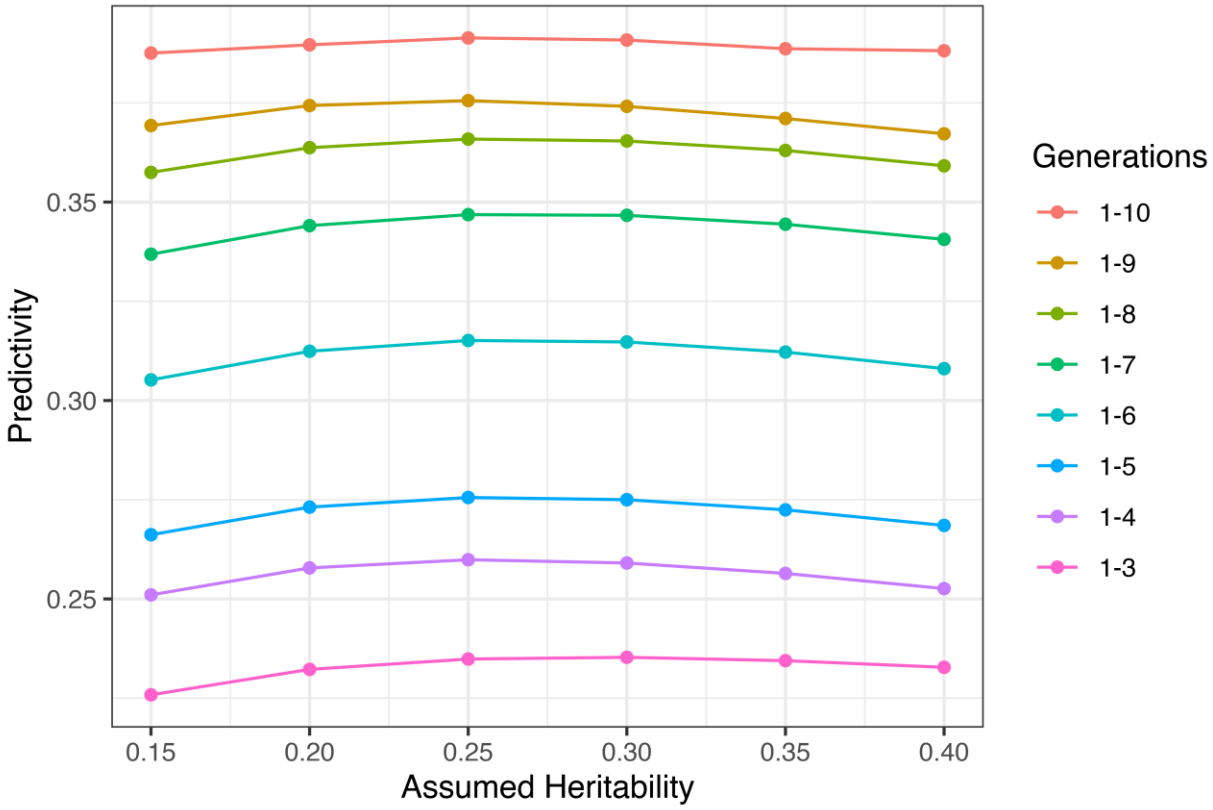

Figure 5.4 Analyzing the MaxPred method by calculating predictivity as $cor(\mathbf{y}_c, \hat{\boldsymbol{u}}_p)$, with $\hat{\boldsymbol{u}}_p$ from AIGREML with partial data for subsets of data with 3 to 10 generations of data and heritabilities equal to 0.15, 0.20, 0.25, 0.30, 0.35, and 0.40.

CHAPTER 6

ESTIMATING GENETIC PARAMETERS OF DIGITAL BEHAVIOR TRAITS AND THEIR RELATIONSHIP WITH PRODUCTION TRAITS IN PUREBRED PIGS[4]

[4] Hollifield, M. K., C. Y. Chen, E. Psota, J. Holl, D. Lourenco, I. Misztal. 2024. *Genetics Selection Evolution*. 56(29). Reprinted here with permission of the publisher.

ABSTRACT

BACKGROUND

With the introduction of digital phenotyping and high-throughput data, traits that were previously difficult or impossible to measure directly have become easily accessible, offering the opportunity to enhance the efficiency and rate of genetic gain in animal production. It is of interest to assess how behavioral traits are indirectly related to the production traits during the performance testing period. The aim of this study was to assess the quality of behavior data extracted from day-wise video recordings and estimate the genetic parameters of behavior traits and their phenotypic and genetic correlations with production traits in pigs. Behavior was recorded for 70 days after on-test at about 10 weeks of age and ended at off-test for 2008 female purebred pigs, totaling 119,812 day-wise records. Behavior traits included time spent eating, drinking, laterally lying, sternally lying, sitting, standing, and meters of distance traveled. A quality control procedure was created for algorithm training and adjustment, standardizing recording hours, removing culled animals, and filtering unrealistic records.

RESULTS

Production traits included average daily gain (ADG), back fat thickness (BF), and loin depth (LD). Single-trait linear models were used to estimate heritabilities of the behavior traits and two-trait linear models were used to estimate genetic correlations between behavior and production traits. The results indicated that all behavior traits are heritable, with heritability estimates ranging

from 0.19 to 0.57, and showed low-to-moderate phenotypic and genetic correlations with production traits. Two-trait linear models were also used to compare traits at different intervals of the recording period. To analyze the redundancies in behavior data during the recording period, the averages of various recording time intervals for the behavior and production traits were compared. Overall, the average of the 55- to 68-day recording interval had the strongest phenotypic and genetic correlation estimates with the production traits.

CONCLUSIONS

Digital phenotyping is a new and low-cost method to record behavior phenotypes, but thorough data cleaning procedures are needed. Evaluating behavioral traits at different time intervals offers a deeper insight into their changes throughout the growth periods and their relationship with production traits, which may be recorded at a less frequent basis.

BACKGROUND

High-throughput phenotyping, digital data recording, and novel traits have recently become topics of interest in animal production. With advancements in technology, phenotypes can be collected with higher accuracy, in greater quantities, and new traits that are difficult or impossible to measure directly can be captured (Brito et al., 2020). Applications include sensors, wearable technology, imaging, video, and audio recording to assess body temperature (Sellier et al., 2014), stress (Lee et al., 2015), disease (Ferrari et al., 2008; Ahmed et al., 2015), behavior (Siegford et al., 2023), or overall health (Sa et al., 2015; Neethirajan et al., 2017).

In pork production, meat quality and quantity are economically relevant traits that are under genetic or genomic selection. Determining the meat characteristics of an animal may not be possible until the peak of production age or slaughter (Neethirajan et al., 2009).]. If a trait that can be measured early in an animal's life is indicative of later production traits, it allows for earlier selection and culling decisions, which can reduce the generation interval. In addition, incorporating phenotypes from progeny at the multiplication or commercial level would benefit nucleus level parents and enhance accuracy of the genomic estimated breeding values (GEBV) of the elite animals. As collecting phenotypes can be costly and labor-intensive, automated data collection via digital phenotyping could increase data collection at a low cost and with more precision than human labor (Neethirajan et al., 2023; Bortoluzzi et al., 2023). Traits that are of interest to capture using digital phenotyping are those that are heritable and that are genetically related to, or that affect an animal's economically relevant production traits.

Animal behavior is one example of such a trait but recording behavior using cameras is challenging due to the difficulty in identifying individual animals. Technologies to obtain automated long-term individualized behavior data include the use of radio frequency identification (RFID; Brown-Brandl et al., 2017), ultra-wideband (Zhuang et al., 2020) and visual fingerprinting (Ravoor et al., 2020) for animal recognition in a pen. However, these methods are fundamentally limited in spatial resolution (RFID and ultra-wideband) and reliability (visual fingerprinting). An alternative approach that provides reliable identification relies on industry-standard ear tags, albeit intermittently, i.e. when the ear tag is exposed to the camera (Psota et al., 2022). Regardless of which method is used, thorough data cleaning is always necessary to ensure that the information captured is realistic and accurate.

Behavior traits that are genetically correlated to production traits can be used for genetic improvement, including activity levels (Obermier et al., 2023), eating patterns (Putz et al., 2019), and management refinement (Berckmans and Norton, 2016). The goal of this study was to create a data quality control procedure and investigate behavior traits that can be captured by digital phenotyping and their phenotypic and genetic correlations with production traits.

MATERIALS AND METHODS

DATASET

The data were provided by PIC (Genus Company, Hendersonville, TN) and included 119,812 day-wise behavior records for 2008 pigs collected between August 26, 2021, and May 23, 2023. All animals were housed on the same farm and belonged to two lines of purebred pigs. Digital behavior phenotypes were extracted from video recordings and included the daily cumulative time each animal spent eating, drinking, lying laterally, lying sternally, sitting, and standing, and the distance traveled. Whereas, standing refers to the raised position which also includes walking or running. The recording period began after the on-test, at about 10 weeks of age, and ended at off-test, for 70 recording days. There were 12 cameras that recorded 14 hours per day (5:00 a.m. to 7:00 p.m.), with one camera per pen and two pens per room had cameras. The two cameras in the same room recorded simultaneously throughout the recording period. The recording group was defined as the animals under the same camera with the same recording start date.

All video was processed by a multi-object tracking algorithm to extract individualized activity data. The first stage of processing consisted of detecting individual pigs using a customized version of the DeepCut pose estimation algorithm (Pishchulin et al., 2016) that detects mid-points, snouts, and right ear tag locations and associated these with individuals. For each detected midpoint, a convolutional neural network (CNN) also estimated the posture of the pig and whether the pig was eating. Snout locations were used to limit the possible locations where eating can take place and to estimate drinking activities based on proximity to the feeder and waterer, respectively.

Once detected, each pig was tracked using Hungarian matching (Kuhn, 1955) to follow detected pigs from one frame to the next, with pigs that were not detected assumed to "stay put" in their previous locations. The most challenging aspect of reliable tracking is maintaining identity, for which a custom ear tag reading method developed by PIC was used (Psota et al., 2022). This method allows tags to be read at low resolution with challenging perspectives, motion blur, noise, and shadows.

Of the 2008 animals with digital behavior records, 1705 had production trait records. The production traits included average daily gain (ADG), back fat thickness (BF), loin depth (LD). All production traits were captured at off-test, at about 20 weeks of age, i.e. at the end of the recording period.

To validate the system's accuracy in determining location, posture, and identification, a trial with 36 randomly selected pigs was conducted (six from each of six pens, with each pen housing 19 pigs). These pigs were distinctly marked for easy identification in video footage. Across five days, 330 annotated images were produced by the tracking algorithm to highlight each pig's location, posture, and ear tag identification. The cross-checker was tasked with first

identifying the pig with specific paint markings in the image. If the pig could not be reliably identified, this image was not included in the analysis. If it could be identified with high confidence, its identity was compared to the known identity from the table of paint-ID correspondences. Its posture and eating/drinking status were also manually recorded and compared to the automated activity detections. Overall, the pigs of interest were all identified in more than 95% of the images and used to evaluate accuracy. Preliminary validation metrics precision, recall, and F1-score indicated an accuracy of correctly annotating location, posture, and identity greater than 97% (Agha et al., 2024).

DATA CLEANING PROCEDURE

Since digital phenotyping is a recently developed and evolving data collection technique, quality control efforts are needed. After analyzing the data patterns, several observations were made that suggested that, under certain conditions, the data may not be reliable. As the equipment was set-up and calibrated, and the farm standard operating procedures and data-extracting algorithms were being created and adjusted, a "learning period" was designated. The "learning period" spanned from the beginning of the recording period, August 21, 2021, to March 17, 2022, and the data collected during this period were discarded from the analyses. After the learning period, the average recording time per day by each camera became more consistent, and the start and end dates for each 70-day recording group became more cyclic than before the learning period.

The data for each pig and each day were summarized into cumulative time spent in each behavior, position, or distance traveled over a 14-hour period. Therefore, days with less than eight hours of recording time were removed as this is not representative of the behavior for the total 14 hours. Days with less than 14 hours and more than 8 hours of recording time were scaled up to 14

hours by dividing the daily record by the number of recording hours and multiplying by 14. The data for culled animals on the day they were extracted from the pen were removed, as the time when an animal was extracted from the pen was not available, so it is not known how many hours to account for the culled animal's activity for that day. The start and end days of the recording period were also removed from the analyses due to the lack of the full 14 recording hours and disruptions from loading and unloading the animals.

After further examination of the data, some records for distance traveled were biologically impossible compared to the time spent standing. For example, one animal stood for only three min and was recorded to have traveled 600 m. After investigation, it was determined that the data-extracting algorithm accumulated meters traveled if the animal was rotating while in a sitting position. To account for this, the daily distance data were truncated to 15 m per min standing, and all daily records that exceeded this ratio were discarded. The data extracting algorithm will be modified for future studies to prevent recording distance while the animal is in the sitting position.

After data cleaning, 77,423 daily records from 1327 animals remained. The average eating time, distance, and recording time per day after the learning period and following data cleaning are shown in Fig. 6.1 a, b and c, and the summary statistics for the behavior traits are in Table 6.1. Since the data in Fig. 6.1 a, b and c are averaged over the recording days, they include animals of various ages and at different stages of the recording period. The peak shown in the average eating time is because, at that time, only young pigs were recorded and they spend more time at the feeder than older pigs. It should be noted that the drinking time behavior measures time spent at the waterer, not the amount of water consumed.

Only animals with off-test production records were used to estimate genetic parameters,

which included 71,999 daily behavior records from 1079 animals, among which 563 were of one line and 516 were of a different line. Summary statistics for the production traits of these animals are in Table 6.2.

MODEL AND ANALYSES

Variance components were estimated using the blupf90+ program (Misztal et al., 2014; Lourenco et al., 2022) that applies a single-trait and a two-trait linear model. The equation for all models can be expressed as:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{W_1 l} + \mathbf{W_2 c} + \mathbf{e}, \quad [1]$$

where $\mathbf{y}$ is the vector of phenotypes, $\boldsymbol{\beta}$ is the vector of the fixed line effects, $\mathbf{u}$, $\mathbf{l}$, and $\mathbf{c}$ are random vectors of additive genetic, common litter, and contemporary group effects, respectively. Contemporary groups were represented by off-test day and year. A pen or camera effect was not included in the model because it was confounded with the litter effect. Elements of $\mathbf{y}$ are related to elements $\mathbf{u}$, $\mathbf{l}$, and $\mathbf{c}$ by incidence matrices $\mathbf{Z}$, $\mathbf{W_1}$, and $\mathbf{W_2}$, respectively, and $\mathbf{e}$ is a random vector of residuals. For single-trait models, the covariance matrices were assumed to be:

$$\text{Var}\begin{bmatrix}\mathbf{u}\\ \mathbf{l}\\ \mathbf{c}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\sigma_u^2 & 0 & 0 & 0\\ 0 & \mathbf{I}\sigma_l^2 & 0 & 0\\ 0 & 0 & \mathbf{I}\sigma_c^2 & 0\\ 0 & 0 & 0 & \mathbf{I}\sigma_e^2\end{bmatrix}, \quad [2]$$

where $\mathbf{A}$ is the numerator relationship matrix, $\mathbf{I}$ is the identity matrix, and σ_u^2, σ_l^2, σ_c^2, and σ_e^2 are variances for the additive genetic, common litter, contemporary group, and residual effects,

respectively.

For two-trait models, the vectors **u**, **l**, and **c**, and **e** were assumed to be distributed as multivariate normal with mean zero and the following covariance structure:

$$\mathrm{Var}\begin{bmatrix}\mathbf{u}\\ \mathbf{l}\\ \mathbf{c}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\otimes\mathbf{G} & 0 & 0 & 0\\ 0 & \mathbf{I}\otimes\mathbf{L} & 0 & 0\\ 0 & 0 & \mathbf{I}\otimes\mathbf{C} & 0\\ 0 & 0 & 0 & \mathbf{I}\otimes\mathbf{R}\end{bmatrix}, \quad [3]$$

where **G** is the additive genetic (co)variance matrix between the two traits, **L** is the common litter (co)variance matrix, **C** is the contemporary group (co)variance matrix, and **R** is the residual (co)variance matrix. Single-trait models were used for heritability estimation for the behavior traits. Two-trait models were used to analyze the relationship between behavior and production traits and to determine the redundancy in the 70 recording days by splitting the recording time into separate periods and determining the relationship of behavior traits in each period with those for the full recording time or with a production trait.

RESULTS AND DISCUSSION

BEHAVIOR TRENDS

All results shown are with data after the learning period and cleaning. Trends in average behavior and posture over the recording period based on the clean data for animals with off-test records are shown in Fig. 6.2 a and b. The data showed a decreasing pattern for eating time, distance traveled, and standing time as the pigs aged. These trends agree with Hyun and Ellis

(2001), who showed that younger pigs eat more meals per day and eat less per meal than older pigs. A slight increasing pattern was seen for time lying laterally and sternally as pigs aged. Figure 6.3 shows the average eating time per recording group over the recording period, where each line is a unique recording group. The substantial variation in behavior between recording groups is shown as in eating time trends in Fig. 6.3. Therefore, it is important to have an adequate data cleaning procedure and statistical model to separate genuine variation between groups from noise. We also observed that the pigs spent more time lying laterally and less time lying sternally in the warmer months than in the cooler months, which agrees with studies from Ekkel et al. (2003) and Huynh et al. (2005). The trends of the average lateral lying time, sternal lying time, and temperature over the recording date are shown in Fig. 6.4. The temperature data were retrieved from the NASA POWER website (https://power.larc.nasa.gov/data-access-viewer/) using the longitude and latitude coordinates of the farm and reflect the average outside air temperature at a height of two meters. Similarly, Aarnink et al. (2001) found that the relative number of pigs that lay laterally increased by 1.8% for each degree Celsius rise in temperature. There was no pattem seen for drinking time as the pigs aged.

BEHAVIOR TRAIT HERITABILITIES AND CORRELATIONS

Table 6.3 shows estimates of phenotypic and genetic correlations, and of heritabilities. Eating time had the highest heritability estimate among the behavior traits, at 0.57, with a standard error of 0.12. The behavior trait with the lowest heritability estimate was laterally lying at 0.19, with a standard error of 0.07. The standard errors of the heritability estimates of the behavior traits were sizeable, ranging from 0.07 to 0.13, likely observed due to the small size of the dataset and the lack of multiple generations with records. The behavior traits with the strongest phenotypic

correlation estimates were standing time and distance (0.67) and laterally lying time and sternally lying time (-0.82). The same trait combinations had the strongest genetic correlation estimates, i.e. 0.93 ± 0.03 for standing time and distance and -0.84 ± 0.04 for laterally lying time and sternally lying time. These estimates are expected, as generally, the animals are in a standing position while mobile and may prefer one lying position over the other, especially if the 70 days of recording time is during consistently hot or cold weather.

REDUNDANCY IN RECORDING TIME

To determine whether all 70 days of recording time were necessary, we analyzed phenotypic correlations of daily and weekly intervals for the same trait and fitted two-trait models with a time interval of a behavior trait as one trait and either the total average of the behavior trait or a production trait as the second trait. In general, estimates of phenotypic correlations between daily and weekly intervals of the same trait became stronger as the recording time progressed and were stronger closer to the end of the recording period compared to the beginning. This suggests that the animals behaved more similarly as they aged and adapted to their environment, which infers redundant information. For example, Fig. 6.5 a and b show estimates of phenotypic correlations between daily and weekly averages for distance traveled; the correlation for weeks 1 and 2 was 0.68, while the correlation for weeks 8 and 9 was 0.82. Days closer to the end of the recording period had stronger correlations than days at the beginning of the recording period. Thus, not all 70 recording days are needed to capture the behavior of the animals or to associate the production traits with behaviors.

For the two-trait models, the behavior traits were split into five intervals: days 1-13, 14-26, 27-40, 41-54, and 55-68. Estimates of phenotypic and genetic correlations were compared between

the two traits and are shown in Figs. 6.6 a and b and 7a and b, respectively. For all traits, generally, the middle intervals had higher correlations with the total average compared to intervals at the beginning and end of the recording time, and estimates of genetic correlations were higher than estimates of phenotypic correlations. For phenotypic and genetic correlations between intervals and total recording period average, eating time had the lowest phenotypic correlation estimate, at 0.73 for the 1- to 13-day interval, and the lowest genetic correlation estimate at 0.95 for the 55-to 68-day interval, respectively. The highest phenotypic correlations were 0.94, for laterally lying and sternally lying, and sitting for the 27- to 40-day interval and 0.94 for sitting for the 41-to 54-day interval. The highest genetic correlation estimates were 1.00 for laterally lying, sternally lying, sitting, and standing for the 27- to 40-day interval, 1.00 for laterally lying, sternally lying, and drinking time for the 41- to 54-day interval, and 1.00 for sitting for the 55- to 68-day interval.

A genetic correlation of 1.00 indicates that the two traits have the same genetic basis. Therefore, if recording is to capture the average behavior of the animals from the on-test period until the off-test period, then the same information, or the most informative data, can be captured during days 27-40 of this period, as this interval has a 1.00 genetic correlation with the total average of the recording period. As storing videos is very costly, recording for the entire 70 days is not necessary, as the behavior of the animal becomes redundant, and the overall behavior patterns can be captured in a span of two weeks. The closer the genetic correlations are to 0, the weaker the relationship between the two traits, indicating that the traits give different genetic information.

RELATIONSHIPS BETWEEN BEHAVIOR AND PRODUCTION TRAITS

The purpose of this study was to determine if a relationship exists between an animal's

behavior and its production performance. If a sufficiently strong relationship exists, the behavior data could predict the production trait phenotypes before the animal's off-test. Two-trait models were fit to estimate the relationship between behavior traits and production traits. To determine if the behavior during a specific time span in the recording period had a stronger relationship with the production traits than the average behavior trait for the entire period, behavior traits measured over five recording intervals were analyzed also in two-trait models with production traits.

The strongest positive genetic and phenotypic correlations between average behavior across the full recording period and production traits were estimated for lateral lying time and ADG (0.50 ± 0.18) and sternal lying time and LD (0.29), respectively (Fig. 6.8 a, b, c). The strongest negative genetic and phenotypic correlations were estimated between distance and ADG (-0.57 ± 0.10 and -0.30, respectively). The strongest positive genetic correlation between average behavior traits for the five time periods and production traits was estimated between the average lateral lying time for the 55-68-day interval and ADG (0.55 ± 0.19). The strongest negative genetic correlation was estimated between average distance for the 55-68-day interval and ADG (-0.70 ± 0.11). Intuitively, time spent lying and distance traveled are expected to have the strongest genetic correlations with ADG, as an animal that expends less energy is expected to grow faster. Obermier et al. (2023) also found that pigs that spent more time lying and that were less active had higher growth rates and greater body weight at a given age. Sitting time was the behavior trait that was estimated to be least phenotypically and genetically correlated with all production traits, while distance was estimated to be the most correlated. As eating and drinking time only consider the amount of time spent at the feeder and waterer and not the quantity of feed or water consumed, it is expected that these traits are not strongly correlated with the production traits. The period with the strongest genetic correlations between behavior and production traits was the 55-68-day

interval, and the period with the weakest correlations was the 27-40-day interval. Therefore, the last two to three weeks before off-test can best capture the relationship between behavior and production traits.

CONCLUSIONS

Digital phenotyping for behavior traits recorded at an individual level provides an opportunity to further understand the association between behavior and production performance. Although behavior and production traits are not highly correlated, this study introduces the possibility of capturing behavior information and its potential association with production. Obtaining additional information on breeding candidates can increase accuracy of (G)EBV, leading to greater genetic improvement. The link between digital behavior data and economically relevant production traits is of interest, and digital phenotyping is a low-cost method to obtain this information on performance. High-throughput phenotyping is a new method for data collection; therefore, extensive quality control measures are needed before implementing the results into evaluations. The results of this study suggest that some pig behaviors, such as standing time, distance traveled, and laterally laying time, are phenotypically and genetically associated with ADG, BF, and LD. The behavior of animals two to three weeks before the off-test date had the strongest genetic correlations with the production traits. Digital phenotyping is promising for enhancing the efficiency, profitability, and rate of genetic gain in pig production.

REFERENCES

Aarnink, A. J., J. W. Schrama, R. J. Verheijen, and J. Stefanowska. 2001. Pen fouling in pig houses affected by temperature. In: Livestock Environment VI, Proceedings of the 6th International Symposium 2001. p 180.

Agha, S., E. Psota, S. P. Turner, C. R. G. Lewis, A. Doeschl-Wilson. 2024. AI-PigNet: Insights into the social interaction of pigs through automated data and social network analysis. Submitted to the 2024 Measuring Behaviour Conference, Aberdeen, Scotland, January 2024.

Ahmed, S., H.-S. Mun, H. Yoe, and C.-J. Yang. 2015. Monitoring of behavior using a video-recording system for recognition of Salmonella infection in experimentally infected growing pigs. Animal 9(1):115-121.

Berckmans, D., and T. Norton. 2016. Precision livestock farming: Examples for poultry. In: Proc. of the XXV World Poultry Congress. p 6-9.

Brito, L. F., H. R. Oliveira, B. R. McConn, A. P. Schinckel, A. Arrazola, J. N. Marchant- Forde, and J. S. Johnson. 2020. Large-scale phenotyping of livestock welfare in commercial production systems: A new frontier in animal breeding. Frontiers in genetics 11:793.

Brown-Brandl TM, Maselyne J, Adrion F, Kapun A, Hessel EF, Saeys W, Van Nuffel A, Gallmann E. Comparing three different passive RFID systems for behaviour monitoring in grow-finish pigs. In Proceedings of the 8th European Conference on Precision Livestock Farming, Nantes, France 2017 Sep 12 (pp. 12-14).

Ekkel, E. D., H. A. Spoolder, I. Hulsegge, and H. Hopster. 2003. Lying characteristics as determinants for space requirements in pigs. Applied Animal Behaviour Science 80(1):19-30.

Ferrari, S., M. Silva, M. Guarino, J. M. Aerts, and D. Berckmans. 2008. Cough sound analysis to identify respiratory infection in pigs. Computers and electronics in agriculture 64(2):318-325.

Huynh, T., A. Aarnink, W. Gerrits, M. Heetkamp, T. Canh, H. Spoolder, B. Kemp, and M. Verstegen. 2005. Thermal behaviour of growing pigs in response to high temperature and humidity. Applied animal behaviour science 91(1-2):1-16.

Hyun, Y., and M. Ellis. 2001. Effect of group size and feeder type on growth performance and feeding patterns in growing pigs. Journal of Animal Science 79(4):803-810.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. Naval research logistics quarterly 2(1-2):83-97.

Lee, J., B. Noh, S. Jang, D. Park, Y. Chung, and H.-H. Chang. 2015. Stress detection and classification of laying hens by sound analysis. Asian-Australasian journal of animal sciences 28(4):592.

Lourenco, D., S. Tsuruta, I. Aguilar, Y. Masuda, M. Bermann, A. Legarra, and I. Misztal. 2022. Recent updates in the BLUPF90 software suite. In: Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP) Technical and species orientated innovations in animal breeding, and contribution of genetics to solving societal challenges. p 1530-1533.

Misztal I, Tsuruta S, Lourenco D, Masuda Y, Aguilar I, Legarra A, et al. Manual for BLUPF90 family of programs. 2014. http://nce.ads.uga.edu/wiki/doku.php?id=application_programs Accessed 21 June 2023.

Neethirajan, S., S. K. Tuteja, S.-T. Huang, and D. Kelton. 2017. Recent advancement in biosensors technology for animal and livestock health management. Biosensors and Bioelectronics 98:398-407.

Neethirajan S, Kemp B. Digital phenotyping in livestock farming. Animals (Basel). 2021;11:2009.

Obermier, D., M. Trenahile-Grannemann, T. Schmidt, T. Rathje, and B. Mote. 2023. Utilizing NU track to Access the Activity Levels in Pigs with Varying Degrees of Genetic Potential for Growth and Feed Intake. Animals 13(10):1581.

Pishchulin, L., E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p 4929-4937.

Psota E, Fitzgerald R, Herring W. Animal ID using human-readable fiducials and simulation-based training. In: Proceedings of the CV4 Animals Workshop in conjunction with the IEEE Computer Vision and Pattern Recognition (CVPR): 28-24 June 2022; New Orleans. 2022.

Putz, A. M., J. C. Harding, M. K. Dyck, F. Fortin, G. S. Plastow, J. C. Dekkers, and P. Canada. 2019. Novel resilience phenotypes using feed intake data from a natural disease challenge model in wean-to-finish pigs. Frontiers in genetics 9:660.

Ravoor, P. C., and T. Sudarshan. 2020. Deep learning methods for multi-species animal re-identification and tracking–a survey. Computer Science Review 38:100289.

Sa, J., M. Ju, S. Han, H. Kim, Y. Chung, and D. Park. 2015. Detection of low-weight pigs by using a top-view camera. In: Proceedings of the fourth international conference on information science and cloud computing (ISCC2015). p 18-19.

Siegford, J. M., J. P. Steibel, J. Han, M. Benjamin, T. Brown-Brandl, J. R. Dórea, D. Morris, T. Norton, E. Psota, and G. J. Rosa. 2023. The quest to develop automated systems for monitoring animal behavior. Applied Animal Behaviour Science 265:106000.

Sellier, N., E. Guettier, and C. Staub. 2014. A review of methods to measure animal body temperature in precision farming. American Journal of Agricultural Science and Technology 2(2):74-99.

Zhuang, S., J. Maselyne, A. Van Nuffel, J. Vangeyte, and B. Sonck. 2020. Tracking group housed sows with an ultra-wideband indoor positioning system: A feasibility study. Biosystems Engineering 200:176-187.

TABLES

Table 6.1 Summary statistics for digital behavior traits after data cleaning

Trait	Mean	Median	SD	Min	Max
Eating time	56.41	54.41	19.55	0.00	165.96
Drinking time	7.25	6.36	4.51	0.0	84.65
Laterally lying time	287.52	285.45	90.81	6.06	718.70
Sternally lying time	359.12	358.20	73.59	79.84	716.75
Sitting Time	22.97	18.06	18.13	0.00	260.19
Standing time	170.28	168.60	52.21	0.00	611.34
Distance (meters)	872.55	827.31	357.47	0.00	3589.54

SD: standard deviation
Traits recorded in time are shown in min

Table 6.2 Summary statistics for production traits

Trait	Mean	Median	SD	Min	Max
ADG, g	710.59	710.77	63.98	510.90	934.00
BF, mm	8.33	8.00	2.30	4.58	17.12
LD, mm	66.34	66.10	5.40	51.50	83.10

ADG: average daily gain; BF: back fat thickness; LD: loin depth; SD: standard deviation

Table 6.3 Estimates of heritabilities and standard errors using single-trait models (diagonal) and of phenotypic (upper diagonal) and genetic correlations with standard errors (lower diagonal) using two trait models

	Eating time	**Drinking time**	**Laterally lying time**	**Sternally lying time**	**Sitting time**	**Standing time**	**Distance (m)**	**ADG**	**BF**	**LD**
Eating time	0.57 ± 0.12	0.17	-0.34	0.04	0.08	0.51	0.18	0.11	0.08	0.09
Drinking time	0.38 ± 0.12	0.38 ± 0.12	0.15	-0.33	0.01	0.23	0.28	-0.03	0.08	-0.13
Laterally lying time	-0.40 ± 0.16	-0.33 ± 0.14	0.19 ± 0.07	-0.82	-0.26	-0.47	-0.06	-0.11	0.00	-0.21
Sternally lying time	-0.41 ± 0.08	-0.43 ± 0.31	-0.84 ± 0.04	0.22 ± 0.08	0.12	-0.07	-0.36	0.20	0.01	0.29
Sitting time	0.01 ± 0.07	0.26 ± 0.09	-0.23 ± 0.08	-0.25 ± 0.19	0.48 ± 0.13	-0.13	-0.09	0.15	-0.03	0.11
Standing time	0.69 ± 0.06	0.62 ± 0.11	-0.72 ± 0.10	-0.62 ± 0.08	-0.48 ± 0.08	0.43 ± 0.11	0.67	-0.16	0.00	-0.09
Distance (m)	0.45 ± 0.09	0.44 ± 0.13	-0.68 ± 0.11	-0.58 ± 0.09	-0.05 ± 1.00	0.93 ± 0.03	0.38 ± 0.10	-0.30	-0.01	-0.23
ADG	0.14 ± 0.15	0.32 ± 0.22	0.50 ± 0.18	-0.10 ± 0.16	0.26 ± 0.16	-0.56 ± 0.11	-0.57 ± 0.10	0.38 ± 0.12	0.29	0.52
BF	0.18 ± 0.07	0.18 ± 0.02	0.19 ± 0.08	-0.04 ± 0.09	0.00 ± 0.08	-0.17 ± 0.07	-0.27 ± 0.09	0.56 ± 0.08	0.53 ± 0.11	0.10
LD	0.04 ± 0.09	-0.07 ± 0.11	0.24 ± 0.12	0.09 ± 0.12	0.03 ± 0.03	-0.37 ± 0.10	-0.48 ± 0.14	0.84 ± 0.08	0.26 ± 0.20	0.30 ± 0.10

ADG: average daily gain; BF: back fat thickness; LD: loin depth;

The behavior traits were averaged over the total recording period. All traits based on time were expressed in

FIGURES

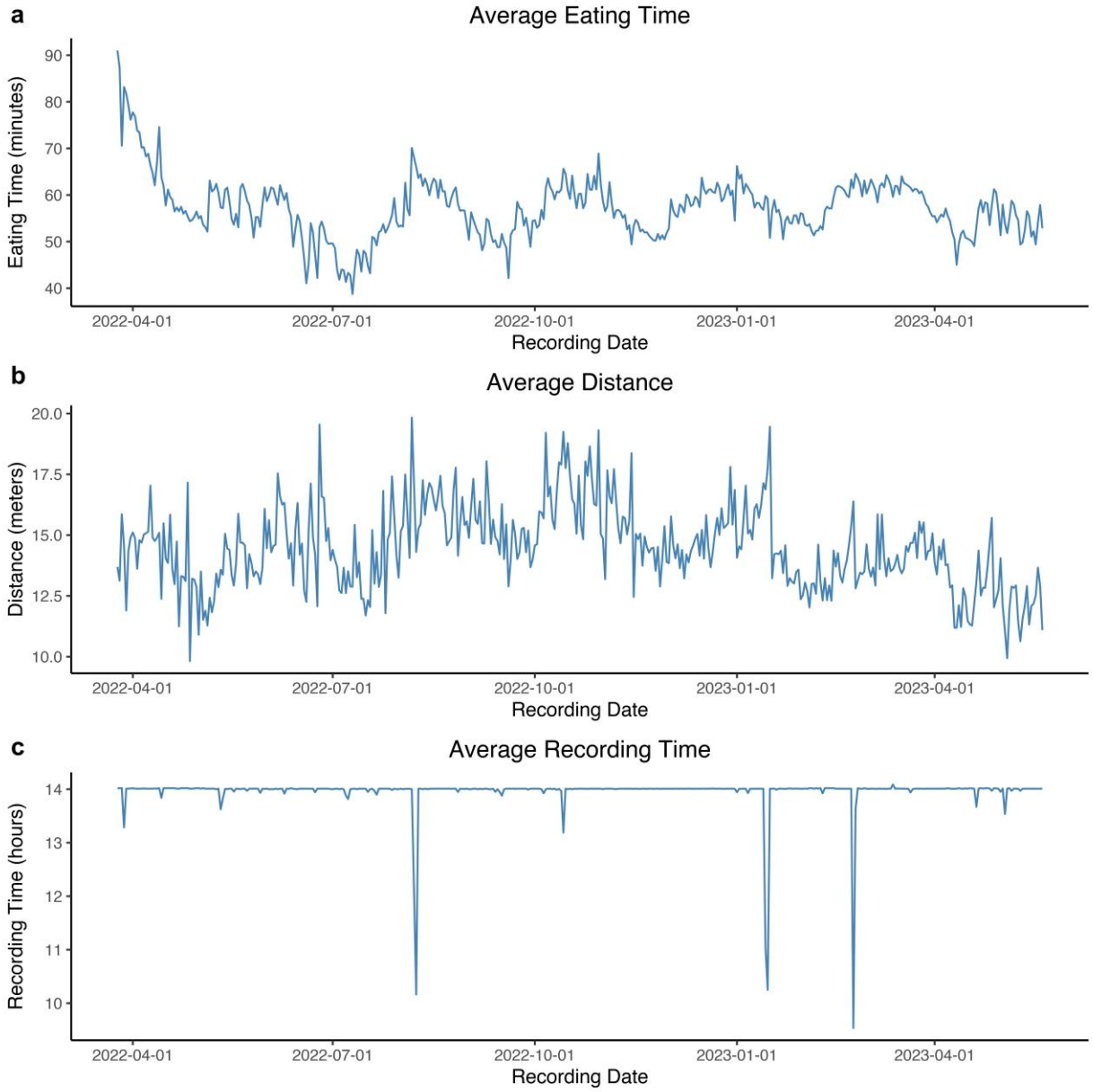

Figure 6.1 Average eating, distance, and recording time per group over time. (a) Average eating time in minutes. (b) Average distance in meters. (c) Average recording time in hours.

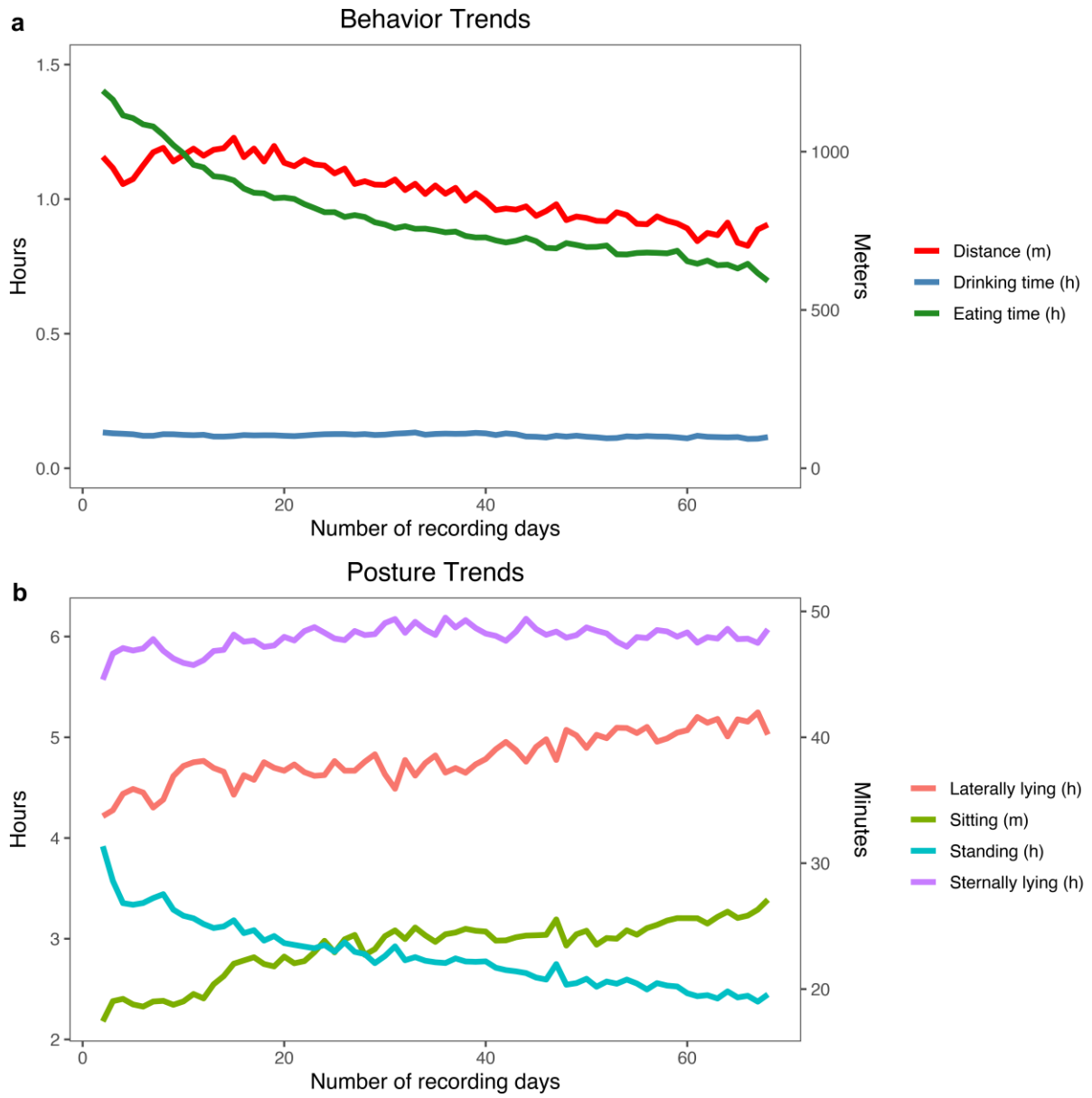

Figure 6.2 Average behavior and posture trends per individual over time. Eating time, drinking time, laterally lying, standing, and sternally lying are shown in h. Sitting is shown in min and distance is shown in m. (a) Behavior trends and (b) Posture trends.

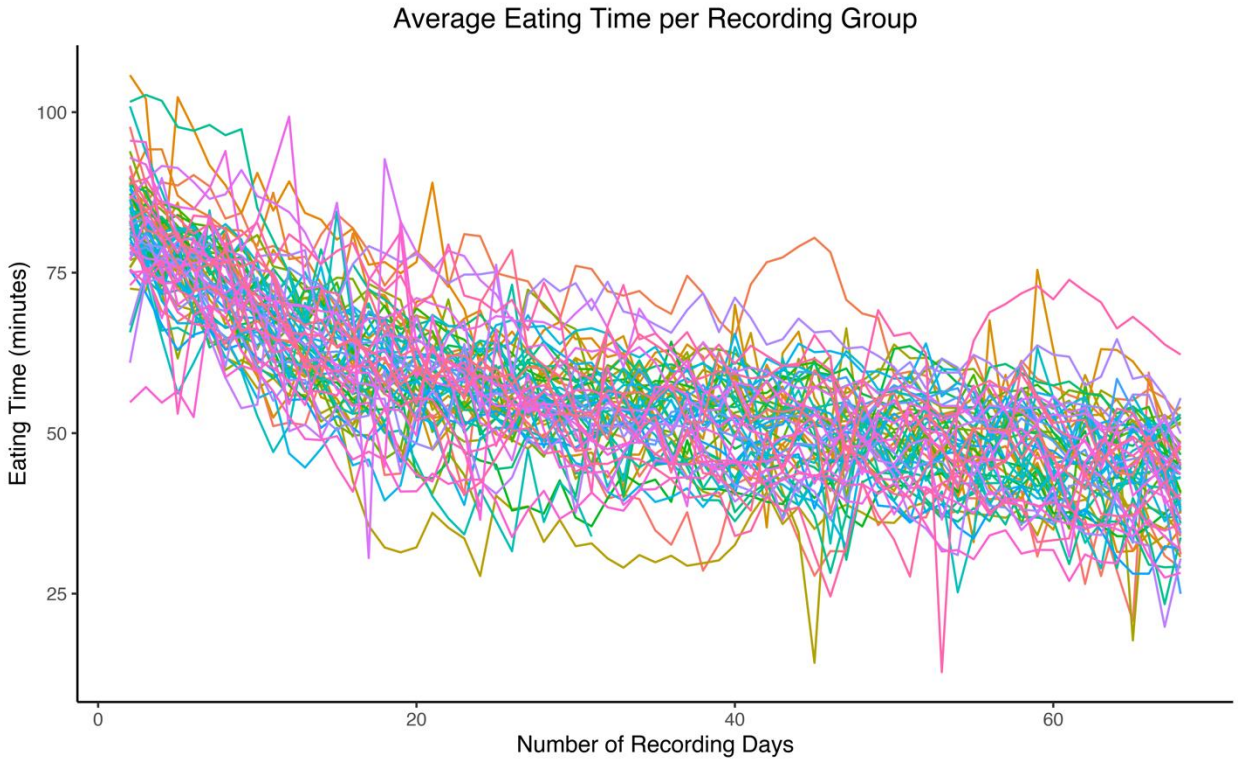

Figure 6.3 Average eating time per group over time.

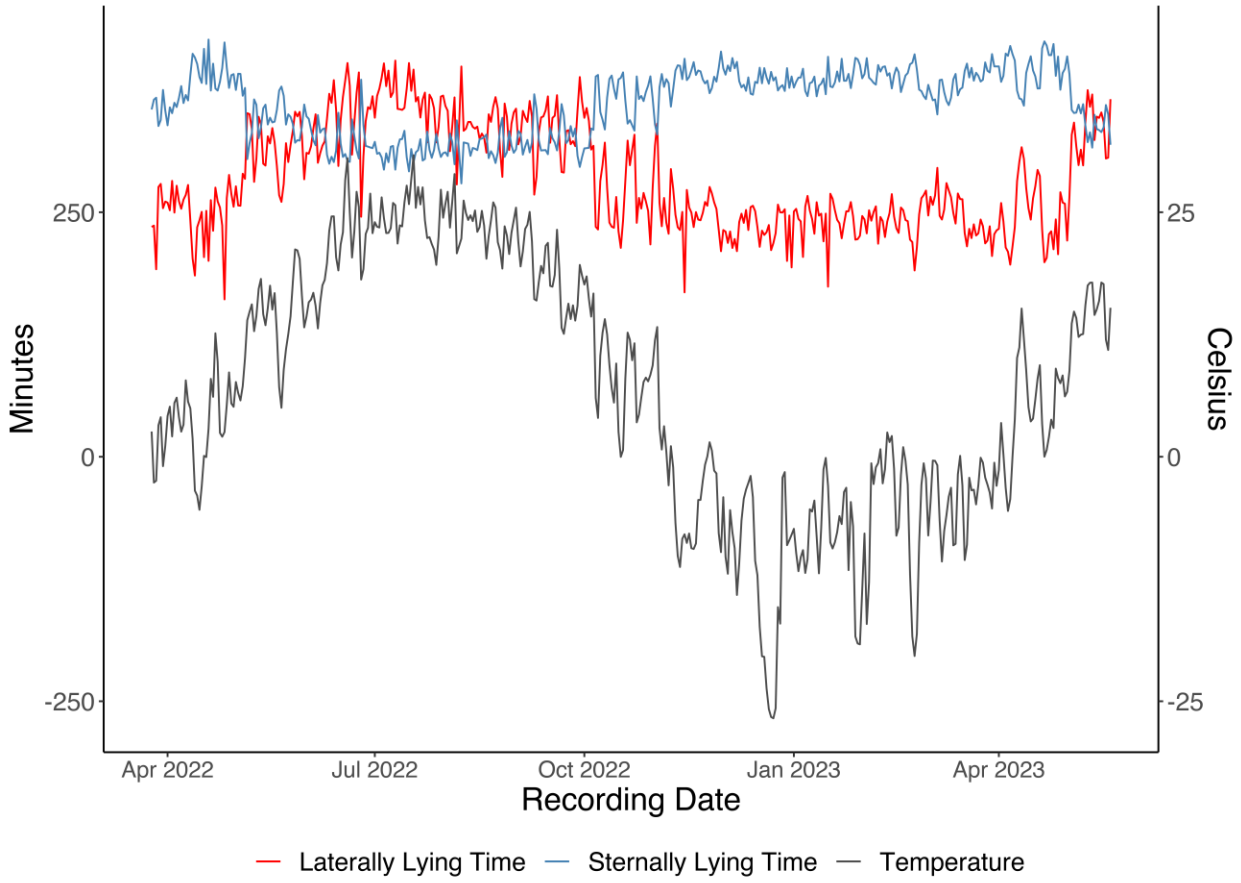

Figure 6.4 Average lateral lying time, sternal lying time, and temperature over time.

The temperature data were obtained from the NASA POWER website (https://power.larc.nasa.gov/data-access-viewer/) using the longitude and latitude coordinates of the farm and were the average air temperatures at a height of two m above the surface of the earth.

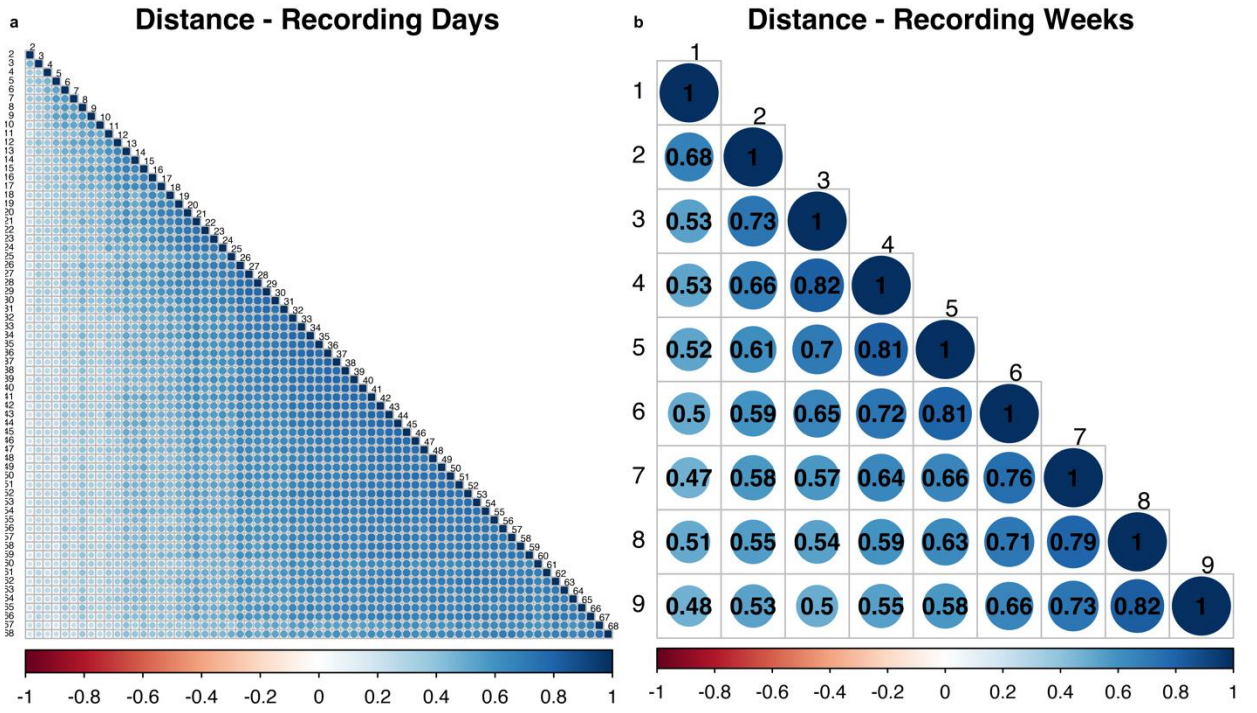

Figure 6.5 Estimates of phenotypic correlations of average daily and weekly distance traveled.

The data used are from the dataset after the data cleaning procedure and for animals with off-test records. (a) Average daily distance traveled and (b) Average weekly distance traveled.

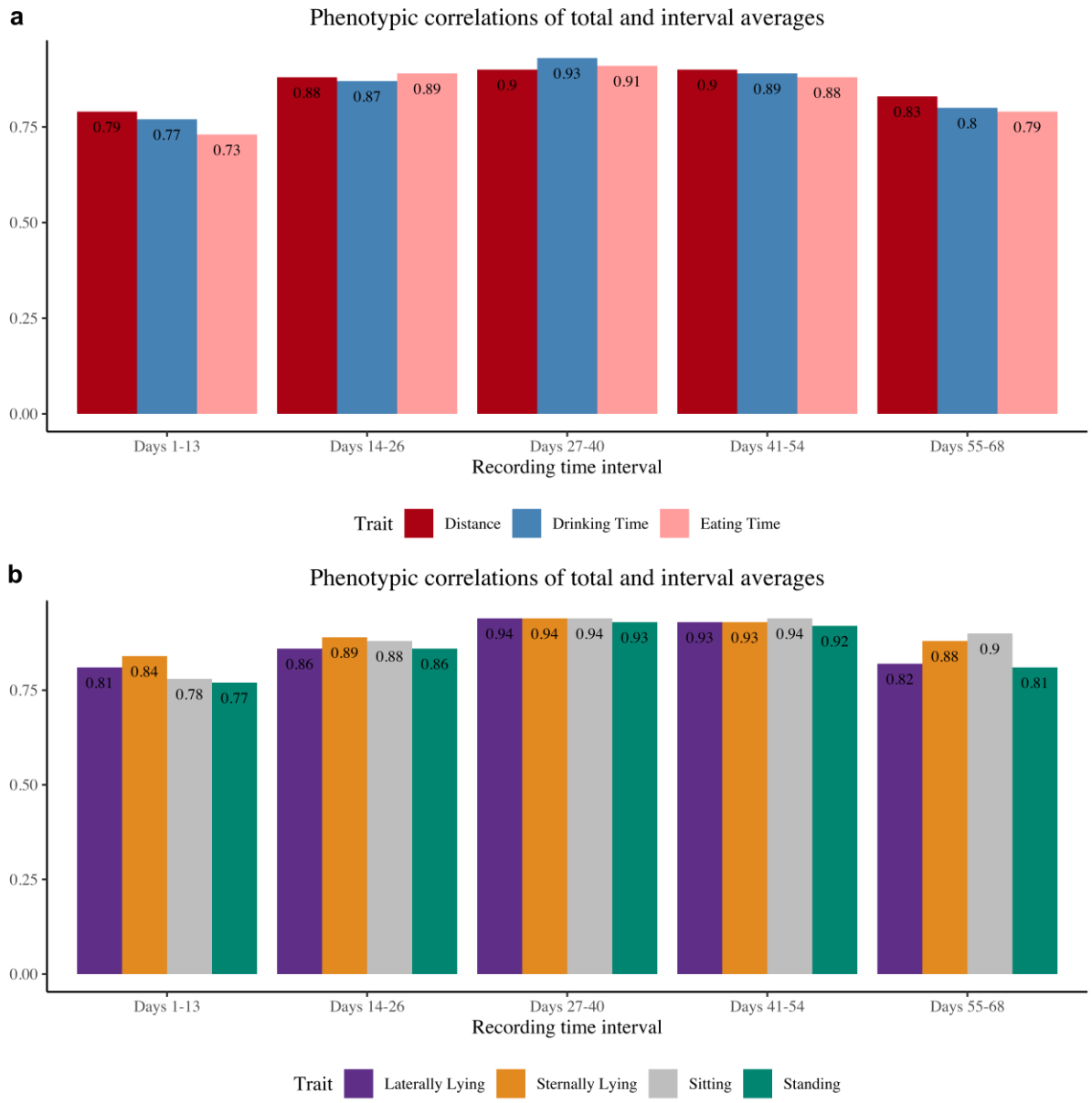

Figure 6.6 Estimates of phenotypic correlations between the total average and the average of each recording time interval for each behavior trait.

All traits are expressed in min, expect for distance which is expressed in m. (a) Traits: distance traveled, drinking time, eating time and (b) Traits: laterally lying time, sternally lying time, sitting time, standing time.

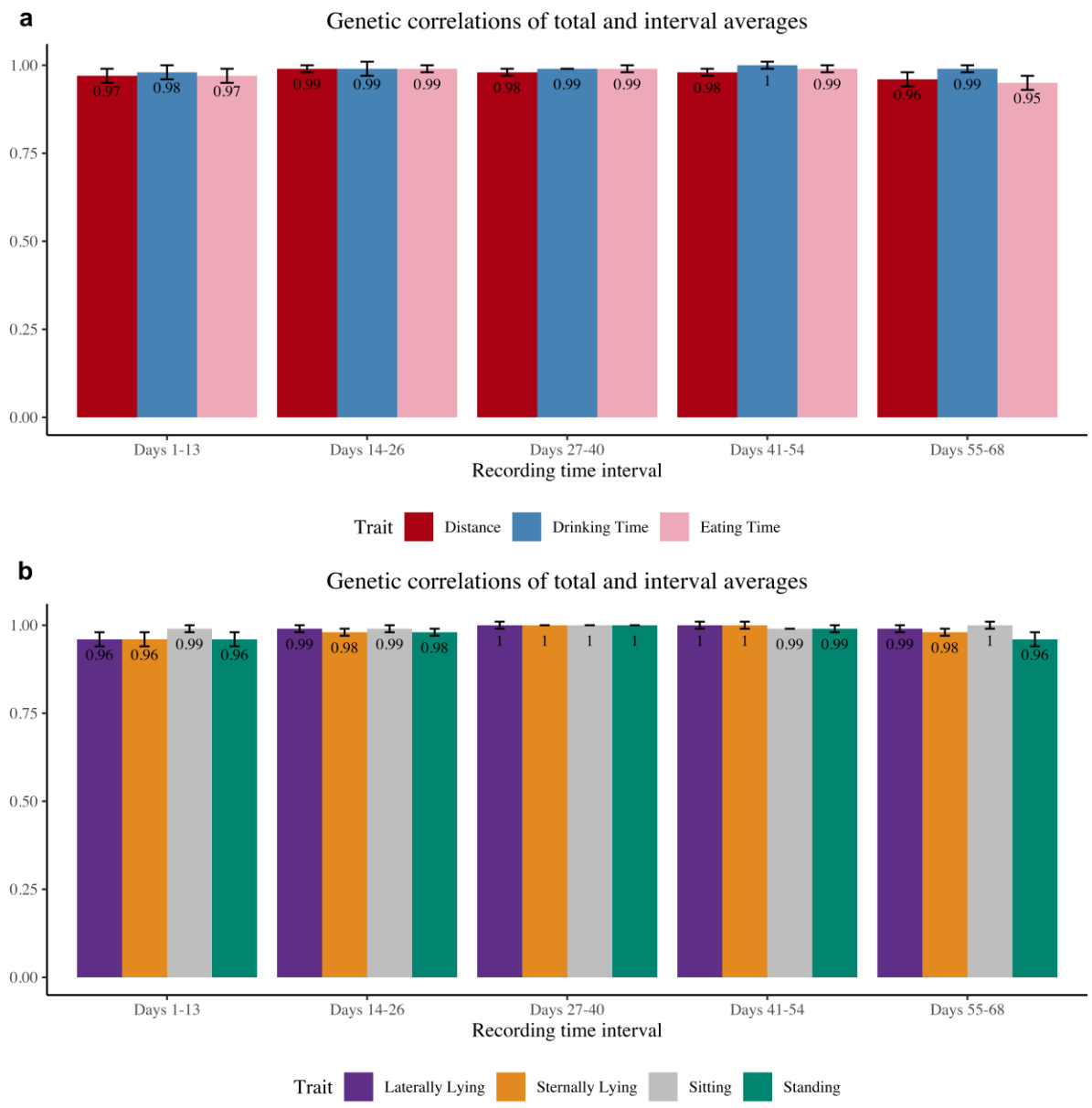

Figure 6.7 Estimates of genetic correlations (standard errors) between the total average and the average of each recording time interval for each behavior trait.

All traits are expressed in min, expect for distance which is expressed in m. (a) Traits: distance traveled, drinking time, eating time and (b) Traits: laterally lying time, sternally lying time, sitting time, standing time.

a

Genetic correlations between ADG and behavior trait time intervals							
Time Period	**Eating Time**	**Drinking Time**	**Lateral Lying**	**Sternal Lying**	**Sitting**	**Standing**	**Distance**
All	0.14 ± 0.23	0.32 ± 0.21	0.50 ± 0.14	-0.10 ± 0.15	0.26 ± 0.15	-0.56 ± 0.11	-0.57 ± 0.10
Days 1-13	0.36 ± 0.17	0.41 ± 1.75	0.48 ± 0.21	-0.06 ± 0.16	0.32 ± 0.26	-0.41 ± 0.10	-0.55 ± 0.14
Days 14-26	0.23 ± 0.18	0.45 ± 0.36	0.52 ± 0.25	-0.01 ± 0.24	0.16 ± 0.15	-0.40 ± 0.10	-0.47 ± 0.12
Days 27-40	0.12 ± 0.21	0.35 ± 0.27	0.49 ± 0.15	-0.22 ± 0.22	0.21 ± 0.15	-0.43 ± 0.09	-0.46 ± 0.11
Days 41-54	-0.05 ± 0.14	0.27 ± 0.23	0.50 ± 0.66	-0.05 ± 0.14	0.30 ± 0.20	-0.51 ± 0.08	-0.63 ± 0.14
Days 55-68	-0.09 ± 0.21	0.05 ± 0.66	0.55 ± 0.19	-0.13 ± 0.16	0.36 ± 0.24	-0.55 ± 0.10	-0.70 ± 0.11

b

Genetic correlations between BF and behavior trait time intervals							
Time Period	**Eating Time**	**Drinking Time**	**Lateral Lying**	**Sternal Lying**	**Sitting**	**Standing**	**Distance**
All	0.18 ± 0.07	0.21 ± 0.11	0.19 ± 0.09	-0.04 ± 0.09	-0.01 ± 0.08	-0.17 ± 0.07	-0.26 ± 0.09
Days 1-13	0.22 ± 0.07	0.29 ± 0.12	0.26 ± 0.10	-0.22 ± 0.13	0.11 ± 0.12	-0.10 ± 0.07	-0.23 ± 0.14
Days 14-26	0.23 ± 0.08	0.37 ± 0.19	0.21 ± 0.16	-0.18 ± 0.16	-0.03 ± 0.09	-0.05 ± 0.07	-0.11 ± 0.10
Days 27-40	0.19 ± 0.08	0.21 ± 0.13	0.20 ± 0.11	-0.08 ± 0.10	-0.01 ± 0.07	-0.14 ± 0.08	-0.21 ± 0.14
Days 41-54	0.11 ± 0.08	0.14 ± 0.10	0.14 ± 0.09	0.05 ± 0.09	-0.02 ± 0.08	-0.22 ± 0.09	-0.34 ± 0.13
Days 55-68	0.05 ± 0.08	0.02 ± 0.16	0.13 ± 0.09	0.14 ± 0.10	-0.04 ± 0.09	-0.28 ± 0.08	-0.37 ± 0.10

c

Genetic correlations between LD and behavior trait time intervals							
Time Period	**Eating Time**	**Drinking Time**	**Lateral Lying**	**Sternal Lying**	**Sitting**	**Standing**	**Distance**
All	0.03 ± 0.10	-0.07 ± 0.12	-0.12 ± 0.11	0.09 ± 0.12	0.20 ± 0.13	-0.38 ± 0.10	-0.47 ± 0.12
Days 1-13	0.19 ± 0.11	0.15 ± 0.13	0.20 ± 0.14	-0.01 ± 0.14	0.27 ± 0.18	-0.25 ± 0.12	-0.34 ± 0.16
Days 14-26	0.09 ± 0.10	0.08 ± 0.14	0.26 ± 0.17	0.10 ± 0.20	0.11 ± 0.13	-0.30 ± 0.12	-0.36 ± 0.15
Days 27-40	-0.01 ± 0.09	-0.12 ± 0.12	0.28 ± 0.12	0.10 ± 0.14	0.12 ± 0.12	-0.38 ± 0.10	-0.43 ± 0.15
Days 41-54	0.00 ± 0.11	-0.10 ± 0.13	0.20 ± 0.15	0.14 ± 0.13	0.22 ± 0.14	-0.42 ± 0.13	-0.58 ± 0.16
Days 55-68	-0.12 ± 0.11	-0.38 ± 0.32	0.25 ± 0.17	0.08 ± 0.14	0.32 ± 0.25	-0.44 ± 0.11	-0.57 ± 0.21

Figure 6.8 Estimates of genetic correlations (standard errors) between production traits, and time intervals or total averages of each behavior trait.

The cells with green circles denote a positive correlation, whereas the cells with red circles denote a negative correlation and the size of the circle indicates the strength of the correlation. The cells without color indicate a range of genetic correlations that passed through 0.0 when the standard error was considered. (a) ADG, (b) BF and (c) LD.

CHAPTER 7

CONCLUSIONS

Continuous improvement in genomic predictions and genetic parameter estimations is needed as the amount of data increases, technology improves, and populations change. Methods to enhance genetic gain can be improved by increasing accuracy, eliminating bias, speeding up computing time, and incorporating more information into genetic evaluations. This dissertation provided a new algorithm for blending the genomic relationship matrix that decreased computing time, investigated the biological understanding of independent chromosome segments in terms of explaining additive genetic variation, tested a heritability estimation method with genomics, which resulted in faster estimation, and explored behavior traits captured by digital phenotyping and their relationship with production traits.

The results showed substantial improvement in genomic prediction and genetic parameter estimation. The findings from the study on method R with genomics show that heritability can be estimated for virtually any size model for which genetic evaluation is feasible. Capturing the genetic variation in a population is possible when all the independent chromosome segments are explained. However, the independent chromosome segments were not well represented in this study. The opportunity to have more information on more traits and more animals exists by incorporating digital phenotyping into production systems. Additionally, with the improved blending algorithm, large genetic evaluations will be obtainable in less time. The optimization of

methods for genomic predictions will enhance the quality of genetic evaluations, leading to more precise selection decisions and increased genetic gain.