GENETICS OF HEAT STRESS IN PIGS WITH FOCUS ON GENOMIC EVALUATIONS USING LARGE NUMBER OF GENOTYPED ANIMALS

by

BRENO DE OLIVEIRA FRAGOMENI

(Under the Direction of Ignacy Misztal)

ABSTRACT

The American pork industry experiences seasonal losses caused by heat stress. Genomic information can help to better identify heat-tolerant animals; however, heat stress evaluation requires complicated models. Single-step GBLUP (ssGBLUP) can be used for genomic selection with complex models and when only a fraction of the animals are genotyped. This method accounts for phenotype, pedigree, and genotype in a unified and simple approach; however, ssGBLUP has a limitation on the number of genotyped animals; it relies on direct inversion of the genomic relationship matrix (**G**); however, inverting a matrix has high computing cost, which creates a bottleneck. The number of genotyped animals is increasing at a fast rate for livestock species, and ssGBLUP would become unfeasible for more than 150,000 genotyped animals. The objective of the first study was to use genomic information to help mitigate problems associated with heat stress in the pork industry. Identifying a threshold for heat stress and including genomic information in the genetic evaluation increased the accuracy of prediction in production traits; therefore, ssGBLUP can be used to help mitigate the impact of heat stress on the US pork industry. The

objective of the second study was to test a recursive algorithm, called algorithm for proven and young animals (APY), to compute the inverse of **G** in an efficient manner. In APY the genotyped population was divided into proven and young, and recursions were based on proven animals. In a simulated study with 25,000 genotyped animals, there was no significant difference between accuracy of GEBV obtained with regular or APY ssGBLUP, which indicate APY can successfully replace the direct inversion of **G**. A third study aimed to compare genomic predictions from regular and APY ssGBLUP for the US Holstein population; 100,000 genotyped animals were used in the study. Correlations of GEBV between the methods were greater than 0.99 when at least 10,000 animals were considered proven in the recursions. In general, genomic information can help mitigate problems due to heat stress in livestock species, and when the amount of genomic information is large, APY should be used in ssGBLUP to remove computing limitations.

INDEX WORDS: Single-step GBLUP, genomic selection, genotype by environment interaction, algorithm for proven and young, genomic recursions

GENETICS OF HEAT STRESS IN PIGS WITH FOCUS ON GENOMIC EVALUATIONS USING LARGE NUMBER OF GENOTYPED ANIMALS

By

BRENO DE OLIVEIRA FRAGOMENI

M.V.D., Universidade Federal da Bahia, Brazil, 2009

M.S., Universidade Federal de Minas Gerais, Brazil, 2012

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

2015

GENETICS OF HEAT STRESS IN PIGS WITH FOCUS ON GENOMIC EVALUATIONS USING LARGE NUMBER OF GENOTYPED ANIMALS

By

BRENO DE OLIVEIRA FRAGOMENI

Major Professor: Ignacy Misztal

Committee: Romdhane Rekaya
J. Keith Bertrand
Kent Grey

Eletronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2015

DEDICATION

To my parents Luiz and Katia.

ACKNOWLEDGEMENTS

I would like to show my sincere appreciation for:

Dr. Ignacy Misztal for taking me on as his student, but more importantly for trusting and guiding me during my PhD, and for letting me be a part of the very important projects I worked on. It was truly enlightening to be a part of this research group.

All the committee members for accepting to be a part of my dissertation: especifically to Dr. Romdhane Rekaya for the great classes and suggestions during the seminars, Dr. Keith Bertrand for guiding me as a TA of his course, and Dr. Kent Gray for the internship and opportunity to learn at SPG (Smithfield Premium Genetics).

Dr. Daniela Lourenço, Dr. Ignacio Aguilar and Dr. Shogo Tsuruta for all the suggestions, corrections and thoughts shared with me. As well as, to all the colleagues from the Animal and Dairy Science department and Animal Breeding and Genetics group: there are so many, it is impossible to cite all of them here, but in special to El Hamidi, Rafael Medeiros, Dennis Richardson, Sreten Andonov, Yutaka Masuda, Ivan Pocrnic and Heather Bradford.

To all the friends I made during this time in Athens, GA. And to all my friend and professors in Brazil, in special to Dr. Thereza Bittencourt, Dr. José Aurélio Bergmann, Dr. Fabio Toral, and Dr. Luiz Fernando Massa (in memoriam) who made themselves present even from far away.

To my family, for the love and support, especially my parents, Luiz and Katia, and my wife, Mariana.

Thank you all very much for everything.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

With the advances in genomics during the past decade, both time and cost for genotyping individuals were drastically reduced. Dense single nucleotide polymorphism (SNP) chips are now more affordable and accessible to industry and producers. When a fraction of a livestock population is genotyped, accounting for that information in genetic evaluations allows for more accurate evaluations; hence, young animals can be selected earlier, reducing the generation interval and increasing genetic change. Since genomic information became available, two main methods were developed to account for it: 1) multistep, where results from the traditional genetic evaluation are used as input to estimate individual marker effects; after that genomic values are calculated as the sum of the effects for markers observed in each animal; and finally, genomic values are blended with parent average to compound the genomic EBV. 2) single-step genomic BLUP, where the genomic information is combined with phenotypes and pedigree in a unique genetic evaluation. Single-step genomic BLUP (ssGBLUP) is the method of choice because of simplicity of use; it is just a BLUP that was modified to account for genomic information. Therefore, ssGBLUP uses the same models as in traditional evaluations, including multiple-trait, random regression, and repeatability models.

Because of economic impacts and the growing interest in environmental and animal welfare issues, seasonal losses due to heat stress are a big concern for the livestock industry. A

sustainable, economical, and straightforward way of dealing with this issue is to select robust animals, which have better performance and are less stressed in extreme environments. Genomic information is now available for nearly all livestock species and can help to more accurately identify the best animals for a specific objective like heat tolerance. Yet, genetic evaluation for heat stress relies on complex statistical models, such as multiple-trait or random regression. Single-step GBLUP has been reported as the most appropriate method for genomic evaluations, especially in such situations.

Because of favorable and promising results obtained with genomic selection and further reduction in genotyping costs and time, the number of animals with genomic information has been increasing drastically, especially in the last three years. This increase has been so large that for some populations, US Holstein being the best example, all genotyped animals cannot be included in ssGBLUP evaluations because of computational limitations. Single-step GBLUP relies on direct inversion of the genomic relationship matrix (**G**) and the pedigree-based relationship matrix among genotyped animals ($\mathbf{A}_{22}$), but inverting a matrix has a cubic computing cost. Consequently, in order use genomics to improve predictions for traits of interest, including performance during heat stress in virtually any population, a more efficient algorithm is needed for obtaining the inverse of the genomic relationship matrix. Therefore, the objectives of these studies were: 1) to use ssGBLUP to perform genomic evaluation for populations experiencing heat stress; 2) to test an algorithm that reduces costs of obtaining the inverse of the genomic relationship matrix in simulated data; 3) to use the aforementioned algorithm for genomic evaluation in a USA Holstein population and compare results with those from regular ssGBLUP.

CHAPTER 2

LITERATURE REVIEW

Using Genomic Information into Genetic Evaluation

The performance of an individual is a sum of its genotype, the environmental conditions, and the interaction between the two. Performance was traditionally improved by selecting individuals based on their genetic merit predicted from phenotypes and pedigree. Harvestein et al. (2003) showed that the considerable differences in size between a chicken line from 1957 and a commercial chicken line in 2001 primarily resulted from genetics. Within the same line of thought, Hill (2008) presented a review on the effects of genetic improvement for different traits in several species.

After the first draft of the human genome project in 2001, there was a promise that genomic information would became available for livestock species and this information would help to improve traditional genetic evaluations using phenotypes and pedigree (Meuwissen et al., 2001). Naturally, with the advances in genotyping techniques and cost reduction, genomic information can now be widely used in a form of dense single nucleotide polymorphism (SNP) map. Dense map information accounts for linkage disequilibrium (LD) between SNP markers and a possible causative gene or quantitative trait loci (QTL). Meuwissen et al. (2001) suggested methods for marker assisted selection, also known as genomic selection (GS), which consist in estimating

genomic breeding values (GEBV) using information on markers in LD with QTL located across the entire genome. The advantages of genomic selection include higher accuracy for estimating breeding values, especially when not many phenotypes are available (Meuwissen et al., 2001)), shorter generation intervals (Konig et al., 2009) and reduced costs for progeny testing (Schaeffer, 2006).

Meuwissen et al. (2001) suggested three different methods to predict GEBV, and each method assumes a different prior distribution for marker variance. BayesA assumed an inverted chi-square distribution for the variance of each marker. Alternatively, BayesB used a similar prior for variances, but allowed the distribution to have a large proportion of markers with zero effect. The third method assumed a prior of normal distribution and constant variance for all markers in a BLUP-like approach[1] .

The aforementioned Bayesian methods were reviewed by Gianola et al. (2009) and modifications regarding the priors were proposed. In order to overcome limitations stated by Gianola et al. (2009), Habier et al (2011) proposed a method called BayesC, which was similar to BayesB except all markers have the same variance. A method that estimates the probability of markers with null effect was called BayesCπ. Several authors proposed additional methodologies or modifications to existing ones, such as Bayesian Lasso (de los Campos et al., 2009), or variable selection (Verbyla et al., 2009).

The BLUP-like genomic method described in Meuwissen et al. (2001) was later described as GBLUP by Habier et al. (2007) and VanRaden (2008). This model is equivalent to a traditional

[1] In Meuwissen et al. (2001) it was named BLUP, but in this document it will be referred to as GBLUP.

BLUP where the numerator relationship matrix (**A**) is substituted by a realized relationship matrix, based on SNP markers (VanRaden, 2008):

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

where Z is a matrix of genotypes (-1 and 1 for homozygous and 0 for heterozygous) centered by twice the observed frequency of the second allele as a difference from 0.5, and $2\sum p_i(1-p_i)$ is a scaling factor, which makes **G** to be analogous to **A**.

Different definitions of **G** were proposed including different scaling factors (Gianola et al., 2009) and individual markers weights (Leutenegger et al., 2003, Amin et al., 2007, VanRaden 2008). The genomic relationship matrix is non-positive definite when the number of markers is smaller than number of animals, or in the presence of clones (animals sharing the same markers). Because of this limitation, usually **G** is blended with a small portion of **A** causing the matrix to be invertible and can also account for polygenic variability (VanRaden, 2008). Another strategy for a full-rank **G** is to add a constant to the diagonal, so $\mathbf{G} = \mathbf{G} + (\mathbf{I}\lambda)$. For blending the usual notation is:

$\mathbf{G}^* = (\alpha\mathbf{G} + \beta\mathbf{A}) + I\lambda$, and $\alpha = 1-\beta$

In general, the above methods can directly utilize only genotyped animals. Phenotypes on ungenotyped animals can be considered indirectly, by creating pseudo-phenotypes for genotyped animals. The most common pseudo-phenotypes are de-regressed breeding values (Garrick et al., 2009) and daughter yield deviations (DYD) (VanRaden and Wiggans, 1991).

Genetic Evaluation with genotyped and ungenotyped animals – Single-Step

Generally, an entire livestock population cannot be genotyped because of cost, availability of animals (culling, dead, foreign animals, etc.), and computation burden. Then, a method that jointly considers information on genotyped and ungenotyped animals was developed. Misztal et al. (2009) proposed a single-step method were **A** was augmented by the genomic relationship matrix. Legarra et al. (2009) proposed to condition the genetic value of ungenotyped animals (index 1) on the genetic value of genotyped animals (index 2); so, the ungenotyped animals could benefit from the genotyped animals through pedigree relationships. The observed or realized relationship matrix (**H**) that describes the joint variance of genetic values for genotyped and ungenotyped was presented as:

$$\mathbf{H}=\begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}+\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} & \mathbf{G} \end{bmatrix}$$

The inverse of **H** replaces the inverse of **A** in the BLUP mixed model equations, and BLUP turns into single-step genomic BLUP (ssGBLUP) as shown by Misztal et al. (2009):

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{H}^{-1}\boldsymbol{\lambda} \end{bmatrix}\begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix}=\begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

where **b** and **u** are vectors of fixed effects and random animal effect, **X** and **Z** are the incidence matrices for the effects contained in **b** and **u**.

Although **H** is difficult to be obtained, its inverse is very simple and was derived by Christensen and Lund (2010) and Aguilar et al. (2010):

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

Among several advantages, ssGBLUP is suitability for multiple trait evaluations, uses raw phenotypes instead of pseudo-phenotypes, and avoids double counting of phenotypic and pedigree information (VanRaden and Wright, 2013, Legarra et al., 2014).

In ssGBLUP $\mathbf{G}$ is scaled to match $\mathbf{A_{22}}$ for a better conditioning of $\mathbf{H}^{-1}$, which improves convergence rate, reduces bias, and increases accuracies. Including all the possible scaling parameters, $\mathbf{H^{-1}}$ is represented as:

$$\mathbf{H}^{-1}=\mathbf{A}^{-1}+\begin{bmatrix}0 & 0\\ 0 & \tau((\alpha\mathbf{G}+\beta\mathbf{A}_{22})+\mathbf{I}\lambda)^{-1}-\omega\mathbf{A}_{22}^{-1}\end{bmatrix}$$

where α and β are used to make **G** positive definite; λ is used to address the non-random genotyping strategies due to selection (Vitezica et al., 2011). Additionally, $\mathbf{G^{-1}}$ and $\mathbf{A}_{22}^{-1}$ can be scaled by τ and ω, respectively, to adjust for different pedigree length for genotyped animals (Misztal et al., 2013); changing ω to lesser values reportedly improved convergence rates and controlled bias (Tsuruta et al, 2013).

Overall, ssGBLUP has been successfully implemented in different species in several countries, such as the US (Tsuruta et al., 2013) and Israeli (Lourenco et al., 2014a) dairy cattle, beef cattle (Lourenco et al., 2015), broiler chicken (Chen et al., 2011), and pigs (Lourenco et al., 2014b), usually with similar or better accuracies than multistep methods.

Heat Stress and Genomic Evaluation

Single-step GBLUP can be applied to problems with complicated modeling and scenarios where other GS strategies are hard to apply. A good example for that is when genotype-by-environment interaction needs to be accounted for by genetic evaluations, which is the case of selection for heat tolerance. The heat stress is a major concern in the industry because it causes losses estimated between $1.69 and $2.36 billion per year (St-Pierre et al., 2003). The losses may increase in the future because of constant selection for production traits and the antagonism between production and heat tolerance; as stated by Misztal and Lovendal (2012, p. 291): "*The more productive a genotype the more sensitive it is to deviation from its optimal environment*". Research has been conducted in experimental conditions to account for the impact of heat stress in different species and populations; however, predicting the effect of heat stress on livestock production in practical conditions is still necessary (Renaudeau et al., 2011). Once the actual impact of heat stress is well described, the problem can be addressed by identifying superior sires and dams that will transmit genetics to tolerate hot environments to the next generation.

A heat load function needs to be defined in order to predict the effect of heat stress on animals' performance. The actual measurement of heat stress of an animal could be performed by rectal temperature or skin surface temperature, temperature near or inside the farm, or other individual measurements, but this information is expensive and/or complicated to measure in large populations. Therefore, the impact of heat stress can also be measured on regularly recorded traits, such as milk production, body weight, and reproductive traits. A linear regression of phenotypes on a temperature index can then be fitted in order to predict the relationship between performance and weather conditions. According to Bohmanova et al. (2007) combining relative humidity with

temperature index in a temperature humidity index (THI) increases the prediction power of heat stress models. Depending on the trait and species, the length of exposure to stressful conditions can also affect the heat load function. Dairy cattle milk production in a given day is influenced by temperature and humidity in the previous 1 or 2 days (Ravagnolo et al, 2000); on the other hand, final weight in pigs can be influenced by temperature and humidity from 10 weeks before the data recording day (Zumbach et al, 2008a).

Once THI is computed, a heat load function can be calculated. The heat load function intends to measure the amount of heat exceeding a given threshold. This threshold value can be interpreted as the limit temperature of the thermic comfort zone:

$$HL = \max(0, THI - THI_T)$$

where THI was the observed THI value for a given day, THI_T was a threshold of THI value, and HL the heat load value. Once the heat load is defined, it can be addressed in a genetic analysis in mainly two different ways:

1) In multiple-trait model approach, phenotypes collected under heat stress condition are addressed as a unique trait, which is correlated with phenotypes collected during cooler weather conditions. Heritability may differ by trait, and the genetic correlation between traits may be less than unity. This method was used by Zumbach et al. (2008b) for modelling heat stress for carcass weight in pigs. This approach has limitations when there are few observations for the trait during heat stress or when animals with phenotypes in different environments are not well tied through the pedigree, for example.
2) In reaction norm analysis, the environmental effect is treated as continuous, and phenotypes are regressed on this continuous scale. In animal breeding and genetic evaluation context, the

reaction norm is best used if fitted with a random regression model (Schaeffer, 2004). Random regression was tested for evaluation of US dairy cows by Ravagnolo and Misztal (2000) and Aguilar et al. (2009), in final weight for pigs in the US by Zumbach et al. (2008b), and for production and fertility traits in Nordic dairy cattle by Kolmodin et al. (2002), among other studies. For this type of model, it is often necessary to use a large quantity of data to accurately estimate genetic values and parameters.

Genomic selection, as mentioned before, can help to increase accuracy of evaluation and to reduce generation interval, which is beneficial for the livestock industry; therefore, using genomic information for heat stress evaluation is expected to help to better identify heat tolerant animals. Moreover, the two main approaches commonly used to address heat stress and genotype-by-environmental interactions depend on complex models with repeated measurements, multiple traits, and complex (co)variance structure. The suitable methodology of choice for genomic evaluation in such cases should be ssGBLUP.

Single-Step GBLUP with a Large Number of Genotyped animals

Because of promising results, successful implementation by the industry, and dramatic reductions in genotyping cost, genomic selection has become the standard method for genetic evaluations. Consequently, the number of animals with genomic information available had an explosive increase in last few years (Cooper et al., 2014). In dairy cattle, more than 950,000 US Holsteins had been genotyped in the United States as of November 2015 (CDCB; https://www.cdcb.us/Genotype/cur_freq.html). Additionally, more than 80,000 Angus cattle have already been genotyped in US (Lourenco et al., 2015). The current implementation of ssGBLUP

has a limitation regarding the number of genotyped animals that can be used for evaluations with a maximum of 150,000 genotypes (Aguilar et al., 2013); therefore, ssGBLUP could not be used for the genomic evaluation of economically important traits, including heat stress, in US Holstein and in the near future for American Angus as well.

This limitation exists because ssGBLUP relies on the direct inversion of the genomic relationship matrix (**G**), which has a cubic computing cost based on the number of genotyped animals. Several methods where suggested to avoid inverting **G**: Legarra and Ducroq (2012) proposed to use unsymmetric equations; Fernando et al. (2013) suggested a SNP model with imputation of genotypes for ungenotyped animals; and Legarra and Misztal (2008), and Liu et al. (2014) proposed fitting a model with SNP effects for genotyped animals. None of these methods could be implemented due to convergence problems, high computing costs, or programming difficulties.

Recently, Misztal et al. (2014) suggested a recursive method to create $\mathbf{G}^{-1}$ without explicitly inverting **G**. The algorithm was based on dividing the genotyped population into two categories, proven and young animals, and had similar procedures as in the method proposed by Henderson (1976) and Quaas (1988) to create $\mathbf{A}^{-1}$. This algorithm was called APY (algorithm for proven and young animals).

According to Henderson (1976) and Quaas (1988), the breeding value of an animal *i* is given by the average of its parents' breeding values plus a random deviation known as Mendelian sampling:

$$u_i = \frac{u_{si} + u_{di}}{2} + \varphi_i$$

where: u_i = breeding value for animal *i*, u_{si} and u_{di} = breeding values for sire and dam of animal *i*, respectively, and φ_i = Mendelian sampling for animal *i*. As a result, the breeding value of an animal is conditioned solely on the breeding value of its parents:

$$u_i | \, u_1, u_2, \ldots, u_{i-1} = u_i | u_{si}, u_{di}$$

Consequently, the inverse of the average relationship matrix is given by the following recursion:

$$\mathbf{A}^{-1} = (\mathbf{I} - \mathbf{P})' \, \mathbf{M}^{-1} (\mathbf{I} - \mathbf{P})$$

Where: $\mathbf{A}^{-1}$ = the inverse of average relationship matrix, $\mathbf{I}$ = identity matrix, $\mathbf{M}^{-1}$ = the inverse of Mendelian sampling diagonal matrix; and $\mathbf{P}$ = matrix that relates animals to their parents and has 2 non-zero elements (for animals' sire and dam).

When genotypes are available, the conditional distribution changes because the genomic covariance among individuals is built under identical by state rules. In this way, animals that share the same alleles are related. Therefore, breeding values of a given genotyped animal are conditioned to the breeding values of all previous genotyped animals. When all animals are genotyped, i.e. GBLUP, the solutions for a given animal will be given by (Misztal, 2014):

$$u_i = \sum_{j=1}^{i-1} p_{ij} u_j + \varepsilon_i$$

Where: p_{ij} = relationship between animals *i* and *j*, and ε_i = Mendelian sampling.

In GBLUP, animals without phenotypes (called young for this recursion) do not contribute to other animals' breeding values; animals with records were called proven. Therefore, the

genotyped population can be split into proven and young, and the breeding value for animal *i* could be written as:

$$u_i = \sum_{j=young} p_{ij}u_j + \sum_{j=proven} p_{ij}u_j + \varepsilon_i$$

However, in GBLUP genotyped animal without phenotypes do not contribute and u_i can be simplified to:

$$u_i = \sum_{j=proven} p_{ij}u_j + \varepsilon_i$$

In this way, the breeding value of an animal is conditioned to the breeding value of all proven animals; the inverse of the genomic relationship matrix can be written as:

$$\mathbf{G}^{-1} = (\mathbf{I} - \mathbf{P})' \mathbf{M}^{-1} (\mathbf{I} - \mathbf{P})$$

The matrix **P** does not have to be dense anymore, which will lead to a dramatic reduction in computational costs, without compromising predictions in GBLUP. Finally, Misztal (2014) presented the following formula for APY $\mathbf{G}^{-1}$:

$$\mathbf{G}^{-1}_{apy} = \begin{bmatrix} \mathbf{G}^{-1}_{pp} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{G}^{-1}_{pp}\mathbf{G}_{py} \\ \mathbf{I} \end{bmatrix} \mathbf{M}^{-1}_{g} \begin{bmatrix} -\mathbf{G}_{py}\mathbf{G}^{-1}_{pp} & \mathbf{I} \end{bmatrix}$$

Where: $\mathbf{G}^{-1}_{pp}$ = the inverse of genomic relationship matrix among proven animals, $\mathbf{G}_{py}$ = partition of genomic relationship matrix relating proven and young animals, and $m_{g,i} = g_{i,i} - \mathbf{G}_{ip}\mathbf{G}^{-1}_{pp}\mathbf{G}_{pi}$.

The inverse of genomic relationship matrix, if well implemented, can have linear and quadratic increasing costs for young and proven animals, respectively, which would allow

inversion of this matrix for a large number of animals. If there is small or no loss in GEBV accuracy for ssGBLUP using APY instead of the direct inversion of **G**, ssGBLUP could be the method of choice for genetic evaluations with millions of genotyped animals.

REFERENCES:

Aguilar, I., A. Legarra, S. Tsuruta, and I. Misztal. 2013. Genetic evaluation using unsymmetric single step genomic methodology with large number of genotypes. Interbull Bull. 47:222–225.

Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128:422–428.

Aguilar, I., I. Misztal, and S. Tsuruta. 2009. Genetic components of heat stress for dairy cattle with multiple lactations. Journal of dairy science, 92(11):5702-5711.

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta, and T.J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science, 93(2):743-752.

Amin, N., C. M. van Duijn, and Y. S. Aulchenko. 2007. A genomic background based method for association analysis in related individuals. PLoS ONE 2:e1274.

Bohmanova, J., I. Misztal, and J.B. Cole. 2007. Temperature-humidity indices as indicators of milk production losses due to heat stress. Journal of dairy science, 90(4):1947-1956.

Chen, C. Y., I. Misztal, I. Aguilar, S. Tsuruta, T. H. E. Meuwissen, S. E. Aggrey, T. Wing, and W. M. Muir. 2011. Genome-wide marker-assisted selection combining all pedigree phenotypic

information with genotypic data in one step: an example using broiler chickens. J. Animal Sci. 89:23-28.

Christensen, O. and Lund, M. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol., 42(1):2.

Cooper, T. A., Wiggans, G. R., & VanRaden, P. M. 2014. Including cow information in genomic prediction of Holstein dairy cattle in the US. Proceedings of 10th WCGALP, Vancouver, Canada.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K.,and Cotes, J. M. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics, 182(1):375-385.

Fernando, R.L., J.C.M. Dekkers, and D. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Gen. Sel. Evol. 46:50.

Garrick, D. J., Taylor, J. F., & Fernando, R. L. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses.Genet Sel Evol, 41(55), 10-1186.

Gianola D, Delos Campos G, Hill WG, Manfredi E, Fernando R: Additive genetic variability and the Bayesian alphabet. Genetics 2009, 183:347–363.

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

Habier D, Fernando RL, Kizilkaya K, Garrick DJ: Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 2011, 12:186.

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

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. 2011. Extension of the Bayesian alphabet for genomic selection. BMC bioinformatics, 12(1):186.

Havenstein, G. B., P.R. Ferket, and M. A. Qureshi. 2003. Growth, livability, and feed conversion of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets. *Poultry Science* 82(10):1500-1508.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator rela-tionship matrix used in prediction of breeding values. Biometrics, 32(1):69-83.

Hill, W. G. 2008. Estimation, ectiveness and opportunities of long term genetic improvement in animals and maize. Lohmann Information, 43(1):3-20.

Kolmodin, R., Strandberg, E., Madsen, P., Jensen, J., & Jorjani, H. 2002. Genotype by environment interaction in Nordic dairy cattle studied using reaction norms. Acta Agriculturae Scandinavica, Section A-Animal Science, 52(1):11-24.

Konig, S., Simianer, H., and Willam, A. 2009. Economic evaluation of genomic breeding programs. J. Dairy Sci., 92(1):382-391.

Legarra, A. and V. Ducrocq. 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in single-step best linear unbiased prediction. J. Dairy Sci. 95:4629–4645.

Legarra, A., Aguilar, I., and Misztal, I. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci., 92(9):4656-4663.

Legarra, A., O.F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. Livest. Prod. Sci. 166:54–65.

Leutenegger, A.-L., B. Prum, E. Genin, C. Verny, A. Lemainque, F. Clerget-Darpoux, and E. A. Thompson. 2003. Estimation of the inbreeding coefficient through use of genomic data. Am. J. Hum. Genet. 73:516–523.

Liu, Z., M.E. Goddard, F. Reinhardt, and R. Reents. 2014. A single-step genomic model with direct estimation of marker effects. J. Dairy Sci. 97:5833–5850.

Lourenco, D. A. L., I. Misztal, S. Tsuruta, I. Aguilar, E. Ezra, M. Ron, A. Shirak, and J. I. Weller. 2014a. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. J. Dairy Sci. 97:1742-1752.

Lourenco, D. A. L., I. Misztal, S. Tsuruta, I. Aguilar, T. J. Lawlor, S. Forni, and J. I. Weller. 2014b. Are evaluations on young genotyped animals benefiting from the past generations? J. Dairy Sci. 97:3930-3942.

Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, and I. Misztal. 2015. Genetic evaluation using single-step genomic BLUP in American Angus. J. Anim. Sci. doi:10.2527/jas2014-8836.

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

Misztal I., Legarra A., Aguilar I. (2009) Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci., 92, 4648–4655.

Misztal, I. and Lovendal P. 2012. Genotype by Environment Interactions in Commercial Populations. Page 291 in Environmental Physiology of Livestock. Collier R. J. and J. Collie L. Wiley-Blackwell

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

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

Misztal, I., Z. G. Vitezica, A. Legarra, I. Aguilar, and A. A. Swan. 2013. Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed. Genet. 130: 252-258.

Quaas, R. L. 1988. Additive genetic model with groups and relationships. J. Dairy Sci. 71:1338–1345.

Ravagnolo, O., & Misztal, I. 2000. Genetic component of heat stress in dairy cattle, parameter estimation. Journal of Dairy Science, 83(9):2126-2130.

Renaudeau, D., Gourdine, J. L., & St-Pierre, N. R. 2011. A meta-analysis of the effects of high ambient temperature on growth performance of growing-finishing pigs. Journal of Animal Science, 89(7):2220-2230.

Schaefer, L. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim.Breed. Genet., 123(4):218-223.

Schaeffer, L. R. 2004. Application of random regression models in animal breeding. Livestock Production Science, 86(1), 35-45.

St-Pierre N.R., B. Cobanov, and G. Schnitkey. 2003. Economic losses from heat stress by US livestock industries. Journal of Dairy Science 86:Suppl E52 - 77

Tsuruta, S., I. Misztal, and T. J. Lawlor. 2013. Genomic evaluations of final score for US Holsteins benefit from the inclusion of genotypes on cows. J. Dairy Sci. 96:3332-3335.

Tsuruta, S., I. Misztal, and T. J. Lawlor. 2013. Short communication: Genomic evaluations of final score for US Holsteins benefit from the inclusion of genotypes on cows. J. Dairy Sci. 96: 3332-3335.

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

VanRaden, P. M., & Wiggans, G. R. 1991. Derivation, calculation, and use of national animal model information. Journal of Dairy Science, 74(8):2737-2746.

VanRaden, P.M. and J.R. Wright. 2013. Measuring genomic pre-selection bias in theory and in practice. Interbull Bull. 47:147–150.

Verbyla, K. L., Hayes, B. J., Bowman, P. J., and Goddard, M. E. 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet. Res., 91(05):307-311.

Vitezica, Z. G., Aguilar, I., Misztal, I., & Legarra, A. 2011. Bias in genomic predictions for populations under selection. Genetics research, 93(05):357-366.

Zumbach, B., I. Misztal, S. Tsuruta, J. Holl, W. Herring, and T. Long. 2007. Genetic correlations between two strains of Durocs and crossbreds from differing production environments for slaughter traits. J. Anim. Sci. 85:901–908.

Zumbach, B., I. Misztal, S. Tsuruta, J.P. Sanchez, M. Azain, W. Herring, and M. Culbertson. 2008a. Genetic components of heat stress in finishing pigs: parameter estimation. Journal of animal science, 86(9):2076-2081.

Zumbach, B., I. Misztal, S. Tsuruta, J.P. Sanchez, M. Azain, W. Herring, and M. Culbertson. 2008b. Genetic components of heat stress in finishing pigs: Development of a heat load function. Journal of animal science, 86(9):2082-2088.

CHAPTER 3

GENOMIC SELECTION AS A TOOL TO MITIGATE SEASONAL LOSSES IN SWINE PRODUCTION DUE TO HEAT STRESS[1]

[1] B.O. Fragomeni, D.A.L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, Huang. Y., Grey., K. and I. Misztal. To be submitted to Journal of Animal Science[1]

ABSTRACT

ABSTRACT: The purpose of this study was to analyze the impact of seasonal losses due to heat stress in different environments and breed combinations, and to evaluate the ability of the genomic information to improve accuracy of estimated breeding values. Data were available for two different swine populations: purebred Duroc animals raised in Texas and North Carolina, and commercial crosses of Duroc and F1 females (Landrace x Large White) raised in Missouri and North Carolina; pedigree provided links for animals from different states. Genotypes were available for 8,000 purebred animals. Traits were off-test weight for purebred and hot carcass weight for crossbred animals. Weather data was collected at airports located close to the farms, and a heat load function was calculated based on temperature humidity indices. Non-genetic analysis were by regressions of phenotype on heat load or heat load x year. Genetic analyzes were either by a single-trait model ignoring the heat stress, by a multiple-trait model with traits based on regular and "hot" seasons, or by a reaction norm model based on heat load. Breeding values were predicted by BLUP or single-step genomic BLUP (ssGBLUP), when genomic information was included. Variance components were estimated by AIREML. For crossbred animals, the coefficient of determination with heat load was 0.34 and 0.21 for Missouri and North Carolina, respectively, and increased to 0.67 and 0.63 with heat load x year. For purebred animals, the coefficient of determination with heat load alone was close to 0, and increased to 0.44 and 0.46 with heat load x year for North Carolina and Texas, respectively. Heritability for crossbred animals was 0.18 in single-trait models and 0.26 in multiple-trait models, with similar equivalent heritabilities in reaction norm models. The correlations between regular and "hot" traits were close to 1.0 for purebreds and around 0.75 for the crossbreds. Realized accuracies for regular and "hot" traits in crossbreds were 0.41 and 0.36 with BLUP, respectively, and 0.44 and

0.38 with ssGBLUP. The effect of heat stress is greater in crossbreds than in purebreds and varies from year to year. The use of the genomic information increases the accuracy of prediction, but the increase with the current genotyping structure is small.

Key words: Single-step GBLUP, genomic selection, genotype by environment interaction, heat stress

INTRODUCTION

Seasonal impacts in livestock production due to heat stress are observed in different species all over the globe. The pork industry is especially affected because pig physiology is not adapted to dissipate heat by sweating or respiration. In the USA pork industry alone, a loss of $299 million is estimated due to heat stress (St-Pierre et al., 2003). Heat stress can also negatively impact animal welfare, possibly affecting the public perception of farmed pigs. Losses due to heat stress are observed for several traits and include smaller carcass value and poor reproductive performance over heat load conditions. Also, heat stress causes lower feed intake (Collin et al., 2001) and reduced muscle/increased fat composition (Bridges et al., 1998), although the mechanisms for the different tissue conformation during heat stress are still not clear (Pearce et al., 2013). Technological advances were made in order to improve cooling strategies in pig farms, however, the impact of heat stress is still present. The amount of heat produced by an animal from an improved genetic line is higher than that produced by an "old genetics animal" (Brown-Brandl et al., 2001).

Bloemhof et al. (2012) demonstrated that reproductive traits are affected by heat load in pigs from Portugal and Spain. Pigs originated from an efficient "Dutch" line were more susceptible to

heat stress than adapted pigs with “Spanish” origin. Bloemhof et al. (2013) estimated genetic parameters for heat stress effects on farrowing rate in sows. Animals with higher productivity tended to be less heat tolerant; however, selection for heat tolerance was possible.

Zumbach et al. (2008a) described a heat load function for body weight in pigs in the USA based on THI averaged over time. The best function was a sum of THI over 70 degrees during the past 10 weeks of life. Zumbach et al. (2008b) also showed a possibility of genetic evaluation for heat stress. Two models were considered: multiple-trait and a reaction norm, which is equivalent to random regression. While the multiple-trait model was simpler, the random regression model allowed for better accounting of heat stress. The study pointed out that the accuracy of evaluation for heat tolerance was low due to a limited number of phenotypes under heat stress.

The availability of high-density DNA markers led to development of methods that could utilize that information in animal breeding (Meuwissen et al., 2001). With genomic information, it is possible to evaluate animals with higher accuracy and earlier in life; reducing generation interval and increasing genetic gain (Schaeffer, 2006). Methods initially developed for commercial genomic evaluation were based on multiple step procedures (Meuwissen et al., 2001) and could not easily be applied to more complex models. Aguilar et al (2010) and Christensen and Lund (2010) developed a method called single-step genomic BLUP (ssGBLUP), which is BLUP with a relationship matrix that combines the pedigree and the genomic information. Single-step GBLUP can be applied for same models as BLUP including those applied for studies in heat stress.

The first objective of the present study was to determine the extent of heat stress in purebred nucleus and commercial crossbred animals in different states in US. The second objective was to determine the utility of the genomic information to identify heat tolerant individuals.

MATERIAL AND METHODS

Data

Data were available for purebred Duroc animals from nucleus farms and for crossbred animals from commercial farms. Crossbred animals were crosses of Duroc sires and F1 Landrace x Large White dams. Animals with conflicts in data, outliers (out of the range of 4 standard deviations), and without weather information available for the data recording (weigh or slaughter) date were removed from the dataset.

Data from purebred animals were collected at farms in North Carolina (NC) and Texas (TX). Phenotypes were available for body weight, collected at age (mean ± standard deviation) of 170±5.19 days in NC from 2003 to 2014 and 168±5.97 days in TX from 2005 to 2014.

Data for crossbred animals were collected in packing plants in North Carolina (NC) and Missouri (MO). Phenotypes were available for hot carcass weight, collected at age (mean ± standard deviation) of 189±13.8 in NC from 2009 to 2014 and 181±11.7 in MO from 2012 to 2014.

Pedigree and Genomic Information

Pedigree file was available for 313,121 Duroc animals and 227,043 crossbred animals. For crossbred animals, dams were identified but dams' pedigrees were unavailable. Even though data was available from two states in each dataset, there was a strong relationship among animals; 30% of all animals had siblings in different states, and approximately the same

percentage of sires had progeny in two states. The detailed information about the number of records in each farm is included in Table 1.

Single nucleotide polymorphism (SNP) information was available for 8,232 purebred Duroc animals. In total 4 different SNP chips were used; 4,251 animals were genotyped for 10K chip (Genomic Profiler 10k BeadChip, GeneSeek-Neogen, USA), 2,803 animals had information for 60K (Infinium PorcineSNP60 v2 BeadChip, Illumina Inc., USA), 1,022 had information for 60KV1 chip (Infinium PorcineSNP60 v1 BeadChip, Illumina Inc., USA) and 160 animals had information for 70k chip (PorcineSNP80 BeadChip, GeneSeek-Neogen, USA). Animals were imputed to 60KV1 chip, using software Beagle 3.3.2 (Browning and Browning, 2009) from October 2011.

Weather Data

The R package "WeatherData" (Narasimhan, 2014) was used to collect daily information from weather stations near the farms. Weather information was also available from "High Plains Regional Climate Center" for MO and "State Climate Office of North Carolina" for NC, As correlations of temperatures and humidity between the two sources were above 0.99, only data from the "WeatherData" package was used.

In the "WeatherData" package, the weather information can be either from public weather stations or from airports; However, only airport information was used in this study because of completeness of data, and because correlations between station and airport were always above 0.98 (0.99 for average of 30 days intervals). For NC farms, Wilmington airport was the closest one, for TX farms, Pampa airport, and for MO Des Moines airport (Iowa) was the closest one.

Temperature humidity index (THI) was calculated for each day following the formula (National Research Council, 1971):

$$THI = t - (0.55 - (0.0055 \times rh)) \times (t - 14.5)$$

where t is the observed maximum daily temperature in Celsius , rh is the observed minimum daily humidity, in a 0 to 100 scale and THI is the temperature humidity index for a given day. THI was transformed to Fahrenheit scale.

Once THI was computed, a heat load function was calculated, following the formula:

$$HL = \max(0, THI - THI_T),$$

where THI was the observed temperature humidity index for a given day, THI_T was a threshold of THI value, and HL the heat load value. The heat load function intended to measure the amount of heat exceeding a given threshold. This threshold value can be interpreted as the maximum range of the thermic comfort zone. For testing purposes, this threshold assumed values from 60 to 77 degrees.

Heat Load Regression

Once heat load values were calculated, a linear regression was fitted in order to find a relationship between the heat load and the phenotype. Data was split into breed (purebred and crossbred) and then into states. Next, average weight (or hot carcass weight) was calculated for each day available, and summarized in a file with date, mean weight, and HL. THI values of 60, 65, 70 and 77 degrees were fitted for heat load. Average HL within 30, 50, and 70 days before weighing (or slaughter) date were considered to account for cumulative effects of heat during the period prior to data collection.

Three different linear models were fitted within each subgroup: 1) regression of phenotypes on heat load; 2) regression of phenotypes on heat load nested in year; and 3) regression of phenotypes on heat load separately for each year. Additionally, data was also summarized by average weight (or carcass weight) for groups of HL value, and linear and quadratic regressions were fitted to analyze the relationship between overall weight (or HCW for crossbreds) and HL. The HL groups were between 0 and maximum HL for each data set, and a different group was specified every 0.5 degree of HL.

Genetic Analyses

Analyses were carried separately for purebred and crossbred animals, under similar models. Data were first analyzed in a single-trait animal model without heat stress information:

$$y_{ijkl} = cg_i + gender_j + litter_k + b * age + animal_l + e_{ijkl}$$

where y_{ijkl} = phenotype (offtest weight for purebred analysis and hot carcass weight for crossbred) of animal *l* in contemporary group *i* of gender *j* in litter *k*; *age* = age of the animal in weigh in (or slaughter) date in days, *b* = regression coefficient for age; and e_{ijkl} = residual effects. The model can be written in matrix notation as:

$$\mathbf{y = Xb + Wl + Za + e}$$

where y is a vector of phenotypes, **X** is the incidence matrix of fixed effects contained in *b*: contemporary group, gender and age as a covariate; **W** is the incidence matrix of random litter effects contained in *l*; **Z** is the incidence matrix of random animal effects contained in *a*, and *e* is a vector or random residuals. Variances were:

$$var\begin{pmatrix} a \\ l \\ e \end{pmatrix} = \begin{bmatrix} \mathbf{A}\sigma_a^2 & 0 & 0 \\ 0 & \mathbf{I}\sigma_l^2 & 0 \\ 0 & 0 & \mathbf{I}\sigma_e^2 \end{bmatrix}$$

where **A** is the numerator relationship matrix, **I** is an identity matrix, and $\sigma_a^2, \sigma_l^2, and\ \sigma_e^2$ are variances for additive genetic direct, litter, and residual effect, respectively.

In a multiple trait model, trait one was body weight (or hot carcass weight) in heat stress conditions defined as average THI for 30 days was above a threshold of 70 or 78 degrees, and trait two were the remaining observations. Trait one could be called "under heat stress" and trait two could be called as either "regular" or "non-heat stress". In matrix notation, the multiple-trait model was the same as the single-trait model but the variances were:

$$var\begin{bmatrix} l_{HS} \\ l_{NHS} \end{bmatrix} = \mathbf{I} \otimes \begin{bmatrix} \sigma_{l_{HS}}^2 & 0 \\ 0 & \sigma_{l_{NHS}}^2 \end{bmatrix},$$

$$var\begin{bmatrix} a_{HS} \\ a_{NHS} \end{bmatrix} = \mathbf{A} \otimes \begin{bmatrix} \sigma_{a_{HS}}^2 & \sigma_{aHS,NHS} \\ \sigma_{aNHS,HS} & \sigma_{a_{NHS}}^2 \end{bmatrix}, \text{ and}$$

$$var\begin{bmatrix} e_{HS} \\ e_{NHS} \end{bmatrix} = \mathbf{I} \otimes \begin{bmatrix} \sigma_{e_{HS}}^2 & 0 \\ 0 & \sigma_{e_{NHS}}^2 \end{bmatrix}$$

where $\sigma_{i_{HS}}^2, \sigma_{i_{NHS}}^2, \sigma_{i_{HS,NHS}}$ are variances of effect *i* of a trait under heat stress, a trait not under heat stress, and covariance between trait under heat stress and not under heat stress, respectively.

The third model used only for crossbred animals was a random regression model:

$$y_{ijkl} = cg_i + gender_j + litter_k + b * age + \sum_{q=1}^{2} a_{lq} * z_{lq} + \sum_{q=1}^{2} c_{iq} * z_{iq} + e_{ijkl}$$

where y_{ijkl} = hot carcass weight of animal *l* in contemporary group *i* of gender *j* in litter *k*; *age* = age of the animal in slaughter date in days, *b* = regression coefficient for age, a_{lq}= random regression coefficient *q* on heat load for additive genetic effect on animal *l*, z_{lq}= *q*th order polynomial for animal *l*, c_{lq}= regression coefficient *q* on heat load for fixed regression for contemporary group *i*, z_{iq}= *q*th order polynomial for contemporary group *i*, and e_{ijkl} = residual effects.

Two different regression coefficients were fitted in the random regression model: 1) ordinary polynomial (intercept and linear), which is equivalent to a reaction norm model, and b-spline orthogonal polynomials (linear splines with one central knot). The variances were:

$$var\begin{bmatrix} a_{q1} \\ a_{q2} \end{bmatrix} = A \otimes \begin{bmatrix} \sigma^2_{q1} & \sigma_{q1,q2} \\ \sigma_{q2,q1} & \sigma^2_{q2} \end{bmatrix}$$

where $a_{qi}$is the *i*th random regression coefficient; intercept and linear effect of HL or linear spline, with a central knot at 7 degrees of HL.

For litter and residual variance, variances were the same as for the single-trait model.

Variance components for all the models were estimated by average information restricted maximum likelihood methods (AIREML) using the AIREMLF90 software (Misztal et al., 2002). For breeding value estimation the BLUPF90 software (Misztal et al., 2002) was used. Finally, genomic estimated breeding values were estimated by ssGBLUP (Aguilar et al 2010; Christensen and Lund, 2010), which consisted in replacing the inverse of numerator relationship matrix ($\mathbf{A}^{-1}$), in the BLUP mixed model equations, by the inverse of the realized relationship matrix ($\mathbf{H}^{-1}$):

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

where $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix, and $\mathbf{A}_{22}^{-1}$ is the inverse of the numerator relationship matrix for genotyped animals.

For crossbreds, accuracies were the correlation between "true" and estimated breeding value for sires in a testing population. True breeding value had different definitions, all of them were calculated using a dataset that included all crossbred phenotypes available, hereby called complete dataset; those definitions were: 1) TBV-E: EBV with complete dataset; 2)TBV-G: GEBV with complete dataset; 3) PYD-E: Progeny yield deviation (PYD) using pedigree BLUP; 4) PYD using ssGBLUP. For cross-validation purposes, EBV were calculated with incomplete data, where phenotypes were removed for animals born in 2013 and later. Sires in the test population were born in 2012 and later, with recorded progeny and with EBV reliability above 0.75 in the complete dataset; in this way, no phenotypic information of progeny of sires in the test population was available when estimated breeding values were computed. PYD was calculated according to VanRaden et al. (1991) with complete data.

For purebred animals, the ability to predict future performance was measured by the correlation between the phenotype corrected for the fixed effects (y-Xb) and the breeding value for genotyped animals born in 2013 and 2014. Prediction accuracy was calculated as correlation divided by the square root of the heritability of the trait (Legarra et al., 2008).

RESULTS

Descriptive statistics of data across states and breeds are shown in Table 1. On purebred animals small fluctuations on weight were observed, additionally a higher proportion of animals on THI above 70 and 77 were observed in Texas when compared to North Carolina. For crossbred data,

animals in North Carolina tend to be slaughtered older than in Missouri, even though their hot carcass weight is on average lighter. A higher proportion of animals raised in higher temperatures was observed in Missouri.

THI averaged for 30 days had a similar pattern during the summer in NC, MO, and TX (Figure 1). During the winter the behavior was fairly different. An arbitrary line at the value of THI equal 75 degrees was traced to favor visual inspection of data.

When a regression of weight on heat load across time was fitted, the best value for R^2 was achieved for average heat load on 30 days before data collection and above 70 degrees (Figure 2), in both states for crossbred animals. For purebred animals it was not possible to find any reasonable trend or predictive power in those equations (Figure 3). In an extra regression fitted with heat load effect nested in year, R^2 improved remarkably, for both breeds and across states (Figures 4 and 5); best values were still for average HL of 30 days and above 70 degrees. For purebred animals an extra regression was fitted with year effect only (Figure 6), which achieved R^2 values slightly lower than heat load nested in year (Figure 5). When data was divided into years and regression was independently fitted (Figure 7), a very similar pattern from Figure 5 was observed (this analysis was only performed for purebred animals).

Finally, linear and quadratic regressions of heat load on weight were fit. A decay in body weight was observed as heat load increased in both farms for crossbred animals (Figure 8), but for purebred animals it was only noticeable in NC (Figure 9). The best fit was achieved with quadratic regression.

Heritability estimates from a single-trait model was 0.24 for purebred and 0.18 for crossbred animals. For the two-trait model that accounted for heat stress, an increase in heritability for the

trait under heat stress was observed when the threshold was above 78 degrees in crossbreds (Table 2). When the threshold for THI was lower than 78 degrees or analyses were performed in purebred data, heritability was similar in both traits or slightly lower for the heat stress trait. Genetic correlations were lower when higher threshold for heat stress were tested.

For linear random regression models, similar variances were observed; which led to initial decrease in heritability and posterior increase, after a minimum at heat load equal to 7; as both results were similar only results with b-splines are shown (Figure 10). Likewise, genetic correlations had similar values in both analyses, where two clusters were observed (Figure 11).

Prediction accuracy for single-trait analysis was higher for genomic EBV (GEBV) than EBV (Table 3). With two-trait analysis, there was a trend of GEBV to achieve higher accuracy than EBV, but it changed as the definition of TBV for the trait under heat stress changed. For the trait under regular weather conditions, prediction accuracy of GEBV was always higher than of EBV. For the random regression model, accuracies were higher for GEBV when TBV was calculated with genomic information (Figure 12); an increase of EBV accuracy was observed as heat load also increased when TBV was based on traditional analysis. In general, accuracy increased with heat load. For purebred animals, the predictability increased from 0.19 with traditional BLUP to 0.60 with ssGBLUP.

DISCUSSION

The graphs of temperature humidity indexes suggest that there are periods of possible heat stress during the summer in all investigated states and years. Bohmanova et al. (2006) described THI affecting production in dairy cattle, with a similar behavior from what is described here. Huyn et

al. (2005) showed physiological changes in fattening pigs starting at 22ºC and Quiniou et al. (2001) showed that 24ºC was the lower critical temperature; these results slightly differ from 21ºC (THI=70) in the present study, mainly because we considered a 30 days average THI instead of the actual temperatures. Zumbach et al. (2008a) in a similar study modeled the heat load curve with THI above 65. Values in the present study are slightly different from the literature, however, they are in the range described by different authors; nevertheless it was expected as temperature and humidity were not collected on farm sites. Also, information about the actual indoor climate the animals were exposed to was not available, but it is likely to be highly correlated to data collected from airport or weather stations, yet THI from on-site data might be lower due to cooling strategies used inside the buildings.

It was possible to observe a decrease in hot carcass weight (crossbred data) during the summer when phenotype is plotted over time: there is a valley coinciding with each summer (Figure 2). On the other hand, it was not possible to find any visual association with the purebred data because other effects might be masking the heat load effect. For crossbred animals it was possible to predict phenotypes based only on heat load values. Best fit in equations, measured by R^2, was calculated by using the average heat load for 30 days (results for 50 and 70 days were not shown). This finding differs from those of Zumbach et al. (2008a), where the best period was 70 days.

The threshold temperature for calculating heat load with the highest R^2 was 70 degrees. For purebred data, HL effect alone could not predict phenotypes, and R^2 values were close to 0 in all analyses (Figure 3). Nesting HL within year helped to increase the R^2 value of all regressions for both crossbreds (Figure 4) and purebreds (Figure 5), indicating that the effect of heat stress

varies by year. For crossbred animals, when HL was nested within year, 65 degrees had a better fit for animals in MO, however, differences from other HL definitions were negligible. Findings for crossbreds are similar to those by Zumbach et al. (2008a) with a different heat load function and length. In Zumbach et al. (2008a) and in the present study, the differences in R^2 between different thresholds and length of heat load were modest. For purebreds it was possible to predict the weight loss based on HL nested within year, but it is not clear whether the fit was due to the interaction of year and HL or year alone.

Year effect was fitted alone for purebred animals (Figure 6) to compare with HL nested within year. There was a substantial difference when HL was included in the model, but for some years there was a positive effect of HL, which was not expected. In order to isolate this effect, a different regression of phenotype on HL was fitted for each year. It can be observed that in some years, (2005, 2008, 2011, and 2013 in Texas and 2004, 2008, and 2013 in North Carolina), HL had a positive or null effect (Figure 7); therefore it is still not possible to affirm that there is a consistent heat stress effect acting on purebred animals across different years; furthermore, some of the negative heat load effects in Figure 7 cannot be considered as heat stress effect, but instead an artifact of the model or due to unknown circumstances.

As expected, a decrease in HCW for crossbred animals was observed as HL increased (Figure 8). The slope was sharper when HL was above 3 degrees; it explains the best fit of quadratic regression in comparison to linear regression. For purebreds there was a difference in results from TX and NC: for the former, once again, no heat stress or relationship between HL and body weight was observed, on the other hand, for the latter, the trend was similar to crossbred's trend. This was not expected, given that the prior analysis did not find an unequivocal heat stress effect

for purebreds. It is unclear whether heat stress effect is masked on animals from Texas or if there is some spurious association for North Carolina data; genetics should be similar across farms and THI information does not differ much.

Animals from higher production lines (e.g., purebreds) are expected to be more susceptible to environmental and heat stress effects (Bloemhof et al. 2013). However, the impact of heat stress can be voided by management practices in the nucleus farms. Farms in different states may have different cooling strategies or technologies, which may be reasons for the differences between TX and NC for purebred animals.

Heritabilities from single-trait analysis for HCW were higher than those described by Dufrasne et al. (2013) and similar to those computed by Zumbach et al. (2007). For body weight in purebred Duroc lines, heritabilities were in the range described by Zumbach et al (2007). For the multiple-trait analysis accounting for heat stress in crossbreds, an increase was observed in heritability for the trait under heat stress, similar to what was observed by Zumbach et al. (2008b). Working with heat stress in dairy cattle, Ravagnolo et al. (2000) and Aguilar et al. (2009) showed that traits under heat stress tend to have higher additive genetic variance, probably because selection is based on the trait under regular weather conditions. Bloemhof et al. (2012), showed high variability between daughter-groups of sires in response to high temperatures, which can also be related to the observed increase in heritability for the trait under heat stress. The genetic correlation in this case was positive and moderate to high, but still indicating a difference in the genetic component underlying the phenotypes.

The magnitude of stress should be taken into account when creating the “heat” and “non-heat” traits. In this study we found that using 78 degrees was more appropriate than 70 because the

stress caused on phenotypes at 70 degrees was not of high enough magnitude; lower thresholds did not show differences in heritability and had higher genetic correlation. For purebred animals heritability was not different when observations were divided in two traits, and genetic correlations were very high. With multiple trait model, observations having lower and higher level of heat stress (say 1 and 10 degrees above the threshold) are treated the same.

Random regression approach takes into account differences in the magnitude of heat stress. Subsequently it is possible to better account for the effect of heat stress and plot heritability/correlations as a function of THI above the threshold. In dairy cattle population, Ravagnolo et al. (2000) and Aguilar et al (2009) had similar trends in random regression models for heat stress, where heritability increased with THI. Genetic correlations described by Ravagnolo et al (2000) had a similar pattern with the findings in the present study. Models using splines and linear regressions showed similar curves, but the model with splines is more flexible.

When genomic information was included, the realized accuracies in the single-trait model increased; the increase was greater in purebreds. In general, the increase in accuracy with the genomic information can be large for genotyped but small for ungenotyped animals (Lourenco et al., 2015). In this study only purebred animals were genotyped.

Realized accuracies in the multiple-trait heat stress model were expected to be lower for the heat stress trait as less information was available. However, the opposite was found. This may be due to definition of true breeding values. When little phenotypic information is available, the accuracies are based on similar parent averages. Therefore, TBV for heat stressed animals needs to be treated carefully. Especially when data is scarce, prediction accuracy for heat stress trait may not be reliable. PYD seems to be the most reliable source for computing accuracy as it

reduces the overestimation of accuracy on the trait under heat stress, especially when EBV from complete data is used as the definition of TBV.

In the random regression model, the realized accuracies increased with HL (results not shown). This increase was observed in both definitions of TBV and was even stronger for BLUP when TBV is considered EBV with complete data. This supports the claim of overestimation of accuracy under heat stress conditions. It was also evident that the correlation between EBV and GEBV increased with the heat load. There was a sharp increase in the correlation when HL was above 7 degrees, which is the critical temperature for heat stress according to the two trait model. As observations with HL above 7 degrees are few, the results might be an artifact of b-splines.

CONCLUSION

We found an evidence of heat stress in both purebred and crossbred animals. The effect of heat stress is stronger in commercial farms and its magnitude changes from year to year. The best way to analyze the pig data under heat stress is by a reaction norm model, possibly using splines. While the accuracy of selection for heat stress is moderate due to relatively few animals under severe heat stress, that accuracy can be improved with availability of the genomic data.

REFERENCES

Aguilar, I., I. Misztal, and S. Tsuruta. 2009. Genetic components of heat stress for dairy cattle with multiple lactations. Journal of dairy science, 92(11), 5702-5711.

Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128:422–428.

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta, and T.J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science, 93(2), 743-752.

Bloemhof, S., A. Kause, E.F. Knol, J.A.M. Van Arendonk, and I. Misztal. 2012. Heat stress effects on farrowing rate in sows: genetic parameter estimation using within-line and crossbred models. Journal of animal science 90(7), 2109-2119.

Bloemhof, S., P.K. Mathur, E.F. Knol, and E.H. van der Waaij. 2013. Effect of daily environmental temperature on farrowing rate and total born in dam line sows. Journal of animal science 91(6), 2667-2679

Bohmanova, J., I. Misztal, and J.B. Cole. 2007. Temperature-humidity indices as indicators of milk production losses due to heat stress. *Journal of dairy science*, 90(4), 1947-1956.

Brown-Brandl, T. M., Eigenberg, R. A., Nienaber, J. A., & Kachman, S. D. (2001). Thermoregulatory profile of a newer genetic line of pigs. Livestock Production Science, 71(2), 253-260.

Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics, 84(2), 210-223.

Dufrasne, M.; Misztal, I.; Tsuruta, S.; Holl, J.; Gray, K. A.; Gengler, N. (2013) Estimation of genetic parameters for birth weight, preweaning mortality, and hot carcass weight of crossbred pigs J. Anjim. Sci. Vol. 91 no. 12: 5565-5571

Huynh, T. T. T., A. J. A. Aarnink, M. W. A. Verstegen, W. J. J. Gerrits, M. J. W. Heetkamp, B. Kemps, and T. T. Canh. 2005. Effects of increasing temperatures on physiological changes in pigs at different relative humidities. J. Anim. Sci. 83:1385–1396

Lourenco, D. A. L., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015. Accuracy of estimated breeding values with genomic information on males, females, or both: an example in broiler chicken. Genet. Sel. Evol. 47: 56

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

Misztal I., Tsuruta S., Strabel T., Auvray B., Druet T., Lee D. H. (2002). "BLUPF90 and related programs (BGF90)," in Proceedings of the 7th World Congress on Genetics Applied to Livestock Production Montpellier, Communication No. 28–27

Narasimhan, R. (2014). weatherData: Get Weather Data from the Web. R package version 0.4.1.

http://CRAN.R-project.org/package=weatherData

National Research Council (1971) A guide to environmental research on animals Natl. Acad. Sci., Washington, DC

Ravagnolo, O., & Misztal, I. (2000). Genetic component of heat stress in dairy cattle, parameter estimation. Journal of Dairy Science, 83(9), 2126-2130.

Schaefer, L. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim.Breed. Genet., 123(4):218-223.

Quiniou, N., Noblet, J., Van Milgen, J., & Dubois, S. (2001). Modelling heat production and energy balance in group-housed growing pigs exposed to low or high ambient temperatures. British Journal of Nutrition, 85(01), 97-106.

VanRaden, P. M., & Wiggans, G. R. (1991). Derivation, calculation, and use of national animal model information. *Journal of Dairy Science*, *74*(8), 2737-2746.

Zumbach, B., I. Misztal, S. Tsuruta, J. Holl, W. Herring, and T. Long. 2007. Genetic correlations between two strains of Durocs and crossbreds from differing production environments for slaughter traits. J. Anim. Sci. 85:901–908.

Zumbach, B., I. Misztal, S. Tsuruta, J.P. Sanchez, M. Azain, W. Herring, and M. Culbertson. 2008a. Genetic components of heat stress in finishing pigs: parameter estimation. Journal of animal science, 86(9), 2076-2081.

Zumbach, B., I. Misztal, S. Tsuruta, J.P. Sanchez, M. Azain, W. Herring, and M. Culbertson. 2008b. Genetic components of heat stress in finishing pigs: Development of a heat load function. Journal of animal science, 86(9), 2082-2088.

TABLES

Table 1.1 – Descriptive statistics for purebred Duroc animals in North Carolina and Texas and crossbred Duroc X F1(Landrace x Large white) in North Carolina and Missouri

	Purebred Duroc		Crossbred Duroc x F1	
State	North Carolina	Texas	North Carolina	Missouri
N	151,336	55,897	141,756	86,435
Weight (kg)	117.3 (13.0)	115.0 (13.0)	92.68 (9.5)	95.1 (8.2)
Age	169.9 (5.2)	168.3 (5.9)	188.63 (13.9)	180.5 (10.7)
N THI > 70	75,457	41,122	64,319	22,855
N THI > 78	22,897	10,759	31,723	5,858

N = number of animals, N THI > 70 and N THI > 78 are the number of animals with data collected with THI value above the mentioned numbers.

Table 1.2 – Heritabilities and genetic correlations for purebred and crossbred data obtained by AIREML in single-trait analysis and two multiple trait analysis, with heat stress definition of 70 or 78 degrees Celsius.

	Purebred		Crossbred	
	Heritability	Genetic Correlation	Heritability	Genetic Correlation
Single trait	0.24(0.01)	----------	0.18(0.01)	----------
HS70	0.23(0.01)	0.98(0.01)	0.18(0.01)	0.77(0.02)
NHS70	0.25(0.01)		0.22(0.01)	
HS78	0.23(0.01)	0.99(0.01)	0.26(0.02)	0.72(0.03)
NHS78	0.24(0.01)		0.19(0.01)	

For multiple traits: HS = Heat stress, NHS = non-heat stress,

Table 1.3 – Accuracies as correlation between true breeding value (TBV) and (genomic) estimated breeding value ((G) EBV) for single-trait and multiple-trait analyses. Where TBV definitions were EBV with complete data (TBV-E), GEBV with complete data (TBV-G), progeny yield deviation using genomic (PYD-G) and with traditional BLUP (PYD-E). For the multiple-trait results, trait1 (Hot) is the trait under heat stress and trait 2 (regular) is the trait under regular weather conditions.

		TBV-G	TBV-E	PYD-G	PYD-E
Single Trait	EBV	0.34	0.28	0.29	0.24
	GEBV	0.43	0.33	0.39	0.30
Trait1 Hot	EBV	0.58	0.56	0.41	0.36
	GEBV	0.63	0.52	0.49	0.38
Trait 2 Regular	EBV	0.55	0.52	0.47	0.41
	GEBV	0.61	0.51	0.55	0.44

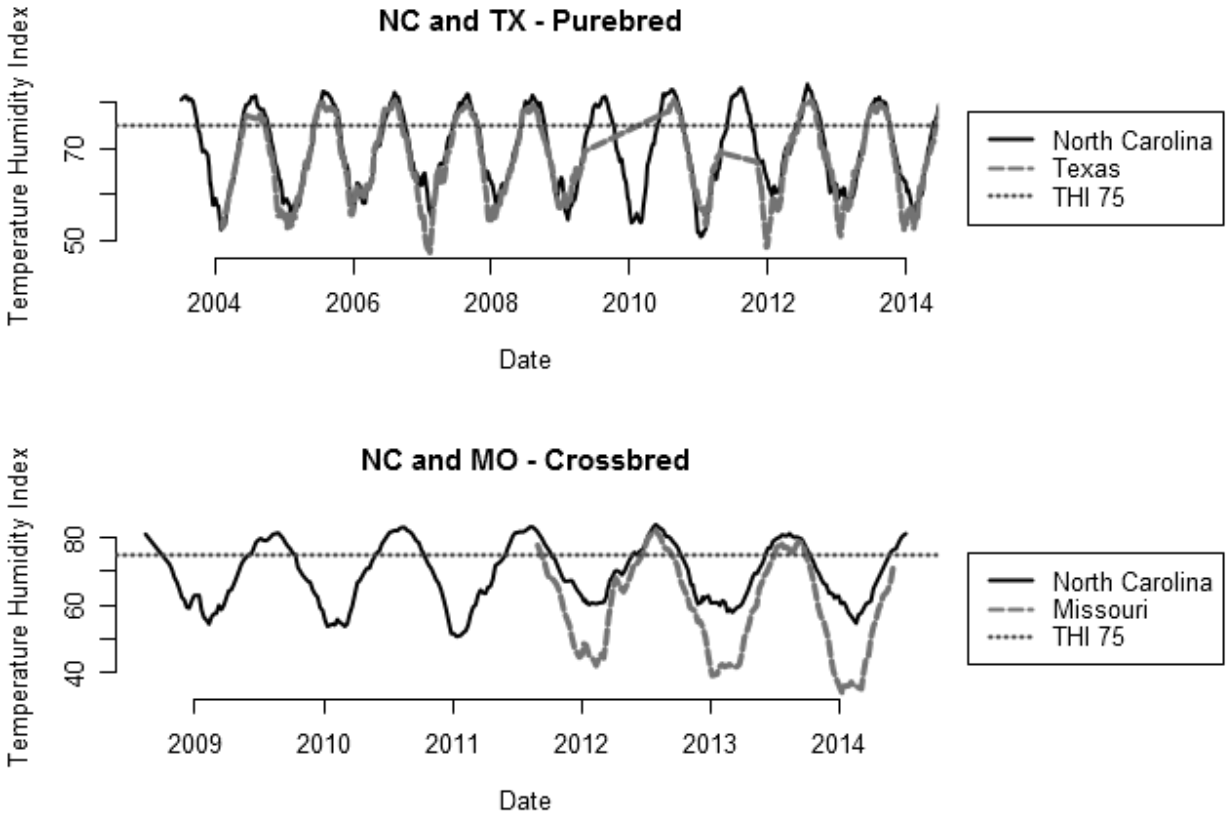

Figure 1.1 – 30 day average temperature humidity index across airport weather stations when data was available, a) for purebred animals in North Carolina and Texas; and b) for crossbred animals in North Carolina and Missouri

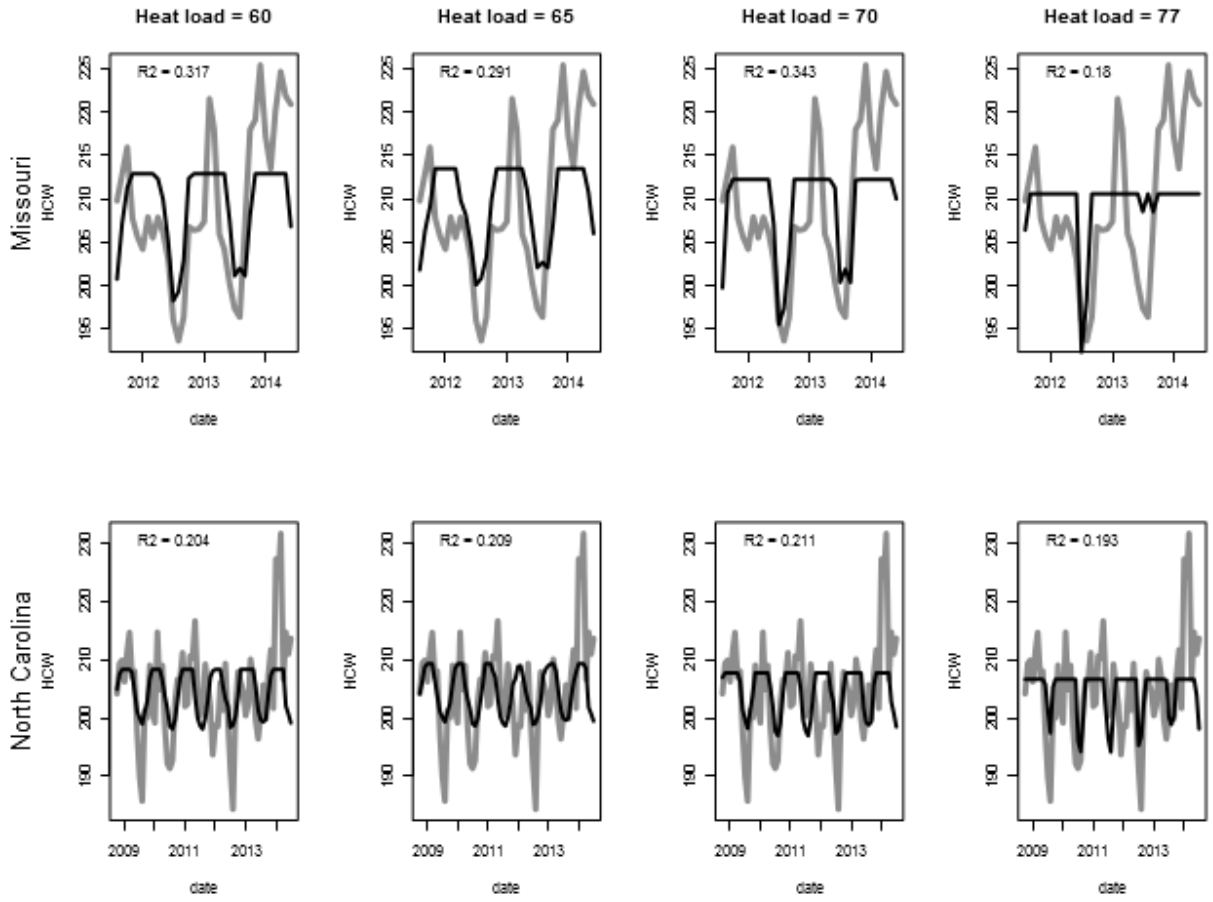

Figure 1.2 – Observed and predicted hot carcass weight by regression of phenotype on heat load, across time for crossbred animals in North Carolina (a) and Missouri (b)

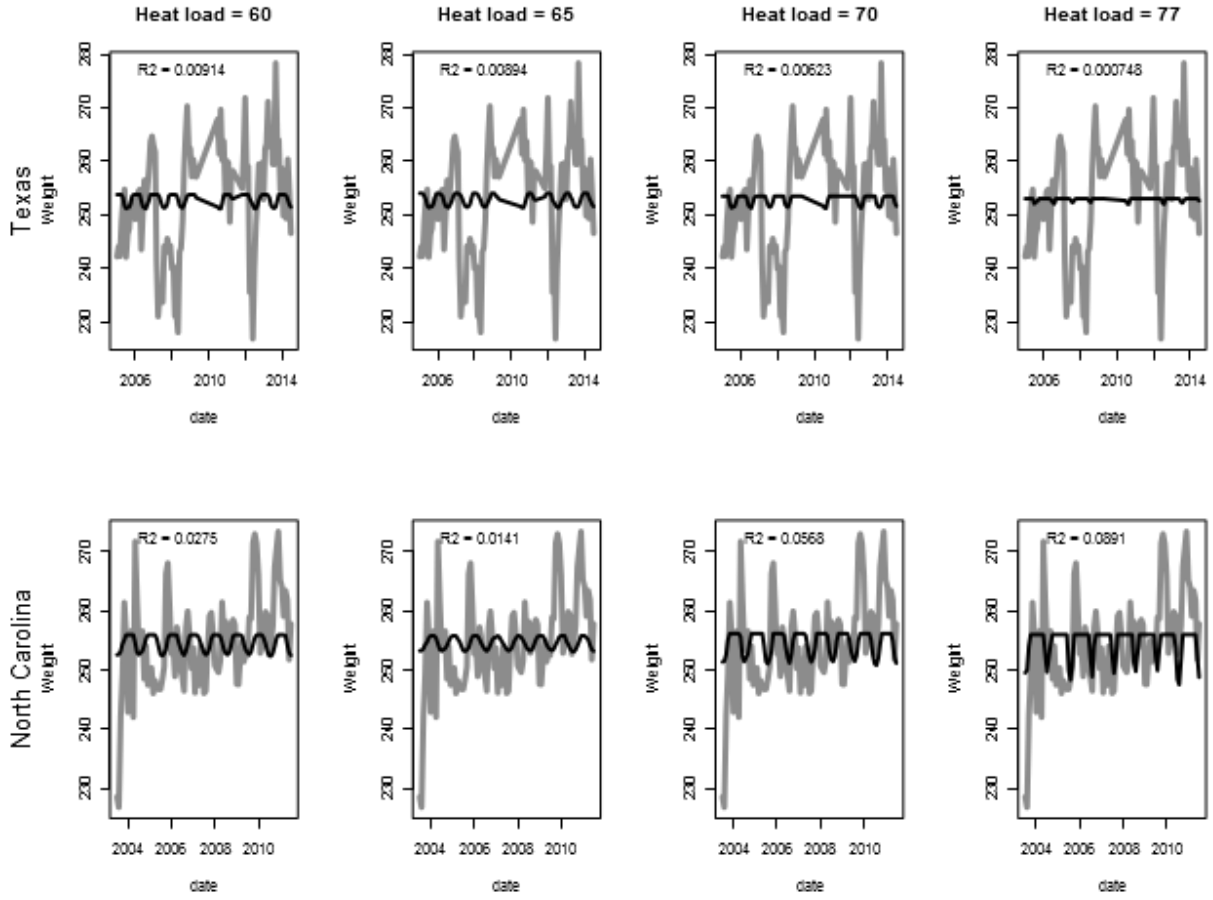

Figure 1.3 – Observed and predicted offtest body weight by regression of phenotype on heat load, across time for purebred Duroc animals in North Carolina (a) and Texas (b)

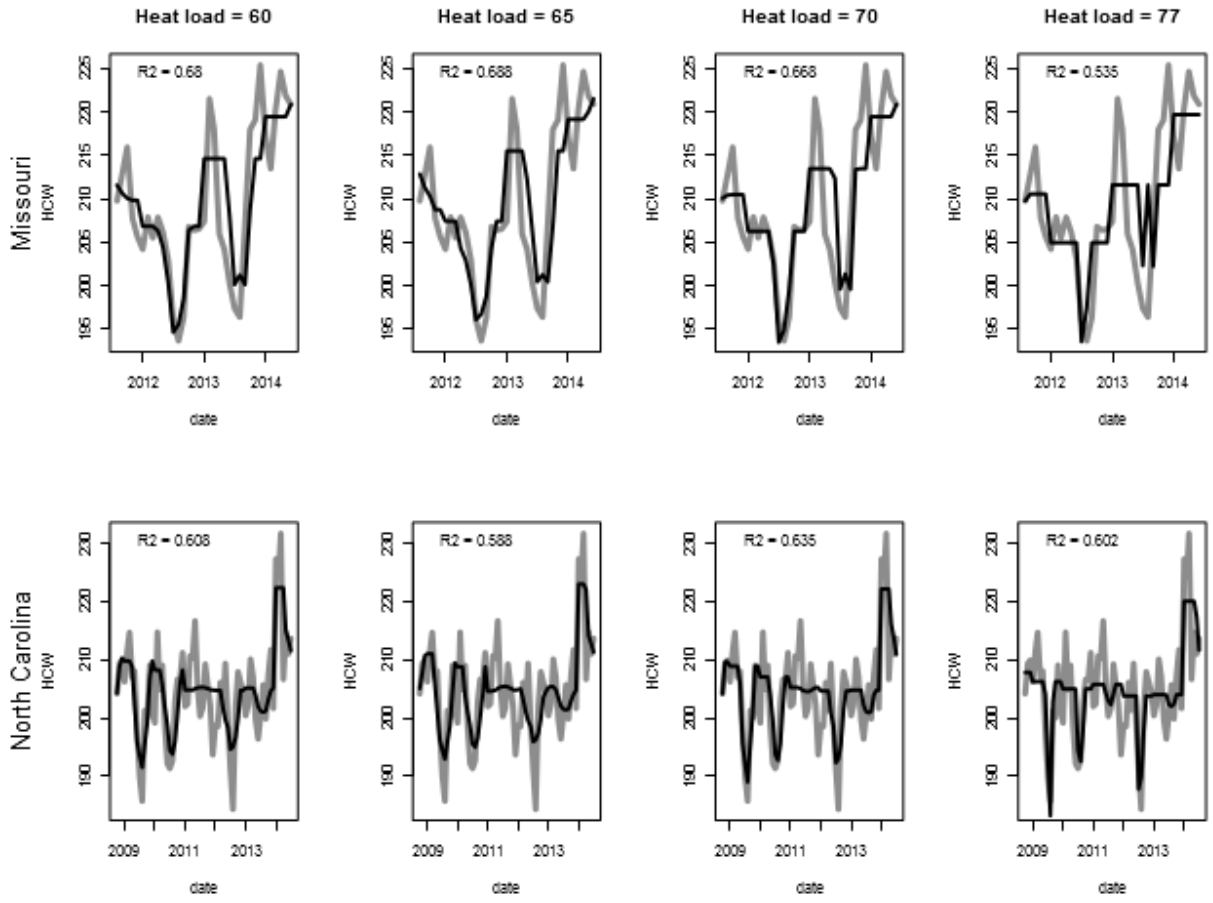

Figure 1.4 – Observed and predicted hot carcass weight by regression of phenotype on heat load nested in year, across time for crossbred animals in North Carolina (a) and Missouri (b)

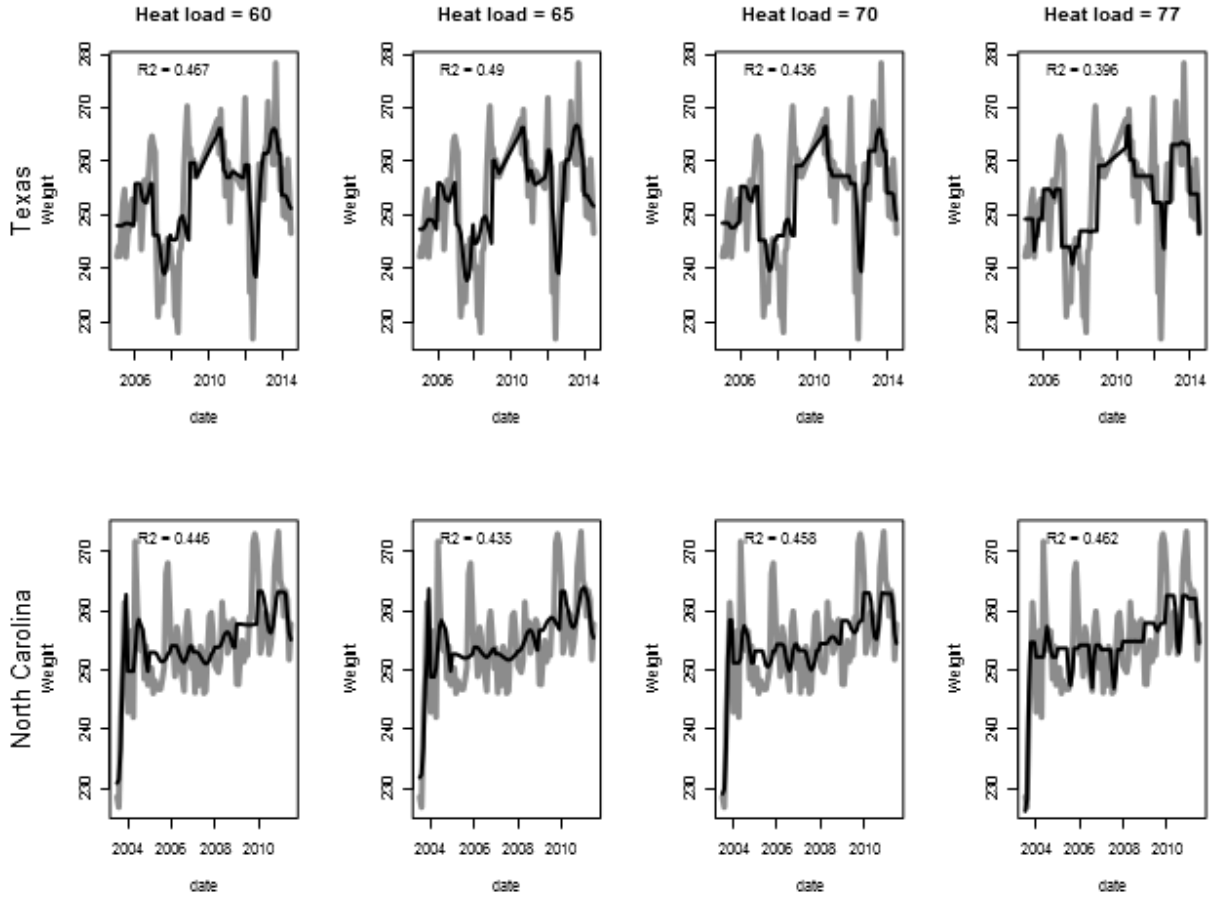

Figure 1.5 – Observed and predicted offtest body weight by regression of phenotype on heat load nested in year, across time for purebred Duroc animals in North Carolina (a) and Texas (b)

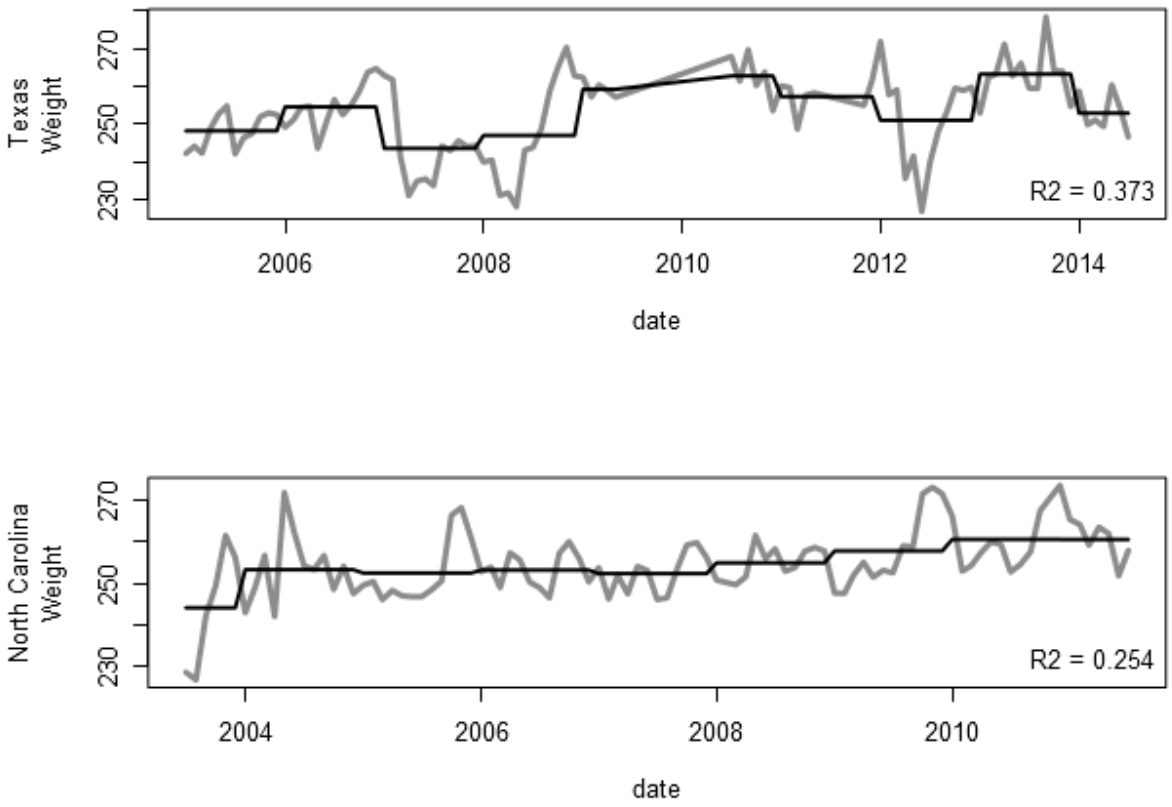

Figure 1.6 – Observed and predicted offtest body weight by regression of phenotype on year, across time for purebred Duroc animals in North Carolina (a) and Texas (b)

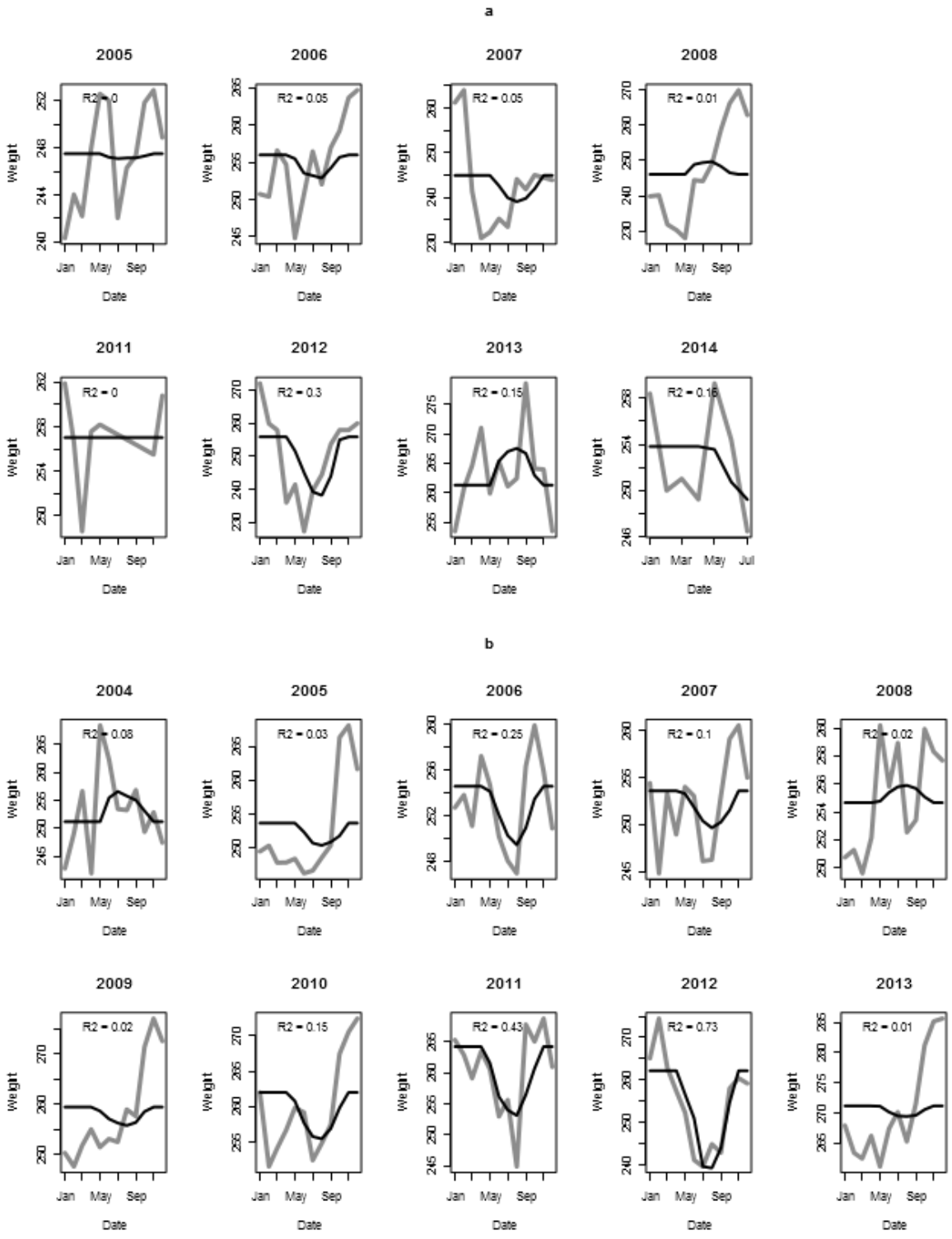

Figure 1.7 – Observed and predicted offtest body weight by regression of phenotype on heat load, across time, within year, for purebred Duroc animals in North Carolina (a) and Texas (b)

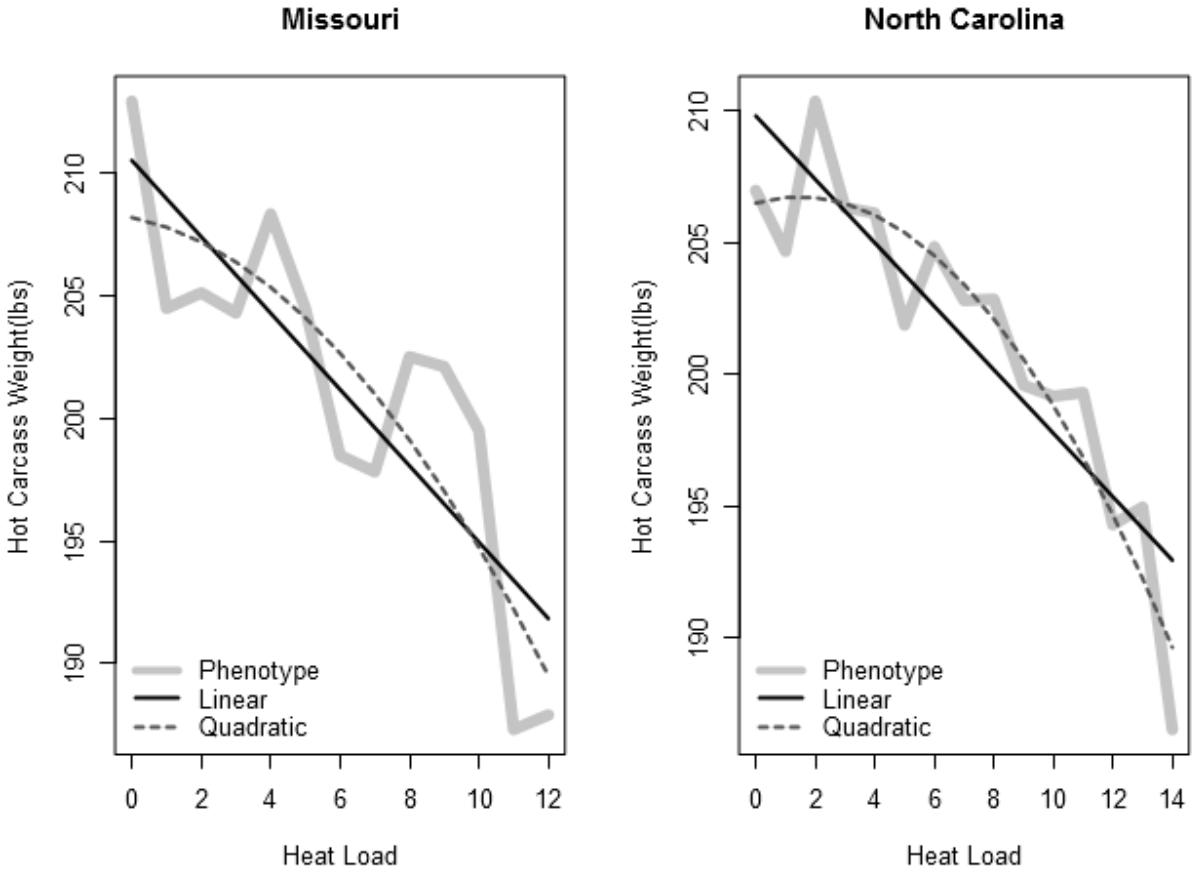

Figure 1.8 – Observed and predicted hot carcass weight by regression of phenotype on heat load, for crossbred animals in North Carolina (a) and Missouri (b)

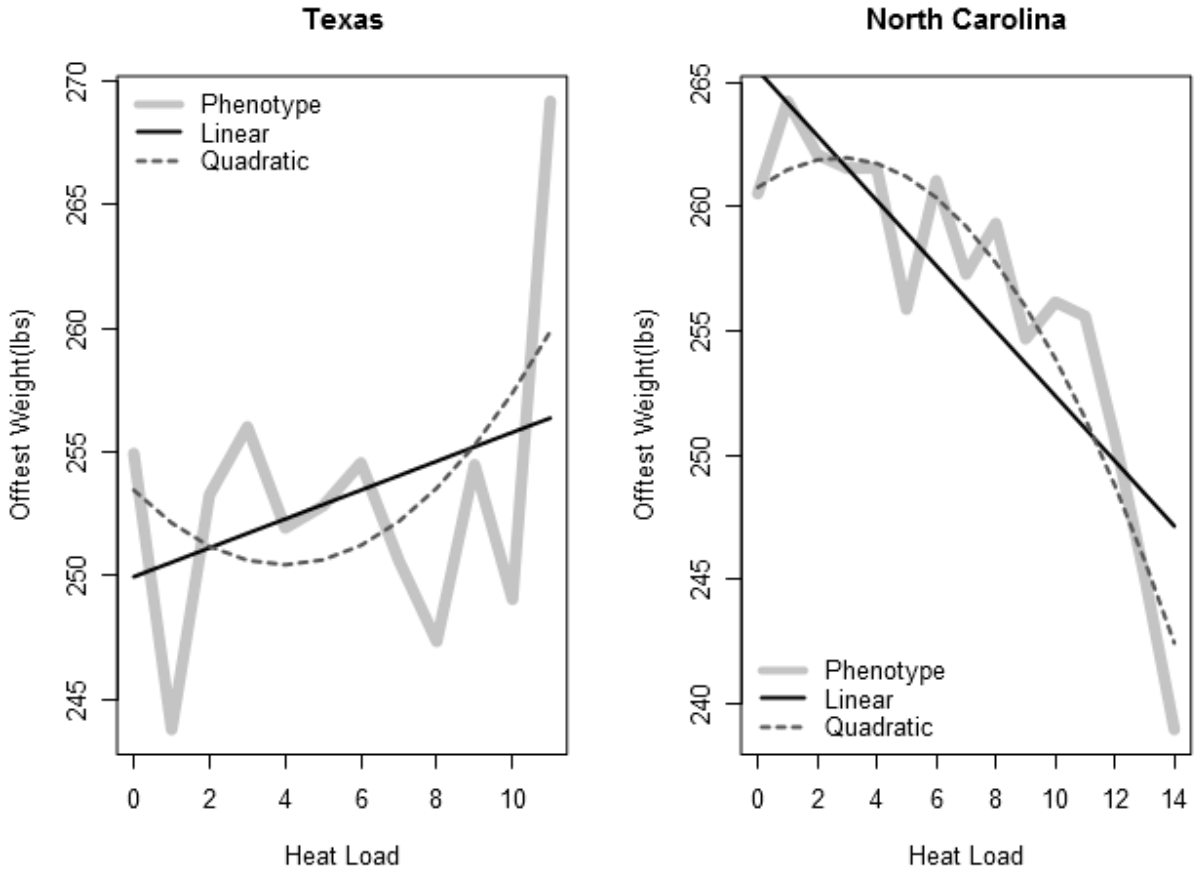

Figure 1.9 – Observed and predicted offtest body weight by regression of phenotype on heat load, for purebred Duroc animals in North Carolina (a) and Texas (b)

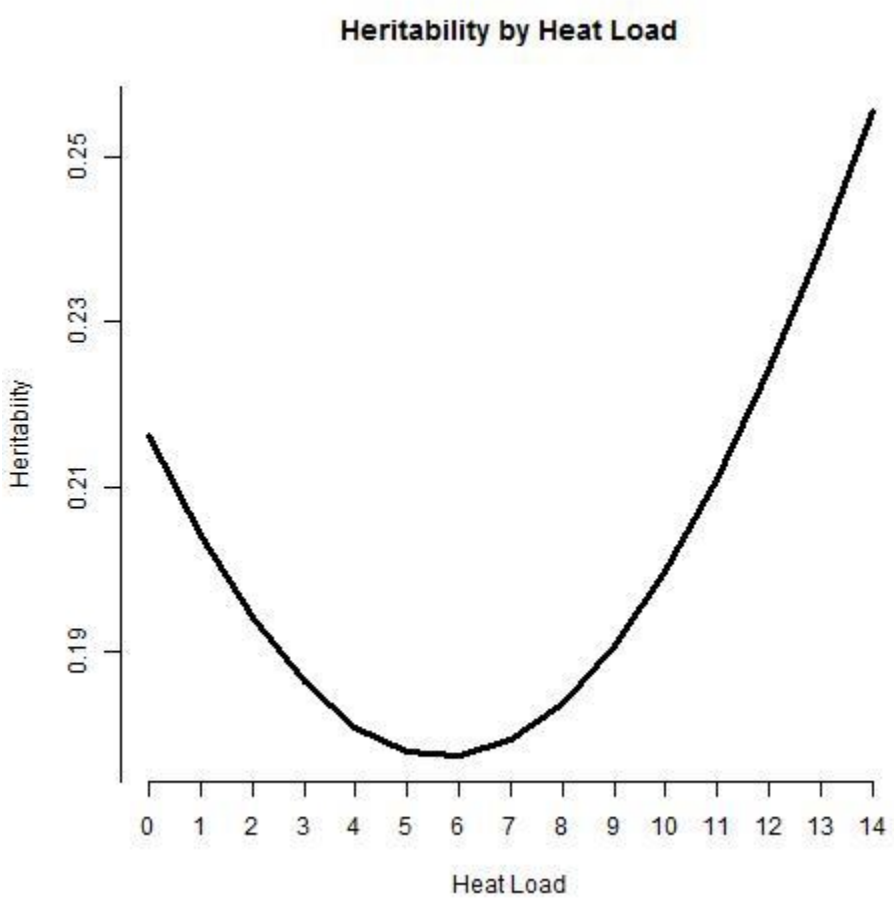

Figure 1.10 – Heritability of hot carcass weight as a function of heat load in a random regression model fitting 2 linear splines.

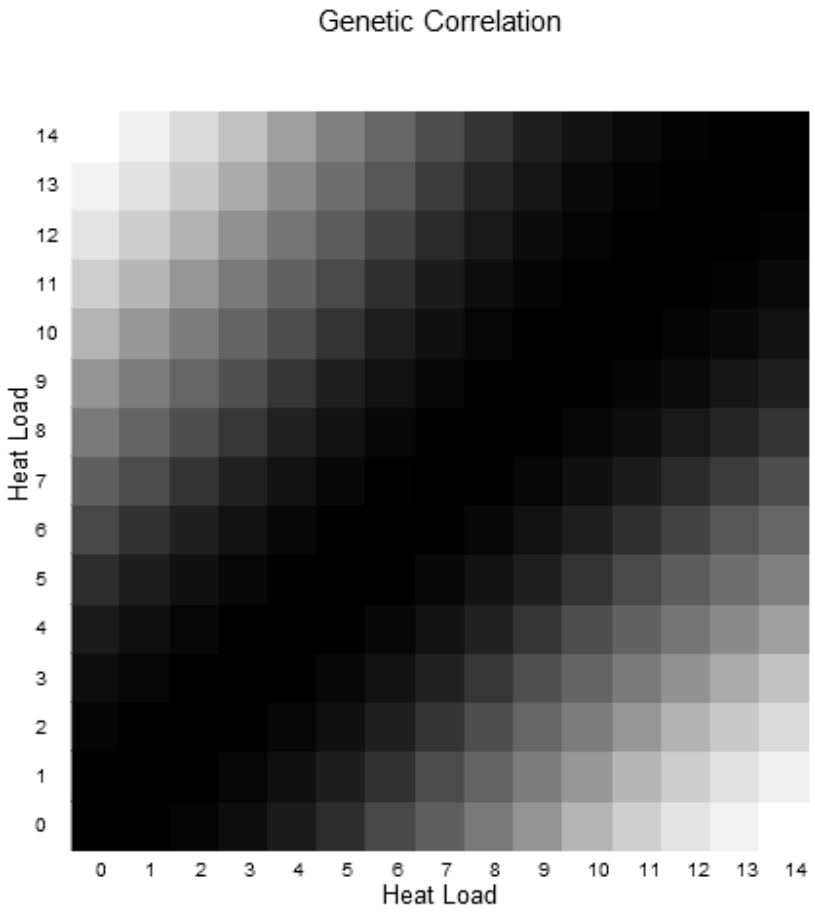

Figure 1.11 – Genetic correlation among heat load values for random regression model fitting 2 linear splines.

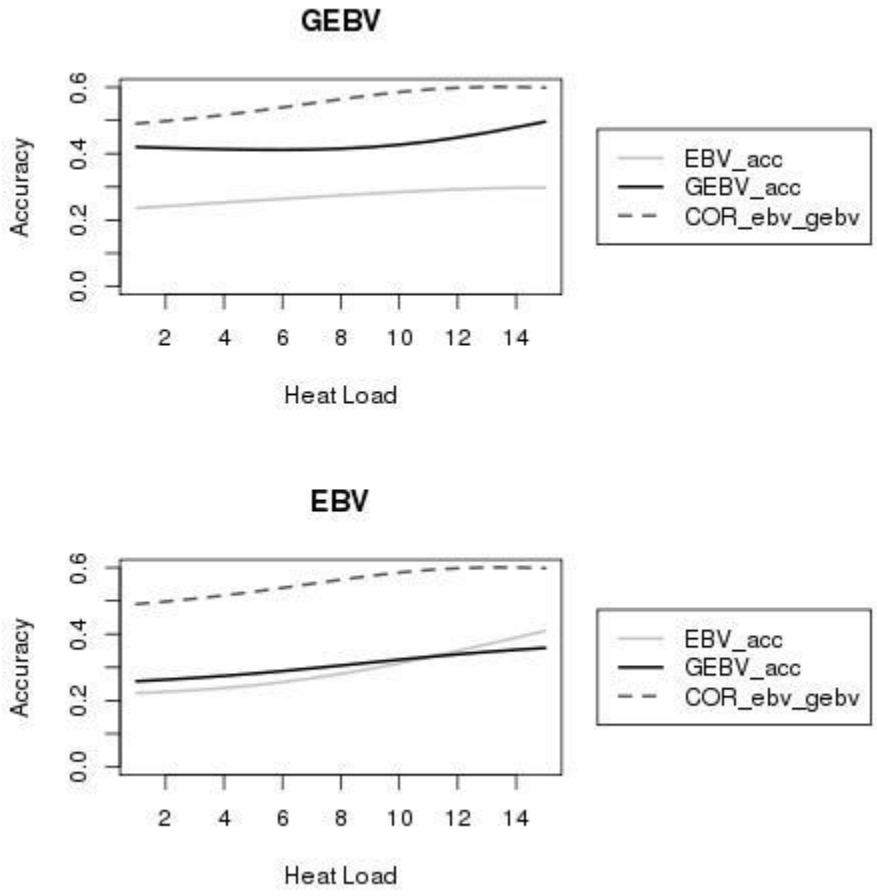

Figure 1.12 – Accuracy as correlation between (genomic) estimated breeding value and true breeding value (TBV) for a1) TBV based on EBV, ordinary linear polynomial, a2) TBV based on EBV, two linear splines, b1) TBV based on GEBV, ordinary linear polynomial, b2) TBV based on GEBV, two linear splines.

CHAPTER 4

Use of Genomic Recursions and Algorithm for Proven and Young Animals for Single-Step Genomic BLUP Analyses – A Simulation Study[1]

[1] B.O. Fragomeni, D.A.L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, and I. Misztal. Journal of Animal Breeding and Genetics. 132(5):340-345. Reprinted here with permission of the publisher.

ABSTRACT

The purpose of the present study was to examine accuracy of genomic selection via single-step genomic BLUP (ssGBLUP) when the direct inverse of the genomic relationship matrix (G) is replaced by an approximation of G^{-1} based on recursions for young genotyped animals conditioned on a subset of proven animals, termed algorithm for proven and young animals (APY). With the efficient implementation, this algorithm has a cubic cost with proven animals and linear with young animals. Ten duplicate datasets mimicking a dairy cattle population were simulated. In a first scenario, genomic information for 20k genotyped bulls, divided in 7k proven and 13k young bulls, was generated for each replicate. In a second scenario, 5k genotyped cows with phenotypes were included in the analysis as young animals. Accuracies (average for the 10 replicates) in regular EBV were 0.72 and 0.34 for proven and young animals, respectively. When genomic information was included, they increased to 0.75 and 0.50. No differences between genomic EBV (GEBV) obtained with the regular G^{-1} and the approximated G^{-1} via the recursive method were observed. In the second scenario, accuracies in GEBV (0.76, 0.51, and 0.59 for proven bulls, young males, and young females, respectively) were also higher than those in EBV (0.72, 0.35, and 0.49). Again, no differences between GEBV with regular G^{-1} and with recursions were observed. With the recursive algorithm, the number of iterations to achieve convergence was reduced from 227, to 206 in the first scenario and from 232 to 209 in the second scenario. Cows can be treated as young animals in APY without reducing the accuracy. The proposed algorithm can be implemented to reduce computing costs and to overcome current limitations on the number of genotyped animals in the ssGBLUP method.

Key words - single step method, genomic selection, genetic evaluation

INTRODUCTION

The single-step genomic BLUP (ssGBLUP; Aguilar et al. 2010; Christensen and Lund, 2010) emerged as the preferred method for genomic evaluation. Advantages of the ssGBLUP method include simplicity, no double counting of phenotypic information, and possibly, resistance to biases due to pre-selection (VanRaden and Wright, 2013; Fernando et al., 2013; Liu et al., 2014; Patry and Ducrocq, 2011). A straightforward implementation of ssGBLUP requires explicit creation and inversion of the genomic (**G**) and pedigree ($\mathbf{A}_{22}$) relationship matrices among genotyped animals (Aguilar et al, 2010). However, the inversion has a cubic cost and quadratic storage with the number of genotyped animals. The current ssGBLUP method has a soft limit on software of about 150k genotyped animals (Aguilar et al., 2013). Several approaches were proposed to overcome this limit, such as solving unsymmetric equations (Legarra and Ducrocq, 2012), using a SNP-only model with imputation of ungenotyped animals (Fernando et al., 2013), and fitting a model with SNP effects for genotyped animals (Legarra and Misztal, 2008; Liu et al. 2014). Still, these models have convergence problems, high computing costs, or difficulty for programming.

Recently Misztal et al. (2014) proposed an approximation to the inversion of **G** based on genomic recursions. They presented an algorithm for proven and young animals (APY), where genomic EBV (GEBV) of a young genotyped animal is conditioned on GEBV for ancestors of the young animal. In this algorithm, the submatrix with relationships among proven animals is inverted directly, whereas all other coefficients of $\mathbf{G}^{-1}$ for young animals are calculated recursively. Therefore, the $\mathbf{G}^{-1}$ obtained from APY has an L shape, with dense blocks among proven and between proven and young, but has only diagonal elements for young animals. As the inversion is

still needed for the subset of proven animals, the cost of APY is cubic with the number of proven animals, but can be linear with the number of young animals if preconditioned conjugate gradient (PCG) algorithm (Tsuruta et al., 2001) is used with an iteration-on-data technique. Consequently, the inverse of **G** with APY can be calculated for millions of young genotyped animals.

In a real population, genotypes may be available for high accuracy proven animals, animals with only own records (e.g., cows with no progenies), and young animals with no records and no progenies. While the number of high accuracy animals is limited, the number of animals with own records can be very high. For instance the total number of genotyped US Holsteins was already over 600k for the December/2014 evaluation, but only 25k were proven bulls with high accuracy breeding values (https://www.cdcb.us/Genotype/cur_density.html). If only young animals (i.e. no progeny or records) are treated as "young" and the other animals with records and/or progeny are treated as "proven", APY will have a high computing cost, but when only high accuracy animals are treated as "proven", it will have a low cost. The purpose of this study is to evaluate the APY algorithm with a simulated dataset mimicking a dairy cattle population with different groups of animals treated as proven and young.

MATERIALS AND METHODS

SIMULATION

Ten duplicate populations were simulated using the software QMSim (Sargolzaei and Schenkel, 2009). For each population, about 1.1 million animals were simulated over 20 generations. Phenotypes (single record) for a trait with heritability of 0.3 were available for females in the last 10 generations. In order to mimic a dairy cattle population, a male: female mating ratio of 1/25 was assumed; the mating was considered to be random in a closed population. Sire and

dam replacement rates were 90% and 30%, respectively, and animals were selected based on high EBV. In order to simulate elite cows in embryo transfer programs, the litter size was set to 5 or 10 progenies per litter with a probability of 0.02 or 0.01, respectively; otherwise it was set to 1.

The simulated genomic data consisted of 45,000 biallelic SNP markers evenly distributed along 29 chromosomes with a total length of 2319 cM, which mimicked the bovine genome without sex chromosomes. A total of 450 biallelic and randomly distributed QTL affected the trait, with effects sampled from a gamma distribution with shape parameter of 0.4 and scale 0.006.

Genomic information was available only for animals in the past 5 generations. In the first scenario, all males with progenies were selected to be the proven genotyped animals that totaled 7200 animals in each replicate of the simulated population. The average number of daughters per genotyped sire was between 4 and 50 with an average of 16. Additionally, 12,800 genotyped males were randomly selected from the last generation as young animals. In the second scenario, an extra 5000 genotyped females with phenotypes were randomly selected from the last generation as young animals, with the total of 25,000 animals with genomic information.

ANALYSES

Traditional (BLUP) and genomic (ssGBLUP) evaluations were performed with the simulated datasets, including phenotypes from females in the last 10 generations, complete pedigree information from all 20 generations, and 20,000 (or 25,000 in the second scenario) genotyped animals. Genomic analyses included the regular ssGBLUP using the direct inversion of $\mathbf{G}$ and $\mathbf{A}_{22}$, and the ssGBLUP using APY to construct $\mathbf{G}^{-1}$ ($\mathbf{G}^{-1}_{APY}$). All the analyses were done for each replicated population using the following model:

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

where y is the observation vector, μ is a vector of fixed effect (overall mean), **a** is the vector of additive animal effect, **e** is the vector of random residual effect and **Z** is the incidence matrix for the random effect in **a**.

It was assumed that $\mathbf{a} \sim N(0, \mathbf{H}\sigma_a^2)$, in which σ_a^2 is the additive genetic variance and **H** is the matrix that combines the numerator relationship matrix (**A**) and **G** (Legarra et al., 2009), with the inverse given by (Aguilar et al., 2010):

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

where $\mathbf{A}_{22}^{-1}$ is the numerator relationship matrix for the genotyped animals. The **G** matrix was blended with a small percentage of the $\mathbf{A}_{22}$ matrix before the inversion, to ensure positive definiteness. This blending was performed as $\alpha\mathbf{G}+\beta\mathbf{A}_{22}$ with values of α=0.95 and β=0.05. Also, **G** was scaled to match the average of $\mathbf{A}_{22}$, since the matrices are in different bases; the first one is in the genotyped population, and the second one is in the base population. The scaling is done by adding the average difference between **G** and $\mathbf{A}_{22}$ to **G** matrix, which takes into account the effect of non-random genotyping caused by selection (Vitezica et al., 2011).

All ssGBLUP computations were performed using the BLUP90IOD program (http://nce.ads.uga.edu/wiki/BLUPmanual), modified to account for genomic information (Aguilar et al., 2011). Accuracy of traditional and genomic evaluations was calculated by the correlation between estimated and simulated breeding values. Tukey's honest significant difference test (Tukey, 1949) was used to locate the pairwise difference between sample means. The difference in correlations means, expressed in standard deviation (SD) units, was also calculated in a way of comparing results from BLUP, ssGBLUP, and APY.

GENOMIC RECURSIONS

As shown by Misztal et al. (2014), the recursion for the additive genetic effect of animal i (u_i) can be written as:

$$\mathbf{u_i|u_1 \ldots u_{i-1}} = \sum_{\mathbf{j=1}}^{\mathbf{i-1}} \mathbf{p_{ij} u_j} + \boldsymbol{\varepsilon_i} \qquad \textbf{[1]}$$

where:

$$\mathbf{p}_{i,1:1-1} = \mathbf{g}_{i,1:i-1}\left(\mathbf{G}_{1:i-1,1,i-1}\right)^{-1},$$

$$\mathbf{M}_{i,i} = m_i = var(\varepsilon_i) = g_{i,i} - \mathbf{p}_{i,1:i-1}\mathbf{g'}_{i,1:i-1} .$$

$$\mathbf{G} = \{g_{ij}\}$$

Then, the inverse of **G** can be created using a formula as in Henderson (1976) and Quaas (1988):

$$G^{-1} = (I-P)'M^{-1}(I-P) = T'M^{-1}T \qquad [2]$$

where

$$\mathbf{P} = \{p_{ij}\}$$

THE APY ALGORITHM.

Animals in several generations can be partitioned into the categories proven and young, so the recursions in [1] would be:

$$u_i|u_1u_2,\ldots,u_{i-1} = \sum_{j \in \text{"proven"}} p_{ij}u_j + \sum_{j \in \text{"young"}} p_{ij}u_j + \varepsilon_i \qquad [3]$$

In GBLUP (BLUP when the left hand side consists only of **G**), the contributions from young animals are 0, so we can rewrite [3] as:

$$u_i|u_1 u_2,\ldots,u_{i-1} = \sum_{j \in \text{"proven"}} p_{ij} u_j + \varepsilon_i \qquad [4]$$

Simplifying the recursions in [2] by using the information in [4] will lead to the APY algorithm:

$$\mathbf{G}_{\mathbf{apy}}^{-\mathbf{1}} = \begin{bmatrix} \mathbf{G}_{\mathbf{pp}}^{-\mathbf{1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{\mathbf{pp}}^{-\mathbf{1}}\mathbf{G}_{\mathbf{py}} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{\mathbf{g}}^{-\mathbf{1}} \begin{bmatrix} -\mathbf{G}_{\mathbf{py}}\mathbf{G}_{\mathbf{pp}}^{-\mathbf{1}} & \mathbf{I} \end{bmatrix}$$

where

$$m_{g,i} = g_{i,i} - \mathbf{G}_{\mathbf{ip}}\mathbf{G}_{\mathbf{pp}}^{-\mathbf{1}}\mathbf{G}_{\mathbf{pi}}$$

and $\mathbf{G}_{\mathbf{pp}}^{-\mathbf{1}}$, $\mathbf{G}_{\mathbf{py}}$, $\mathbf{G}_{\mathbf{yp}}$, $\mathbf{G}_{\mathbf{ip}}$, **and** $\mathbf{G}_{\mathbf{pi}}$ are subsets of $\mathbf{G}^{-1}$ and **G.**

Thus, the APY algorithm only requires the inversion of **G** for animals treated as proven, assuming no contributions from young animals. Savings in memory and computations are due to a fact that $\mathbf{G}_{yy}$ does not have to be computed or stored and G_{apy}^{-1} is a sparse matrix.

RESULTS

In the first scenario (Figure 1), where just males were genotyped, for proven animals, accuracies (mean ± standard deviation) for EBV with traditional BLUP (no genomic information) (0.72±0.01) were significantly lower ($p<0.0001$, and difference of 3 SD units) than for GEBV with ssGBLUP (0.75±0.01) and APY (0.75±0.01); no differences were found between GEBV with the regular ssGBLUP and APY ($p>0.05$, and no difference in the means). For young animals, the pattern of accuracies was the same: accuracy for EBV with traditional BLUP had a significantly ($p<0.0001$, and 7 SD units) lower value (0.35±0.02), and no significant differences ($p>0.05$, and 1 SD unit difference) were found between GEBV with ssGBLUP (0.49±0.02) and APY (0.51±0.02). The increment in accuracy for adding genomic information for young animals was larger than for proven animals.

In the second scenario (Figure 2), where extra 5,000 genotyped females with a single record were considered young, the accuracy of EBV with traditional BLUP was still the lowest (significantly different from GEBV with ssGBLUP at $p<0.0001$, and 3, 8, and 3 SD units of difference in the means) for the three classes (0.72±0.01, 0.35±0.01, and 0.50±0.02 for proven bulls, young males, and young females, respectively). No significant differences ($p>0.05$, no difference in the means for the first group, and 1 SD unit in difference for the last two) were observed between GEBV with ssGBLUP (0.76±0.01, 0.51±0.02, and 0.59±0.03 for proven bulls, young males, and young females, respectively) and APY (0.76±0.01, 0.53±0.02, and 0.57±0.02 for proven bulls, young males, and young females, respectively).

The method used to invert the **G** matrix affected the number of rounds to achieve convergence in the ssGBLUP. In the first scenario, with regular $\mathbf{G}^{-1}$, the average (±SD) number of rounds was 226.5 (±7.1), whereas this number was reduced to 205.8 (±5.01) with the $\mathbf{G}^{-1}_{\mathbf{apy}}$ matrix. In the second scenario, this number was 209.4 (±6.27) with the $\mathbf{G}^{-1}_{\mathbf{apy}}$ and 231.8 (7.34) with the regular $\mathbf{G}^{-1}$. In both scenarios, the difference in numbers of rounds between regular $\mathbf{G}^{-1}$ and $\mathbf{G}^{-1}_{\mathbf{apy}}$ was significant ($p<0.001$).

DISCUSSION

The gain in accuracy with the addition of the genomic information followed the trends of North American Holstein bulls' reliabilities, as shown in Habier et al. (2007). In their study, young animals had a large increase in accuracy due to the inclusion of genomic information, whereas the increase for proven bulls was small. The similarity between GEBV in the regular ssGBLUP and APY for the first scenario was expected, as young animals do not contribute substantial information for the predictions. In fact, GEBV with the APY algorithm were marginally more

accurate, although the difference was not statistically different. In the second scenario, GEBV with the APY algorithm were also superior for young animals but slightly inferior for cows, again with no significant differences. Subsequently, use of $\mathbf{G}^{-1}_{\mathbf{apy}}$ as compared to $\mathbf{G}^{-1}$ does not seem to affect GEBV much, and in fact, may even increase accuracies for young animals. Such an outcome may have a rational as **G** calculated as in VanRaden (2008) is usually singular and requires blending with $\mathbf{A}_{22}$ to facilitate inversion. Different levels of blending have a negligible effect on accuracy (Misztal et al., 2010) indicating that the main purpose of blending in ssGBLUP is improving numerical properties of **G** and indirectly $\mathbf{G}^{-1}$.

Both results suggest that considering only a subset of animals in the recursion is sufficient, and subsequently, including all the animals in the recursion may be redundant. In the Henderson (1976) algorithm for inverting the numerator relationship matrix, the recursion ignores all animals but parents without any loss of accuracy for two reasons. First, the recursion operates on the inverse and not on the original relationships. Second, a progeny is expected to inherit 50% of genes from each parent, with other relatives contributing through parents only. Including more than two animals in the recursion would add unnecessary coefficients and possibly decrease the accuracy of EBV. The regular **G** computed in this study is an approximation.

The accuracy of EBV for a young animal in BLUP depends on EBV accuracies of parents only, whereas the accuracy of GEBV is a function of accuracies of all the other genotyped animals. In a study by Daetwyler et al. (2010), the last accuracy was approximated as a function of independent chromosome segments (ICS), with the number of segments for commercial populations < 10,000 (Ricards, 2014). Assuming an infinitesimal model (or GBLUP), values of all ICS contain all the genomic information in the population from which GEBV of any animal in the population can be derived. It is possible that proven (or base) animals in APY in fact indirectly

relate to the number of ICS. As long as the number of base animals is greater than the number of ICS, the APY algorithm would result in accurate GEBV. It is hard to estimate the number of ICS in this study, however, due to similarity of breeding structure to a dairy population it is likely that ICS < 20,000, a number used in this study. A future study will address the concept of ICS in the APY algorithm.

An important decision is which animals should be included in the proven population. For dairy cattle, the first approach is to include animals who have reliable breeding values in the national program into the reference population (De Roos et al., 2007; VanRaden et al., 2009). The second approach is to increase the reference population by including genomic information for cows, which resulted in minimal gain in accuracy (Cooper et al., 2014). In our study, the first approach was mimicked by selecting just bulls with progeny to be in the proven population. If animals are just proxies for ICS, the choice of animals treated as proven may be less important or unimportant.

As the $\mathbf{G}_{apy}^{-1}$ matrix is sparser than the regular $\mathbf{G}^{-1}$ matrix, the number of iterations to achieve convergence was expected to decrease, which was observed in the present study. Also, the storage for nonzero elements in $\mathbf{G}_{apy}^{-1}$ is lower than with regular $\mathbf{G}$; for a matrix of 50k animals but only 10k treated as proven, the memory requirement is about 5 times lower. This may be beneficial in variance component estimation with genotypes where $\mathbf{G}^{-1}$ needs to be stored explicitly. Masuda et al. (2014) showed that AI REML can be run successfully in a reasonable time when the number of genotyped animals is 20k and the sparse matrix inversion is by supernodal techniques. With $\mathbf{G}_{apy}^{-1}$, this limit can increase a few fold with little impact on quality of estimates.

The real impact of the APY formula is for solving large systems of equations with the iteration-on-data and PCG algorithm (Tsuruta et al, 2001). In this case, the implicit $\mathbf{G}_{apy}^{-1}$ is not needed but only its product by a vector. Then we can also rewrite the APY formula to avoid building the complete **G** matrix and use the marker information for young animals. Assuming **G**=**ZZ**'/q, where **Z** is a matrix of genotypes and q is a normalizing constant, the APY algorithm can be alternatively expressed as:

$$\mathbf{G}_{apy}^{-1} = \begin{bmatrix} \mathbf{G}_{pp}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{pp}^{-1}\mathbf{Z}_p\mathbf{Z}_y/q \\ \mathbf{I} \end{bmatrix} \mathbf{M}_g^{-1} \begin{bmatrix} -\mathbf{Z}_y\mathbf{Z}_p/q\mathbf{G}_{pp}^{-1} & \mathbf{I} \end{bmatrix}$$

and

$$m_{g,i} = g_{i,i} - \mathbf{z}_i'\mathbf{Z}_p\mathbf{G}_{pp}^{-1}\mathbf{Z}_p\mathbf{z}_i/q^2$$

In this case, **Z** can be stored with low precision to lower memory requirements; savings can be up to 8 times when **G** is stored as double-precision and **Z** as scaled one-byte integer. Also, the costs become approximately linear with the number of genotyped animals.

CONCLUSIONS

Genomic breeding values obtained with the proposed genomic recursions are accurate for proven and young genotyped animals in a population with a structure that mimics a dairy cattle population. Using the algorithm for proven and young animals to invert the genomic relationship matrix can reduce the number of iterations to achieve convergence in a single-step genomic BLUP analysis. When using the proposed algorithm, animals with a single record and/or with a single progeny can be treated as young without harming the evaluations. The positive results in the present study should be validated with field data before implementing the algorithm in national evaluations.

REFERENCES

Aguilar, I., Legarra, A., Tsuruta, S., and Misztal, I. (2013). Genetic evaluation using unsymmetric single step genomic methodology with large number of genotypes. *Interbull Bulletin.* 47, 222–225.

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., Lawlor, T. J. (2010). *Hot topic*: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J Dairy Sci.*, 93, 743–752.

Aguilar, I., I. Misztal, A. Legarra, S. Tsuruta. (2011) Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *J. Anim. Breed. Genet.*, 128, 422–428.

Christensen, O. F., Lund. M.S. (2010) Genomic prediction when some animals are not genotyped. *Gen. Sel. Evol*., 42, 1–8.

Cooper, T.A., G. R. Wiggans, and P.M. VanRaden. (2014) Including cow information in genomic prediction of Holstein dairy cattle in the US. In proc. 10th WCGALP, Vancouver, Canada.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, 185, 1021-1031.

De Roos, A.P.W., Schrooten, C., Mullaart, E., Calus, M.P.L., Veerkamp, R.F. (2007) Breeding Value Estimation for Fat Percentage Using Dense Markers on *Bos taurus* Autosome 14. *J Dairy Sci.*, 90, 4821–4829.

Fernando, R. L., Garrick, D., Dekkers, J. C. M. (2013). Bayesian regression method for genomic analyses with incomplete genotype data. In *Proc. of the 64th Annual Meeting of the*

European Federation of Animal Science, Nantes, France.

Habier, D., Fernando, R. L., Dekkers, J. C. (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177, 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, 69–83.

Legarra, A., Misztal, I. (2008) Technical note: Computing strategies in genome-wide selection. *J. Dairy Sci.*, 91, 360–366.

Legarra, A., Aguilar, I., Misztal, I. (2009) A relationship matrix including full pedigree and genomic information. *J Dairy Sci.*, 92, 4656–4663.

Legarra, A., Ducrocq, V. (2012) Computational strategies for national integration of phenotypic, genomic, and pedigree data in single-step best linear unbiased prediction. *J. Dairy Sci.* 95, 4629–4645.

Liu, Z., M.E. Goddard, F. Reinhardt, Reents, R. (2014) A single-step genomic model with direct estimation of marker effects. *J. Dairy Sci.* 97, 5833–5850.

Lund, M.S., Su, G. (2009) Genomic selection in the Nordic countries. *Interbull bulletin*, 39, 39–42.

Masuda, Y., Baba, T., Suzuki, M. (2014). Application of supernodal sparse factorization and inversion to the estimation of (co)variance components by residual maximum likelihood. *J. Anim. Breed. Genet.*, 131, 227–236.

Misztal, I., Aguilar, I., Legarra, A., Lawlor, T. J. 2010. Choice of parameters for single-step genomic evaluation for type. *J. Dairy Sci.*, 93(Suppl. 1):166.

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

Patry, C., Ducrocq, V. (2011) Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *J. Dairy Sci*.94, 1011–1020.

Quaas, R.L. (1988) Additive genetic model with groups and relationships. *J. Dairy Sci.,* 71, 91-98.

Ricard, A. (2014) Is heterozygotie at the "Gait Keeper" Gene an Advantage for the Trotteur Français?. In proc. 10th WCGALP. Vancouver, Canada

Sargolzaei, M., Schenkel, F.S. (2009) QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, 25, 680–681.

Tsuruta, S., Misztal, I., Strandén. I. (2001) Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. *J. Anim. Sci*., 79, 1166–1172.

Tsuruta, S., Misztal, I., Aguilar, I., & Lawlor, T. J. (2011) Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. *J. Dairy Sci.*, 94, 4198–4204.

Tukey, J.W. (1949). Comparing individual means in the analysis of variance. *Biometrics*. 5:99–114.

VanRaden, P.M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91, 4414–4423.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., Schenkel, F.S. (2009) *Invited Review* Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.,* 92, 16–24.

VanRaden, P.M., Wright, J.R. (2013) Measuring genomic pre-selection bias in theory and in practice. *Interbull Bulletin*. 47, 147–150.

Vitezica, Z. G., Aguilar, I., Misztal, I., Legarra, A. (2011) Bias in genomic predictions for populations under selection. *Genet. Res. (Camb.)*, 93, 357–366.

FIGURES

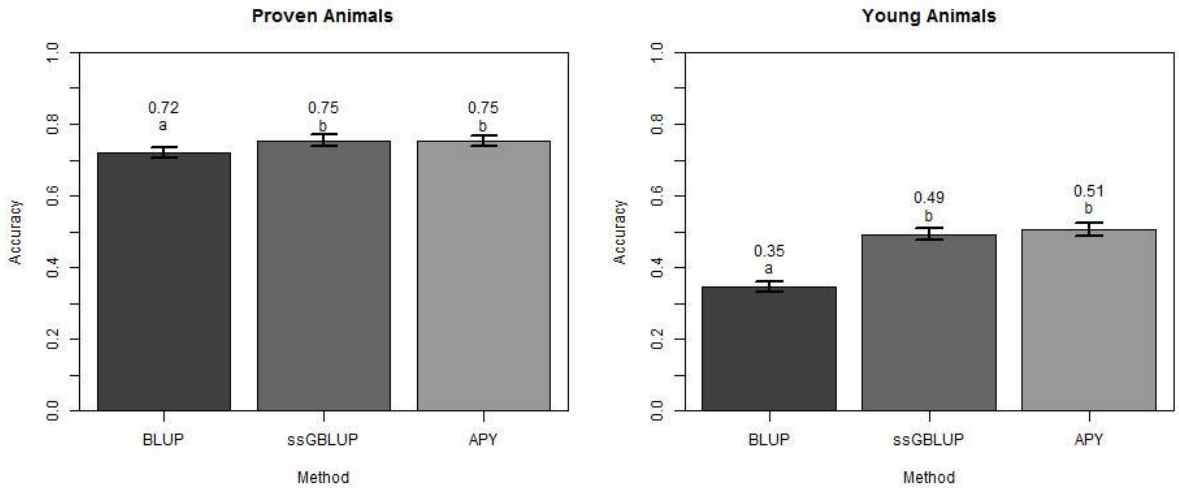

Figure 4.1. Accuracy means (±standard deviation) on proven and young animals for traditional (BLUP) and genomic evaluations using regular ssGBLUP (ssGBLUP) and ssGBLUP with genomic recursion to invert G matrix (APY) in the first scenario.

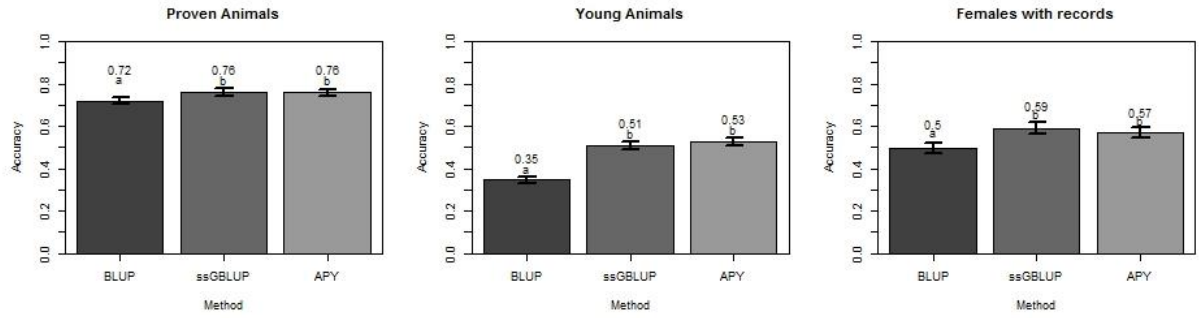

Figure 4.2. Accuracy means (±standard deviation) on proven animals, young, and females for traditional (BLUP) and genomic evaluations using regular ssGBLUP (ssGBLUP) and ssGBLUP with genomic recursion to invert G matrix (APY) in the second scenario.

CHAPTER 5

HOT TOPIC: USE OF GENOMIC RECURSIONS IN SINGLE-STEP GENOMIC BLUP WITH A LARGE NUMBER OF GENOTYPES [1]

[1] B.O. Fragomeni, D.A.L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. J. Lawlor, and I. Misztal. Journal of Dairy Science.98 (6) :4090-4094. Reprinted here with permission of the publisher.

ABSTRACT

The purpose of this study was to evaluate the accuracy of genomic selection in single-step genomic BLUP (ssGBLUP) when the inverse of the genomic relationship matrix (G) is derived by the APY (Algorithm for Proven and Young). This algorithm implements genomic recursions on a subset of "proven" animals. Only a relationship matrix for animals treated as "proven" needs to be inverted, and extra costs of adding animals treated as "young" are linear. Analyses involved 10,102,702 final scores on 6,930,618 Holstein cows. Final score, which is a composite of type traits, is popular trait in the U.S. and was easily available for this study. A total of 100k animals with genotypes were used in the analyses and included 23k sires (16k with more than 5 progenies), 27k cows, and 50k young animals. Genomic EBV (GEBV) were calculated with a regular inverse of **G**, and with the **G** inverse approximated by APY. Animals in the "proven" subset included only sires (23k), sires + cows (50k), only cows(27k), or sires with more than 5 progenies(16k). The correlations between GEBV with APY and regular GEBV of genotyped animals were 0.994, 0.995, 0.992, and 0.992, respectively. Later, animals in the "proven" subset were randomly sampled from all genotyped animals in sets of 2k, 5k, 10k, 15, and 20k; each sample was replicated four times. Respective correlations were 0.97 (5k sample), 0.98 (10k sample) and 0.99 (20k sample), with minimal difference between samples of same size. Genomic EBV with APY are accurate when the number of animals used in the subset is between 10k and 20k, with little difference between the ways of creating the subset. Due to approximately a linear cost of APY, the ssGBLUP with APY can possibly support any number of genotyped animals without affecting accuracy.

(Key words: single-step method, genomic selection, genomic recursion)

INTRODUCTION

Single-step genomic BLUP (ssGBLUP) (Aguilar et al., 2010; Christensen and Lund, 2010) emerged as a simple yet accurate tool for genetic evaluations. Its main advantages compared to multistep methods are simplicity, no double counting, and resistance to pre-selection bias (Vitezica et al., 2011; VanRaden and Wright, 2013; Legarra et al., 2014). As originally defined, ssGBLUP uses classical BLUP mixed equations extended with the inverse of the genomic (**G**) and pedigree ($\mathbf{A}_{22}$) relationship matrices for genotyped animals. With algorithms as described in Aguilar et al. (2011), the cost of obtaining these matrices is cubic, and currently there is a soft limit of about 150k genotyped animals in the model; however, there are over 600k genotyped animals available for US Holsteins (https://www.cdcb.us/Genotype/cur_density.html). Several approaches were proposed to overcome such a limit (Legarra and Ducrocq, 2012; Liu et al., 2014; Fernando et al., 2014) but either they have convergence problems, or are expensive and hard to program and use with data and a variety of models such as multiple trait or random regressions.

Faux et al (2012) attempted to extend the rules used in creation of the numerator relationship matrix to approximate the inverse of **G**. Their method was based on incomplete Cholesky factorization where only genomic relationships between close relatives were considered. However, the approximation was not accurate enough, and steps proposed to increase that accuracy were expensive.

Recently Misztal et al. (2014) proposed a method based on genomic recursion, where genomic breeding value (GBV) of a new genotyped animal is conditioned on GBV of all the previous genotyped animals. One of their proposed algorithms was called APY (Algorithm for

Proven and Young animals). This algorithm conditioned "young" animals on a small subset of "proven" animals. The APY algorithm has a cubic cost with the number of animals treated as proven and linear cost with the animals treated as young; direct inversion is required for only a small portion of **G** composed by relationships among animals treated as proven. This algorithm was tested with simulated data and with US Holsteins data (Fragomeni et al., 2014). In simulations, accuracies with APY were close to those with direct inverted **G** even when some animals with records were treated as young. This suggests that the definition of "proven" is not critical and this subset may not need to be composed by parents, animals with records, or possess any other special requirement. In US Holsteins with genotypes on 15k proven bulls and 60k young bulls, the correlations of GEBV obtained through APY and regular method were >0.99.

In real data sets, genotyped animals include bulls and cows. While the number of proven bulls is limited and rises slowly (~ 2000/year for US Holsteins), the number of cows with genotypes can be very high. The purpose of this study is to evaluate the accuracy of GEBV with APY for US Holsteins considering genotypes of bulls and cows and treating various groups of animals as proven and young.

MATERIAL AND METHODS

GENOMIC RECURSIONS

The recursion for the additive genetic effect of animal i (u_i) can be written as (Misztal et al., 2014):

$$u_i|u_1 \ldots u_{i-1} = \sum_{j=1}^{i-1} p_{ij} u_j + \varepsilon_i,$$

where u is an additive genetic effect, p relates animals to all previous individuals, and ε is the error term. Calculations can proceed as:

$$\mathbf{p}_{i,1:i-1} = \mathbf{g}_{i,1:i-1}(\mathbf{G}_{1:i-1,1:i-1})^{-1},$$
$$\mathbf{M}_{i,i} = m_i = \text{var}(\varepsilon_i) = g_{i,i} - \mathbf{p}_{i,1:i-1}\mathbf{g}'_{i,1:i-1},$$

where **M** is a diagonal matrix of genomic Mendelian sampling and **G** = $\{g_{ij}\}$ is a genomic relationship matrix. Then, the inverse of **G** can be created using a formula as in Henderson (1976) and Quaas (1988):

$$\mathbf{G^{-1}} = (\mathbf{I} - \mathbf{P})'\mathbf{M^{-1}}(\mathbf{I} - \mathbf{P}) = \mathbf{T}'\mathbf{M^{-1}}\mathbf{T}$$

where **T** is a triangular matrix, **P** = $\{p_{ij}\}$, and **I** is an identity matrix; if many of its elements are very small they can be set to 0 and $\mathbf{G}^{-1}$ may be computed at a low cost.

THE APY ALGORITHM

In genomic recursions, contributions from proven and young animals can be separated as:

$$u_i|u_1, u_2, \ldots, u_{i-1} = \sum_{j \in \text{"}proven\text{"}} p_{ij} u_j + \sum_{j \in \text{"young"}} p_{ij} u_j + \varepsilon_i$$

However, the contribution of information from young animals to other genotyped animals is 0 in GBLUP because young animals do not get information from data. Then, neglecting these contributions:

$$u_i|u_1,u_2,\ldots,u_{i-1} = \sum_{j\in"\,proven"} p_{ij}u_j + \varepsilon_i$$

As shown in Misztal et al. (2014), the simplified recursions lead to a new formula for an approximate inverse of G called APY (algorithm for proven and young):

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{G}_{pp}^{-1} + \mathbf{G}_{pp}^{-1}\mathbf{G}_{py}\mathbf{M}_g^{-1}\mathbf{G}_{yp}\mathbf{G}_{pp}^{-1} & -\mathbf{G}_{pp}^{-1}\mathbf{G}_{py}\mathbf{M}_g^{-1} \\ -\mathbf{M}_g^{-1}\mathbf{G}_{yp}\mathbf{G}_{pp}^{-1} & \mathbf{M}_g^{-1} \end{bmatrix}$$

$$m_{g,i} = g_{ii} - \mathbf{G}_{ip}\mathbf{G}_{pp}^{-1}\mathbf{G}_{pi}$$

where $\mathbf{G}_{pp}$ is a subset of $\mathbf{G}$ relating proven animals, $\mathbf{G}_{py}$ relates proven and young animals, $\mathbf{G}_{ip}$ relates the i^{th} young animal with all proven animals, and $\mathbf{M}_g$ is a diagonal matrix. While this algorithm results in the same GEBV for GBLUP as the regular inversion of G^{-1}, for ssGBLUP the APY algorithm leads to an approximation, as a young genotyped animal may provide ties to ungenotyped ancestors. This happens if at least one of its parents is not genotyped.

The APY $\mathbf{G}^{-1}$ is a sparse matrix with non-zero elements forming an L shape, with only a diagonal for the submatrix due to young animals; the only direct inversion required is for $\mathbf{G}_{pp}$. Whereas the regular $\mathbf{G}^{-1}$ requires quadratic storage and cubic computations, the APY $\mathbf{G}^{-1}$ requires quadratic storage and cubic computations only for animals treated as proven, and linear storage and computations for animals treated as young. When the number of animals treated as proven is a small fraction of all animals, the APY $\mathbf{G}^{-1}$ has approximately a linear cost and can provide large savings in memory and especially in computing time.

FIELD DATA

To check the quality of this approximation for $\mathbf{G}^{-1}$ we tested it using real data. Phenotypic data included 11,626,576 records for final score on 7,093,380 cows, with 10,709,878 animals in the pedigree provided by Holstein Association USA Inc. (Brattleboro, VT). Final score is a weighted linear combination of five major breakdown score for type traits in dairy cattle, and was chosen for this study because of availability of records. Genotypes on 42,503 SNP markers were available for 569,404 animals. However, in order to have comparisons with the regular ssGBLUP where direct inversion of **G** is used, analyses involved only 100,000 of the genotyped animals, which is the limitation of ssGBLUP for the available computer. Thus, genotypes were considered for all 23,174 bulls with progeny information, all 27,215 cows with records (hereinafter termed "cows"), and additionally 49,611 young animals.

ANALYSES

Initially, GEBV were calculated using the regular ssGBLUP which applies direct inversion for **G**. Secondly, GEBV were calculated using APY to obtain $\mathbf{G}^{-1}$ recursively ($\mathbf{G}^{-1}_{APY}$) with several different definitions for proven animals: only sires; sires and cows; only cows; sires with more than 5 progenies including sons and daughters. Thirdly, previous analyses were repeated with "proven" animals randomly sampled from the group of all 100k genotyped animals in sets of 2k, 5k, 10k, 15k, and 20k animals; the sampling was replicated 4 times. Evaluations for final score were done using a single trait model as described in Tsuruta et al. (2002). All analyses were conducted with blup90iod2 (http://nce.ads.uga.edu/wiki/BLUPmanual) program with modifications as in (Aguilar et al., 2011). The quality of approximations was assessed by

correlations between GEBV for the almost 50k young animals obtained from ssGBLUP using direct inversion of full **G** (regular ssGBLUP) and ssGBLUP using approximated $\mathbf{G}^{-1}$ from the APY algorithm.

RESULTS AND DISCUSSION

Table 1 summarizes runs with regular and APY ssGBLUP when the subset of animals treated as "proven" were sires, sires + cows, cows, and sires with > 5 daughters. For all subsets, the correlations of GEBV obtained with a regular and APY algorithms are > 0.99. In all cases except when cows were treated as proven, the convergence rate was close to a regular run, indicating good computing properties. The smallest set of proven animals with good predictive ability was sires with more than 5 daughters (16,434 animals). Treating more animals as proven, i.e., including sires with less than 5 progeny, only marginally affected the correlations. Computing an inverse for 16k animals (assuming cubic algorithm for inversion) costs about 200-fold less than for 100k animals and would costs 4000-fold less for 600k animals.

Surprisingly good correlations were observed with only cows treated as proven although the convergence rate was affected, but was still much better than with ssGBLUP with unsymmetric equations constructed to avoid the inverse of **G**. (Aguilar et al., 2013). This means that the original definition of animals as young and proven is not necessarily important for accuracy of GEBV, only the number of animals in $\mathbf{G}_{pp}$ matters. To tests this hypothesis, 2k, 5k, 10k, 15k, and 20k animals were chosen randomly from all bulls and cows and treated as proven in the APY algorithm. Rounds to convergence increased with the subset size but were lower than with the regular algorithm. This suggest that $\mathbf{G}^{-1}$ by APY is well numerically conditioned. The

correlations of GEBV with the regular and APY algorithms ranged from >0.94 for 2k animals to >0.99 for 20k animals, with very small variations among the replicates (Table 2). This means that the choice of animals in $\mathbf{G}_{pp}$ is mostly arbitrary.

Initially, the last statement seems hard to believe, however, recursions generate very similar inverses regardless of the order of animals. The Single Step modifies the pedigree relationship matrix (**A)** towards a realized relationship matrix (**H)**. Possibly, to obtain a good **H** only a good sample of genotyped animals is needed, and several such samples may exist.

To test whether the presence of sires and cows is crucial for good properties of APY, an extra set included 20k animals selected randomly only from young animals. The correlations of GEBV for this set were slightly lower than with complete random 20k choice and similar to a 15k random sample. Also, the convergence rate was slightly worse. In general, we expect better properties of APY when animals treated as "proven" are well related to animals treated as "young". While proven sires are well related to the general population, cows and young animals may be less so.

The Henderson's algorithm for creating the inverse of the numerator relationship matrix ($\mathbf{A}^{-1}$) is based on younger animals conditioned on older animals (Henderson, 1976). In such a case, each recursion has at most two nonzero elements, each with a value of 0.5 and due to a parent. However, an identical $\mathbf{A}^{-1}$ can be derived with animals in the reverse order (see Appendix in Misztal et al., 2014). In such a case, the number of nonzero elements in each recursion can be greater than 2 and they can take different values. Assume the following genomic recursion, where the additive genetic effect of an animal *i* is conditioned on the first *m* animals:

$$u_i|u_1,\ldots,u_{i-1} = \sum_{j=1}^{\min(i-1,m)} p_{ij}u_j + \varepsilon_i(\mathrm{m}),$$

where $\varepsilon_i(m)$ is the error term. While the error term should be smaller with larger m, apparently the reduction of $\varepsilon_i(m)$ for m >10k is small. In an alternate interpretation, the inverse of **G** created with APY is becoming more accurate as m increases; with small improvements beyond m >10k.

The limited number of animals required in the recursion (< 20k) suggests that the genomic information for a population has a limited dimensionality (< 20k). Nearly all genomic information from a reference population is usually assumed to be accounted by SNP solutions with a medium size chip (~50k). However, many SNPs are correlated. Pintus et al. (2013) found that 15,207 principal components extracted from matrices based on 39,555 SNP markers explained 99% of the genetic variation. Thus the real dimensionality of the SNP information may be ~15k. Alternately, when the number of QTL is high, the accuracy of GBLUP is dependent on the number of independent chromosome segments, with the number of the segments usually < 10k (Daetwyler et al., 2010). Further research will determine whether the limits based on the recursion, eigenvalues, and chromosome segments are related through equivalent models.

The US Holstein population is very homogenous. In other species, populations may be more diverse and a larger subset may be needed. Lourenco et al. (2015) applied APY to genetic evaluation of US Angus for 3 traits with 52k genotyped animals. Using 4k and 8k subsets generated 84% and 97% of gains in accuracy over BLUP compared to a regular ssGBLUP. A detailed analysis on the number and choices of animals treated as "proven" in APY will be a topic for a separate study. Further investigations will also look at whether specific subgroups of animals are invariant to the selection of the subset of animals defined as "proven".

The original derivation of the APY algorithm was based on labeling animals in the recursion as proven. Since the algorithm works with any sufficiently large subset of animals in the recursion, the designation of "proven" or "young" may no longer be relevant. In particular, the animals can be decomposed into a base genomic relationship group (b) and a conditional genomic relationship group (c), e.g., with relevant matrices G_{bb} and G_{bc}.

In this paper we focused on accuracy of ssGBLUP with $\mathbf{G}^{-1}$ calculated by APY. In practical implementations, important issues will be memory requirements and computing costs for a large number of genotyped animals. Assume a total of n=500k genotyped animals, recursion on m=20k animals, and double precision half-storage. The amount of memory necessary for APY $\mathbf{G}^{-1}$ is approximately 80 Gbytes (n*m*8=20k*500k*8bytes) or 8% of 1 Tbyte ($n^2/2*8$=500k*500k*8 bytes /2) required for a regular half-stored $\mathbf{G}^{-1}$. As current servers have memory capacity in the order of terabytes, the memory requirements will not limit the APY algorithm. Computing APY $\mathbf{G}^{-1}$ would require approximately $m^3+2m^2(n-p)$ operations, or about 0.3% operations (n^3) for a regular $\mathbf{G}^{-1}$. Another issue is efficient computations of $\mathbf{A}_{22}^{-1}$. In separate analyses (results not provided), computing this matrix using formulas similar to Stranden and Mantysaari (2014) took negligible time and memory.

CONCLUSIONS

Inverse of a genomic relationship matrix can be approximated with the APY algorithm where actual inversion is applied only to a small subset of genotyped "proven" animals and an approximate inversion by recursion is applied on "young" animals. The approximation is very accurate when the number of animals in the subset is 10k or greater while storage and computing

costs can be dramatically lower. The choice of animals in the subset is arbitrary as various definitions including random choices provide similar accuracy. The convergence rate is superior to conventional inversion. Costs of APY inversions with a larger number of animals are approximately linear making the algorithm potentially suitable for any number of genotypes. Single-step GBLUP with APY may be suitable for models with any number of genotyped animals.

ACKNOWLEDGEMENTS

We gratefully acknowledge the very helpful comments by the two anonymous reviewers and insightful suggestions by Jenny Price. This research was supported by grants from Zoetis (Kalamazoo, MI), Cobb-Vantress Inc. (Siloam Springs, AR), Smithfield Premium Genetics (Rose Hill, NC), American Angus Association (St. Joseph, MO), Holstein Association USA (Brattleboro, VT), Pig Improvement Company (Hendersonville, PIC), and by Agriculture and Food Research Initiative Competitive Grants no. 2015-67015-22936 from the US Department of Agriculture's National Institute of Food and Agriculture. AL thanks INRA metaprogram SelGen and projects X-Gen and GenSSeq.

REFERENCES

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743–752.

Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128:422–428.

Aguilar , I., A. Legarra, S. Tsuruta, and I. Misztal. 2013. Genetic evaluation using unsymmetric single step genomic methodology with large number of genotypes. Interbull Bull. 47:222–225.

Christensen, O.F., and M.S. Lund. 2010. Genomic predictions when some animals are not genotyped. Gen. Sel. Evol. 42:2.

Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J.A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021-1031.

Faux, P., N. Gengler, and I. Misztal. 2012. A recursive algorithm for decomposition and creation of the inverse of the genomic relationship matrix. J. Dairy Sci. 95:6093‑6102.

Fernando, R.L., J.C.M. Dekkers, and D. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Gen. Sel. Evol. 46:50.

Fragomeni, B.O., I. Misztal, D.A.L. Lourenco, S. Tsuruta, Y. Masuda, and T.J. Lawlor. 2014. Use of genomic recursions and algorithm for proven and young animals for single-step genomic BLUP analyses with a large number of genotypes. In proc. 10th WCGALP, Vancouver, Canada.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69‑93.

Lourenco, D.A.L., S. Tsuruta, B. Fragomeni, I. Aguilar, Y. Masuda, J.K. Bertrand, I. Misztal. 2015. Genomic evaluation by single-step GBLUP in Angus. J. Animal Sci. (*In submission*).

Legarra, A. and V. Ducrocq. 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in single-step best linear unbiased prediction. J. Dairy Sci. 95:4629–4645.

Legarra, A., O.F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. Livest. Prod. Sci. 166:54–65.

Liu, Z., M.E. Goddard, F. Reinhardt, and R. Reents. 2014. A single-step genomic model with direct estimation of marker effects. J. Dairy Sci. 97:5833–5850.

Pintus, M. A., E.L. Nicolazzi, J.B.C.H.M. Van Kaam, S. Biffani, A. Stella, G. Gaspa, C. Dimauro and N.P.P. Macciotta. 2013. Use of different statistical models to predict direct genomic values for productive and functional traits in Italian Holsteins. J. Anim. Breed. Genet. 130:32-40 .

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

Quaas, R. L. 1988. Additive genetic model with groups and relationships. J. Dairy Sci. 71:1338-1345.

Standen, I. and E.A. Mantysaari. 2014. Comparison of some equivalent equations to solve single-step GBLUP. In proc. 10th WCGALP, Vancouver, Canada.

Tsuruta, S., I. Misztal, L. Klein, and T. J. Lawlor. 2002. Analysis of age specific predicted transmitting abilities for final scores in Holsteins with a random regression model. J. Dairy Sci. 85:1324–1330.

VanRaden, P.M. and J.R. Wright. 2013. Measuring genomic pre-selection bias in theory and in practice. Interbull Bull. 47:147–150.

Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet. Res. 93:357–366.

TABLES

Table 5.1. Correlations between genomic EBV with regular and APY ssGBLUP for young genotyped animals and rounds to convergence for different subsets of animals used in recursions

Definition of subset	**Animals in subset**	**Correlation**	**Rounds to convergence**
All	100,000	1.000	567
Sires	23,174	0.994	432
Sires + cows	50,389	0.995	428
Cows	27,215	0.992	797
Sires > 5 progenies	16,434	0.992	415

Table 5.2. Ranges of correlations between genomic EBV with regular and APY ssGBLUP for young genotyped animals and rounds to convergence when different numbers of randomly sampled animals were used in the subset for recursions

Number of proven animals	Correlation	Rounds to convergence
2,000	0.943–0.944	351–357
5,000	0.971–0.972	354–367
10,000	0.985	391–403
15,000	0.989–0.990	411–480
20,000	0.992–0.993	416–425
20,000*	0.989–0.990	552–556

*Proven were randomly sampled from the group of young animals.

CHAPTER 6

CONCLUSIONS

It is possible to identify impacts of heat stress on phenotypes in a commercial crossbred swine population. The trait under heat stress can be more heritable than under mild conditions because of higher genetic variance and lack of selection. Therefore, robust animals can be identified and selected for breeding. Including genomic information in a genetic evaluation for heat stress can increase accuracy of prediction, helping to mitigate seasonal losses.

In order to extend mitigation of heat stress impacts to different populations with more than a hundred thousand genotyped animals, an efficient method for inverting the genomic relationship matrix is required in single-step GBLUP. The algorithm for proven and young animals gives the same accuracy as the direct inversion of the genomic relationship matrix, and can be used in single-step genomic evaluations for large genotyped populations.

For a large genotyped population (US Holstein), correlations between genomic EBV from single-step with regular inversion and with the algorithm for proven and young are close to 1 when the number of proven animals is at least 10,000, independent on the choice of proven. The number of proven animals may be a function of effective population size. The computing cost of this new algorithm increases linearly with the number of young animals, enabling single-step GBLUP to work with millions of genotyped animals.

APPENDIX A

Changes in variance explained by top SNP windows over generations for three traits in broiler chicken [1]

[1] Breno de Oliveira Fragomeni, Ignacy Misztal, Daniela Lino Lourenco, Ignacio Aguilar, R. Okimoto, William M Muir Front Genet. 5:332. Reprinted here with permission of the publisher.

ABSTRACT

The purpose of this study was to determine if the set of genomic regions inferred as accounting for the majority of genetic variation in quantitative traits remain stable over multiple generations of selection. The data set contained phenotypes for five generations of broiler chicken for body weight, breast meat, and leg score. The population consisted of 294,632 animals over 5 generations and also included genotypes of 41,036 SNP for 4,866 animals, after quality control. The SNP effects were calculated by a GWAS type analysis using single step genomic BLUP approach for generations 1 to 3, 2 to 4, 3 to 5, and 1to 5. Variances were calculated for windows of 20 SNP. The top ten windows for each trait that explained the largest fraction of the genetic variance across generations were examined. Across generations, the top 10 windows explained more than 0.5% but less than 1 % of the total variance. Also, the pattern of the windows was not consistent across generations. The windows that explained the greatest variance changed greatly among the combinations of generations, with a few exceptions. In many cases, a window identified as top for one combination, explained less than 0.1% for the other combinations. We conclude that identification of top SNP windows for a population may have little predictive power for genetic selection in the following generations for the traits here evaluated.

(Keywords: Genomic Selection, genome-wide association study, QTL, ssGBLUP, gene identification.)

INTRODUCTION

Past studies of genomics in livestock usually focused either on best estimation of breeding values (Calus, 2010) or on identification of major SNP (Goddard and Hayes, 2009). For the latter, the purpose is exploring associations between SNP and phenotypes to better understand the genetic architecture of a trait or to use identified major SNP for genetic selection. With important SNP identified, the selection can be performed with simple tests for a few SNP.

Genetic selection using major SNP is successful if they explain a sizeable portion of the genetic variation and if their effects change little over time. Earlier simulation studies showed that LD identified in one generation decays very slowly over generations (Meuwissen, et al., 2001, Solberg et al., 2009). However, under strong selection the decay is much faster (Muir, 2007). Therefore, newer studies advocate continuous genotyping and recalculation of SNP effects (Wolc et al., 2011, Sonesson and Meuwissen, 2009, Habier et al., 2007). While the selection pressure would act on the largest QTLs, it is not clear how this would impact the identification and estimation of values for the top SNP that may indicate presence of QTLs.

Identification of an individual SNP linked to a QTL is difficult because of the high collinearity of SNPs. SNPs may be in LD with a QTL so windows of consecutive SNPs can capture the effect of a QTL better than a single SNP (Habier et al, 2011). Also, SNP segments are useful to discriminate important effects from statistical noise (Sun et al., 2011). Bolormaa et al. (2010) looked at SNPs within 1 Mbp intervals. Peters et al. (2012) used windows of 5 adjacent SNP. In a simulation study, effects of individual QTL were best explained by the combined effect of 8 adjacent SNP (Wang et al., 2012). The optimal window size may also be a function of effective population size (Goddard, 2008).

There is a shortage in studies searching for stability of marker effects across generations in production traits for broiler chicken. Despite this, in a layer population, Wolc et al. (2012) found that 1Mbp SNP windows with large effects had consistent effects across generations, but windows that explained little variance of the trait were not validated. If a window effect is constant across generations or subsets of population, it can be indicative of a causative gene on that trait; however, if the effect is not robust, it can correspond to an unstable, sample-specific association that is not expected to provide good out-of-sample predictions.

One common issue on genome association studies is the large number of false positive gene discovery. Information from the chicken QTL database (Hu et al., 2013) shows a large number of QTL described—2,467 for growth traits, 68 for meat quality traits, and 28 for conformation—but few of these have been validated or reproduced by other studies. This can be observed not only in chicken, but in studies on all livestock species. In this way, GWAS results should be carefully interpreted before considering an association as a causative effect. A possible causative effect should be easily accessed in further assays considering similar population structure.

The purpose of this study was to identify SNP windows that explain major portions of genetic variance and see if those values are preserved during a course of selection for growth in chicken.

MATERIALS AND METHODS

The data was provided by Cobb-Vantress Inc. (Siloam Springs, AR). A total of 294,632 phenotypes from a pure line of broiler chicken collected across five consecutive generations (G1, G2, G3, G4, and G5) were used in this study. This was the sire line, selected mainly for growth rate, meat yield, feed conversion and livability, and secondarily for reproduction traits. The numerator relationship matrix included 297,017 animals. For the first two generations, animals

were selected for genotyping based on body weight and conformation scores; leg defects were very unlikely. The remaining animals (from G3 to G5) were randomly selected for genotyping. The number of animals in each generation are shown in Table 1. The number of observations, means, and standard deviation for all the traits are shown in Table 2.

Initially, genotype information from 4,922 animals in a chip with 57,635 SNPs was available (Groenen et al., 2011). The genomic data was subject to a quality control (QC) before the analysis. This QC removed SNPs with minor allele frequency < 0.05, with call rates < 0.9, and monomorphic SNPs. It also removed genotypes with call rates < 0.9. After QC, the genotype file had 4,866 animals genotyped for 41,036 SNPs.

SNP solutions were estimated by ssGWAS (genome-wide association study using a single-step BLUP approach) (Wang et al., 2012; Dikmen et al., 2013). In this methodology, the data was initially analyzed by a multi-trait single-step genomic BLUP (ssGBLUP; Aguilar et al, 2010) with the same model as used for BLUP analyses (Chen et al., 2011). Effects in the model included sex, contemporary group, animal additive, and maternal permanent environmental effects. Concerning the genomic information, the genomic relationship matrix **(G)** was scaled for the average of the numerator relationship matrix for the genotyped animals ($\mathbf{A}_{22}$), which took into account the effect of non-random genotyping caused by selection (Vitezica et al., 2011). Subsequently, EBV for genotyped animals (GEBV) were converted to SNP effects and weights of SNP effect were refined iteratively. The procedure followed the S1 scenario described in Wang et al. (2012), with GEBV computed once and SNP weights refined through 3 iterations. The equation for predicting SNP effects using weighted genomic relationship matrix was (Wang et al., 2012):

$$\hat{\mathbf{u}} = \mathbf{DZ}'[\mathbf{ZDZ}']^{-1}\hat{\mathbf{a}}_{\mathbf{g}}$$

In which: $\hat{\mathbf{u}}$ is the vector with estimated SNP marker effects, **D** is a diagonal matrix of weights for variances of SNP effects, **Z** is a matrix relating genotypes of each locus to each individual, and $\hat{\mathbf{a}}_{\mathbf{g}}$ is the additive genetic effect for genotyped animals.

The individual variance of SNP effect (the same as in **D**) was estimated as (Zhang et al., 2010):

$$\hat{\sigma}^2_{u,i} = \hat{u}_i^2 2p_i(1 - p_i)$$

In which: $\hat{u}_i^2$ is the square of the *i*th SNP marker effect, p_i is the observed allele frequency for the second allele of the *i*th marker in the current population.

When windows of *n* adjacent SNPs were used; the variances attributed to them were calculated by summing the variance of the next *n* SNPs, for each SNP. Next, the combination that contained the highest values for exclusive windows was chosen to avoid double counting. It could happen that some windows had less than *n* SNPs if they were between two windows explaining more variance or in a window at the end or beginning of a chromosome. However, those smaller windows do not explain significant part of the variance.

The analyses were performed in four scenarios: complete data set; only genotypes and phenotypes from generations G1, G2 and G3; generations G2, G3 and G4; and from generations G3, G4 and G5. Numerator relationship matrix was complete in all scenarios. All ssGWAS computations were performed using the BLUPF90 family programs (Misztal et al., 2002) modified to account for genomic information (Aguilar et al., 2010).

The choice for ssGWAS was due to its ability to support phenotypes from ungenotyped animals directly, to handle multiple trait models, and to avoid spurious solutions on SNP effects due to sampling. Sampling in Bayesian alphabet family models is strongly dependent on priors and may produce spurious SNP estimates (Gianola et al., 2009, van Hulzen et al., 2012). Comparing GWAS

models in a simulated population, Wang et al. (2012) showed that ssGWAS was the most accurate method to capture the effect of potential QTLs; windows of SNP effects were used in their study.

RESULTS

Preliminary results showed small individual SNP variances for all three traits, with just a few SNPs explaining more than 0.5% of the variance of the trait (Figure 1). Experiments with different SNP window sizes exhibited large noise with small sizes and absence of peaks with large sizes. Subsequently, windows of 20 SNP were chosen as a reasonable size.

The variance explained by each SNP window is shown in Figures 2-4 (corresponding to body weight, breast meat, and leg score, respectively); also, the 10 largest points were marked with a red vertical line. It is possible to see that all those traits are mainly affected by many regions with small effects, with few regions that explain more variance. These regions tended to change across the generations, but some of them retain a consistent value among the top 10 regions in all the scenarios, even though, the variance explained by those windows did not contribute significantly to the genetic variability of the trait.

For body weight, there were three regions that persisted among the top 10 in all the scenarios (Figure 2). Although these top three regions have been described before, the percentage of variance explained was small; only one region was above 2.5% and all the others were below 1.6%. The total variance explained by the top 10 windows summed up to 7.63%.

For breast meat, two regions were consistent among the scenarios (Figure 3). The window with larger effect for this trait explained 1.14% of the total variance, in the subset containing generations

3 to 5. The other windows explained at most 1%. The total variance explained by the top 10 windows was 6.26%.

For leg score, the value of just one region was constant across the analysis in chromosome 7 (Figure 4), the variance explained by this windows was 1.12% in the subset containing generations 3 to 5. All the other windows explained less than 1% of the genetic variance for this trait. The total variance explained by the sum of the top 10 windows was 6.01%.

DISCUSSION

In our study, the three persistent regions observed for body weight could be related with QTLs previously described in the literature. The region in chromosome 1 was consistent with the one described by Carlborg et al. (2003) that associated this with a QTL responsible for body weight. The region in chromosome 4 can be related with those found by Carlborg et al. (2004), Ikeobi et al. (2004), and Ankra-Badu et al. (2010), all of whom detected a QTL for body weight in this region. The region in chromosome 14 was close to that described by Jennen et al. (2004) and Carlborg et al. (2003) for body weight. For breast meat, the region in chromosome 3 was close to those reported by Ikeobi et al. (2004) and Uemoto et al. (2009) for pectoralis muscle mass, and to those found by Gao et al. (2011) for chest width. The other region, in chromosome 8, was related by Ikeobi et al. (2004) to the pectoralis muscle mass trait. For leg score, the region in chromosome 7 had no relationship with any QTLs described previously in the literature for this trait in chicken. Nevertheless, there is a sequence of homeobox genes in the region around 16Mbp in the same chromosome in the chicken genome. These homeobox genes (HOXD4, HOXD8, HOXD9, HOXD11, HOXD12, and HOXD13) are related with regulation of anatomical development, and

might have a relationship with the leg disease score (Hillier et al., 2004). Thus, the findings in the current research are in concordance with Hayes and Goddard (2010), that a small number of markers with validated associations would explain a small portion of the genetic variance in the trait.

Wolc et al. (2012) found that for egg traits in layer chicken most of the SNPs with large effect were consistent across six generations, in both training and validating datasets. These findings could not be supported by the present results. Even though variances from three windows for body weight, two for breast meat, and one for leg disease score in the present study were stable across generations, for the other regions the results were different; it is possible that the lack of regions with larger effect on these traits, as illustrated in Figures 2 to 4, is the reason for the difference in findings. Another possible reason is the method used by the aforementioned authors; they used the BayesB method, which assumes large effect for a few markers and is highly influenced by the prior information (Gianola et al., 2009, van Hulzen et al., 2012). In addition, the generation interval in layer chicken is a few times longer than in broiler chicken so their generations may have been overlapping. Yet, the genetic architecture could be different among the traits in the present study and in the aforementioned work.

Large changes in the variance explained by SNP windows could be indirectly due to small effective population size and subsequent low number of independent chromosome segments. According to Daetwyler et al. (2008) and supported by Goddard (2009), the number of such segments (q) is equal to $2N_eL/log(4N_eL)$, where N_e is the effective population size and L is the length of chromosome in Morgans. Assuming N_e=50 (lower range showed in Andreescu et al., 2007) and L=39, q=435. Subsequently there are > 100 SNP per 1 chromosome segment, if we apply the formula to this dataset. This causes collinearity and possibly a high variance inflation factor for

the estimators, amplified by changes to the effective population size during the selection. While 435 segments suggest that 435 SNP could explain nearly all variation, this is not so as the boundary between segments is fluid.

Meuwissen et al. (2001) have found a small decay in accuracy as the relationship between prediction and training generations decreases in a simulation study. According to the authors this decrease was small enough to maintain the success of breeding schemes after 6 generations without re-estimation of SNP effects, however, their simulation assumed random mating. Also in a simulation study, Sonensson and Meuwissen (2009) found that re-estimating the genomic effects in every generation can maintain the accuracy of the predictions of breeding values constant. Solberg et al. (2009) also found a decrease in accuracy in further generations. They observed that with a denser panel the decay was smaller, which is probably a consequence of a higher LD between the markers and the simulated QTL. All above mentioned studies did not simulate selection in the data.

Muir (2007) showed that directional selection caused a great decline in accuracy of GEBV, demonstrating that high accuracies in the training generations were not maintained in future generations under selection. This can be a sign that the LD between marker and QTL can be lost across generations under selection, and can result in the changes observed in the present study. Alternatively, the QTL with largest effects are rapidly fixed by selection leaving SNPs with small effects remaining. In a real dataset from layer chicken, Wolc et al. (2011) demonstrated that the decay in accuracy was large enough to require a retrain of the model in every generation. Accurate estimations of genomic breeding value depend on the consistency of LD between markers and QTLs across generations (Calus 2009), as well as proper SNP effect estimation. The LD is created and maintained by the selection process, among other factors (Lynch and Walsh, 1998). On the

other hand, if a change in the allele frequency of two different loci is observed, which can be caused by selection, the LD between them can decrease (Calus, 2009). The results shown in those studies clearly display a loss of genomic prediction accuracy due to the decay of LD. This could also be extended to GWAS, and the negative impact LD decay might have on the accuracy of associations. The variation in the estimates of SNP variance in the present study can be related with those findings, because using values estimated in a different generation would lead to low predictive power if they are not constant.

The small values for SNP effect and percentage of variance explained that were obtained in this study can be related to the findings on Muir et al. (2008). The authors found significant absence of rare alleles in commercial chicken lines. Such findings were related to high inbreeding and consequently to a considerable number of alleles missing, which will reduce the allelic and genetic variability. This narrowed genetic variability can result in weaker associations for the markers, since important alleles could be lost in the process.

The short-term decay in accuracy depends more on the decrease of genomic relationships captured by markers rather than on LD (Habier et al., 2007). Therefore, the accuracy of genomic evaluation is mainly controlled by genomic relationships (Daetwyler et al., 2012, Wientjes et al., 2013). In particular, Daetwyler et al. (2012) found that 86% of the accuracy in genomic selection was retrieved by using SNP from a single chromosome. Subsequently, windows with large effects in Manhattan plots may be an artifact of relationships and not due to LD. The reason why the accuracy does not collapse completely in further generations is that some LD still persists over time, even though selection process and divergence can erode LD. Thus, the observed changes in the SNP effects across the generations in the present study can be a consequence of the changes in the relationship structure across different generations more than decay in LD.

CONCLUSION

Except for a few regions, the variation explained by the top SNP windows changes over generations. Therefore, even if SNP windows with large variance are detected in a particular data set, their usefulness for genomic selection over many generations is limited. The variance explained by an individual window is not enough to lead selection decisions based on the top regions for the studied traits.

ACKNOWLEDGEMENT

This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2009-65205-05665 from the USDA National Institute of Food and Agriculture. We would like to acknowledge Miguel Perez-Enciso for the useful comments made on this paper and Cobb-Vantress Inc. for providing the data.

REFERENCES

Andreescu, C., Avendano, S., Brown, S.R., Hassen, A., Lamont, S.J., Dekkers, J.C. (2007). Linkage disequilibrium in related breeding lines of chickens. *Genetics.* 177, 2161–2169. doi: 10.1534/genetics.107.082206.

Ankra-Badu, G., Shriner, D., Le Bihan-Duval, E., Mignon-Grasteau, S., Pitel, F., Beaumont, C., Duclos, M.J., Simon, J., Porter, T.E., Vignal, A., Cogburn, L.A., Allison, D.B., Yi, N., Aggrey, S. (2010). Mapping main, epistatic and sex-specific QTL for body composition

in a chicken population divergently selected for low or high growth rate. *BMC genomics*. **11**:1. 107.

Bolormaa, S., Pryce, J. E., Hayes, B. J., Goddard, M. E. (2010). Multivariate analysis of a genome-wide association study in dairy cattle. Journal of dairy science. **93**:8. 3818-3833.

Calus, M. P. L. (2010). Genomic breeding value prediction: methods and procedures. *Animal.* **4**:2 157-164.

Carlborg, O., Hocking, P. M., Burt, D. W., & Haley, C. S. (2004). Simultaneous mapping of epistatic QTL in chickens reveals clusters of QTL pairs with similar genetic effects on growth. *Genetical Research.* **83**:3. 197-209.

Carlborg, Ö., Kerje, S., Schütz, K., Jacobsson, L., Jensen, P., & Andersson, L. (2003). A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome research*. **13**:3. 413-421.

Chen, C. Y., Misztal, I., Aguilar, I., Tsuruta, S., Aggrey, S. E., Wing, T., & Muir, W. M. (2011). Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: An example using broiler chickens. *Journal of animal science*. **89:**1. 23-28.

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

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

Dikmen, S., Cole, J.B., Null, D.J., Hansen, P.J. (2013). Genome-Wide Association Mapping for Identification of Quantitative Trait Loci for Rectal Temperature during Heat Stress in Holstein Cattle. *PLoS ONE* **8**:7. doi:10.1371/journal.pone.0069202

Gao, Y., Feng, C. G., Song, C., Du, Z. Q., Deng, X. M., Li, N., & Hu, X. X. (2011). Mapping quantitative trait loci affecting chicken body size traits via genome scanning. *Animal genetics*. **42**:6. 670-674.

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

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

Goddard, M. E., & Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics. **10**:6. 381-391.

Groenen, M. A., Megens, H. J., Zare, Y., Warren, W. C., Hillier, L. W., Crooijmans, R. P., Vereijken, A., Okimoto, R., Muir, W.M., Cheng, H. H. (2011). The development and characterization of a 60K SNP chip for chicken. *BMC genomics*, **12**:1, 274-283.

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

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. BMC bioinformatics. **12**:1. 186.

Hayes, B.J., Bowman, P.J., Daetwyler, H.D., Kijas, J.W., van der Werf, J.H. (2012). Accuracy of genotype imputation in sheep breeds. *Animal Genetics* **43**:1.72-80.

Hayes, B., Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. *Genome* **53**:11. 876-883.

Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., ... & Dodgson, J. B. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. **432**:7018. 695-716.

Hu, Z. L., Park, C. A., Wu, X. L., & Reecy, J. M. (2013). Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic acids research.* **41**:D1. D871-D879.

Ikeobi, C. O. N., Woolliams, J. A., Morrice, D. R., Law, A., Windsor, D., Burt, D. W., Hocking, P. M. (2004). Quantitative trait loci for meat yield and muscle distribution in a broiler layer cross. *Livestock Production Science*. **87**:2. 143-151.

Jennen, D. G., Vereijken, A. L., Bovenhuis, H., Crooijmans, R. P., Veenendaal, A., Van der Poel, J. J., Groenen, M. A. (2004). Detection and localization of quantitative trait loci affecting fatness in broilers. *Poultry science*. **83**:3. 295-301.

Lynch, M., Walsh, B. (1998). Genetics and analysis of quantitative traits. Sunderland: Sinauer Associates.

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

Misztal, I., A. Legarra, I. Aguilar. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of dairy science* **92**:9. 4648-4655.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., Lee, D.H. (2002). BLUPF90 and related programs (BGF90). Proc. 7th World Congr. Genet. Appl. Livest. Prod. Montpellier, France. Communication No. 28–07.

Muir, W.M. (2007). Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics* 124, 342–355.

Peters, S. O., Kizilkaya, K., Garrick, D. J., Fernando, R. L., Reecy, J. M., Weaber, R. L., Silver, G.A., Thomas, M. G. (2012). Bayesian genome-wide association analysis of growth and yearling ultrasound measures of carcass traits in Brangus heifers. Journal of animal science. **90**:10. 3398-3409.

Solberg, T.R., Sonesson, A.K., Woolliams, J.A., Odegard, J, Meuwissen, T.H. (2009) Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. *Genet Sel Evol.* **41**:53 (2009). doi: 10.1186/1297-9686-41-53.

Sun, X., Habier, D., Fernando, R.L., Garrick, D.J., Dekkers, J.C.M. (2011). Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian methods. *BMC Genetics* **5**:(Suppl 3):S13 doi:10.1186/1753-6561-5-S3-S13

Sonesson, A.K., Meuwissen. T.H.E. (2009). Testing strategies for genomic selection in aquaculture breeding programs. *Genet Sel Evol* **41**:1 37. doi: 10.1186/1297-9686-41-37

Uemoto, Y., S. Sato, S. Odawara, H. Nokata, Y. Oyamada, Y. Taguchi, S. Yanai O. Sasaki, H. Takahashi, K. Nirasawa, E. Kobayashi. (2009) Genetic mapping of quantitative trait loci affecting growth and carcass traits in F2 intercross chickens. *Poultry science*. **88**:3. 477-482.

van Hulzen, K.J., Schopen, G.C., van Arendonk, J.A., Nielen, M., Koets, A.P., Schrooten, C., Heuven, H.C. (2012). Genome-wide association study to identify chromosomal regions associated with antibody response to Mycobacterium avium subspecies paratuberculosis

in milk of Dutch Holstein-Friesians. *Journal of dairy science,* **95**:5. 2740-2748. doi: 10.3168/jds.2011-5005.

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

Wang, H., Misztal, I., Aguilar, I., Legarra, A., Muir, W.M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. *Genetics Research* **94**:02. 73-83. doi: 10.1017/S0016672312000274.

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

Wiggans, G. R., VanRaden, P. M., Cooper, T. A. (2011) Technical note: Adjustment of all cow evaluations for yield traits to be comparable with bull evaluations. *Journal of dairy science*. **95**:6. 3444–3447

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

Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D. J. & Zhang, Q. (2010). Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS One*. **5**:9. doi:10.1371/journal.pone.0012648

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

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., Lawlor, T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science*. **93**:2. 743-752.

Wolc, A., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., Preisinger, R., Habier, D., Fernando, R., Garrick, D. J., Hill, W. G., Dekkers, J. C. M. (2012). Genome-wide association analysis and genetic architecture of egg weight and egg uniformity in layer chickens. *Animal genetics*. **43**:s1. 87-96.

Muir, W. M., Wong, G. K. S., Zhang, Y., Wang, J., Groenen, M. A., Crooijmans, R. P., Megens, H.J., Zhang, H., Okimoto, R., Vereijken, A., Jungerius, A., Albers, G.A.A., Lawley, C.T., Delany, M.E., MacEachern, S., Cheng, H. H. (2008). Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *Proceedings of the National Academy of Sciences*. **105**:45. 17312-17317.

Daetwyler, H. D., Kemper, K. E., Van der Werf, J. H. J., & Hayes, B. J. (2012). Components of the accuracy of genomic prediction in a multi-breed sheep population. *Journal of animal science*. **90**:10. 3375-3384.

TABLES

Table A.1 – Number of animals with phenotypes and genotypes in each generation.

Generation	Phenotypes	Genotypes
G1	95,770	1,142
G2	72,795	1,165
G3	66,241	754
G4	52,808	801
G5	7,018	1,004
Total	294,632	4,866

Table A.2 – Number of observations, mean, and standard deviation for the three traits.

Trait	Observations	Mean	Standard Deviation
Body Weight	294,632	92.66	17.2
Breast Meat	75,377	45.68	7.22
Leg Score	294,632	1.17	0.38

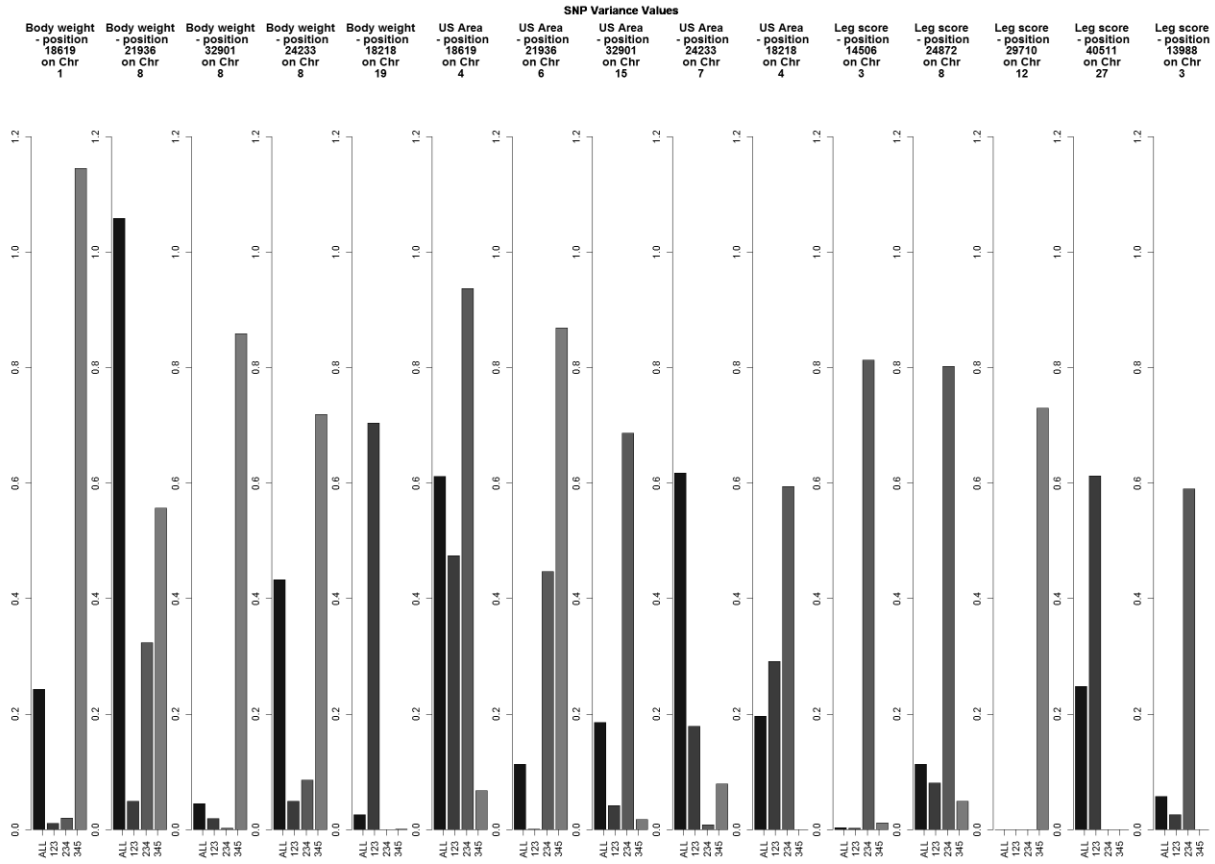

Figure A.1 Variance explained by the top 5 individual SNPs based on the combined results for all datasets for each trait.

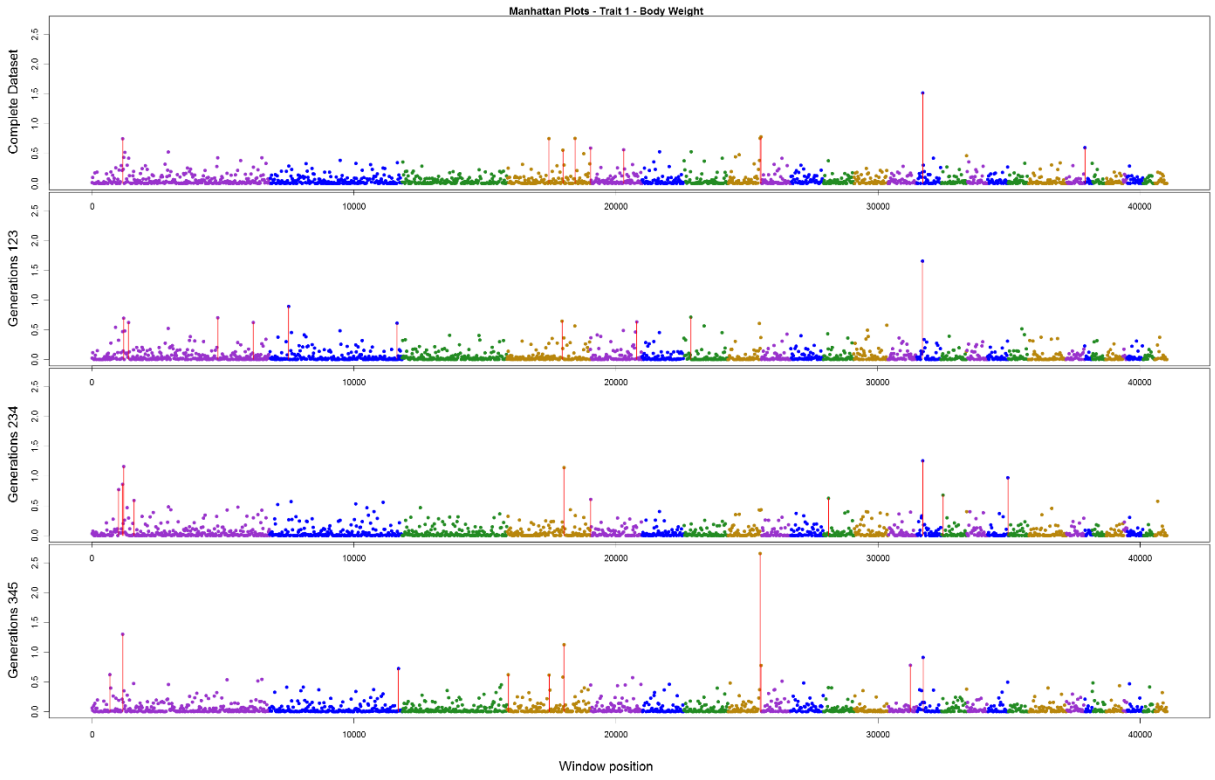

Figure A.2 Manhattan plots for percentage of variance explained for Body Weight, performed for all the data set, and the subsets of generations.

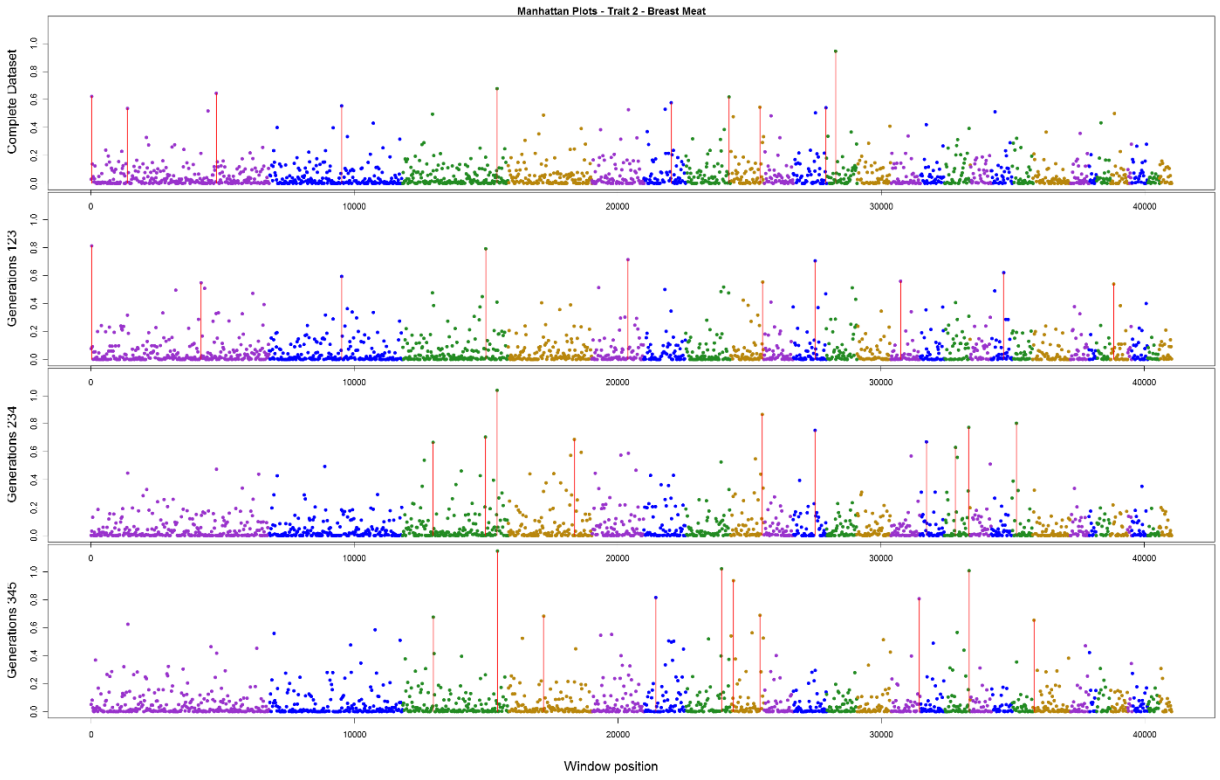

Figure A.3 Manhattan plots for percentage of variance explained for Breast Meat, performed for all the data set, and the subsets of generations.

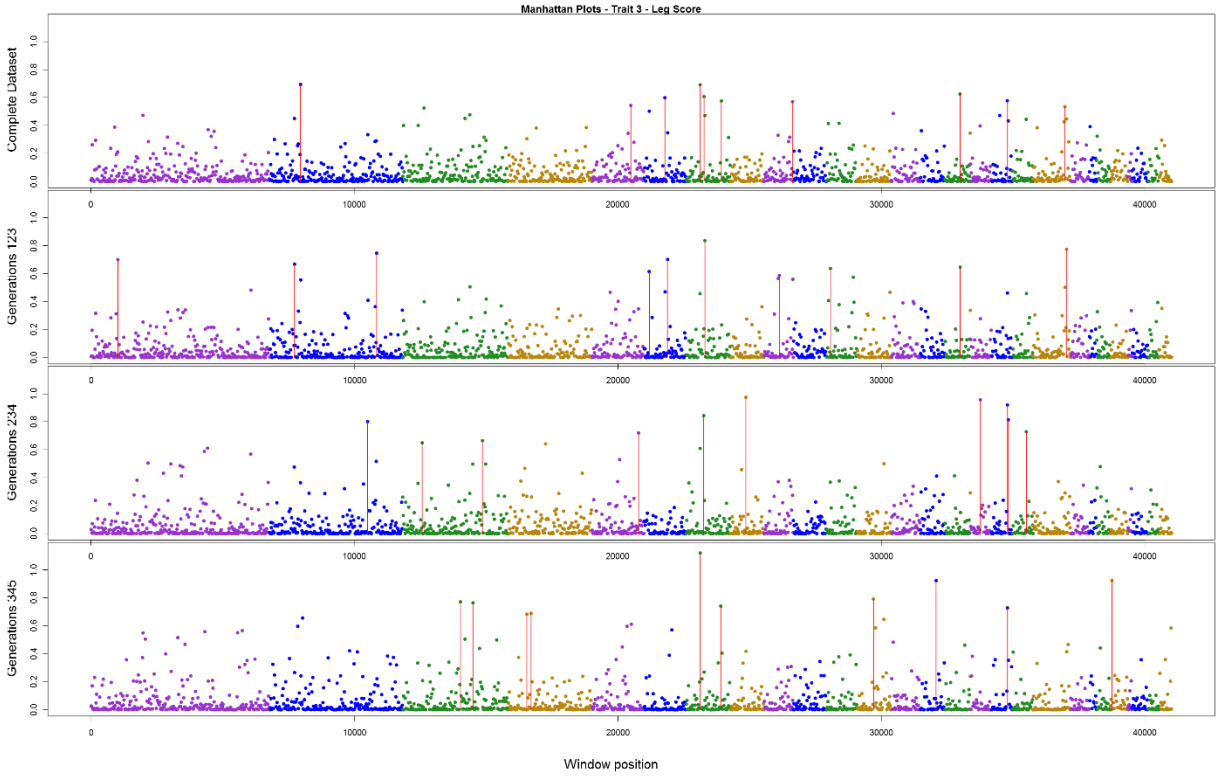

Figure A.4 Manhattan plots for percentage of variance explained for Leg Score, performed for all the data set, and the subsets of generations.