EFFECT OF MISIDENTIFICATION AND MULTIPLE POPULATIONS ON THE GENOMIC RELATIONSHIP MATRIX AND GENOMIC EVALUATIONS

by

REGINA MARIE SIMEONE

(Under the Direction of J. KEITH BERTRAND)

ABSTRACT

Mislabeled genotyped animals can impact genomic analysis, as they give rise to false relationships within the population. Diagonal elements of the genomic relationship matrix (**G**) may be a useful indicator of mislabeled animals if **G** is scaled using current allele frequencies. Simulated data were used to find theoretical diagonal elements of **G** and field data were used to evaluate the utility of diagonal elements to separate a small and large population of animals in a genotyped chicken dataset. The effect of mislabeled animals on the accuracy of genomic predictions was also evaluated. When the diagonal elements are centered close to 1.00, mislabeled animals may have incorrectly scaled and abnormally large diagonal elements. Use of diagonal elements of **G** can identify animals from secondary populations; populations must be of unequal size or have different allele frequencies. Presence of mislabeled animals negatively affected genomic evaluations through loss of accuracy.

INDEX WORDS: Allele frequency, Chicken, Genomic relationship matrix, Genomic selection, Single-step procedure

EFFECT OF MISIDENTIFICATION AND MULTIPLE POPULATIONS ON THE GENOMIC RELATIONSHIP MATRIX AND GENOMIC EVALUATIONS

by

REGINA MARIE SIMEONE

BA, Swarthmore College, 2006

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

MASTER OF SCIENCE

ATHENS, GEORGIA

2010

EFFECT OF MISIDENTIFICATION AND MULTIPLE POPULATIONS ON THE GENOMIC RELATIONSHIP MATRIX AND GENOMIC EVALUATIONS

by

REGINA MARIE SIMEONE

Major Professor: J. Keith Bertrand

Committee: Ignacy Misztal
Shogo Tsuruta

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2010

DEDICATION

To my family.

You have been with me every step of the way.

ACKNOWLEDGEMENTS

I would like to thank Dr. Ignacy Misztal, who has worked tirelessly with me in developing this study and the analysis. I am consistently amazed at the amount of dedication and time that he gives to each of his students and I know that I could not have completed this thesis without his guidance. In addition to being a source of academic support, Dr. Misztal has always shown me kindness and encouraged me outside of the computer lab. It has been a pleasure and an honor to work with such a talented and patient individual.

I would like to thank Dr. J. Keith Bertrand, for believing that I could succeed in this program and for always having a kind word or piece of advice to share.

Dr. Shogo Tsuruta has always been available for me to ask questions and favors, even when he has been at his busiest. I would like to thank him for his calm demeanor and encouragement throughout these two years, as well as for agreeing to help with my defense at the final hour.

Drs. Ignacio Aguilar, Chin-Yi Chen, and Selma Forni have always been happy to help me with the smallest programming or theory questions and my day-to-day work was made much easier because of their presence.

Ryan Davis is an excellent editor and showed great patience with the many drafts of my papers. The same goes to Lee Tittsworth, who endured multiple drafts of multiple papers.

Finally, I would like to thank Ryan Davis, Jamie Williams, and Joy King. They have been constant sources of support and laughter and their encouragement has seen me through many difficult days.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

The goal of animal breeding is to improve the genetic potential of animals for economically important traits. This has traditionally been done through selection programs that make use of vast pedigrees and data collection on traits of interest and implement best linear unbiased prediction (BLUP) methodologies. Further advances in animal breeding are now possible through use of a technique known as genomic selection. Genomic selection uses information from dense single nucleotide polymorphism (SNP) marker maps to obtain estimates of SNP effects on traits of interest. In theory, genomic selection can increase the rate of genetic gain, increase the accuracy of prediction, and reduce generation intervals. SNP estimates can be used to obtain genomic estimated breeding values that have accuracies close to traditional pedigree-based predictions. Alternatively, SNP markers can be used to construct a genomic relationship matrix (**G**), which is a matrix of realized, as opposed to expected, relationships. **G** can provide more accurate relationships than the additive relationship matrix (**A**) and may be an acceptable substitute for **A** in genotyped populations. It is also possible to combine **G** with **A** in the case of mixed populations of genotyped and phenotyped animals.

Successful implementation of the genomic relationship matrix requires error-free phenotypic and genotypic datasets. Errors in datasets will lead to false predictions or misleading relationships. While observed parent-progeny genotype conflicts can be used to identify possible genotyping errors on the molecular level, errors that lead to

incorrect population structure, such as mislabeling of genotyped animals, are more difficult to detect. Incorrect population structure can give rise to spurious associations between SNPs and quantitative trait loci (QTL) or can lead to an incorrect estimation of allele frequencies in the population. Methods to detect mislabeled animals within a population are needed.

The presence of mislabeled animals in a genotyped dataset may impact the construction of **G** and negatively affect the evaluation of animals. The purpose of this work was threefold: to evaluate the utility of **G** in identifying mislabeled animals in a genotyped population, to explore the effect on prediction of the presence of mislabeled genotyped animals in a genomic evaluation, and to explore the effect on prediction of evaluating a multi-line population using different allele frequencies to scale **G**.

CHAPTER 2

LITERATURE REVIEW

Improved management techniques, selection programs, and an increased understanding of quantitative genetics have brought about tremendous changes in animal breeding. The field continues to develop as individual breeders and large companies search for newer and more efficient ways to breed high producing and economically valuable animals. Quantitative genetics has become an invaluable tool in animal breeding and is being utilized in national evaluation systems for a variety of species; moreover, advanced technological and molecular methods are being incorporated into traditional breeding programs to augment the already powerful statistical procedures available. Improvement of quantitative genetics techniques is crucial to the continued development of the industry, as knowledge of animal genetics decreases the cost of production, increases animal welfare, and increases the accuracy of selection.

Animal breeding programs are centered on mating livestock with the highest genetic potential in order to produce offspring with even higher genetic potential. Traditional animal breeding makes use of animal relationships through extensive pedigrees; from these pedigrees, it is possible to assign value to animals in terms of their ability to transmit important genes to their offspring. More advanced technologies, however, have emerged. The mapping of various genomes has allowed the fields of molecular and quantitative genetics to converge. The future of animal breeding may now lay in evaluations that make use of detailed genomic information.

Traditional Animal Breeding and Beyond

The goal of any genetic improvement program is to maximize the rate of increase of economically important traits thought to be controlled by an animal's genetic composition (Gianola, 2000). Traits of economic importance in animal breeding, such as milk production, meat quality, and egg size tend to be quantitative in nature. A quantitative trait is one controlled by more than one gene, meaning that the phenotypes of the trait tend to follow a normal distribution. Selection for these traits is a complicated endeavor (Bourdon, 2000).

Animal breeding has traditionally consisted of a model relating one or more traits of interest to parameters thought to have an effect on the phenotypes. These effects can be genetic, modeled as random effects, or environmental, modeled as fixed effects (Gianola, 2000). Gianola (2000) presented important landmarks in traditional animal breeding that have enabled breeders to predict genetic merit in selection candidates. Breeders generally consider an animal's breeding value (BV) for a trait to be of utmost importance. A breeding value is the sum of the additive effects of an animal for a particular trait (Falconer and Mackay, 1996). Perhaps one of the most widely used methods in animal breeding is that of best linear unbiased prediction, or BLUP, which was developed by C. R. Henderson in 1973. BLUP is a mixed model equation that maximizes the correlation between true and predicted breeding values and minimizes prediction error in order to estimate effects and predict the estimated breeding value (EBV) (Mrode, 2005). The equation is as follows:

$$\mathbf{y}=\mathbf{X}\mathbf{b}+\mathbf{Z}\mathbf{a}+\mathbf{e}\,,$$

where **b** is a vector of fixed effects, **a** is a vector of random effects, and **X** and **Z** are incidence matrices. This has become known as a mixed model equation (MME) and is represented as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix},$$

where **R** and **G** are variance-covariance matrices ($\text{var}(e) = \mathbf{I}\sigma_e^2 = \mathbf{R}$ and $\text{var}(a) = \mathbf{A}\sigma_a^2 = \mathbf{G}$). The numerator relationship matrix (**A**), which designates the expected additive genetic relationships among individuals in a pedigree, is vital to this equation (Mrode, 2005). Other significant landmarks include methods to estimate variance components. One of the most widely used is that of restricted maximum likelihood (REML). Bayesian methodologies have also substantially impacted animal breeding through applications of Markov chain Monte Carlo (MCMC) methods such as Gibbs sampling. Statistical methods in animal breeding extend beyond linear models into models for categorical traits (Gianola, 2000).

These statistical methods have led to unprecedented levels of accuracy in the prediction of animal merit, allowing for substantial gain in traits of interest. For example, from 2000 to 2009, milk production in the United States has seen a 13% increase in kilograms produced per year (Figure 2.1). Similarly, extensive gains have been seen in meat production in broiler chickens (Figure 2.2). Uses of progeny testing and parent averages have lead to noteworthy gains in animal production; limitations, however, do exist. Hayes et al. (2009b) have outlined some of the hurdles of traditional animal breeding. First, progeny testing to obtain high accuracy sires is a lengthy process, particularly in cattle. Obtaining an EBV with 75% accuracy in a young bull takes approximately five years (Schaeffer, 2006). Second, although relatively easy to

implement with traits of moderate to high heritability, traditional selection is less successful for traits with low heritability, such as fertility (Hayes et al., 2009b; Hoglund et al., 2009). Third, selection has focused almost exclusively on production traits resulting in important secondary traits such as fertility and health, which often have a negative correlation with production traits, experiencing little gain (Hoglund et al., 2009). Sex-limited traits are also problematic, particularly for traits expressed in females, due to selection being more focused on male animals than female. Similar problems exist for traits expressed late in life or after death, such as carcass quality (Heuven et al., 2009; Hoglund et al., 2009). Fourth, the genetic effects of a trait of interest are often confounded with environmental effects, making accurate prediction difficult (Hayes et al., 2009b). Quantitative geneticists and animal breeders have worked to address these issues and increase the genetic gain possible in selection.

Despite significant gains in animal breeding through the use of BLUP and animal breeding programs, greater gains can be achieved. The availability of molecular information has opened new avenues for genetic improvement. The sequence of the human genome was mostly completed by 2003 (Schmutz et al., 2004). Sequences of the bovine and chicken genome soon followed in 2004; all are continually being updated (Liu et al., 2009; Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution, 2004). The availability of such rich genetic information has spawned research into molecular genetics as a source of improvement in animal breeding. Genes are the building blocks of traits of interest; selecting directly on available genes is an exciting prospect.

Quantitative Trait Loci and Molecular Markers

Quantitative trait loci (QTL) are areas of the genome thought to control the genes that may influence a trait of interest (Beuzen et al., 2000). Making selection decisions based on QTL is a tempting prospect; these loci, however, are generally difficult to find and expensive to map. Additionally, quantitative traits of interest have complicated expression and are mostly under the control of many QTL. For these reasons, research has focused on the use of markers to select for QTL without knowing what or where the QTL actually are.

The use and availability of molecular markers depend on the trait of interest. In humans, markers are generally used in family and association studies for human diseases. In livestock, markers are generally used for production and secondary traits of interest. Restriction fragment length polymorphisms (RFLPs) are nucleotide changes in the genome that result in the creation or deletion of a restriction site. If a DNA sequence is cleaved in an inappropriate location, fragments that are either too long or too short will be detected through the use of gel electrophoresis (Beuzen et al., 2000). These fragments were the first type of genetic marker used in the identification of polymorphisms throughout the genome. RFLPs are informative yet problematic because many mutations that may affect traits do not result in a restriction site change and cannot be identified by gel electrophoresis. For this reason, creating a dense map of DNA markers using RFLPs is difficult and time consuming and use has faded in lieu of more appropriate markers (Vignal et al., 2002).

Microsatellite markers quickly replaced RFLPs as a way to identify polymorphisms (Vignal et al., 2002). A microsatellite marker is made up of nucleotide

repeats; many of these occur throughout the genome. Microsatellites are informative because they are polymorphic and abundant, but they are susceptible to mutation and results are generally not reproducible between laboratories (Vignal et al., 2002).

Single nucleotide polymorphisms (SNPs) have become popular for use in genetic marker studies. An SNP is a single base change at a locus within the genome; such occurrences are prevalent in human and livestock. For example, in humans one SNP occurs every 100 bases (Beuzen et al., 2000). SNPs are stably inherited from parent to offspring and are ideal for studies that make use of family information. Moreover, SNPs are suitable for high-throughput genetic techniques that run numerous samples simultaneously. SNP markers can be used to make dense molecular maps. These maps are necessary for genetic selection or gene identification schemes using molecular markers. Although SNPs are less variable than microsatellites, they are easier to trace from parent to offspring and their presence is clear. Whereas microsatellite markers can have numerous repeats and may be more informative than SNPs, they are problematic in that they are not abundant enough within various genomes to be of great use in association studies (Bahram and Inoko, 2007). SNP markers can be used to rapidly and inexpensively genotype populations.

The availability of prevalent and inexpensive SNP markers has encouraged research in the use of SNPs to identify disease genes in humans or for selection purposes in animals. Initial work in humans and livestock involved searches for genes associated with traits of interest. Two approaches exist for such studies: candidate gene approach and genome scanning. The rationale behind the candidate gene approach is that quantitative genetic variation in a phenotype is due to a functional mutation of a gene.

The candidate gene approach tests probable genes for linkage to a disease or trait of interest (Zhu and Zhao, 2007). Such an approach, though informative, is restricted in that it requires prior knowledge of genetic architecture to identify genes of interest. The requirement of prior knowledge makes the candidate gene approach subjective and increases the chance that researchers will miss important genes. Moreover, replication of results is generally low and there is no guarantee that a phenotype is controlled by only one QTL or gene (Zhu and Zhao, 2007).

Despite its limitations, the candidate gene approach has been successful in identifying some major genes in both humans and livestock. Grisart et al. (2002) were able to use positional cloning to map the gene DGAT1, which has been shown to have a substantial impact on milk composition and fat content in dairy cattle. Similarly, Drogemuller et al., (2001) mapped three RFLPs having an effect on litter size in German pig lines. In humans, the candidate gene approach has been able to map genes related to risk factors for rheumatoid arthritis (Kurreeman et al., 2007), breast cancer (Pharoah et al., 2007), and other common diseases and addictions.

Genome wide association studies (GWAS) approach gene identification differently. In GWAS, a dense set of SNPs in a genome are tested for association with genetic variation of a disease or quantitative traits of interest in livestock (Hirschhorn and Daly, 2005). With GWAS, entire genomes are scanned for linkage with a disease. Such studies therefore have the capacity to identify traits of both Mendelian and quantitative nature. Typically, GWAS uses a case-control structure in which allele frequencies of individuals with the trait of interest are compared to those without the trait of interest (Pearson and Manolio, 2008). GWAS is susceptible to a number of pitfalls, including

sensitivity to genotyping errors (Pearson and Manolio, 2008), the difficulty of separating gene and environmental contributions to a phenotype (Maher, 2008), and spurious associations due to population structure or selection bias (Wang et al., 2005).

Marker Assisted Selection

Marker assisted selection (MAS) was one of the first attempts to use information from genetic markers for selection purposes. Dekkers (2004) defines MAS as "direct selection on genes or genomic regions that affect economic traits." As the name suggests, MAS relies on genomic markers, namely SNPs, associated with traits of interest. Dekkers describes three types of observable loci available for MAS: direct markers, linkage equilibrium (LE) markers, and linkage disequilibrium (LD) markers. A direct marker is an actual gene or functional mutation that directly affects the trait of interest. These types of markers are generally adequate for simply inherited traits, but are difficult to detect and causality is not easy to prove (Dekkers, 2004). LE markers are SNP loci in population-wide LE with a functional mutation or QTL; LE is defined as the random association between alleles at two or more loci (Falconer and Mackay, 1996). LE markers are easily detected using breed crosses or large half-sibling families and require sparse marker maps. To select on LE markers, one must ensure that the phase of the alleles between two populations is the same (Dekkers, 2004). LD markers are SNP loci in population-wide LD with a QTL; LD is defined as the nonrandom association between alleles at two or more loci (Falconer and Mackay, 1996). LD markers are generally close to a gene of interest. Testing for LD is the method used in candidate gene and GWAS studies. Both direct and LD markers allow for selection across populations, whereas LE

markers may not have that capability due to phase differences between alleles in multiple populations (Dekkers, 2004).

Guillaume et al. (2008) conducted a simulation study to compare the reliability of genetic values of young animals obtained with and without marker information for milk, fat, and protein yields for two populations from 2004 and 2006. Forty-five microsatellite markers were used to follow 14 QTL. They found that MAS-obtained EBVs were better predictors than parent average predictors; moreover, MAS performed better in the 2006 population than it did in the 2004 population, likely due to the presence of more genotype records. MAS also saw an increase in accuracy when more progeny were genotyped and had records. The group suggested, however, that the superiority of MAS would likely be reduced when faced with many QTL of small effects as opposed to a few QTL with large effects.

Goddard and Hayes (2007) summarized several factors that govern the success of MAS. If the existing EBVs for a trait of interest are already high, MAS will provide very little gain. Substantial gain would instead be expected if MAS was used for traits that are difficult to select in the traditional manner. The proportion of genetic variance actually explained by the QTL markers will also dictate genetic progress. Marker effect estimation must be accurate as well, or selection may have small or negative effects.

Though MAS methodologies initially seemed promising, the expected gains were not observed in practical applications. One reason for this includes recombination between markers and QTL, which breaks down the relationship between markers and QTL. If QTL are close enough to markers, a recombination event is rare, even after multiple generations; therefore, markers must be very close to QTL for MAS to be

successful (Boichard et al., 2006). Much of the excitement surrounding marker assisted selection was due to the possible gains for lowly heritable traits, like fertility traits. Unfortunately, such traits are not completely understood and searching for markers is a nearly impossible feat (Misztal, 2006). Many traits of interest are also affected by environmental factors, epistatic interactions, or dominance interactions, which may confound marker effects or change the expression of a QTL (Misztal, 2006).

QTL and Linkage Disequilibrium

Marker assisted selection depends on accurately mapped QTL explaining a large portion of genetic variation. Without such QTL, MAS is not an adequate selection tool. Hayes and Goddard (2001) performed a meta-analysis of information from dairy cattle and pig mapping experiments regarding QTL controlling growth, carcass and meat quality in pigs, and fat percentage, protein yield, fat yield, and milk yield in dairy cattle. They found that the distribution of QTL in both pig and dairy cattle is leptokurtic – there are many QTL with small effects and few with large effects (Figure 2.3). This implies that most phenotypic variation, particularly for quantitative traits, is the result of numerous QTL and no one QTL can determine a phenotype. Moreover, it has been noted that mapping is an inexact science and it is difficult to map QTL with much precision without using massive family studies; QTL are therefore mapped to very large confidence intervals, making MAS difficult to implement effectively (Hayes, 2010).

Markers in LD with QTL may alleviate some of the problems of imprecise mapping. LD is used extensively in MAS because it is easier to detect than functional mutations and phase does not matter; marker LD can thus be used in multiple

populations. The r^2 measurement is currently widely implemented for LD studies in multiple species. The formula to obtain r^2 is:

$$r^2 = \frac{D^2}{freq(A1) \times freq(A2) \times freq(B1) \times freq(B2)},$$

in which D is equal to:

$$D = freq(A1\_B1) \times freq(A2\_B2) - freq(A1\_B2) \times freq(A2\_B1),$$

where Freq(A1_B1) is the frequency of the A1_B1 haplotype in the population (Hayes, 2010). The unit r^2 ranges from 0 (no LD between A1 and B1) to 1 (complete LD between A1 and B1). It is currently the preferred measurement of LD between markers and QTL (Hayes, 2010).

The amount of LD present in a genome varies among species and populations and can be caused by numerous factors, namely, mutation, migration, selection, and small population size. In livestock, small population size is generally considered the main cause of LD (Hayes, 2010). While marker to QTL LD is difficult to map, marker to marker LD in a species is a good estimate for the extent of marker to QTL LD in a genome. Heifetz et al. (2005) evaluated the marker to marker LD between commercial lines of layer chickens using microsatellite markers. They found that significant LD existed between markers separated by less than 5 cM. Additionally, LD was conserved between generations, suggesting that LD would not have to be re-estimated every generation to implement selection. LD was, however, specific to each line, making it difficult to use the results from one line in another. Andreescu et al. (2007) evaluated LD among nine breeding lines on two chromosomes of broiler chickens. They reported that LD extended over shorter distances than previously reported in livestock and that LD is consistent between closely related lines. Moreover, LD declined as expected, as the distance

between markers increased. Sargolzaei et al. (2008) used 10,000 SNPs across the genome to analyze LD in Holsteins. The group found that LD decayed rapidly at distances greater than 100 kb but within 100 kb, substantial LD occurred ($r^2 = 0.58$); within 10 kb, the mean r^2 was 0.73. Despite these results, it was suggested that a denser SNPs map was necessary to obtain reliable associations. It should be noted that despite the species used, LD decays over distance and over the course of numerous generations.

Another question of interest is whether LD markers in one breed can be used in another; the ability to use the same markers for multiple breeds would be efficient in terms of cost and time because SNP effects would not have to be estimated for each population. Goddard et al. (2006) examined the r^2 between Angus and Holstein cattle using 9,323 SNPs. They found that when markers are close together (< 10 kb apart), LD exists between breeds; however, at greater distances, LD was not consistent and the phase between markers in both breeds is often reversed. This means that the same marker might be favorable for one breed but unfavorable for the other (Goddard et al., 2006). It was concluded that LD existed in ancestral populations of Angus and Holstein prior to the species' divergence; the LD between SNPs residing within 10 kb of each other is still maintained between divergent populations because the recombination distance was so small and little breakdown of LD occurred over time.

The amount of LD between Holstein-Friesian, Jersey, and Angus cattle was compared by de Roos et al. (2008). As genomic distance increased, LD decreased rapidly in all populations (Figure 2.4). As expected, the LD phase was different between multiple populations particularly at larger genomic intervals; dense marker maps were suggested for any selection program because markers were more likely to be in LD with QTL. The

authors suggested that while a map of 50,000 markers was appropriate for selection within breeds, selection between breeds would ultimately require a map of 200,000 to 300,000 markers to ensure distances of 10 to 15 kb between markers. High throughput genotyping chips from Illumina and Affymetrix allow rapid analysis of entire genomes with dense maps of 50,000 to 60,000 SNP (Li et al., 2008); denser maps are currently being researched. Until such dense maps are created, markers effects are only being estimated for single-breed selection programs.

Genomic Selection

Following the disappointing performance of direct selection strategies, Meuwissen et al. (2001) proposed the genomic selection (GS) procedure, which exploits LD between known markers within a genome. GS assumes that with a dense marker map, markers will be sufficiently close to some unknown QTL to be in LD. Moreover, rather than estimating the marker effects one by one, marker effects are estimated simultaneously, thus avoiding having thousands of parameters to estimate with many fewer phenotypic observations and equations. One of the major benefits of GS is that knowledge of the exact location or function of a QTL is not necessary. Markers are correlated with positive effects on quantitative traits across all families and can be used for selection without establishing phase (Meuwissen et al., 2001). Meuwissen et al. tested GS by simulating a genome of 1,000 cM with haplotype markers evenly interspersed every 1 cM and analyzing it using multiple methods: least squares regression, BLUP, BayesA, and BayesB.

Least squares regression is a stepwise approach in which genes are added to the model only if they significantly improve the fit of the model. No assumptions are made

regarding the distribution of SNP effects, which are treated as fixed (Hayes, 2010; Meuwissen et al., 2001). First, a single SNP regression is performed using the model: $\mathbf{y}=\mu\mathbf{1}_n+\mathbf{X}_i\mathbf{g}_i+\mathbf{e}$, in which $\mathbf{y}$ is a vector of data, $\boldsymbol{\mu}$ is the overall mean, $\mathbf{g}_i$ is a vector of SNP effects, and $\mathbf{X}$ is a design matrix allocating records to genotypes. Second, the mode $\mathbf{y}=\mu\mathbf{1}_n+\sum_i \mathbf{X}_i\mathbf{g}_i+\mathbf{e}$ is used to compute genomic estimated breeding values (GEBVs). The least squares approach is problematic because the significance level of the SNPs may be too lenient and because SNPs effects may be overestimated due to multiple testing (Hayes, 2010).

By assuming that all SNP effects are random and that all genes explain an equal amount of variance, BLUP methods can be used to estimate SNP effects (Meuwissen et al., 2001). The BLUP method used the model $\mathbf{y}=\mu\mathbf{1}_n+\sum_i \mathbf{X}_i\mathbf{g}_i+\mathbf{e}$ to estimate SNP effects, $\mathbf{g}_i$; $\mathbf{X}$ is a matrix allocating marker genotypes to phenotypes. In Henderson's mixed model equations, this becomes $\begin{bmatrix}\hat{\mu}\\ \hat{\mathbf{g}}\end{bmatrix}=\begin{bmatrix}\mathbf{1}_n'\mathbf{1}_n & \mathbf{1}_n'\mathbf{X}\\ \mathbf{X}'\mathbf{1}_n & \mathbf{X}'\mathbf{X}+\mathbf{I}\lambda\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{1}_n'\\ \mathbf{X}'\mathbf{y}\end{bmatrix}$, in which λ is equal to σ_e^2/σ_g^2. A challenge for the implementation of BLUP includes choosing the value for σ_g^2, the SNP variance; one choice is estimating the total additive genetic variance and then dividing that number by the total number of SNPs. Once SNP estimates are obtained, GEBVs are obtained using the equation $\text{GEBV}=\mathbf{X}\hat{\mathbf{g}}$. BLUP is problematic in that the chromosome segment with the largest variance will be overestimated, though not as much as when least squares methods are used. Additionally, if some QTL have a large effect on a phenotype, BLUP methods are likely not appropriate (Hayes, 2010; Meuwissen et al., 2001).

Bayesian methods account for differences in SNP variability by assuming, *a priori*, that there are some QTL segments with large effects and others with little to no effect. Methods BayesA and BayesB treat SNP effects as random but allow the variance explained by each SNP to fluctuate according to some distribution (Meuwissen et al., 2001). Both methods analyze the model at the level of the data in the same way as the BLUP analysis; at the level of the SNP effects, the methods differ. Using BayesA Meuwissen et al. assumed the prior distribution of the SNP variance ($\sigma^2_{g_i}$) as a scaled inverted chi-square, $\chi^{-2}(v,s)$, in which v is the degrees of freedom and S is a scaling parameter. BayesB assumes many SNP markers that will not be in LD with QTL and will therefore not contribute to the genetic effect (Meuwissen et al., 2001). BayesB assumes that most SNPs markers will have a variance of 0 and that the rest will have a variance following a scaled inverted chi-square:

$$\sigma^2_{g_i} = 0 \qquad \text{with probability } \pi$$

$$\sigma^2_{g_i} \sim \chi^{-2}(v, S) \qquad \text{with probability } (1-\pi).$$

The advantage of the Bayesian approaches is that they may give a more realistic approximation of the distribution of SNPs effects, allowing for better and more accurate GEBV estimation.

Meuwissen et al. used each of the four methods to estimate the true breeding value (TBV) based on haplotype effect and correlated them with GEBVs, which the authors called EBVs. The correlation between true and estimated BVs represented the accuracy of selection on GEBVs; bias of the models was also taken into account by completing a regression of TBV on EBV (Table 2.1). They found that the least squares method had the lowest accuracy and the most bias; BLUP gave reasonable accuracies and

much less bias than least squares. The Bayesian methods gave the highest accuracy and the least amount of bias, with BayesB slightly outperforming BayesA (Table 2.1). This groundbreaking work has paved the way for numerous simulation and field data studies in GS.

Genomic evaluations have been performed in a number of species with the results generally suggesting that GS increases the predictive ability and accuracy of selection. Legarra et al. (2008) showed this with a GS experiment in a population of laboratory mice. They found that a model including only genomic information outperformed both a pedigree model and a model including genomic and pedigree information; the combined model was likely outperformed due to collinearity of SNP and additive effects. Muir (2007) examined the accuracy of prediction of GEBVs compared to those of traditional BLUP through simulation. He found that GEBVs will work well for traits, even those of low heritability, provided that enough training generations are used to estimate marker effects. Muir also stated that GS methods require phenotypic records and that breeding programs cannot rely solely on genotypic data; for that reason, scrupulous records must be maintained.

Numerous GS studies have been completed in dairy cattle and GS has been included in national evaluation procedures since 2009 (USDA). Studies have focused on production traits, such as fat percentage (de Roos et al., 2007), and fertility traits, such as days open, as these are historically difficult to select and improve (Hoglund et al., 2009). Moreover, comprehensive reviews of GS in dairy have been completed by VanRaden et al. (2009) and Hayes et al. (2009b). Current findings suggest that BLUP methods, while outperformed in the Meuwissen et al. (2001) paper by Bayesian methods, may perform

only slightly worse or even outperform Bayesian when used with real data instead of simulated data (Hayes et al., 2009b). One likely cause of this difference is that most traits may fit the BLUP model better than the Bayesian models: traits will likely be under the control many QTL of small effect rather than few QTL of large effect (Hayes et al., 2009b).

While most reports suggest that using the densest SNP panel available will give greater accuracies (Hayes et al., 2009b), studies have been completed to test whether subsets of SNP panels can also provide accurate measures. Subset panels would be worthwhile, as they are less expensive and therefore could be more widely used than dense panels. Gonzalez-Recio et al. (2009) examined food conversion rate in broiler chickens and observed that preselecting for the most informative SNPs, using nonparametric methods, was more effective than traditional BLUP or Bayesian methods using the entire genome.

SNP imputation is also an important technology being developed for use with SNP subsets. SNP imputation uses data from family and population based algorithms to build a complete SNP dataset using a smaller subset (Weigel et al., 2010). The accuracy of such imputation was investigated by Weigel et al. (2010) in Jersey cattle using different sized SNP subsets with and without population information. They found that accurate imputations were possible, particularly when a population is known. This has great implications for future work, as the method of choosing selection candidates could drastically change. Subset SNP panels could be used to genotype dams, which typically undergo less strenuous selection, and dense genotypes could be used to genotype sires, ultimately finding the best combinations of animal mating for optimal gain. The use of

SNP imputation could reduce cost as well increase the efficiency of GS programs provided that SNP subsets are appropriate for the populations and traits being studied (Weigel et al., 2010).

Despite the exciting promise of GS, real challenges still lie ahead. The cost of genotyping animals is a hurdle to producers, particularly for traits like fertility, which require large numbers of genotypes and phenotypes (Dekkers, 2004; Hayes et al., 2009b). Additionally, animals will need to be genotyped repeatedly because LD between markers and QTL will likely break down after several generations (Goddard, 2009). Animal scientists must be attentive to alternative methods of GS, such as using subset SNP chips, which could save time and money if developed for breeds and specific traits of interest. Other challenges include incorporating GS techniques into current parent-average based evaluation systems and using GS to evaluate multiple breeds.

The Genomic Relationship Matrix

In addition to excitement over the estimation and use of SNP effects in genomic breeding programs, research has also focused on using genomic information to construct genomic relationship matrices that replace traditional additive relationship matrices. This was suggested by Nejati-Javaremi et al. (1997) in order to increase the accuracy of selection by including more information about animal relationships. With more animals being genotyped for denser SNP chips, using the genomic relationship to predict animal effects is becoming a realistic possibility.

The additive relationship matrix, **A**, uses pedigree data to calculate expected probabilities of the genes shared among relatives. Such expected relationships have been used extensively in genetic improvement programs. However, they do not capture all of

the relationships within a population. First, pedigrees make the unlikely assumption that animals in the base population are unrelated (VanRaden, 2007). Second, **A** is a matrix of *expected* relationships; actual relationships may be more or less depending on the Mendelian sampling (VanRaden, 2007). The genomic relationship matrix, **G**, measures the amount of genes that are identical by descent between animals in a pedigree and calculates exact fractions of shared genes. With the SNP maps currently available, accurate relationship matrices can be constructed.

The genomic relationship matrix can be substituted for the additive relationship matrix in the traditional mixed model equations (Habier et al., 2007). Genotyped animals can obtain more accurate predictions due to better relationships shown in **G**. According to Habier et al. (2007), the accuracy of GEBVs is nonzero using **G** even if there is no LD in the population. This is promising because weak LD or the assumption of LD when no LD exists detracts from genetic accuracy and gain. VanRaden (2007; 2008) has provided methods to calculate **G** using markers and allele frequencies. Matrix **M** allocates genotypes to animals in the form of 0, 1, and 2 for homozygous first allele, heterozygous, and homozygous second allele, respectively. Matrix **P** contains the frequencies of the second allele at each locus; columns of **P** are equal to $2\mathbf{p_i}$. Matrix **Z** is equal to **M** – **P**. Finally, matrix **G** is equal to $\frac{\mathbf{ZZ'}}{2\sum \mathbf{p_i}(1-\mathbf{p_i})}$. Division by $2\sum \mathbf{p_i}(1-\mathbf{p_i})$ makes **G** analogous to **A** (VanRaden, 2007). This means that diagonal element of **G** should be close to 1 and off diagonals should be close to 0. This derivation is also known as a kinship matrix in human genetics (Amin et al., 2007; Astel and Balding, 2009).

There are other methods available to construct **G**. The second method weights markers by reciprocals of their expected variance: $\mathbf{G=ZDZ'}$, where **D** is a diagonal matrix

with $\mathbf{D}_{ii}=\frac{1}{\mathbf{m}[2\mathbf{p}_i(1-\mathbf{p}_i)]}$ (VanRaden, 2008). This method was first proposed for human genetics studies (Amin et al., 2007; Leutenegger et al., 2003). The third method regresses **MM**' on **A** to obtain **G**: $\mathbf{MM}'=g_0\mathbf{11}'+g_1\mathbf{A}+\mathbf{E}$, in which g_0 is the intercept, g_1 is the slope, and **E** includes differences of true from expected fractions of DNA as well as measurement error (VanRaden, 2008).

Substitution of **G** into the mixed model equations is straightforward. The data can be modeled as $\mathbf{y}=\mathbf{Xb}+\mathbf{Zu}+\mathbf{e}$, where **Xb** is the mean, **Z** allocates genotypes to records, **u** contains additive effects for each marker, and **e** is a random error vector with variance equal to $\mathbf{R}\sigma_e^2$ (VanRaden, 2008). Summing **Zu** over all marker loci gives the vector of breeding values, **a**, with $V(a) = G\sigma_a^2$ (VanRaden, 2008). In mixed model notation, this becomes:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{G}^{-1}\sigma_e^2/\sigma_a^2 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}.$$

In addition to more exact relationships, use of the genomic relationship matrix is preferable when then number of genotyped animals is less than the number of markers, which is usually the case. It is computationally more efficient than estimating SNP effects and may also provide more accurate GEBVs (Hayes, 2010).

The accuracy of GEBVs obtained using **G** has been studied extensively. Hayes and Goddard (2008) presented a study in which they compared the prediction of breeding values using traditional and **G** derived prediction equations through a simulation and Angus datasets. They found that estimates of heritability were closer to true heritabilities when using **G** rather than **A**. Additionally, they found that reducing the number of

markers used to create **G** reduced heritability estimates. This suggests that denser SNP panels detect more relationships than sparse panels. Another study by Hayes et al. (2009c) describes the increase in accuracy through the use of **G** over **A**, even for animals that do not have phenotypes. The authors note, however, that increased numbers of phenotypes, greater heritability, and family records increase accuracy of breeding values, as is the case when using traditional pedigree matrices. Lowly heritable traits, such as fertility traits, require more observations on the genetic and phenotypic level to make substantial progress. This is particularly true if they are inversely related to production traits (Hoglund et al., 2009). Despite its limitations, use of the genomic relationship matrix represents a feasible alternative to multistep SNP estimation or traditional breeding approaches.

The scaling factor used to construct **G** is of primary importance. According to VanRaden (2007; 2008), scaling **G** by the factor $2\sum \mathbf{p}_i(1-\mathbf{p}_i)$ makes it analogous to **A**. Additionally, **G** is constructed to give more weight to rare alleles. Moreover, animals homozygous for many rare alleles will have higher inbreeding coefficients (VanRaden, 2008). VanRaden states that the frequencies used should come from the unselected base population, because these are likely to be in Hardy-Weinberg equilibrium. In a simulation study, VanRaden found that base frequencies provided more reliable genomic predictions than did the simple frequencies estimated from the population (2008). An algorithm has been suggested by Gengler (2007) to obtain estimates of these frequencies. Other options include using a constant value for allele frequencies, such as 0.5, or using frequency estimates from the current population. Gianola et al. (2009) have also suggested an algorithm to modify the denominator used to scale **G**, which assumes that the allele

frequencies in the base population are not independent. The results of these studies suggest that the choice of allele frequency to scale **G** is dependent on the dataset used.

It may be possible to use genomic relationships in multi-breed evaluations. Harris and Johnson (2010) evaluated a mixed population of 8,706 Holstein-Friesian, Jersey, and Friesian x Jersey crossbred bulls, of which 5,212 were genotyped, using a genomic relationship matrix and integrating results with a national evaluation. Since **G** is scaled using allele frequencies, Harris and Johnson noted that covariance between relatives in such a population should consider differences in allele frequencies between breeds. **G** was therefore constructed so that allele frequencies did not have be taken into account using a regression technique:

$$\mathbf{ZZ}' = b_1 \mathbf{11}' + b_2 \mathbf{A} + \mathbf{E},$$

which does not require estimates of allele frequencies (VanRaden, 2008). To take breed into account, the regression equation was generalized into a multiple regression:

$$\mathbf{ZZ}' = \sum_{k \le l} b_{1(kl)} \mathbf{J}_{(kl)} + \sum_{k \le l} b_{2(kl)} \mathbf{K}_{(kl)} + \mathbf{E},$$

where $\mathbf{J}_{(kl)}$ and $\mathbf{K}_{(kl)}$ are breed specific matrices (Harris and Johnson, 2010). **G** was constructed such that diagonals were partitioned into the breed fractions of the animals to account for difference in variance between breeds. When breed effect was not taken into account, diagonal elements of **G** were distorted, with some animals having elements less than 1.00. When **G** was partitioned to take breed into account, diagonal elements for all animals appeared centered on 1.00. Taking breed effect into account and blending results with parent average information increased the reliability of unproven bulls over the reliability using only genomic relationships (Harris and Johnson, 2010).

Another approach in genomic evaluations when some animals are not genotyped is the use of the combined pedigree-genomic relationship matrix, **H**. Despite increased accuracy using **G**, the available data is limited due to small genotyped populations. Using only genotyped animals excludes a large number of observations and relationships among ungenotyped animals. While using the multistep procedure to predict SNP effects is effective and allows the use of all the data, accuracy can be lost in the procedure. Combining genomic and pedigree relationships in one matrix allows the use of all available data in a single step; moreover, it can easily be implemented in current evaluation systems. Misztal et al. (2009) proposed a single-step procedure that augments the traditional **A** matrix with relationships from genotyped animals in the **G** matrix:

$$\mathbf{H}=\begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{G}\end{bmatrix}=\mathrm{A}+\begin{bmatrix}0 & 0\\ 0 & \mathbf{G}-\mathbf{A}_{22}\end{bmatrix}.$$

This idea was expanded by Legarra et al. (2009) to include a joint relationship matrix based on pedigree and genomic relationships:

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

The joint relationship matrix was developed independently by Christensen and Lund (2010). The inverse of **H** is then easy to obtain:

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

with $\mathbf{A}_{22}^{-1}$ equal to the inverse of the pedigree based relationship matrix for genotyped animals. Aguilar et al. (2010) evaluated final score in Holstein cattle using **H**. In the analysis, several methods were used to scale **G**: constant allele frequency of 0.5, frequencies from the base population, and frequencies from the current population.

Genomic predictions were the most accurate and least biased when the constant 0.5 was used along with an additional factor, λ, that scaled the difference between the genomic and pedigree information by multiplying $\mathbf{G}^{-1}$ and $\mathbf{A}^{-1}$ by a constant:

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

The value of λ is variable and depends on the population being examined; when λ is equal to 0, no genomic information is used and when λ is equal to 1, all genomic information is used.

There have been other studies in the use of **H**. Forni et al. (2010), in a GS study of litter size in a pig population using the single-step procedure, found that variance component estimates were inflated if **G** was not scaled correctly, particularly for small datasets. Different methods were used to scale **G**: the constant 0.5, the average minor allele frequency, the current allele frequencies, the current allele frequencies multiplied by a constant to scale **G** to **A**, and the Gianola correction. Results of Forni et al. (2010) indicate that **G** should be scaled to resemble **A** in order to have similar variance components; breeding values were not as affected by the scaling of **G**. Chen et al. (2010) compared the single step procedure to both traditional evaluation and the multistep BayesA procedure. While the single step procedure outperformed both pedigree and multistep procedures for one line of broiler chickens, it did not for a second line. The study highlighted the importance of phenotypic data in evaluations, and also suggested strong selection may influence the results of genomic evaluation (Chen et al., 2010).

Technical Considerations

The quality of SNP genotypes is crucial to genomic evaluations; errors and uninformative SNPs can decrease the accuracy of evaluation. Uninformative or unreliable

SNPs should be eliminated, reducing computational effort (Wiggans et al., 2009). Certain restrictions have been suggested to select informative and error-free SNP genotypes (Wiggans et al., 2009). Some of the more successful restrictions have been made official, as in the case of USDA dairy evaluations. SNPs that show no variation or have a minor allele frequency (MAF) less than 0.02 will not contribute to the evaluation and should be removed. As more animals are genotyped, it will be possible to include SNPs with lower MAF. SNPs that show large deviation from Hardy-Weinberg equilibrium should also be removed. If a SNP is highly correlated with another SNP, it is likely that the two will be in LD with the same marker and will give the same information; using both SNPs is unnecessary and one can be removed (Wiggans et al., 2009). Careful storage and evaluation of DNA samples ensures that quality SNP genotypes are obtained (Wiggans et al., 2010). Parent-progeny conflicts can be used to determine if a genotyping error has occurred, both in terms of laboratory error or mislabeling.

Genotyping errors represent a common and serious problem in genomic evaluation schemes. A genotyping error occurs when an animal's actual genotype and the genotype determined through molecular analysis do not correspond (Bonin et al., 2004). In GWAS, even laboratory errors rates of 0.5 to 1 % can mask important associations (Abecasis et al., 2001). In GS schemes, genotyping errors can reduce the accuracy of the prediction (Wiggans et al., 2009; 2010). Genotyping errors can occur at any stage in the genotyping process and include errors in sampling, DNA extraction, analysis, and scoring. Human error can also contribute to genotyping errors. For that reason, careful SNP selection and sample management are crucial for successful and accurate genomic analyses.

Another important component of GS is the population from which genotypes are drawn. Differences in allele frequencies between populations will impact the scaling factor of **G** and can be seen as changes in the expected diagonal elements of **G** (Harris and Johnson, 2010). In human genetics, population structure can give rise to false associations in diseased individuals, particularly in GWAS studies, because population structure will cause apparent LD between unlinked loci (Astle and Balding, 2009; Hirschhorn and Daly, 2005). Several methods have been developed to correct for false association in human genetics.

The transmission disequilibrium test (TDT) tests for systematic differences between genotypes of affected children and expected genotypes due to Mendelian inheritance; alleles responsible for disease will be over-transmitted (Astle and Balding, 2009). TDT uses heterozygous parents of affected children to obtain a χ^2 statistic testing whether an allele is transmitted more than it should be if no linkage with the disease exists. A major limitation of the TDT is obtaining enough families with heterozygous parents and affected children and so it is not usually used for population studies (Astle and Balding, 2009).

Genomic control (GC) is another method to correct for population structure. Devlin and Roeder (1999) posited the GC method to detect associations due to population heterogeneity. Marker information is used to adjust for inflation in test statistics due to population. The GC method was extended to quantitative traits by Bacanu et al. (2002). Amin et al. (2007) showed that the method could be used to compensate for errors in genealogy or pedigree through the use of a kinship matrix. To do so, a constant, λ, is used to reduce inflation of test statistics caused by false association due to population structure

(Amin et al., 2007; Astle and Balding, 2009). An important limitation of GC is that differentiating when significant linkage is due to the underlying population structure and when there is actual linkage is not possible; all test statistics are scaled by λ and can thus have reduced power to find actual association (Astle and Balding, 2009).

Structured association divides populations into clusters based on allele frequencies and then combines association evidence within each cluster (Price et al., 2006). These procedures are not generally executed in animal breeding programs since researchers focus on genome-wide simultaneous estimation of effects or the construction of **G**, rather than identifying individual QTL of interest. Structured association can be used, however, to find unknown population structure, provided populations have divergent enough allele frequencies (Pritchard et al., 2000). This method was successfully able to determine the population structure of a 20 breed chicken population, and could even distinguish between lines of closely related breeds (Rosenberg et al., 2001).

It is important to be aware of the population being used in animal evaluation. Different populations will have alleles in different phases, which is why SNP estimates for one breed cannot be used in another (Goddard et al., 2006; Hayes et al., 2009a). Moreover, the alleles in one population will likely exist at different frequencies than the same alleles in other populations due to selection pressures (Falconer and Mackay, 1996). For this reason, unknown population structure or admixture can be problematic. Differences in allele frequencies between populations may be a useful tool as genomic technologies advance and available SNP panels become denser.

Summary

Through the use of traditional evaluation techniques combined with cutting edge computational and molecular considerations, animal breeding and genetics is a rapidly growing field. Despite substantial gains made through traditional selection schemes using parent averages and the pedigree relationship matrix, even greater gains can be achieved using recently accessible genomic information. Genomic selection in animal breeding can reduce cost, reduce generation intervals, increase accuracy, and increase the rate of genetic gain. Worldwide selection programs have already been established that make extensive use of genomic information.

As the technology to genotype animals becomes cheaper, a larger population of animals will become available for genomic evaluation. This availability will increase the accuracy of single nucleotide polymorphism marker estimates and provide richer and more complete information about genomic relationships. As these datasets continue to grow, so does the chance of genotyping error or data mismanagement. Unintentional population admixture or unknown population structure will affect the estimation of single nucleotide polymorphism effects and the accurate construction of the genomic relationship matrix. Efficient and reliable methods of detecting errors at the molecular and population level are needed to ensure the uncompromised quality of genomic evaluations.

REFERENCES

Abecasis, G. R., S. S. Cherny, and L. R. Cardon. 2001. The impact of genotyping error on family-based analysis of quantitative traits. Eur. J. Hum. Genet. 9: 130-134.

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

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

Andreescu, C., S. Avendano, S. R. Brown, A. Hassen, S. J. Lamont, and J. C. M. Dekkers. 2007. Linkage Disequilibrium in Related Breeding Lines of Chickens. Genetics. 177: 2161-2169.

Astle, W. and D. J. Balding. 2009. Population structure and cyrptic relatedness in genetic association studies. Stat. Sci. 24: 451-471.

Bacanu, S. A., B. Devlin, and R. Kathryn. 2002. Association studies for Quantitative traits in structured populations. Genet. Epidemiol. 22: 78-93.

Bahram, S. and H. Inoko. 2007. Microsatellite markers for genome-wide association studies. Nat. Rev. Genet. 8.

Beuzen, N. D., M. J. Stear, and K. C. Chang. 2000. Molecular markers and their use in animal breeding. Vet. Journal. 160: 42-52.

Boichard, D., M. Fritz, M. N. Rossignol, F. Guillaume, J. J. Colleau, and T. Druet. 2006. Implementation of marker-assisted selection: Practical lessons from dairy cattle. Proc. 8th World Congr. Genet. Appl. Livest. Prod., Commun. 22-11. Instituto Prociencia, Belo Horizonte, Brazil.

Bonin, A., E. Bellemain, P. Bronken Eidesen, F. Pompanon, C. Brochmann, and P. Taberlet. 2004. Invited review: How to track and assess genotyping errors in population genetics studies. Mol. Ecol. 13: 3261-3273.

Bourdon, R. M. 2000. Understanding animal breeding. 2nd ed. Prentice Hall, Upper Saddle River, NJ.

Chen, C. Y., I. Misztal, I. Aguilar, S. Tsuruta, T. H. E. Meuwissen, S. E. Aggrey, and W. M. Muir. 2010. Genome wide marker assisted selection in chicken: Making the most of all data, pedigree, phenotypic, and genomic in a simple one step procedure. Proc. 10th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany.

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

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics. 179: 1503-1512.

de Roos, A. P. W., C. Schrooten, E. Mullaart, M. P. L. Calus, and R. F. Veerkamp. 2007. Breeding value estimation for fat percentage using dense markers on *Bos taurus* autosome 14. J. Dairy Sci. 90: 4821-4829.

Dekkers, J. C. M. 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim Sci. 82: E313-328.

Devlin, B. and K. Roeder. 1999. Genomic Control for Association Studies. Biometrics. 55: 997-1004.

Drogemuller, C., H. Hamann, and O. Distl. 2001. Candidate gene markers for litter size in different German pig lines. J. Anim Sci. 79: 2565-2570.

Falconer, D. S. and T. Mackay. 1996. Introduction to Quantitative Genetics. 4th ed. Longman, New York.

Forni, S., I. Aguilar, I. Misztal, and N. Deeb. 2010. Genomic relationships and biases in the evaluation of sow litter size. Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany.

Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. animal. 1: 21-28.

Gianola, D. 2000. Statistics in Animal Breeding. J. Am. Stat. Assoc. 95: 296-299.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando. 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics. 183: 347-363.

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

Goddard, M. E., B. Hayes, H. McPartlan, and A. J. Chamberlain. 2006. Can the same genetic markers be used in multiple breeds? Proc. 8th World Congr. Genet. Appl. Livest. Prod., communication no. 22-16. Belo Horizonte, MG, Brasil.

Goddard, M. E. and B. J. Hayes. 2007. Genomic selection. J. Anim. Breed. Genet. 124: 323-330.

Gonzalez-Recio, O., D. Gianola, G. Rosa, K. Weigel, and A. Kranis. 2009. Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. Gen. Sel. Evol. 41: 3.

Grisart, B., W. Coppieters, and F. Farnir. 2002. Position candidate cloning of a QTL in Dairy Cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12: 222-231.

Guillaume, F., S. Fritz, D. Boichard, and T. Druet. 2008. Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle. Genet. Sel. Evol. 40: 91-102.

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

Harris, B. L. and D. L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy. Sci. 93: 1243-1252.

Hayes, B. 2010. Lecture notes: Genomic evaluation short course. Col. State. Univ., Fort Collins, CO. Jan. 11-15, 2010.

Hayes, B., P. Bowman, A. Chamberlain, K. Verbyla, and M. Goddard. 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Gen. Sel. Evol. 41: 51.

Hayes, B. and M. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution. 33: 209 - 229.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009b. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92: 433-443.

Hayes, B. J. and M. E. Goddard. 2008. Technical note: Prediction of breeding values using marker-derived relationship matrices. J. Anim Sci. 86: 2089-2092.

Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009c. Increased accuracy of artificial selection by using the realized relationship matrix. Gen. Res. 91: 47-60.

Heifetz, E. M., J. E. Fulton, N. O'Sullivan, H. Zhao, J. C. M. Dekkers, and M. Soller. 2005. Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. Genetics. 171: 1173-1181.

Heuven, H., R. van Wijk, B. Dibbits, T. van Kampen, E. Knol, and H. Bovenhuis. 2009. Mapping carcass and meat quality QTL on Sus Scrofa chromosome 2 in commercial finishing pigs. Genet. Sel. Evol. 41: 4.

Hirschhorn, J. N. and M. J. Daly. 2005. Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6: 95-108.

Hoglund, J. K., B. Guldbrandtsen, G. Su, B. Thomsen, and M. S. Lund. 2009. Genome scan detects quantitative trait loci affecting female fertility traits in Danish and Swedish Holstein cattle. J. Dairy Sci. 92: 2136-2143.

Kurreeman, F. A. S., L. Padyukov, R. B. Marques, S. J. Schrodi, M. Seddighzadeh, G. Stoeken-Rijsbergen, A. H. M. Van der Helm-van Mil, C. F. Allaart, W. Verduyn, J. Houwing-Duistermaat, L. Alfredsson, A. B. Begovich, L. Klareskog, T. W. J. Huizinga, and R. E. M. Toes. 2007. A candidate gene approach identifies the TRAF1/C5 region as a risk factor for rheumatoid arthritis. PLoS Medicine. 4: e278-1524.

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

Legarra, A., C. Robert-Granie, E. Manfredi, and J.-M. Elsen. 2008. Performance of Genomic Selection in Mice. Genetics. 180: 611-618.

Li, C., M. Li, J.-R. Long, Q. Cai, and W. Zhen. 2008. Evaluating cost efficiency of SNP chips in genome-wide association studies. Genet. Epidemiol. 32: 387-395.

Liu, Y., X. Qin, X.-Z. Song, H. Jiang, Y. Shen, K. J. Durbin, S. Lien, M. Kent, M. Sodeland, Y. Ren, L. Zhang, E. Sodergren, P. Havlak, K. Worley, G. Weinstock, and R. Gibbs. 2009. Bos taurus genome assembly. BMC Genomics. 10: 180.

Leutenegger, A.-L., B. Prum, E. Génin, 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.

Maher, B. 2008. Personal genomes: The case of the missing heritability. Nature. 456: 18-21.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics. 157: 1819-1829.

Misztal, I. 2006. Challenes of application of marker assisted selection - a review. Anim. Sci. Pap. Rep. 24: 5-10.

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

Mrode, R. A. 2005. Linear Models for the Prediction of Aniaml Breeding Values. 2nd ed. CABI, Cambridge, MA.

Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124: 342-355.

Nejati-Javaremi, A., C. Smith, and J. P. Gibson. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. J. Anim Sci. 75: 1738-1745.

Pearson, T. A. and T. A. Manolio. 2008. How to interpret a genome-wide association study. JAMA. 299: 1335-1344.

Pharoah, P. D. P., J. Tyrer, A. M. Dunning, D. F. Easton, B. A. J. Ponder, and S. Investigators. 2007. Association between common variation in 120 candidate genes and breast cancer risk. PLoS Genet. 3: e42.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38: 904-909.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 155: 945-959.

Rosenberg, N. A., T. Burke, K. Elo, M. W. Feldman, P. J. Freidlin, M. A. M. Groenen, J. Hillel, A. Maki-Tanila, M. Tixier-Boichard, A. Vignal, K. Wimmers, and S. Weigend. 2001. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics. 159: 699-713.

Sargolzaei, M., F. S. Schenkel, G. B. Jansen, and L. R. Schaeffer. 2008. Extent of linkage disequilibrium in Holstein cattle in North America. J. Dairy Sci. 91: 2106-2117.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123: 218-223.

Schmutz, J., J. Wheeler, J. Grimwood, M. Dickson, J. Yang, C. Caoile, E. Bajorek, S. Black, Y. M. Chan, M. Denys, J. Escobar, D. Flowers, D. Fotopulos, C. Garcia, M. Gomez, E. Gonzales, L. Haydu, F. Lopez, L. Ramirez, J. Retterer, A.

Rodriguez, S. Rogers, A. Salazar, M. Tsai, and R. M. Myers. 2004. Quality assessment of the human genome sequence. Nature. 429: 365-368.

Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. 2004. Nature. 432: 695-716.

VanRaden, P. M. 2007. Genomic measures of relationship and inbreeding. Interbull. 37: 33-36.

VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91: 4414-4423.

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

Vignal, A., D. Milan, M. SanCristobal, and A. Eggen. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genet. Sel. Evol. 34: 275-305.

Wang, W. Y. S., B. J. Barratt, D. G. Clayton, and J. A. Todd. 2005. Genome-wide association studies: Theoretical and practical concerns. Nature. 6: 109-118.

Weigel, K. A., C. P. Van Tassell, J. R. O'Connell, P. M. VanRaden, and G. R. Wiggans. 2010. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J. Dairy Sci. 93: 2229-2238.

Wiggans, G. R., T. S. Sonstegard, P. M. VanRaden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor, F. S. Schenkel, and C. P. Van Tassell. 2009. Selection of single-

nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92: 3431-3436.

Wiggans, G. R., P. M. VanRaden, L. R. Bacheller, M. E. Tooker, J. L. Hutchison, and T. A. Cooper. 2010. Selection and management of DNA markers for use in genomic evaluation. J. Dairy Sci. 93: 2287-2292.

Zhu, M. and S. Zhao. 2007. Candidate gene identification approach: Progress and challenges. Int. J. Biol. Sci. 3: 420-427.

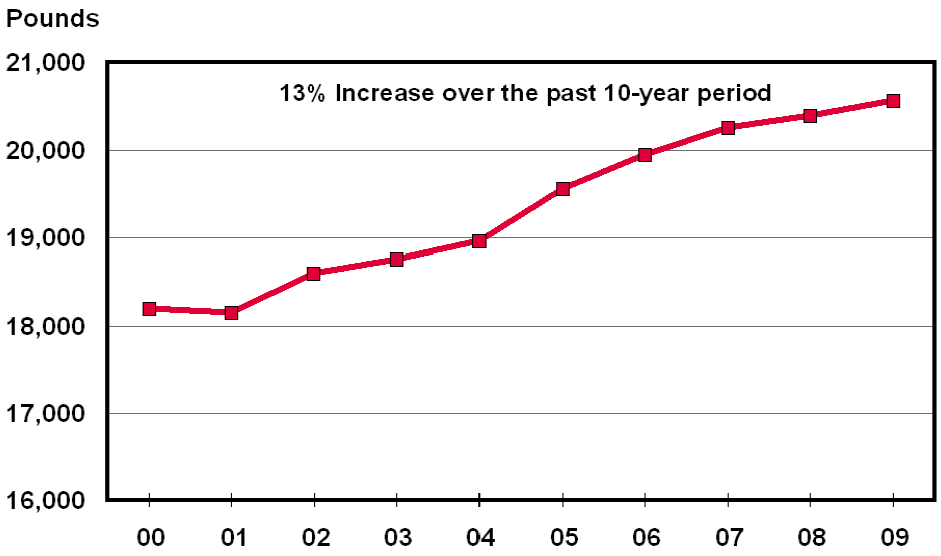

Figure 2.1: Increase in milk production by cow per year from 2000-2009 (USDA, April 2009)

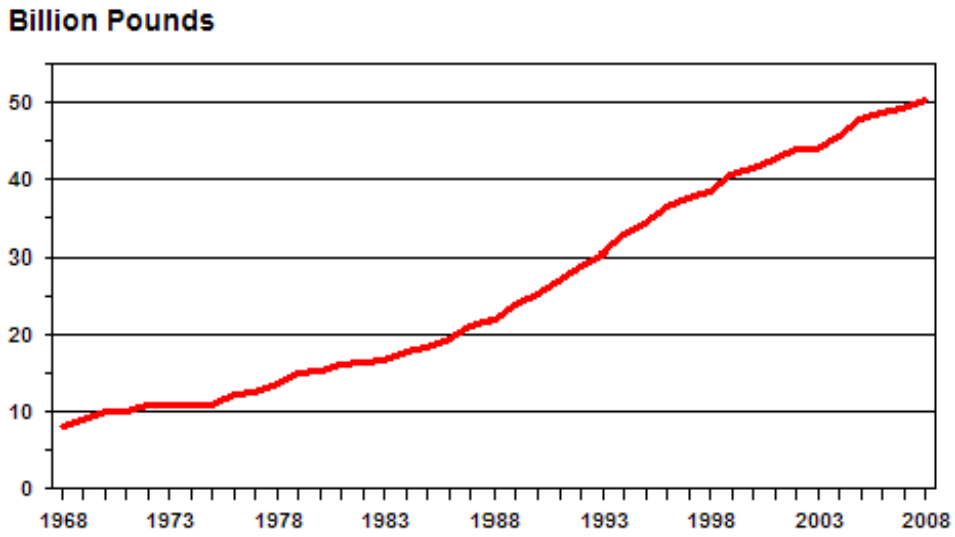

Figure 2.2: Increase in Broiler pounds of meat produced in United States, 1968-2008 (USDA, May 2009)

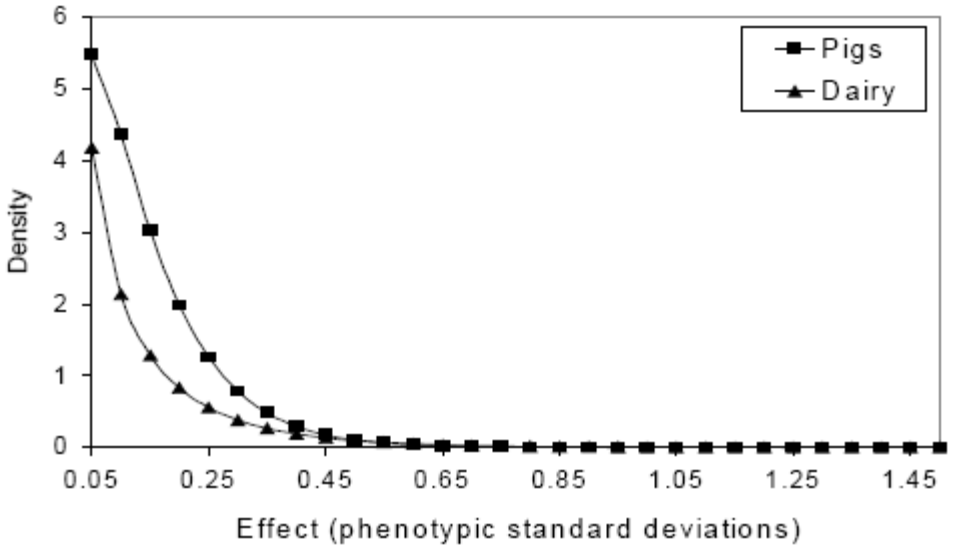

Figure 2.3: Distribution of QTL effects. The distribution was obtained through a meta-analysis of pig and dairy cattle QTL mapping experiments (Hayes and Goddard, 2001). The gamma distributions are moderately leptokurtic and suggest that there are few QTL with large effect on the phenotype and many QTL with small effects.

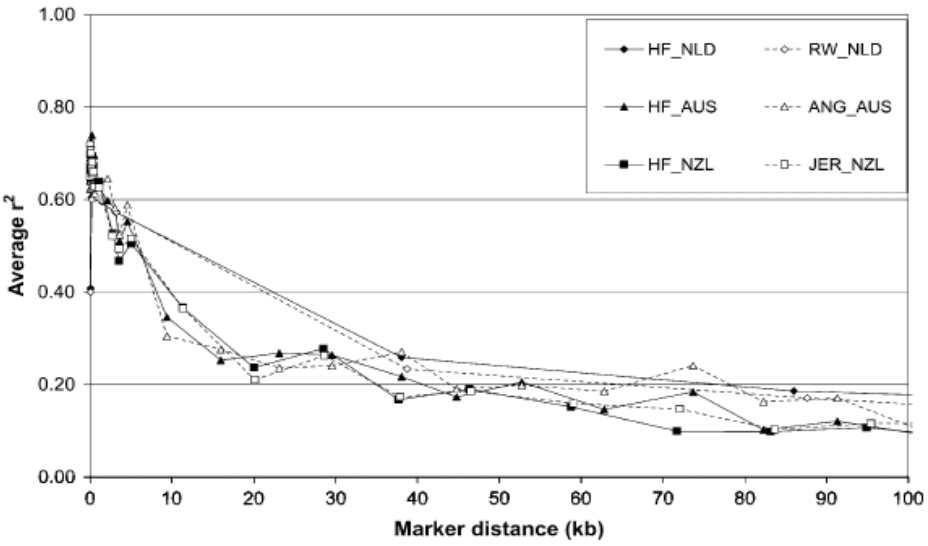

Figure 2.4: Average linkage disequilibrium as a function of genomic distance. Linkage disequilibrium was measured for subsets of Holstein-Friesian, Jersey, and Angus cattle (de Roos et al., 2008). The plots show that linkage disequilibrium decreases as genomic distance increases.

Table 2.1: Comparison of estimated vs. true breeding values using different models for genomic selection (Meuwissen et al., 2001). The correlations reflect the accuracy of selection while the regression methods indicate the amount of bias of the methods (should be close to 1).

	$r_{\text{TBV;EBV}}$ + SE	$b_{\text{TBV.EBV}}$ + SE
LS	0.318 ± 0.018	0.285 ± 0.024
BLUP	0.732 ± 0.030	0.896 ± 0.045
BayesA	0.798	0.827
BayesB	$0.848 + 0.012$	$0.946 + 0.018$

CHAPTER 3

EVALUATION OF THE UTILITY OF THE GENOMIC RELATIONSHIP MATRIX AS A DIAGONOSTIC TOOL TO DETECT MISIDENTIFIED GENOTYPED ANIMALS IN A BROILER CHICKEN POPULATION[1]

[1] Simeone, R., I. Misztal, I. Aguilar, and A. Legarra. To be submitted to *Journal of Animal Breeding and Genetics*

Abstract

The objectives of this study were to explore simulated distributions of diagonal elements of the genomic relationship matrix (**G**), to evaluate the utility of **G** as a diagnostic tool in detecting different populations within a genomic dataset, and to evaluate the effect of misidentified genotyped animals on the accuracy of genomic evaluation. Populations of 10,000 animals with three (six) generations were simulated with 60,000 SNP varying in frequency at each locus between 0.02 and 0.98, 0.1 and 0.9, 0.25 and 0.75, and 0.45 and 0.55. **G** was scaled using current allele frequencies. Diagonal elements of **G** averaged 1.00 ± 0.01 (0.03) and ranged from 0.84 through 1.36. Mixed populations were simulated with three or six generations. A 7,000 animal population with frequencies of second alleles ranging from 0.02 through 0.98 was combined with a 1,750 or 7,000 animal population with frequencies of second alleles ranging from 0.0 through 1.0. Two peaks were seen in distributions, representing each population. With three generations and 1,750 animals added, the first peak had a mean of 0.94 ± 0.004 and the second had a mean of 1.84 ± 0.02; with 7,000 animals, the first peak had a mean of 1.17 ± 0.01 and the second had a mean of 1.19 ± 0.004. Results for six generations were similar. Field data for body weight at six weeks was provided by Cobb-Vantress for broiler chickens. Genotype data were available for 3,285 animals genotyped for 57,636 SNP. Analysis used a combined genomic and pedigree relationship matrix; **G** was scaled using current allele frequencies. The distribution of diagonal elements was multimodal; 3,195 animals had elements ranging from 0.54 to 1.19, 88 animals had elements ranging from 1.73 to 2.09, and 1 animal had an element of 3.23. Animals with diagonal elements greater than 1.5 were identified as coming from another chicken line or as having low call

rates. Slight improvements in accuracy when misidentified animals were removed were observed. Analysis of diagonal elements of **G** may separate multiple populations. **G** may be a useful diagnostic tool to help identify misidentified animals or secondary populations.

Key words: allele frequency, broiler chickens, diagonal elements, genomic relationship matrix

Introduction

Genomic selection has been used successfully in dairy cattle and other species (Chen et al., 2010; Hayes et al., 2009a; VanRaden et al., 2009). Genomic proofs for young animals may be almost as accurate as progeny tested animals (Meuwissen et al., 2001; Schaeffer, 2006). Initially, genomic predictions were derived by estimating the effects of SNP markers; however, predictions by genomic BLUP with a genomic relationship matrix (**G**) may have similar or increased accuracy (Hayes et al., 2009b).

The genomic relationship matrix may be viewed as a matrix of realized relationships. When constructed from large SNP arrays, **G** is likely to reflect real relationships better than a pedigree-based relationship matrix (**A**) because it contains information on unrecorded pedigrees and on the Mendelian sampling (Hayes et al., 2009b). **G** should be scaled so that averages of diagonal and off-diagonal elements are similar to **A** (VanRaden, 2008). Scaling can be done using current allele frequencies and estimates of variance are similar to those obtained through traditional evaluation (Forni et al., 2010).

Allele frequencies have been used to identify population structure in admixed populations (Pritchard et al., 2000; Rosenberg et al., 2001). Harris and Johnson (2010)

have indicated that multi-breed populations will have animals with distorted diagonal elements if breed is not taken into account. Another method of identification may therefore be to use the diagonal elements of **G** to separate populations. It is unknown how small secondary populations will be expressed in the diagonal elements of **G** or how they might impact predictions in genomic analysis. Moreover, the theoretical distribution of diagonal elements of **G** for single and multiple populations is still unknown. The objectives of this study were to explore by simulation the distributions of the diagonal elements of **G** for single and multiple populations, to evaluate the utility of **G** as a diagnostic tool in the detection of different populations within a genomic dataset, and to evaluate the effect of misidentified genotypes on the accuracy of genomic evaluation.

Materials and Methods

Simple Theoretical Distribution of Diagonal Elements of G

Assume that all genotyped loci are present in the same frequency, p = 0.5. If genotypes are coded as 0, 1, and 2, $\mathbf{M} = \begin{cases} 0 \\ 1 \\ 2 \end{cases}$ and $\mathbf{P} = \begin{cases} 1 \\ 1. \\ 1 \end{cases}$ Matrix **Z** is then equal to **M** – **P**; for each possible genotype, $\mathbf{Z} = \begin{cases} -1 \\ 0 \\ 1 \end{cases}$ and under Hardy-Weinberg equilibrium, genotype frequencies are $\begin{matrix} aa \\ aA \\ AA \end{matrix} = \begin{cases} 1/4 \\ 1/2 \\ 1/4 \end{cases}$. Then, $E(z^2) = 1/2$ and $Var(z^2) = 1/4$. It follows that $E\left(\sum_{i=1}^{n} z_i^2\right) = n/2$ and $Var\left(\sum_{i=1}^{n} z_i^2\right) = n/4$, where i = 1 to the number of SNP. Therefore, $\sum z_i^2 \approx N(n/2, n/4)$; scaling z_i^2 to be analogous to **A** gives $\sum z_i^2/q \approx N(1, 1/n)$. This suggests that the diagonal elements of **G** should be distributed on 1.0 with a sharp peak

and variance decreasing with increased number of SNP. Theoretical distributions of diagonal elements of **G** for 60,000 SNP assuming equal allele frequencies at each loci are presented in Figure 3.1. With differing allele frequencies at each locus, the distribution will likely widen but remain normally distributed with a mean of 1.0. Multiple peaks would not be expected

Simulated Data

Populations were simulated to find the distribution of the diagonal elements of **G**. Each population had either three or six generations, of which 10 % were male and had a total of 10,000 animals. The three-generation population consisted of one unrelated base population of 3,000 animals and two more generations of 3,000 and 4,000 animals. The six-generation population consisted of one unrelated base population of 1,000 animals, and five more generations of 1,000, 1,500, 1,500, 2,000, and 3,000 animals. Mating was random; dams could have multiple progeny, and sires and dams could be used as parents in multiple generations.

The genome was simulated with 60,000 SNP. The frequency of the second allele at each locus varied between 0.02 and 0.98, 0.1 and 0.9, 0.25 and 0.75, and 0.45 and 0.55, for a total of eight populations. Progeny genotypes were created by selecting one of each of the parental chromosomes. Recombination rate between loci varied between 0.02 and 0.1.

Four mixed populations were also simulated with either three or six generations. First, a 7,000 animal pedigree was simulated with the frequency of the second allele ranging from 0.02 through 0.98 (G3, G6). Second, a 1,750 or 7,000 animal pedigree was simulated with the frequency of the second allele ranging from 0.0 through 1.0. Third, the

secondary populations were added to the original 7,000 animal populations and resulted in two datasets with either 8,750 animals (G3_1750, G6_1750) or 14,000 animals (G3_7000, G6_7000). For each population, the frequency of the second allele at each locus of the combined population was calculated.

Field Data

Body weight (100g) at six weeks was provided by Cobb-Vantress for broiler chickens in three generations. Details of the dataset and methods used in the analysis are provided in Chen et al. (2010). Data consisted of 3,285 animals genotyped for 57,636 SNP; of these animals, 3,284 had both phenotypes and genotypes. Monomorphic loci or loci with frequency less than 0.02 were removed from that data, leaving 48,006 SNP for the analysis.

For the initial analysis, data were divided into a training set, consisting of all phenotyped and genotyped animals from the first two generations, n=2,485, and a validation set, consisting of phenotyped and genotyped animals from the third generation, n=799. This was known as ALL data. Subsequent analysis involved removing animals with high diagonal elements from the data and regrouping into training (n=2,397) and validation (n=798). This was known as CLEANED data.

Models and Analysis

Body weight was analyzed using an animal model in a single-step procedure as in Chen et al. (2010). The genomic relationship matrix was constructed as in VanRaden et al. (2008) as $\mathbf{G} = \mathbf{ZZ'} / 2\sum p_i(1-p_i)$. Frequencies of the second allele at each locus in the population were calculated to scale **G** for field and simulated data. **G** was computed for each dataset and the distributions of the diagonal elements of **G** were plotted. Means and

standard deviations of the diagonal elements were obtained. A theoretical 99% range of diagonal elements for simulated data was computed by multiplying the SD of the diagonals by 2.567, from the normal distribution. To visualize differences in the genotypes that might cause differences in the diagonal elements of **G**, the distributions of the second allele frequencies were plotted.

Predictions of EBV were made for the validation data using the training data for ALL and CLEANED. Accuracy was defined as the correlation between the predicted breeding value and the true breeding value: $r(\hat{u},u) = {r(\hat{u},u+e)}/{h}$, where h is the square root of the heritability (Chen et al., 2010; Legarra et al., 2008).

Results and Discussion

Diagonal elements of simulated data were examined to find theoretical distributions of animals in three or six generations with genomes simulated using different allele frequencies. The distribution of diagonal elements for six generations of animals, with allele frequencies ranging from 0.02 through 0.98 is shown in Figure 3.2. The mean of this distribution is 1.00 (0.03), corresponding to the mean of the diagonal elements in the traditional relationship matrix, **A**. Diagonal values range from a minimum of 0.88 to a maximum of 1.36; 99% of the diagonal elements should fall within 0.91 and 1.09. Statistics of the diagonal distributions of multiple populations were examined to ensure that the distribution was consistent regardless of population number, generation size, or allele frequency (Table 3.1). Very little variation in the distribution of the diagonal elements of **G** is seen within populations simulated over three or six generations. Animals simulated over three generations showed little spread around 1.0, suggesting minimal inbreeding or change in allele frequencies, which would be expressed

as very low or very high diagonal elements; animals simulated over six generations showed wider spread around 1.0, suggesting slightly more inbreeding and changes in allele frequencies.

While the majority of diagonal elements were very close to 1.0, each population had animals with diagonal elements outside of the 99% range (Table 3.1, Figure 3.2). With three generations, eleven animals had diagonal elements ranging from 1.12 through 1.25 for animals simulated with frequencies between 0.02 – 0.98, 0.10 – 0.90, and 0.45 – 0.55; nine animals had diagonal elements ranging from 1.12 to 1.26 for animals simulated with frequencies between 0.25 – 0.75. Greater numbers of animals had diagonal elements outside of the 99% range with six generations. For animals with frequencies between 0.02 – 0.98, one animal had a diagonal element of 0.88 and 234 animals had elements ranging from 1.10 to 1.36; animals with frequencies between 0.10 – 0.90 had five animals with elements ranging from 0.87 to 0.90 and 221 animals had elements ranging from 1.10 to 1.28; animals with frequencies between 0.25 – 0.75 had two animals with elements of 0.86 and 0.88 and 201 animals had elements ranging from 1.12 through 1.32; animals with frequencies between 0.45 – 0.55 had 17 animals with elements ranging from 0.84 to 0.90 and 207 animals had elements ranging from 1.10 to 1.30. Individuals with low diagonals occurred in the first generation due to animals whose genotypes made them appear to be less related to the rest of the population. Individuals with higher diagonals occurred due to the mating of relatives, including half-sib, sire-progeny, and grandsire-progeny in later generations. These events arose due to the smaller number of males than females in each population, particularly as generation size grew past that of the previous generation.

Further simulation investigated how a secondary population with different allele frequencies from the first population changed the distribution of diagonal elements. Diagonal elements from multiple populations within the same dataset should show different distributions when **G** is scaled using current allele frequencies from the combined populations. Figure 3.3 shows the distributions of the diagonal elements of **G** for each of the four combined-population datasets. The distribution of the diagonal elements is bi-modal for each dataset. The elements of G3_1750 ranged from 0.93 to 0.96, with an average of 0.94 (0.004) for the original 7,000 member population and from 1.82 to 1.96, with an average of 1.84 (0.02) for the subset 1,750 member population; the elements of G3_7000 ranged from 1.18 to 1.21, with an average of 1.19 (0.004) for the original 7,000 member population and from 1.15 to 1.25 for the subset 7,000 member population, with an average of 1.17 (0.01). The elements of G6_1750 ranged from 0.89 to 1.27, with an average of 0.97 (0.05) for the original 7,000 member population and from 1.57 to 1.95, with an average of 1.71 (0.04) for the subset 1,750 member population; the elements of G6_7000 ranged from 1.18 to 1.49, with an average of 1.24 (0.04) for the original 7,000 member population and from 1.03 to 1.36, with an average of 1.10 (0.03) for the subset 7,000 member population.

The diagonal distributions show no overlap between the two populations in G3_1750 and G6_1750; however, the diagonal distributions show slight overlap between the two populations in G3_7000 and G6_7000. This is likely because **G** is scaled using allele frequencies estimated from the complete population. With an equal number of animals from both populations contributing to the allele frequency estimation, there is less difference in scaling between the two populations and the differences are more

difficult to detect than for populations with smaller subsets. Additionally, it may be easier to discern populations when allele frequencies between them are very different. Animals of the same breed but of different lines may be difficult to separate because they may share more alleles in the same frequency than those in the simulation.

The genomic relationship matrix of field data was examined. Figure 3.4 shows the distribution of the diagonal elements of **G**; the distribution was multimodal. Animals with diagonal elements greater than 1.5 were considered to have abnormally large diagonal elements. Of the 3,284 animals, 3,195 had diagonal elements ranging from 0.54 to 1.19, 88 had diagonal elements ranging from 1.73 to 2.09, and 1 had a diagonal element of 3.23. In contrast, diagonal elements for **A** had a mean of 1.00 and ranged from 1.00 to 1.09. In this case, **G** and **A** were not similar in distribution or scale, suggesting that genomic relationships were picking discrepancies in the data that were not identified using expected relationships in **A**. After consultation with the genotyping lab, a problem with analysis was discovered. The animal with a diagonal element of 3.23 was identified as having a low call rate. Three of the remaining 88 animals' genotypes were confirmed but were suspected of having a sampling error. The final 85 animals were identified as being mislabeled and coming from a second line of broiler chickens. It appears that analysis of the diagonal elements of G can detect small problematic or mislabeled animals within a larger dataset.

Table 3.2 shows the statistics of the diagonal elements of **G** for ALL and CLEANED. While the overall mean of the diagonal elements of **G** for the CLEANED dataset increased from 1.03 to 1.10 when problematic genotypes were removed, the

maximum of the CLEANED dataset decreased from 3.23 to 1.20 and the variance decreased from 0.025 to 0.002.

The distributions of the frequency of the second allele at each locus for all animals, animals with diagonal elements of **G** less than 1.5, and animals with diagonal elements of **G** greater than 1.5 are shown in Figure 3.5. The distribution of the differences in allele frequencies at each locus for the animals with diagonal elements above and below 1.5 is also shown. For all animals, the distribution of the second allele frequency had a mean of 0.51 (0.26) and ranged from 0.02 to 0.98 (Figure 3.5a). Upon closer examination, differences in second allele frequencies were identified between the 3,195 animals with diagonal elements below 1.5 and the 89 animals with diagonal elements above 1.5. The distribution of the frequency of the second allele for animals with diagonal elements below 1.5 had a mean of 0.51 (0.26) and was similar to that of the distribution of second allele frequencies for all animals, though it did have a small number of loci with frequencies below 0.02 and above 0.98 (Figure 3.5b); the distribution of the frequency of the second allele for animals with diagonal elements above 1.5 had a mean of 0.52 (0.30) had a large number of animals with second allele frequencies below 0.02 and above 0.98 (Figure 3.5c). The large peaks below 0.02 and above 0.98 of the animals with abnormally large diagonal elements indicate that the frequencies of the second alleles are different than those of the animals with diagonal elements below 1.5. Incorrectly including these animals in the evaluation affected the calculated second allele frequency of the population. Markers were included that should not have been, as seen by loci with no variation in the animals with diagonal elements below 1.5; differences in allele frequencies between the populations averaged across the entire population allowed

some uninformative markers to be included in the evaluation. Since **G** is scaled using the allele frequencies calculated from the population, animals whose genotypes differ greatly from those of the overall population will likely have improperly scaled diagonal elements.

Estimates from training data were used to make predictions on the validation data. A traditional evaluation using only phenotypic information was also completed for comparison. Accuracy estimates based on phenotypic and genomic information for ALL and CLEANED are provided in Table 3.3. Removing the animals with high diagonal elements increased accuracy of the CLEANED evaluation by 0.01. The small increase seen was most likely because the genotypes of the misidentified animals were contributing noise that was averaged over many animals and only slightly affected the predictions.

Genomic breeding values (GEBVs) were predicted for animals in the validation dataset on predictions from the training data using genomic relationships. GEBVs from ALL and CLEANED ranged from -0.84 to 0.93 and -0.86 to 0.96, respectively. The GEBVs obtained for animals using the CLEANED dataset were deviated from those obtained using the ALL dataset. The deviation of GEBVs in SD units is provided in Figure 3.6. Removing suspected misidentified animals from ALL caused a slight increase in breeding values (mean = -0.01 (0.01)). The deviation of CLEANED from ALL is skewed to the left: 50 animals had a deviation in EBV greater than two SD from the mean. Six of these animals showed a decrease in EBV while 44 showed an increase in EBV. However, the correlation between GEBVs in ALL and CLEANED was 1.00,

suggesting that while the magnitude of GEBV may have changed for animals after removal of misidentified individuals, the ranking remains identical.

Accurate genomic evaluations depend on error-free genomic data. Even if no errors exist at the genotype level, the presence of a secondary breed or line mislabeled as belonging to an overall dataset can impact the results by altering the allele frequency of the population (Bonin et al., 2004; Hirschhorn and Daly, 2005). SNP effects calculated from one breed or population of animals do not provide accurate GEBVs for animals of another breed or population (Goddard et al., 2006). This suggests that different breeds of animals or populations that developed independently of each other will likely have different allele frequencies; the phase of the SNP markers may be different between populations (Goddard et al., 2006).

The presence of a small secondary population within a larger population did not greatly impact the genetic evaluation. Even with incorrect animals included in the analysis, the genomic evaluation outperformed the traditional evaluation, likely because **G** captured more relationships than would have been captured by **A** alone (Hayes et al., 2009b). It is likely that adding more phenotypes would have improved the evaluation (Hayes et al., 2009a), but the slight improvement in accuracy observed using CLEANED indicates that the misidentified animals were detrimental and should have been removed prior to the evaluation. A greater number of misidentified animals may have decreased accuracy more or biased the results. It may be possible to evaluate simultaneously multiple, equally sized populations using the estimated allele frequencies of both. Further research must be completed to find the impact of combining genotyped populations on genomic evaluation.

The construction of **G** uses allele frequencies to scale the matrix, making it analogous to **A** (VanRaden, 2008). In this case, allele frequencies of the current population were used, rather than using a constant, such as 0.5 as in Aguilar et al. (2010), or using an algorithm to impute the allele frequencies of the base population as in VanRaden (2008). The simulated datasets indicated that diagonal elements of **G** should be distributed around 1.0, which corresponds to little inbreeding in the population. VanRaden has indicated that using incorrect allele frequencies has an impact on the estimate of genomic inbreeding coefficients (VanRaden et al., 2008). The choice of scaling factor of **G** is important to the evaluation, as an incorrect scaling factor will bias the results (Aguilar et al., 2010; Forni et al., 2010).

Scaling **G** using current allele frequencies makes the genomic relationship matrix population specific. In some cases, frequency differences between many marker alleles of two populations of animals will cause **G** to be scaled incorrectly for individuals in one of the populations. This was seen in simulation and field data. Even a small subset of animals impacted the calculated second allele frequencies of the population enough to include uninformative markers; moreover, the same small subset was detectable when the diagonal elements of **G** were plotted as a second distribution greater than 1.0 for field and simulated data. **G** is easy to compute and diagonal elements simple to isolate. This indicates **G** may be a useful diagnostic tool to help recognize misidentified animals or secondary populations. Further tests may verify the utility and sensitivity of the genomic relationship matrix to identify multiple populations.

Conclusions

The diagonal elements of the genomic relationship matrix for animals from a simulated population had a narrow distribution centered on 1.00. As the number of

generations simulated increased, the distribution of diagonal elements widened. This corresponded with the theory that diagonal elements should be centered on 1.0. Diagonal elements corresponding to animals from different populations may be higher than average. A distribution of diagonal elements with multiple peaks may be due to an admixed population, unreliable sample data, or a low call rate. A second population of misidentified animals may be identifiable in a larger population if the allele frequencies between the two populations are very different. Removal of misidentified animals and genotypes will increase the accuracy of prediction, with the level of increase dependent on the number of removed genotypes.

Acknowledgements

The authors thank Cobb-Vantress for access to data for this study. This study was partially funded by AFRI grants 2009-65205-05665 and 2010-65205-20366 from the USDA NIFA Animal Genome Program.

References

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

Bonin, A., E. Bellemain, P. Bronken Eidesen, F. Pompanon, C. Brochmann, and P. Taberlet. 2004. Invited review: How to track and assess genotyping errors in population genetics studies. Mol. Ecol. 13: 3261-3273.

Chen, C. Y., I. Misztal, I. Aguilar, S. Tsuruta, T. H. E. Meuwissen, S. E. Aggrey, and W. M. Muir. 2010. Genome wide marker assisted selection in chicken: Making the most of all data, pedigree, phenotypic, and genomic in a simple one step procedure. Proc. 10th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany.

Forni, S., I. Aguilar, I. Misztal, and N. Deeb. 2010. Genomic relationships and biases in the evaluation of sow litter size. Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany.

Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal. 1: 21-28.

Goddard, M. E., B. Hayes, H. McPartlan, and A. J. Chamberlain. 2006. Can the same genetic markers be used in multiple breeds? Proc. 8th World Congr. Genet. Appl. Livest. Prod., communication no. 22-16. Belo Horizonte, MG, Brasil.

Harris, B. L. and Johnson, D. L. 2010. Genomic predictions for New Zealand dairy bulls and

integration with national genetic evaluation. J. Dairy Sci. 93: 1243-1252.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009a. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92: 433-443.

Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. Gen. Res. 91: 47-60.

Hirschhorn, J. N. and M. J. Daly. 2005. Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6: 95-108.

Legarra, A., C. Robert-Granie, E. Manfredi, and J.-M. Elsen. 2008. Performance of Genomic Selection in Mice. Genetics. 180: 611-618.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics. 157: 1819-1829.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 155: 945-959.

Rosenberg, N. A., T. Burke, K. Elo, M. W. Feldman, P. J. Freidlin, M. A. M. Groenen, J. Hillel, A. Maki-Tanila, M. Tixier-Boichard, A. Vignal, K. Wimmers, and S. Weigend. 2001. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics. 159: 699-713.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123: 218-223.

VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91: 4414-4423.

VanRaden, P. M., M. E. Tooker, and N. Gengler. 2008. Effects of allele frequency estimation on genomic predictions and inbreeding coefficients. J. Dairy Science. 91(E-Suppl. 1): 506. (Abstr.)

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

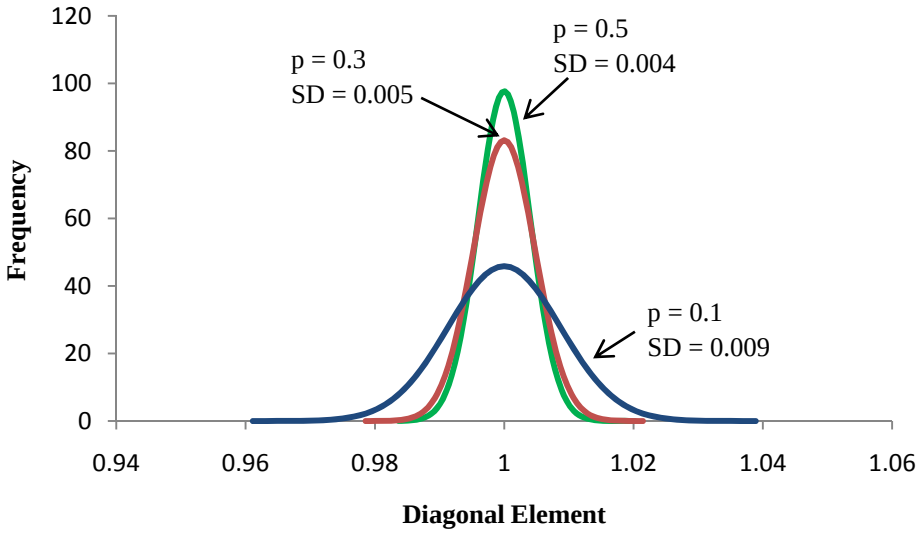

Figure 3.1: Theoretical distribution of diagonal elements of **G** for 60,000 SNP assuming equal allele frequencies (p) at each locus.

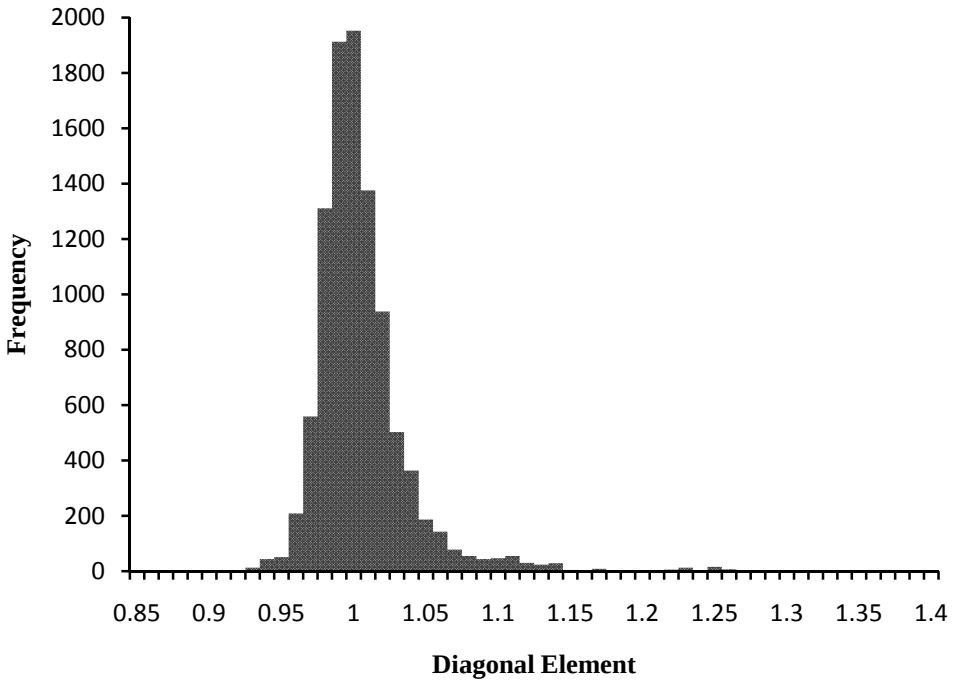

Figure 3.2: Distribution of diagonal elements for six generations of animals (n=10,000). Base population (n=1,000) was assumed to be unrelated; subsequent populations were generated based on parents in previous generations (n=1,000, 1,500, 1,500, 2,000, 3,000). The genome was simulated with 60,000 SNP ranging in frequency between 0.02 and 0.98, with recombination frequency between 0.02 and 0.1.

Table 3.1: Statistics of the diagonal distributions of multiple populations[1]

No. Generations	Allele Frequencies	Mean (SD)	Min	Max	99 % range
3[2]					
	0.02 – 0.98	1.00 (0.01)	0.98	1.25	0.98 – 1.02
	0.25 – 0.75	1.00 (0.01)	0.98	1.26	0.98 – 1.02
	All Frequencies	1.00 (0.01)	0.98	1.26	0.98 – 1.02
6					
	0.02 – 0.98	1.00 (0.03)	0.88	1.36	0.91 – 1.09
	0.10 – 0.90	1.00 (0.04)	0.87	1.28	0.91 – 1.09
	0.25 – 0.75	1.00 (0.04)	0.86	1.32	0.89 – 1.11
	0.45 – 0.55	1.00 (0.03)	0.84	1.30	0.91 – 1.09
	All Frequencies	1.00 (0.03)	0.84	1.36	0.91 – 1.09

[1]Populations were simulated over three or six generations with different allele frequencies. The distributions of the diagonal elements of **G** were then plotted over all generations.

[2] Animals with allele frequencies of 0.02 – 0.98, 0.10 – 0.90, and 0.45 – 0.55 had identical results.

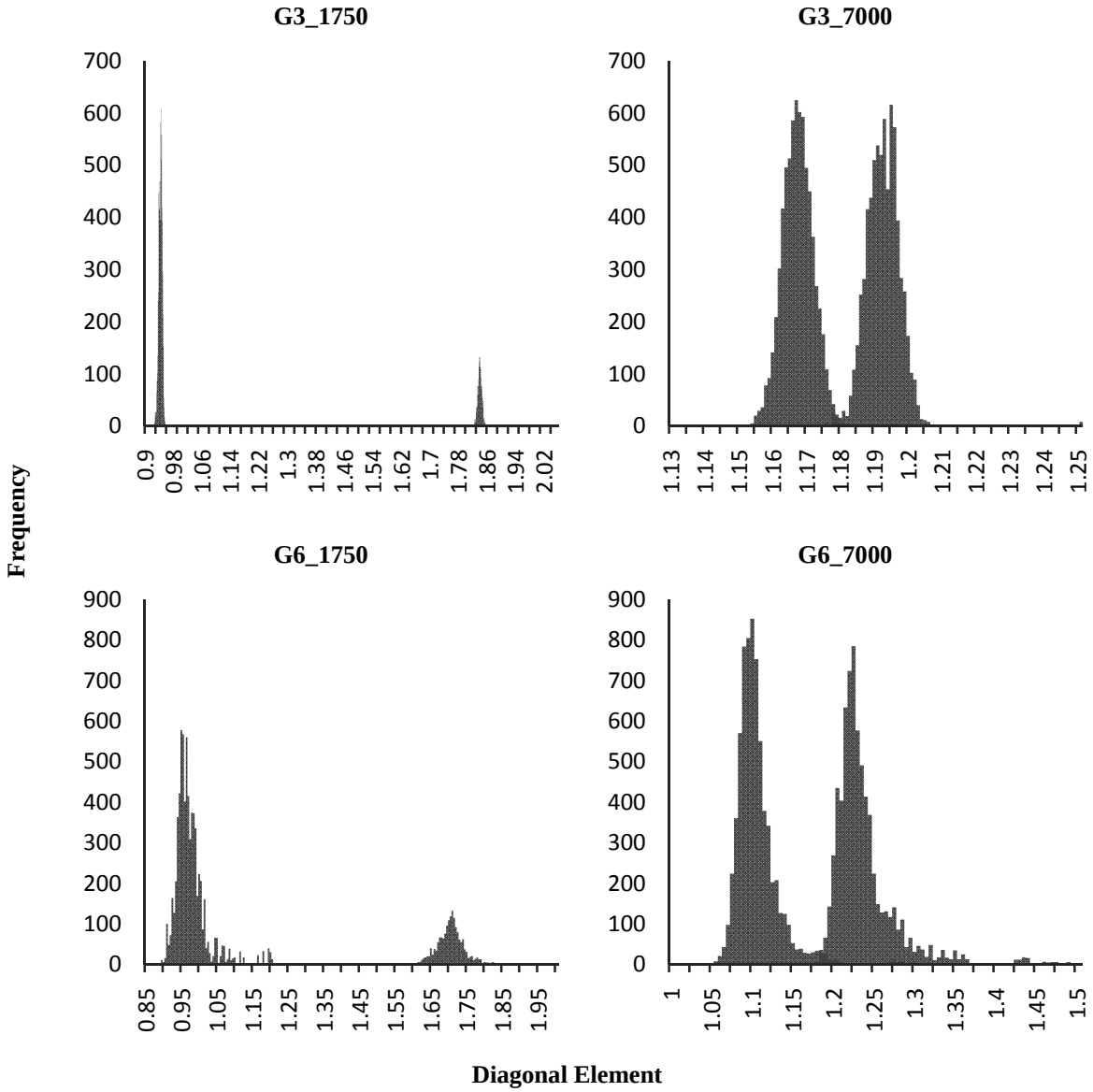

Figure 3.3: Distributions of the diagonal elements of **G** for each of the four combined-population datasets. Distributions are shown on different scales.

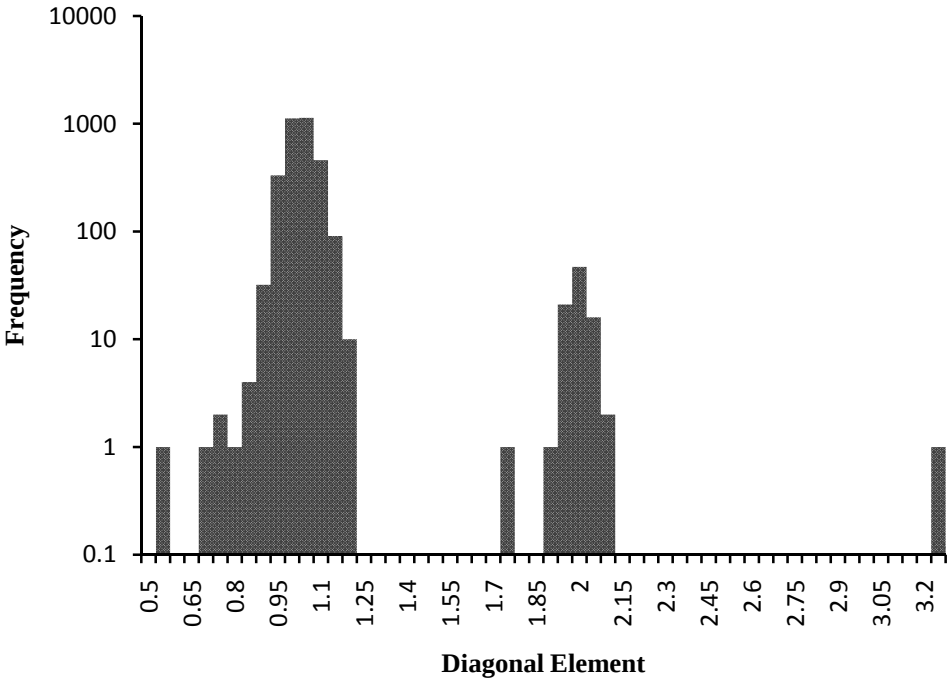

Figure 3.4: Distribution of the diagonal elements of **G** for field data. Results are shown in a logarithmic scale.

Table 3.2: Statistics of the diagonal elements of **G** for the ALL and CLEANED datasets[1]

Dataset	No. of records	Mean	Maximum	Minimum	Variance
ALL	3,284	1.03	3.23	0.56	0.025
CLEANED	3,195	1.10	1.20	0.57	0.002

[1] ALL dataset consisted of 3,284 genotyped animals, 89 of which had diagonal elements greater than 1.5; CLEANED dataset consisted of 3,195 animals that remained after removal of 89 animals with high diagonals

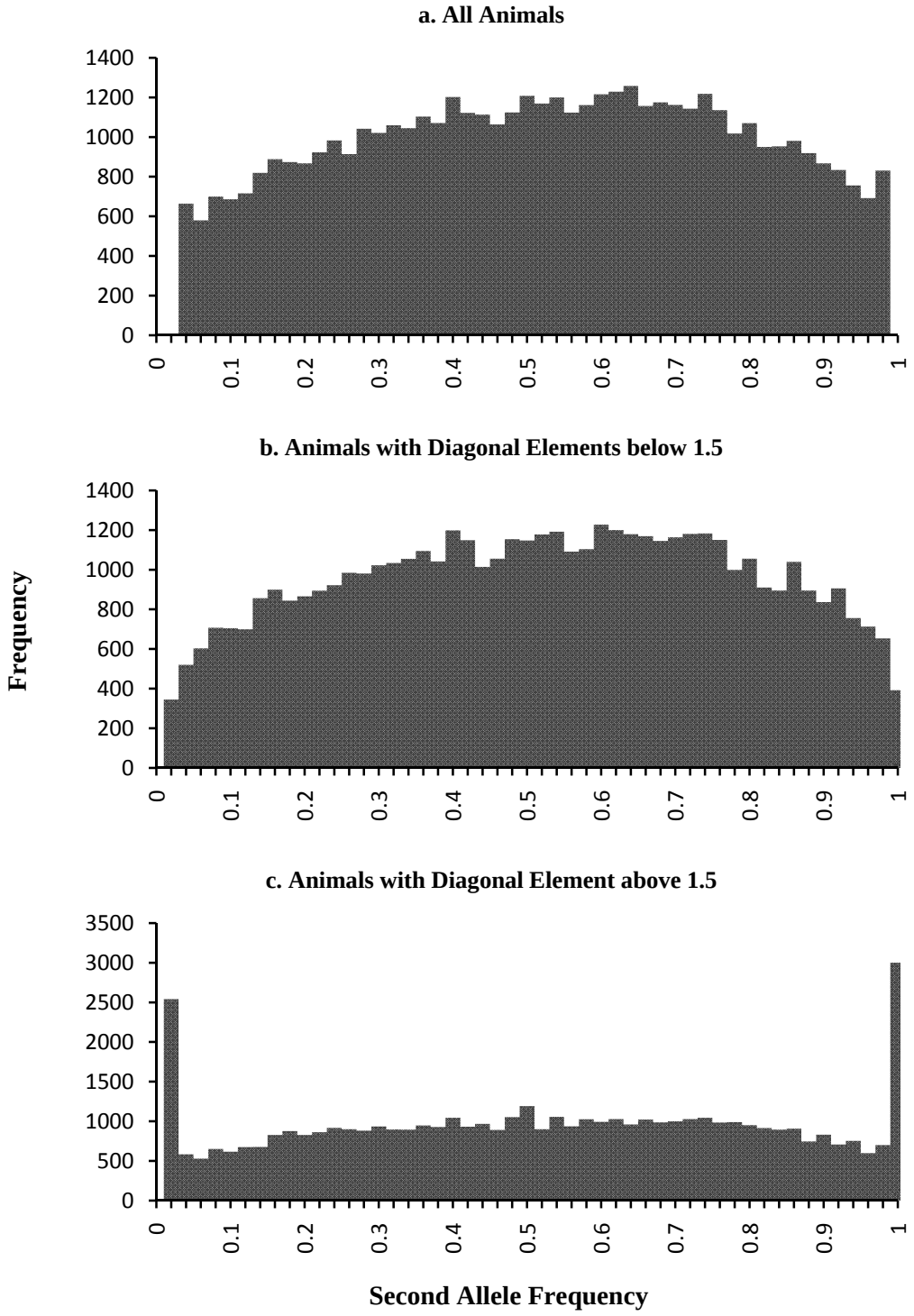

Figure 3.5: Distributions of the frequency of the second allele at each locus for all animals (a), animals with diagonal elements of **G** less than 1.5 (b), and animals with diagonal elements of **G** greater than 1.5 (c).

Table 3.3: Accuracy estimates based on phenotypic and genomic information for ALL and CLEANED[1]

Dataset	Genomic evaluation	Traditional evaluation[2]
ALL	0.64	0.48
CLEANED	0.65	0.48

[1] Accuracy was defined as $r(\hat{u},u+e)/h$, in which h was the square root of heritability

[2] Traditional evaluations were completed using only pedigree and no genomic information

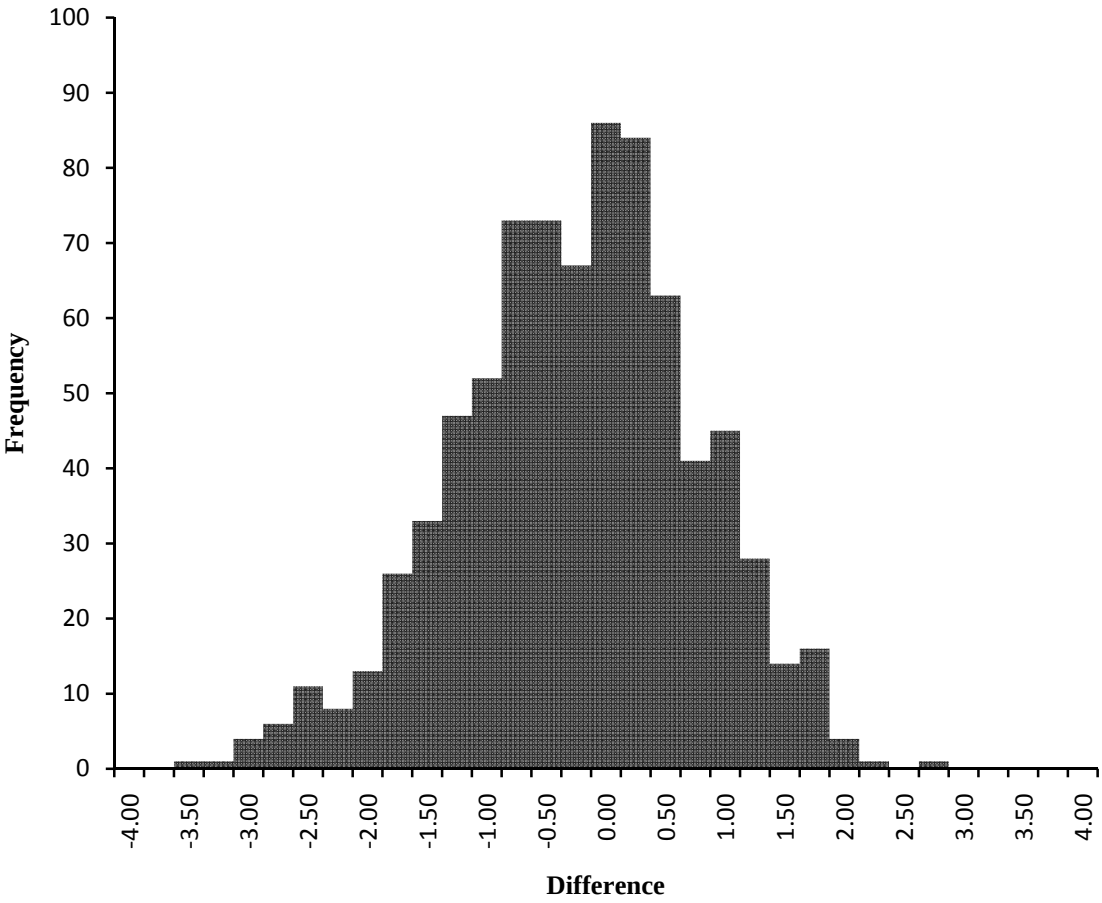

Figure 3.6: Deviation of GEBVs of the CLEANED dataset from GEBVs of the ALL dataset in standard deviation units

CHAPTER 4

EVALUATION OF A MULTI-LINE BROILER CHICKEN POPULATION USING A SINGLE STEP GENOMIC EVALUATION PROCEDURE[1]

[1] Simeone, R., I. Misztal, and I. Aguilar. To be submitted to *Journal of Animal Breeding and Genetics*

Abstract

The objective of this study was to evaluate effects on prediction of analyzing a multi-line chicken population as one line using a combined pedigree-genomic relationship matrix in a single-step procedure. Body weight at six weeks was provided by Cobb-Vantress for two lines of broiler chickens, A and B, each with three generations. Phenotypic records were available for 183,695 and 164,149 broilers and genotypic records were available for 3,195 and 3,001 broilers for lines A and B, respectively. Lines A and B were combined to create a multi-line population and were analyzed using a single-step procedure that combines the additive relationship matrix (**A**) and the genomic relationship matrix (**G**). **G** for the multi-line population was scaled using allele frequencies computed from either line A, line B, the multi-line population, or a constant, 0.5. Diagonal elements of **G** for each animal were isolated. Distributions of the elements were plotted for each of four allele frequencies. When allele frequencies from line A or line B were used, two peaks appeared in the distributions of **G** representing animals from each line. When allele frequencies were calculated from the multi-line population, A and B were indistinguishable. When the constant allele frequency of 0.5 was used, the distributions of A and B overlapped but had distinct peaks. Genomic estimated breeding values (GEBVs) were predicted for all animals in a validation dataset from the multi-line population using allele frequencies from line A, line B, the multi-line population, and the constant allele frequency, 0.5. Use of different allele frequencies in the multi-line evaluation predicted different GEBVs for each population but GEBVs had strong correlations with correctly predicted GEBVs. It may be possible to use **G** as a diagnostic tool to identify population substructure; moreover, **G** may be used to evaluate multiple

populations simultaneously but care should be taken to adjust the matrix appropriately or to scale the difference between **G** and **A** in order to obtain unbiased estimations.

Key Words: allele frequency, chickens, genomic estimated breeding value, genomic relationship matrix

Introduction

The genomic relationship matrix (**G**) is a matrix of realized, as opposed to expected, relationships. When used for genomic selection, it reflects more accurate relationships than the pedigree relationship matrix (**A**) (Hayes et al., 2009b). Genomic selection initially focused on the estimation of SNP effects to obtain genomic estimated breeding values (GEBVs), however, recent work has indicated that using **G** in place of **A** may have similar or increased accuracy (Chen et al., 2010; Hayes et al., 2009b). Combining **A** with **G** in order to obtain predictions for genotyped and ungenotyped animals is also possible (Aguilar et al., 2010; Christensen and Lund, 2010).

G must be scaled so that it is analogous to **A**; that is, diagonal elements should be close to 1.0 and off-diagonal elements close to 0.0 (VanRaden, 2008). If the elements of **G** are larger than those of **A**, predictions and variance estimates may be inflated (Forni et al., 2010). This problem can be resolved by scaling **G** using allele frequencies estimated from the population, and by making further adjustments using a constant to bring elements close to those of **A** (Forni et al., 2010).

G is scaled using allele frequencies estimated from a population and most research has focused on creating **G** with single-breed populations. Different breeds or lines of animals are under different selection pressures; as a result, SNP markers in linkage disequilibrium with quantitative trait loci (QTL) of interest may be in different

phases between breeds (Hayes et al., 2009a). Harris and Johnson (2010) indicated that constructing **G** in a multi-breed population using regression techniques without taking differences between breeds into account led to distortion in the diagonal elements of **G**. If **G** is constructed using a multi-breed or line population, differences in allele frequencies should be taken into consideration (Harris and Johnson, 2010).

How the elements of **G** will change if multiple, equally sized populations are analyzed as one or if a second population is incorrectly added to an existing dataset is unclear. The addition of a second population may alter the predictions of GEBVs because **G** will be scaled differently by the change in allele frequencies and creation of more relationships due to shared alleles. Off-diagonal elements of the matrix may show a greater number or fewer relationships when incorrect allele frequencies are used. The choice of allele frequencies to use in analysis is important to have accurate and unbiased predictions. The objective of this study was to evaluate the effect on prediction of analyzing a multi-line chicken population as one line using the single-step procedure proposed by Aguilar et al. (2010).

Materials and Methods

Data

Body weight at six weeks (100g) was provided by Cobb-Vantress for two lines of broiler chickens: A and B, each with three generations. Phenotypic records were available for 183,695 and 164,149 broilers for lines A and B, respectively. Subsets of these populations were genotyped and genotypic records were available for 3,195 and 3,001 broilers for lines A and B, respectively. Animals from lines A and B were genotyped for the same 57,636 SNPs and frequencies of the second-allele at each locus were calculated

for each line. Allele frequencies at each loci between lines A and B were moderately correlated (r = 0.57), indicating that differences in frequencies at multiple loci existed.

Lines A and B were combined to create a multi-line population. Descriptions of the phenotypic data for all animals and genotyped animals in lines A and B, and the multi-line population are provided in Table 4.1. Lines A and B were analyzed separately with and without genomic data to establish GEBVs and EBVs using the correct population structure and allele frequencies. **G** for the multi-line population was scaled using allele frequencies computed from line A or line B, estimated from the multi-line population, or using the constant 0.5. The multi-line population was split into a training dataset, consisting of all animals from the first two generations, and a validation dataset, consisting of all animals with phenotypes and genotypes from the third generation. The training dataset included 290,632 animals (155,811 from line A and 134,821 from line B) and the validation dataset included 1,597 animals (798 animals from line A and 799 animals from line B).

Model and analysis

A single-trait model was used for the analysis:

$$\mathbf{y}=\mathbf{X}_{\mathrm{b}}+\mathbf{W}_{\mathrm{mp}}+\mathbf{Z}_{\mathrm{u}}+\mathbf{e},$$

in which **y** was a vector of BW observations; **X**, **W**, and **Z** were the appropriate incidence matrices relating observations to animals; **b** was a vector of fixed effects for hatch and sex, **mp** was a vector of random maternal permanent environmental effects, **u** was a vector of random additive genetic effects that integrated polygenic and genomic breeding values (Aguilar et al., 2010), and **e** was a vector of residuals. Variance components for the multi-line population were estimated using REML as $\sigma^2_{\mathrm{mp}}= 0.23$, $\sigma^2_{\mathrm{u}} = 0.83$, and

$\sigma_e^2 = 3.77$, with $h^2 = 0.17$. Analysis was done as in Chen et al. (2010) and used the combined genomic and pedigree relationship matrix, **H**.

The genomic relationship matrix of all genotyped animals was constructed as in VanRaden et al. (2008) as $\mathbf{G} = \frac{\mathbf{ZZ'}}{2\sum p_i(1-p_i)}$ and used second-allele frequencies for line A, line B, the multi-line population, and the constant, 0.5, to scale **G**. Correct scaling of **G** depends on correct estimation of allele frequencies and animals with genotypes that are dissimilar to the calculated allele frequencies will likely have elements of **G** that are scaled inappropriately. Diagonal elements of **G** represent an animal's relationship with itself (inbreeding) (VanRaden, 2008). The distributions of the diagonal elements of **G** for each animal at each allele frequency were plotted. Off-diagonal elements, which represent animals' relationships with other animals, are also affected by the second-allele frequencies used to scale **G** (VanRaden, 2008). Statistics of the off-diagonal elements of **G** were examined to see how the use of different second-allele frequencies affected relationships among animals.

GEBVs were predicted for all animals in the validation dataset from the multi-line population using second-allele frequencies from line A, line B, the multi-line population, and constant 0.5. To find how the use of different allele frequencies affected the predictions, GEBVs from the multi-line population analysis were separated into GEBVs for line A and B and were correlated with the GEBVs obtained when line A and line B were analyzed alone as well as EBVs obtained with no genomic information.

Results and Discussion

To examine how the construction of **G** changed due to differences among the second-allele frequencies, the distributions of the diagonal elements of **G** were plotted after each analysis (Figure 4.1). When the second-allele frequencies from either line A or line B were used, two distinct peaks appeared in the distributions. When the second-allele frequency from line A was used, the mean of the diagonal elements for line A animals was 1.00 (0.04) and ranged from 0.56 through 1.40 while the mean of the diagonal elements for line B animals was 2.04 (0.09) and ranged from 1.61 through 2.17. Similarly, when the second-allele frequency from line B was used, the mean of the diagonal elements for line A animals was 2.16 (0.05) and ranged from 1.65 through 2.34 while the mean of the diagonal elements for line B animals was 1.00 (0.04) and ranged from 0.62 through 1.17. When the second-allele frequency calculated from the multi-line population or the constant 0.5 were used, the distribution of the diagonal elements did not behave the same way. Using multi-line second allele frequencies, lines A and B were indistinguishable from each other. Line A animals had a mean diagonal element of 1.14 (0.03) and ranged from 0.72 through 1.42 while line B animals had a mean diagonal element of 1.13 (0.04) and ranged from 0.77 through 1.53. The overall mean of the diagonal distribution using multi-line allele frequencies was 1.15 (0.05). This value is larger than the expected value of 1.0 when the founder population is assumed to be unrelated, but the single peak in the distribution indicates that it may be possible to use multi-line populations by estimating a combined allele frequency. The large diagonal element, however, may cause inflated or inaccurate breeding values. Using the constant second-allele frequency of 0.5, diagonal elements of lines A and B overlapped but the

distribution was bimodal. Line A animals had a mean diagonal element of 1.42 (0.03) and ranged from 0.89 through 1.52 while line B animals had a mean diagonal element of 1.39 (0.03) and ranged from 1.06 through 1.79.

When the second-allele frequencies of line A were used, the off-diagonals of **G** ranged from -0.18 to 2.15 and had a mean of 0.25 (0.49). When the second-allele frequencies of line B were used, the off-diagonals of **G** ranged from -0.16 to 2.29 and had a mean of 0.29 (0.53). When the second-allele frequencies of the multi-line population were used, the off-diagonals of **G** ranged from -0.36 to 1.21 and had a mean of 0.00 (0.27). When the constant 0.5 was used, the off-diagonals of **G** ranged from -0.11 to 1.45 and had a mean of 0.5 (0.19). Negative off-diagonal elements represent individuals sharing fewer alleles than would be expected given the allele frequencies used to scale **G** (Astle and Balding, 2009). Combining lines A and B created relationships among animals that did not exist due to shared SNP markers among animals from both lines that were likely not identical by descent.

After the evaluation of the multi-line population, GEBVs for animals in the validation dataset were separated into those for animals from line A or line B and then correlated with GEBVs or EBVs obtained from prior analysis using either line A or line B and the correct second-allele frequency for each ($GEBV_A$ and $GEBV_B$, EBV_A and EBV_B). This was completed for each of the four second-allele frequencies used. Statistics for GEBVs estimated for the multi-line population are presented in Table 4.2 and correlations between GEBVs and EBVs are provided in Table 4.3.

GEBVs and EBVs estimated for the single populations were obtained. EBVs predicted with traditional BLUP evaluation had a mean of -0.14 (0.47) for Line A and -

0.28 (0.34) for Line B, while GEBVs predicted using the correct allele frequencies had a mean of 0.07 (0.59) for Line A and 0.00 (0.47) for Line B. GEBVs were slightly inflated compared with EBVs when genomic information was included in the evaluation; moreover, the ranges of GEBVs were larger than those of EBVs for both Line A and B. The correlations between EBVs and GEBVs were 0.74 and 0.61 for line A and B, respectively (Table 4.3).

Use of different scaling factors may bias or change the scale of GEBVs (Aguilar et al., 2010). Mean GEBVs for Lines A and B varied based on second-allele frequency used. Mean GEBVs for animals in either line were closest or equal to $GEBV_A$ or $GEBV_B$ when the correct second allele frequency for Line A or B was used but the ranges of GEBVs were altered. Use of second-allele frequencies that did not correspond to a line resulted in decreased estimates of GEBVs compared to $GEBV_A$ or $GEBV_B$ but means were similar to those estimated without genomic information. Use of the multi-line second allele frequency resulted in inflated (Line A) or deflated (Line B) GEBVs compared to $GEBV_A$ and $GEBV_B$, respectively. Use of 0.5 second-allele frequency resulted in deflated GEBVs for both Line A and Line B compared to GEBVs and EBVs. While mean values indicate differences in GEBV predictions, the correlations between GEBVs from Line A animals and $GEBV_A$ were all 0.97, and correlations between GEBVs from Line B and $GEBV_B$ were all 0.96, indicating that despite differences in predictions, animals were ranked appropriately.

Slight differences were seen between GEBVs from the multi-line population and EBV_A and EBV_B. Correlations between line A GEBVs and line A EBVs were 0.72, 0.72, 0.72, and 0.75 for second-allele frequencies from line A, line B, the multi-line

population, and 0.5, respectively. Similarly, correlations between line B GEBVs and line B EBVs were 0.55, 0.55, 0.56, and 0.59 for second-allele frequencies from line A, line B, the multi-line population, and 0.5, respectively. These correlations were slightly less than those between correctly estimated single-line GEBVs and EBVs.

The prediction of GEBVs is of primary interest to animal breeders. Inclusion of genomic relationship information allows for more accurate predictions by constructing relationships based on shared alleles instead of expected relationships (Hayes et al., 2009b; VanRaden, 2008). Rather than using prediction equations to estimate SNP effects, as in Meuwissen et al. (2001), **G** allows for more accurate relationships and also can be used with ungenotyped populations (Aguilar et al., 2010; Christensen and Lund, 2010). Use of multiple populations in one evaluation has proven difficult due to differences in environment and selection pressures, leading to differences in population allele frequencies. In this case, GEBV estimates were inflated or deflated depending on the second-allele frequency used to scale **G**. Animals do, however, appear to be ranked correctly. It is possible to scale the difference between **G** and **A** in the construction of the combined genomic-pedigree relationship matrix in order to reduce differences in GEBVs and this could prove to be a valuable tool in multi-line evaluations (Aguilar et al., 2010).

Allele frequencies for the same loci vary from population to population. Harris and Johnson (2010) indicated that diagonal element of **G** are distorted when animals of different breeds are analyzed together, even if **G** is constructed by regression methods without using second-allele frequencies. It may be possible to use **G** as a diagnostic tool to identify population substructure. To do so allele frequencies between populations must be very different. If the two populations have similar allele frequencies or similar

numbers of animals exist in each, it may be almost impossible to differentiate between the two using diagonal elements if **G** is scaled using the combined population second-allele frequencies. If some other allele frequency is used in place of the current allele frequency, it may be possible to separate two, equally sized, similar populations. When the constant 0.5 was used to scale **G**, the distribution of diagonal elements overlapped but showed two distinct peaks, indicating that multiple populations may be detectable but not necessarily separable prior to analysis.

Use of **G** avoids the problem of population substructure by estimating relationships rather than SNP effects (Hayes and Goddard, 2008); with a dense enough marker map it may not be necessary to worry about population structure (Toosi et al., 2010). The use of correct allele frequencies, however, is crucial to the construction of **G** because **G** is scaled to be analogous to **A** using allele frequencies (VanRaden, 2008). Animals that are homozygous for rare alleles will tend to have a higher genomic inbreeding coefficient than those who are not (VanRaden, 2007); moreover, allele frequency estimation has more of an effect on genomic inbreeding than on genomic predictions (VanRaden et al., 2008). Differences in allele frequencies between populations can cause animals from one population to be homozygous for alleles that a second population is not. Use of incorrect allele frequencies causes apparent high inbreeding; additionally, increased false relationships among animals that are also incorrectly scaled can inflate or deflate GEBV predictions or cause incorrect ranking of animals in a population.

Conclusions

Evaluation of two populations changed with the allele frequency used to scale the genomic relationship matrix. Using allele frequencies from line A, line B, the multi-line population, or the constant, 0.5, resulted in inflated or deflated genomic breeding values but showed strong correlations with correctly estimated genomic estimated breeding values or estimated breeding values. It may be possible to use the genomic relationship matrix to evaluate multiple populations simultaneously by using the average allele frequency of the mixed population but care must be taken to adjust the genomic relationship matrix appropriately. Evaluation of the diagonal elements of the genomic relationship matrix suggests that populations of equal size and with some similarities in population allele frequency are difficult to separate; populations with greater differences in allele frequencies may be easier to separate.

Acknowledgements

The authors thank Cobb-Vantress for access to data for this study. This study was partially funded by AFRI grants 2009-65205-05665 and 2010-65205-20366 from the USDA NIFA Animal Genome Program.

References

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

Astle, W. and Balding, D. J. 2009. Population structure and cryptic relatedness in genetic association studies. Stat. Sci. 24: 451-471.

Chen, C. Y., I. Misztal, I. Aguilar, S. Tsuruta, T. H. E. Meuwissen, S. E. Aggrey, and W. M. Muir. 2010. Genome wide marker assisted selection in chicken: Making the most of all data, pedigree, phenotypic, and genomic in a simple one step procedure. Proc. 10th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany.

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

Forni, S., I. Aguilar, I. Misztal, and N. Deeb. 2010. Genomic relationships and biases in the evaluation of sow litter size. Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany.

Goddard, M. E., B. Hayes, H. McPartlan, and A. J. Chamberlain. 2006. Can the same genetic markers be used in multiple breeds? Proc. 8th World Congr. Genet. Appl. Livest. Prod., communication no. 22-16. Belo Horizonte, MG, Brasil.

Hayes, B., P. Bowman, A. Chamberlain, K. Verbyla, and M. Goddard. 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Gen. Sel. Evol. 41: 51.

Harris, B. L. and Johnson, D. L. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci. 93: 1243-1252.

Hayes, B. J. and M. E. Goddard. 2008. Technical note: Prediction of breeding values using marker-derived relationship matrices. J. Anim Sci. 86: 2089-2092.

Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. Gen. Res. 91: 47-60.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics. 157: 1819-1829.

Toosi, A., R. L. Fernando, and J. C. M. Dekkers. 2010. Genomic selection in admixed and crossbred populations. J. Anim. Sci. 88: 32-46.

VanRaden, P. M. 2007. Genomic measures of relationship and inbreeding. Interbull. 37: 33-36.

VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91: 4414-4423.

VanRaden, P. M., M. E. Tooker, and N. Gengler. 2008. Effects of allele frequency estimation on genomic predictions and inbreeding coefficients. J. Dairy Science. 91(E-Suppl. 1): 506. (Abstr.)

Table 4.1: Descriptions of the phenotypic data for body weight for all animals and genotyped animals in lines A and B, and the multi-line population [1]

Line		No. of Records	Mean (SD)
A			
	All Animals	183,695	24.50 (3.22)
	Genotyped Animals [2]	3,195	25.12 (2.97)
B			
	All Animals	164,149	23.53 (3.17)
	Genotyped Animals	3,001	23.39 (2.63)
Multi-Line [3]			
	All Animals	347,844	24.04 (3.24)
	Genotyped Animals	6,196	24.28 (2.94)

[1] Phenotypic data for body weight (100g) at 6 weeks existed for two lines of broiler chickens, A and B, over three generations

[2] Genotyped animals represent subsets of lines A and B with both phenotypes and genotypes

[3] Multi-line represents both lines A and B treated as one dataset

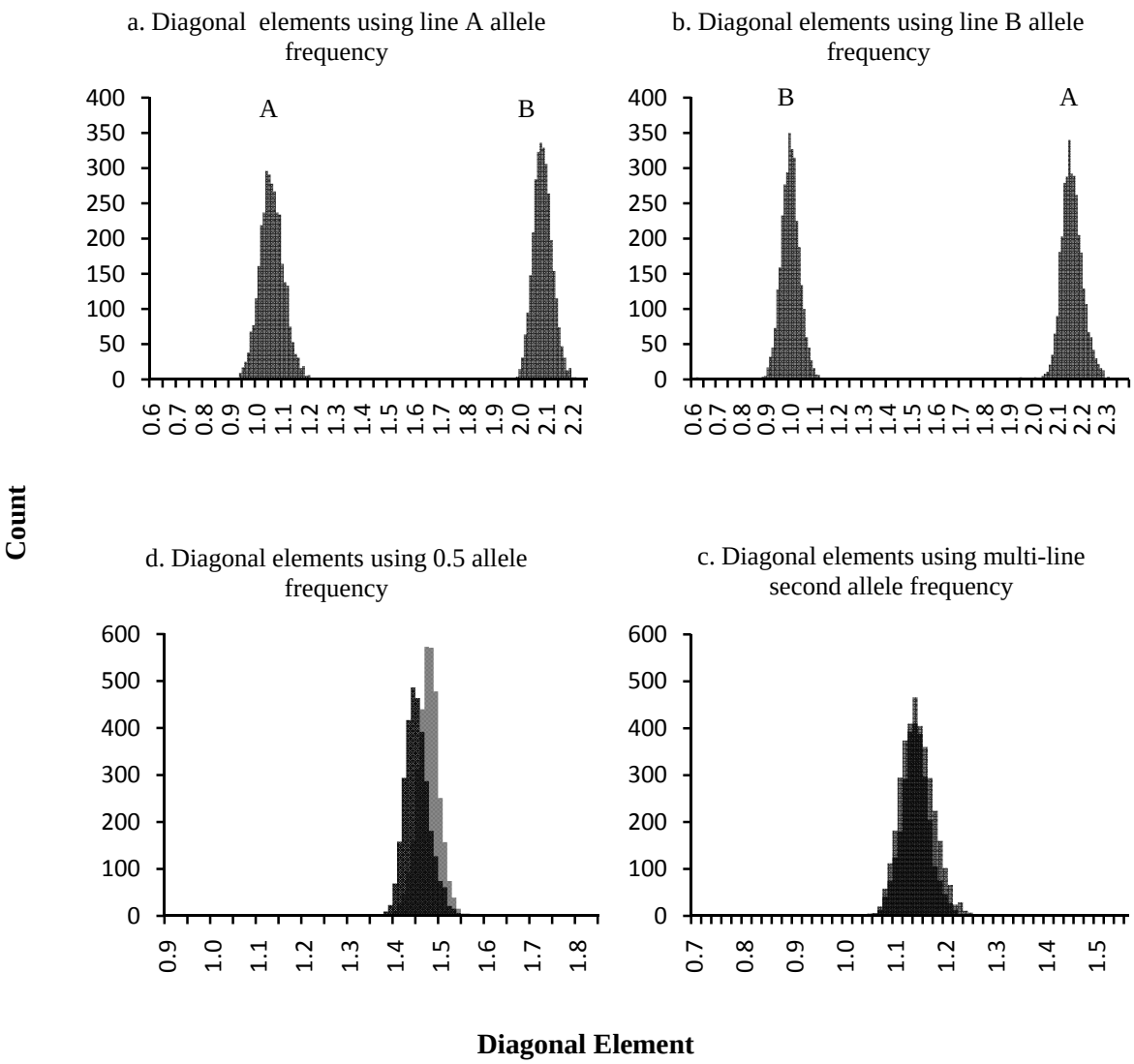

Figure 4.1: Distributions of the diagonal elements of **G** constructed with second-allele frequencies from line A, line B, the multi-line population of A and B, and 0.5. Distributions are shown on different scales.

Table 4.2: Statistics for GEBVs and EBVs estimated for the multi-line population[1]

Population	Mean (G)EBV (SD)	Minimum	Maximum
Line A, Traditional	-0.14 (0.47)	-1.32	1.01
Line A, SSP[2] (A)	0.07 (0.59)	-1.71	1.70
Multi (A)	0.07 (0.60)	-1.75	1.76
Multi (B)	-0.12 (0.60)	-1.96	1.54
Multi (AB)[3]	0.13 (0.58)	-1.64	1.70
Multi (0.5)	-0.18 (0.53)	-1.81	1.29
Line B, Traditional	-0.28 (0.34)	-1.31	0.69
Line B, SSP (B)	0.00 (0.47)	-1.51	1.31
Multi (A)	-0.25 (0.47)	-1.74	0.93
Multi (B)	0.02 (0.49)	-1.54	1.25
Multi (AB)	-0.06 (0.47)	-1.49	1.11
Multi (0.5)	-0.34 (0.42)	-1.64	0.76

[1] The combined population consisted of animals from lines A and B. After genomic evaluation, GEBVs obtained for animals in the third generation were separated into those belonging to A or B. Frequencies used to scale **G** are in parentheses.
calculated; this frequency was used to scale the genomic relationship matrix
[2] Single-step procedure
[3] AB refers to the second-allele frequency calculated from the multi-line population

Table 4.3: Correlations between GEBVs and EBVs for lines A, B, and the Multi-line population using different allele frequencies[1]

Line A (G)EBVs						
	Line A, Traditional	Line A, SSP (A)	Multi (A)	Multi (B)	Multi (AB)	Multi (0.5)
Line A, Traditional	1.00	0.74	0.72	0.72	0.72	0.75
Line A, SSP (A)		1.00	0.97	0.97	0.97	0.97
Multi (A)			1.00	1.00	1.00	0.99
Multi (B)				1.00	1.00	0.99
Multi (AB)					1.00	0.99
Multi (0.5)						1.00

Line B (G)EBVs						
	Line B, Traditional	Line B, SSP (B)	Multi (A)	Multi (B)	Multi (AB)	Multi (0.5)
Line B, Traditional	1.00	0.61	0.55	0.55	0.56	0.59
Line B, SSP (B)		1.00	0.96	0.96	0.96	0.96
Multi (A)			1.00	1.00	1.00	0.99
Multi (B)				1.00	1.00	0.99
Multi (AB)					1.00	0.99
Multi (0.5)						1.00

[1] Allele frequency used to scale **G** is in parentheses

CHAPTER 5

CONCLUSIONS

The results of this study indicate that the genomic relationship matrix may be a useful tool for identifying misidentified animals in a genotyped population; the sensitivity of the diagonal elements of the genomic relationship matrix is dependent on the number of misidentified animals and the similarity of allele frequencies between populations. Simulation studies indicate that the distribution of diagonal elements of the genomic relationship matrix should be centered on 1.00 and widens with increased generations and relatedness. Additionally, analysis of simulated data indicated that scaling the genomic relationship matrix using allele frequencies calculated from a complete population of two lines of animals led to a bi-modal distribution, correctly separating the populations. An analysis of field data indicated that diagonal elements of the genomic relationship matrix were incorrectly scaled for a small subset of misidentified animals from a second line of broiler chickens; moreover, the accuracy of prediction was negatively affected by the presence of mislabeled animals. When two full lines of chickens were analyzed as one with allele frequencies calculated for the complete population, the two lines were indistinguishable; when 0.5 was used to scale the genomic relationship matrix, the distribution of diagonal elements overlapped but was bimodal. When diagonal elements of the genomic relationship matrix are scaled using calculated allele frequencies, identification of multiple populations may be possible as long as one population is smaller than the other or if the two populations have very different allele frequencies.

Analysis of the combined field data indicated that genomic estimated breeding values changed based on the allele frequency used. Use of 0.5 to scale the genomic relationship matrix resulted in reduced genomic estimated breeding values. When the genomic relationship matrix was scaled using allele frequencies from one line or the other, genomic estimated breeding values for animals from the same population were similar to those obtained using only the single step procedure and one population, while animals from the other line had deflated genomic estimated breeding values. Use of the multi-line frequency resulted in inflated (line A) or deflated (line B) predictions compared to genomic estimated breeding values estimated from the single populations. Use of 0.5 resulted in reduced predictions. Despite differences in the mean genomic estimated breeding values, predictions with different allele frequencies were highly correlated with each other and with genomic estimated breeding values obtained in a single-step procedure using only one population. This suggests that use of incorrect allele frequency may alter genomic estimated breeding values but that animals will still be ranked correctly.

It may be possible to evaluate multi-line populations using the genomic relationship matrix provided that differences in allele frequencies are accounted for or that the genomic relationship matrix is scaled appropriately. Use of a genomic relationship matrix with multiple populations may create false relationships between unrelated animals due to loci that are identical by state but not by descent. A denser SNP map may alleviate this problem, as animals between breeds will have fewer identical loci.

Use of genomic data is promising for animal breeding but has not yet reached its full potential. Adequate methods of identifying mislabeled animals are needed in order to

obtain accurate genomic estimated breeding values and the presence of multiple breeds or lines can influence the results of the analysis. The genomic relationship matrix may be useful in the separation of multiple populations but further research is needed to identify its limitations and sensitivity.