2011

Genomic selection of purebred animals for crossbred performance under dominance

Jian Zeng
Iowa State University

Genomic selection of purebred animals for crossbred performance under dominance

by

Jian Zeng

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Animal Breeding and Genetics

Program of Study Committee:

Rohan L. Fernando, Major Professor

Jack C.M. Dekkers

Dorian J. Garrick

Alicia L. Carriquiry

Iowa State University

Ames, Iowa

2011

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, I would give my gratitude to my advisor Dr. Rohan L. Fernando for his guidance, patience and support throughout this research and the writing of this thesis. His insights and words of encouragement have often inspired me and renewed my hopes for conquering academic difficulties. This thesis would not have been completed without the discussions with him. I would also like to thank my committee members for their efforts and contributions to this work: Dr. Jack C.M. Dekkers, Dr. Dorian J. Garrick and Dr. Alicia L. Carriquiry. I very much appreciate the classes that Dr. Jack C.M. Dekkers, Dr. Rohan L. Fernando and Dr. Dorian J. Garrick provided from where I learned a lot about quantitative genetics, linear models and Bayesian theory which directly contributed to this work. I would additionally like to thank my colleagues Dr. Wei He, Toosi Ali, Dr. David Habier, Dr. Kadir Kizilkaya and Dr. Chunkao Wang etc for their help, support and valuable hints about the research, computer programming, learning etc. Especially, I would like to give my special thanks to my parents whose patient love enabled me to complete this work.

CHAPTER 1. OVERVIEW

1.1 Introduction

Genomic selection (GS) is a variant of marker-assisted selection that uses genome-wide single nucleotide polymorphisms (SNP) to predict individual breeding values for selection (Meuwissen et al., 2001; Goddard and Hayes, 2007). Numerous studies have shown encouraging results of applying GS in selection of purebreds (Meuwissen et al., 2001; VanRaden, 2008; Calus et al., 2008; VanRaden et al., 2009; Hayes et al., 2009; Habier et al., 2007, 2010, 2011). However, except in dairy cattle, most of the animals used for livestock production systems are crossbreds with advantages of heterosis (Sheridan, 1981) and breed complementarity (Moav, 1973). For such systems, the breeding goal in purebreds is typically to optimize the performance of crossbred descendents.

Recent studies have shown that GS is also an appealing method to select purebreds for crossbred performance, particularly when the crossbreds are used for training (Dekkers, 2007; Piyasatian et al., 2007; Ibanez-Escriche et al., 2009; Toosi et al., 2010; Kinghorn et al., 2010). As compared to alternative methods that use covariance theory, such as combined crossbred and purebred selection proposed by Wei and van der Steen (1991) and Lo et al. (1993), GS can give substantially greater response to selection (Dekkers, 2007; Piyasatian et al., 2007), lower the rate of inbreeding (Dekkers, 2007; Daetwyler et al., 2007), and it does not require a systematic collection of pedigree that connects crossbreds to purebreds. Moreover, it is not necessary to measure the crossbred phenotypes every generation of GS, because in theory the estimates of SNP effects can be applied through a few generations with only a negligible loss in prediction accuracy (Meuwissen et al., 2001; Habier et al., 2007).

Under dominance, the model proposed by Dekkers (2007) and Kinghorn et al. (2010), which

fits breed-specific substitution effects of SNP alleles (BSAM), outperforms the usual additive model that only fits a common substitution effect for each SNP (Kinghorn et al., 2010). But this may not hold under additive inheritance alone (Ibanez-Escriche et al., 2009), which suggests the difference in linkage disequilibrium (LD) across breeds is not important when breeds are related. Then, the advantage of fitting BSAM observed in the study of Kinghorn et al. (2010) may be primarily due to the dominance effect – as the allele substitution effect is a function of the dominance effect and the allele frequency (Falconer and Mackay, 1996).

If this is true, explicitly including the dominance effect in the GS model can be beneficial to purebred selection for crossbred performance. Further, it is easy to extend Bayesian regression models used for GS to accommodate dominance by fitting within-locus dominance deviations as random effects in the model (Xu, 2003; Lee et al., 2008; Toro and Varona, 2010; Wellmann and Bennewitz, 2010; Zeng et al., 2011). Given that the marker-QTL associations are similar across breeds, the breed-specific allele substitution effects for a purebred parent can be computed by combining the estimates of additive and dominance effects from training with the allele frequencies from the other parental breed of a cross (Kinghorn et al., 2010). Then, the advantages of using this dominance model over BSAM would be 1) breed origin of SNP alleles must be known or inferred for fitting BSAM (Dekkers, 2007; Kinghorn et al., 2010) but are not needed for the dominance model, and 2) the estimates of SNP effects can be successively applied over generations or across breeds with updated allele frequencies to develop prediction equations specific to the breed. Thus, to assess the performance of the dominance model in comparison with the alternatives for GS on purebreds for crossbred improvement constitutes the first objective of this research.

In beef cattle, the commercial populations often consist of admixtures of crossbreds with unknown breed composition. It is known that association studies in such populations can give misleading results due to spurious associations of SNP with trait phenotypes (Rabinowitz, 1997; Flint-Garcia et al., 2003; Hirschhorn and Daly, 2005). Several methods have been proposed to reduce the false association signals generated by population structure (Kennedy et al., 1992; Spielman et al., 1993; Pritchard et al., 2000; Meuwissen et al., 2002; Price et al., 2006; Yu et al., 2006). It has been argued, however, that in Bayesian regression models where all SNP

are fitted simultaneously breed differences are explained by the SNP and therefore problems due to population admixture can be ignored in GS. This has been verified by Toosi et al. (2010) under additive gene action. In this study we will verify if population admixture can be ignored even under dominance gene action. We hypothesize that if the model includes dominance breed differences in crossbreds will be explained by the SNP and therefore admixture can be ignored.

1.2 Research Objectives

The objectives of this study were

- to compare the performance of the additive model, BSAM, and the dominance model in the GS of purebreds for crossbred performance based on: 1) accuracy of prediction in a variety of training populations, and 2) crossbred response to 20 generations of selection in a simulated two-way crossbreeding program;
- to evaluate accuracy of prediction of purebreds for crossbred performance with training on admixed populations, when breed composition was either ignored or considered.

CHAPTER 2. REVIEW OF LITERATURE

2.1 Advantages of Crossbreeding

Crossbreeding is widely used in livestock to produce individuals with superior performance for characters of economic importance. Most of the superiority of crossbred over purebred animals is attributable to hybrid vigor or heterosis, which has been generally found to occur in swine (Johnson, 1980), poultry (Kosba, 1978), sheep (Nitter, 1978), beef (Cundiff and Gregory, 1999), and dairy cattle (Lopez-Villalobos et al., 2000). Even more than heterosis, crossbreeding is sought for breed complementarity, which is to combine different desirable characteristics from pure lines or breeds (Moav, 1973). In addition to the genetic advantages of heterosis and breed complementarity, another commercial benefit of crossbreeding is that the hybrids that are sold for production are not suitable for breeding because the heterosis would not be retained in the descendents of commercial crossbreds.

2.2 Genetic Basis of Heterosis

Heterosis has been under investigation for over a century, but the genetic basis underlying this phenomenon is still controversial (East, 1936; Lippman and Zamir, 2007; Birchler et al., 2010). The mechanism of heterosis was first explained by Shull (1908) and East (1908), who proposed the overdominance hypothesis that attributes heterosis to the superiority of heterozygous genotype over both homozygous genotypes at a single locus. The existence of overdominance has been observed in many traits (Li et al., 2001; Luo et al., 2001; Estelle et al., 2008; Ishikawa, 2009; Boysen et al., 2010; Dagnachew et al., 2011). A possible mechanism for overdominance is pleiotropy, where the gene has two alleles affecting different components of the trait in opposite directions. Thus, the phenotype of a heterozygote, which carries both variants (alleles) of

the gene, would surpass either homozygote (Falconer and Mackay, 1996). Overdominance can arise even when allele effects are additive for each component of the trait in opposite directions but the trait is defined by the multiplicative combination of the components (Wallace, 1968; Falconer and Mackay, 1996). Evidence for overdominance has also been found at the molecular level (Berger, 1976; Comings and MacMurray, 2000; Birchler et al., 2010).

An alternative widely accepted hypothesis, the dominance hypothesis attributes heterosis to the dominance complementation of detrimental recessive alleles at different loci (Bruce, 1910; Davenport, 1908). Suppose one parent with haplotype AAbb, where capital letter represents beneficial dominant allele, is crossed to another parent with haplotype aaBB. Hybridization would then result in a complementation of detrimental effects by dominant alleles at both loci. As a result, the crossbred phenotype would exceed the mean of the parents. The two contending hypotheses use different types of dominance to explain the mechanism of heterosis, however, they are hardly differentiable when loci are linked in repulsion phase, and Jones (1917) termed this phenomenon as pseudo-overdominance.

To reveal the major cause for heterosis, numerous studies have been conducted but the results are conflicting (Pirchner and Mergl, 1977; Xiao et al., 1995; Li et al., 2001; Luo et al., 2001; Frascaroli et al., 2007; Lippman and Zamir, 2007). The relative contribution of these models to the occurrence of heterosis is still obscure, though the dominance hypothesis is somewhat more favored in (Charlesworth and Willis, 2009). For some traits, the occurrence of heterosis may be attributable to the combination of dominance and overdominance with comparable effects (Li et al., 2008). Further, these hypotheses are virtually connected because both rely on the presence of dominance gene action and only differ in the degree of dominance.

Besides dominance or overdominance, the generation of heterosis also depends on the relationship between the parental populations. East (1936) reviewed relevant studies and concluded that heterosis is positively associated with the genetic disparity of the parental populations. Evidence can be found from the fact that plant crosses that typically use highly inbred lines often manifest higher level of heterosis than animal crosses, which are made by different mildly inbred lines or different breeds to avoid a severe loss in fertility (Falconer and Mackay, 1996). Yield advantages in hybrid crops can range from 15% to 50% (Duvick, 1999) while heterosis

in animal crosses are up to about 10% (Johnson, 1980; Kosba, 1978; Cundiff and Gregory, 1999). Falconer and Mackay (1996) comprehensively formulated how the dominance and the difference in gene frequency across parental populations jointly affect the level of heterosis in a cross as follows. Consider a single biallelic locus that has effect a, d, and $-a$ at the dominant homozygous, heterozygous, and recessive homozygous genotypes, respectively. The dominant allele frequency is p in the sires and p' in the dams, while the recessive allele frequency is q in the sires and q' in the dams. Let $y = p - p' = q - q'$, which denotes the difference in gene frequency between sires and dams, so that $p' = p - y$ and $q' = q - y$. Given Hardy-Weinberg equilibrium holds in parental populations and the sires are randomly mated to the dams, the parental means, M_s, M_d, and the crossbred mean, M_{F_1}, are

$$M_s = a(p - q) + 2dpq, \tag{2.1}$$

$$M_d = a(p - q - 2y) + 2d[pq + y(p - q) - y^2], \tag{2.2}$$

$$M_{F_1} = a(p - q - y) + d[2pq + y(p - q)]. \tag{2.3}$$

Thus, the amount of heterosis at this locus, expressed as the difference between the crossbred and the average parental means, is

$$H_{F_1} = M_{F_1} - \frac{1}{2}(M_s + M_d) = dy^2. \tag{2.4}$$

In the absence of epistatic interaction between loci, heterosis is attributable to the additive combination of the effects of the loci that jointly affect the trait.

$$H_{F_1} = \sum dy^2 \tag{2.5}$$

It can be concluded from equation 2.5 that the dispersion in gene frequency between parental populations increases the amount of heterosis in crossbreds. Further, fixing one allele in the sires and the alternate in the dams at each locus would maximize the heterosis. Given the difference in gene frequency between parental populations is constant, the amount of heterosis linearly increases with the degree of positive dominance at each locus. If epistasis is also present, the linearity would be affected, however, the presence of epistasis alone cannot cause any heterosis (Crow and Kimura, 1970; Falconer and Mackay, 1996). Further, most of the studies placed the

epistatic interactions to a secondary or minor role in heterosis (Li et al., 2001; Luo et al., 2001; Li et al., 2008; Estelle et al., 2008), though it may be important to some traits (Meffert et al., 2002; Abasht and Lamont, 2007).

The contributions of gene frequency and non-additive gene action to heterosis implies that the occurence of heterosis is related to the proportion of non-additive, particularly dominance, genetic variations for the trait. It has been found that dominance can explain up to over 10% of phenotypic variance (Wei and van der Werf, 1993; Misztal et al., 1997; Gengler et al., 1997; Culbertson et al., 1998). It is generally more important in lowly heritable traits, such as fertility and fitness, where dominance variance can be twice as large as the additive variance (Hoeschele, 1991; Crnokrak and Roff, 1995). As expected, heterosis is substantially higher for fertility and fitness than for production traits (Gengler et al., 1997; Cundiff and Gregory, 1999). Thus, except in dairy cattle, commercial animals are often produced from crossbred dams that have high levels of fertility due to heterosis (Sheridan, 1981).

2.3 Methods to Select Purebreds for Crossbred Performance

Non-additive effects, albeit the likely basis of heterosis, are usually ignored in genetic evaluations of purebreds for crossbred performance in selection programs. The performance of purebreds is correlated to the crossbred performance depending on the level of additive genetic variability. For traits with significant non-additive variance and therefore potential heterosis, the purebred performance may not be a good predictor of crossbred performance. Furthermore, because nucleus purebreds are typically kept in superior environments whereas commercial crossbreds are exposed to various stresses under field conditions, the prediction of crossbred performance by purebred data may subject to the genotype by environment interactions (Dekkers, 2007). As a consequence, the genetic correlation between purebred and crossbred animals can be as low as 0.4 to 0.7 (Wei and van der Werf, 1995; Lutaaya et al., 2001; Merks and de Vries, 2002).

Thus, the conventional strategy that relies on selection of purebreds or pure lines on their own performance (PLS) is not effective to improve the crossbred performance. A number of methods have been proposed as alternatives to PLS to obtain greater response in crossbreds.

These can be classified into three groups:

- Reciprocal recurrent selection (RRS) where nucleus individuals are selected based on the hybrid performance of their sibs or descendents (Comstock et al., 1949; Bell et al., 1950);
- Combined crossbred and purebred selection (CCPS) where purebreds are selected based on selection index theory or mixed-model procedure that merges performance information of crossbred relatives with that of purebred animals (Wei and van der Steen, 1991; Lo et al., 1993);
- Marker-assisted selection (MAS), or more recently, genomic selection (GS) where purebreds are selected based on the effects of genetic markers or single nucleotide polymorphisms (SNP) estimated using crossbred performance (Dekkers and Chakraborty, 2004; Dekkers, 2007).

RRS that uses crossbred descendents or crossbred sib information as selection criteria can more efficiently exploit non-additive genetic variance than PLS. The practical value of RRS, however, was not as encouraging as expected in most of the experiments (Bowman, 1959; Calhoon and Bohren, 1974; Bell, 1982; Wei and van der Steen, 1991).

CCPS, which can be viewed as a combined method of PLS and RRS, simultaneously exploits additive and non-additive genetic variability (Wei and van der Steen, 1991). Different methods have been developed to implement CCPS. One approach is to treat purebred and crossbred performance as genetically different traits and use selection index theory to estimate the purebred breeding values for crossbred performance (Wei and van der Werf, 1994; Bijma and Arendonk, 1998). Alternatively, genetic evaluations of purebreds for crossbred performance can be obtained by best linear unbiased prediction (BLUP) via Henderson's mixed model equations (Lo et al., 1993, 1997). Although CCPS has been shown to give greater short-term crossbred response (Bijma and Arendonk, 1998), the long-term response in crossbreds will be impaired by the consequent increase of inbreeding rate because it increases the probability of coselection within family (Bijma et al., 2001; Dekkers, 2007). In addition, to implement CCPS requires routine collections of crossbred phenotypes and pedigree that can link crossbred descendents to

their purebred parents, which would increase the investment in the program (Dekkers, 2007). Moreover, it is very difficult to explicitly accommodate dominance in the model for CCPS. Lo et al. (1995) has shown that 25 parameters are needed to model the genotypic variances and covariances between purebreds and crossbreds under dominance, and the model complexity increases as more breeds are involved in the crossbreeding system. These drawbacks have limited the widespread application of CCPS in livestock.

GS proposed by Meuwissen et al. (2001) is an extension of MAS using genome-wide SNP as markers whose effects are treated as random in a mixed linear model. Once the effects of SNP have been estimated from training, they can be applied to predict the breeding values of genotyped animals at an early stage without own phenotypic records available to accelerate genetic progress (Harris et al., 2008). As SNP saturate the genome with high-density, effects of quantitative trait loci (QTL) that underlie the trait are expected to be captured by SNP associated with QTL through population-wide linkage disequilibrium (LD), which is consistent across families. Further, given SNP are linked to QTL, SNP reflect more accurate genetic relationship among genotyped individuals than pedigree by accounting for recombination event of loci and random sampling of gametes (Habier et al., 2007). Thus, pedigree might not be needed for GS. Moreover, it is not necessary to measure the phenotypes every generation of GS, because in theory the estimates of SNP effects can be applied through a few generations with only a negligible loss in prediction accuracy (Meuwissen et al., 2001; Habier et al., 2007). With such advantages, recent studies have shown encouraging results of GS in the selection of purebreds (Meuwissen et al., 2001; VanRaden, 2008; Calus et al., 2008; VanRaden et al., 2009; Hayes et al., 2009; Habier et al., 2007, 2010, 2011) and in purebred selection for crossbred performance (Dekkers, 2007; Piyasatian et al., 2007; Ibanez-Escriche et al., 2009; Toosi et al., 2010; Kinghorn et al., 2010; Mujibi et al., 2011).

In principle, Dekkers (2007) demonstrated that MAS or GS with marker effects derived from the commercial crossbred level led to substantially higher crossbred response and a lower rate of inbreeding compared to CCPS when the estimation of marker effects was accurate. Given that the SNP effects in a crossbred population originate in parental populations from different breeds, the usual additive model for GS that only fits a common substitution effect

for each SNP, however, may not be appropriate. For this reason, Dekkers (2007) and Kinghorn et al. (2010) suggested to use statistical models that accommodate breed-specific effects of SNP alleles to fit crossbred phenotypes (BSAM), and then to apply the estimates in the predictions of genomic breeding values (GEBV) of purebreds for crossbred performance specific to the breed. This method has been called marker-assisted selection for commercial crossbred performance (CC-MAS) in Dekkers (2007) or reciprocal recurrent genomic selection (RRGS) in Kinghorn et al. (2010). The authors also pointed out that the prerequisite for fitting BSAM, however, is that breed origin of SNP alleles must be known or inferred.

The performance of BSAM has been studied by stochastic simulations (Ibanez-Escriche et al., 2009; Kinghorn et al., 2010). Under additive gene action, fitting BSAM is beneficial only when the parental breeds are distantly related and the number of SNP are small relative to the size of the training population (Ibanez-Escriche et al., 2009). Under dominance, Kinghorn et al. (2010) demonstrated a clear advantage of BSAM over the additive model in crossbred response, assuming the estimation of SNP effects was perfect.

2.4 Population Admixture – A Concern in Genomic Selection

Given that the aim of selecting purebreds is to improve crossbred performance, commercial crossbred populations should be used for training in GS to take heterosis and genotype by environment interactions into account. Toosi et al. (2010) reported that training on crossbreds gave 11% more accurate prediction of crossbred genetic merit than training on purebreds. However, commercial populations may consist of admixtures of crossbreds with unknown breed composition, which is common in beef cattle. It has been argued that accuracy of GS with training in such populations may be compromised due to admixture resulting in spurious associations of SNP with trait phenotypes (Rabinowitz, 1997; Flint-Garcia et al., 2003; Hirschhorn and Daly, 2005).

Several methods have been proposed to address this problem due to population admixture (Kennedy et al., 1992; Spielman et al., 1993; Pritchard et al., 2000; Meuwissen et al., 2002; Price et al., 2006; Yu et al., 2006). The transmission/disequilibrium test (TDT), proposed by Spielman et al. (1993) for case-control studies, concludes the significance of a locus effect

from a Chi-square test on the frequency of the parental alleles transmitted to the affected offspring. Kennedy et al. (1992) proposed to use a mixed-model procedure that includes a random polygenic effect in the model in addition to the fixed effect of the locus under consideration. Following Kennedy et al. (1992), Meuwissen et al. (2002) proposed to fit the effect of the locus as random in the model with a identity-by-descent (IBD) probability matrix to combine the information from cosegregation and linkage disequilibrium. A stepwise procedure, proposed by Pritchard et al. (2000), uses unlinked markers to identify population structure and then analyzes the data accounting for the identified structure. Yu et al. (2006) unified different mixed-model methods by fitting a fixed effect to the SNP under testing, a fixed effect for the identified population structure, and a random effect for background polygenes. In contrast to these methods modeling population structure, Price et al. (2006) adopted principle components analysis to remove the effect of the relatedness from the data by adjusting the trait phenotypes and the SNP genotypes based on the genetic relationship.

The impact of population admixture on GS can be considered by explicitly including breed and heterosis effects in the model, following Hill (1982) and Lo et al. (1995)'s method (see Chapter 3). It has been argued, however, that in Bayesian regression models where all SNP are fitted simultaneously breed differences are explained by the SNP and therefore problems due to population admixture can be ignored in GS. Toosi et al. (2010) has shown that prediction accuracy can be as high as 0.8 when breed composition of admixed populations was ignored in training under additive gene action.

CHAPTER 3. METHODS AND PROCEDURES

3.1 Simulations

Results would be presented by two sets of simulations: a set of preliminary simulations with large dominance effects and a set of more realistic simulations. In the preliminary simulations, the dominance variance and heterosis were chosen to be large enough to clearly detect any advantage of including dominance in the model. Then the simulations with more realistic parameters was followed to verify if the advantages observed in the preliminary simulations would still hold.

3.1.1 Preliminary Simulations

3.1.1.1 Genome

In each simulation, a genome was simulated with 300 QTL randomly distributed on an one Morgan chromosome with 3,000 evenly spaced SNP. All loci were biallelic with starting allele frequency of 0.5 and a reversible mutation rate of 2.5×10^{-5}. A binomial map function was used to model recombination with interference on a chromosome (Karlin, 1984).

The QTL additive effect a is defined as half the difference in genotypic value between alternate homozygotes, and the dominance effect d the deviation of the value of the heterozygote from the mean of the two homozygotes, which is set to zero (Falconer and Mackay, 1996). Bennewitz and Meuwissen (2010) synthesized QTL mapping results from many studies in pigs for meat quality and carcass traits, and concluded that an exponential distribution with rate parameter 5.81 appears to be an adequate "generating mechanism" for the absolute values of the detected QTL additive effects. Following their findings, the same distribution was used here to generate the unsigned value of the additive effect for each QTL, and the sign of the

effect was positive or negative with equal probability. Although the dependency of additive and dominance effects has been studied (Kacser and Burns, 1981; Caballero and Keightley, 1994; Bennewitz and Meuwissen, 2010), a consistent relationship has not been observed. Thus, we assumed, for simplicity, that the dominance effects were independent of the the additive effects. For this reason, the absolute values of dominance effects were independently sampled from an exponential distribution that is identical to that used for the additive effects. This is a reasonable choice because a L-shaped distribution, such as an exponential distribution, that reflects high probability for the occurrence of small effects is also a plausible distribution for dominance effects. In order for the trait to manifest positive heterosis, which was assumed to be favorable for the trait, only 20% of the sampled dominance effects were made negative. The resulting distribution of dominance coefficients, defined as the ratio of dominance effects over the absolute values of additive effects, was similar to the distribution discussed in Bennewitz and Meuwissen (2010).

A base population consisting of 500 founders was randomly mated for 1,000 discrete generations to create LD between loci. In generation 1,001, loci with minor allele frequency less than 0.1 were removed from the panel because fixed loci do not contribute to genetic variability. This procedure resulted in a variable number of loci that stay in the panel across replications of the simulation. For uniformity, in each replication, 100 QTL were randomly selected from the rest to define the trait and 1,000 SNP were randomly selected to stay in the panel.

3.1.1.2 Trait

The relative contribution of the additive and dominance effects to the genetic variability of the trait was assumed to be 2 to 1 through scaling the QTL effects. The scaling procedure (shown in APPENDIX A) did not introduce any relationship between the additive and dominance effects. The correlation between the additive and dominance effects has been checked that it was close to zero. After the scaling, on average, there were about 45% partial dominant and 35% over-dominant QTL affecting the trait. The trait phenotypes were simulated by adding a standard normal deviate to the genotypic value of each animal to account for an half of the phenotypic variance. As a result, the broad sense heritability h^2_{bs} of the trait was 1/2

and the narrow sense heritability h^2_{ns} was 1/3 in generation 1,001.

3.1.1.3 Breed formation

Four breeds, A through D, were simulated by randomly sampling 100 animals from generation 1,001 and random mating for 53 more generations to mimic the breed formation in reality. In generation 1,054, the genetic disparity between breeds was primarily due to the LD phase and allele frequencies specific to the breed. Averaged over simulations, the heterozygosity of a given breed was about 0.3, and the mean difference in allele frequency between breeds was also about 0.3. In each simulation, although the same set of QTL characterized the trait, the contribution of QTL effects to the phenotypic variability differed across breeds due to the disparity in gene frequency. Across simulations, the observed values of variance components in a given breed varied due to genetic drift during the 54 generations of random mating following breed seperation. For example, in breed A the observed average value of h^2_{bs} was 0.46 ± 0.06 and of h^2_{ns} was 0.29 ± 0.06, and the difference between breed A and breed B in h^2_{bs} was 0.03 ± 0.09 and in h^2_{ns} was 0.01 ± 0.09.

3.1.1.4 Crossbreeding program

A two-way crossbreeding program with 20 generations of selection was simulated as described below and depicted in Figure 3.1. In generations 1 through 20 of selection, 100 males and 500 females were selected from 1,000 candidates in each parental breed based on their genomic estimated breeding values (GEBV). The selected animals were then randomly mated within-breed to produce the next generation of purebreds, each with size of 1,000. Meanwhile, the 100 selected males in the sire breed were randomly crossed to the 500 selected females in the dam breed to produce 1,000 crossbred descendents. The goal was to improve the crossbred performance through continual selections in both parental breeds given that the SNP effects for the prediction of GEBV were estimated only once in generation 1 and successively applied to the following 19 generations of selection. Here we used 1,000 progeny of breed A (breed B) produced in generation 1,054 as the sire (dam) breed candidates in the first generation of selection. The estimates of SNP effects were obtained in a crossbred AB population consisting

of 1,000 animals that were produced by randomly crossing the same parents of the purebred candidates in the first generation of selection.

Starting from the same set of purebred selection candidates in generation 1, the following 19 generations of selection were repeated 100 times. Further, in order to account for differences in purebreds due to genetic drift, the above process was repeated with eight different sets of purebred selection candidates in generation 1 of selection that initiated the crossbreeding program. As a result, the cumulative response to selection was calculated from a total of 800 replicates (eight simulations each with 100 replicates). A mixed linear model (shown in APPENDIX A) was used for testing if the cumulative response in generation 20 of selection with one GS model is significantly different from that with another.

3.1.1.5 Genomic prediction and validation

A variety of crossbreeds and admixed populations were made based on the four breeds in generation 1,053, where breed A was always used as the sire breed in a cross (Figure 3.2; Table 3.1). The crossbred A(BC) was produced from a three-way cross, where breed A animals were terminal sires. The crossbred (AB)(CD) was produced from a four-way cross to take the advantages of heterosis from the both paternal (AB) and maternal (CD) crossbred lines. An inter-mating of AB animals produced the F2 crossbred $(AB)^2$. Analogous to the the two-way crossbreeding system, the ultimate goal in a three-way or four-way system is to optimize the terminal crossbred animals for livestock production.

It is common in beef cattle that purebred sires are mated to the crossbred dams or the dams with heterogeneous breed composition to produce commercial animals. Our simulation also considered such situations. The crossbreds A(MIX2) were made from mating breed A sires to dams that were a mixture of breed A, breed B and their crossbreds from various crosses and backcrosses. Similarly, the crossbreds A(MIX4) were made from mating breed A sires to dams that were a mixture of the four breeds and their heterogeneous crossbreds (Table 3.1). Breed A and breed B animals were pooled together with equal proportions to make the admixed population A+B. Based on training in this pooled population, sires in breed A were evaluated for crossbred AB performance. The breed composition in these admixed populations

were assumed known without error.

The crossbred AC is a special training population because it helped to explore the performance of GS in purebred A for crossbred AB performance when the SNP information was from another relevant crossbred AC. Finally, the "training on crossbreds for crossbred performance" scenarios were compared to the conventional scenarios of "training on purebreds for crossbred performance" by using SNP estimates from the next generation of breed A or breed B.

To minimize the contribution of pedigree relationships to the genomic prediction, the performance of GS on the various training populations was validated on breed A animals that were in generation 1,051 but were not direct ancestors of the training populations (Figure 3.2). Thus, the training and validation populations were separated by at least five generations. The prediction accuracy was measured as the correlation between true breeding values (TBV) and GEBV of breed A validation individuals for performance of their crossbred descendents in the target population. The training and validation populations that each had size of 1,000 individuals had been simulated for 24 times starting from different base populations.

3.1.2 More Realistic Simulations

It has been found that dominance can account for up to about 10% of the phenotypic variability in livestock and heterosis from an animal cross can be about 10% (Chapter 2). However, in the preliminary simulations, the dominance variance was assumed 1/6, and the heterosis from crossing breed A to breed B was as high as 39.9% averaging over replications.

Ignoring selection, heterosis can be adjusted by the size of dominance effects and the proportion of beneficially directional dominance in simulation, based on Equation 2.5. Therefore, for a more realistic simulation, the dominance genetic variance was lowered to 10% and the additive genetic variance was raised to 40% to retain the h^2_{bs} as 0.5. The proportion of positively directional dominance effects was decreased from 80 to 75% to reduce the magnitude of heterosis. From 24 simulations, eight that matched realistic parameters were selected to evaluate the response to selection and genomic prediction accuracy as described in the preliminary simulations. In the selected simulations, the average heterosis was 12.2% and the mean of h^2_{bs} was 0.49 ± 0.08 and of h^2_{ns} was 0.39 ± 0.07 in breed A.

3.2 Statistical Models

In the following, we consider models that either explicitly account for breed composition or allow breed differences to be implicitly modeled through the SNP.

3.2.1 Modeling Breed Effects and Heterosis

Hill (1982) and Lo et al. (1995) have shown the theory for modeling the genotypic means in multibreed population under dominance gene action. This theory was used here with minor modifications to model the breed effects and heterosis accounting for heterogeneous breed composition in admixed populations.

The genetic value for an animal i in any training population is the sum of the genotypic values of all QTL defining the trait:

$$G_i = \sum_{t=1}^{m} G_{S_i^t D_i^t} \tag{3.1}$$

where $m = 100$ is the total number of QTL, S_i^t and D_i^t are the alleles at QTL t inherited from the sire and dam. Except in population A+B, any individual in the training populations was a progeny of the sire from breed A or from crossbred AB. Thus, the breed origin for S_i^t allele can only be breed A or B. Given all individuals were diploid, there were in total 8 possible ways to specify the breed origin for alleles S_i^t and D_i^t of animal i:

$$N_1 : S_i^t \in A, D_i^t \in A;$$

$$N_2 : S_i^t \in A, D_i^t \in B;$$

$$N_3 : S_i^t \in A, D_i^t \in C;$$

$$N_4 : S_i^t \in A, D_i^t \in D;$$

$$N_5 : S_i^t \in B, D_i^t \in A;$$

$$N_6 : S_i^t \in B, D_i^t \in B;$$

$$N_7 : S_i^t \in B, D_i^t \in C;$$

$$N_8 : S_i^t \in B, D_i^t \in D.$$

Let N_r with $r \in \{1, ..., 8\}$ denote the event that the QTL alleles have r^{th} class of breed origin and $\Pr(N_r)$ the probability of N_r. Then, the expected value of G_i from Equation 3.1

can be written as:

$$\begin{aligned} \mathrm{E}(G_i) &= \sum_{t=1}^{m} \mathrm{E}(G_{S_i^t D_i^t}) \\ &= \sum_{t=1}^{m} \Big[\sum_{r=1}^{8} \mathrm{E}(G_{S_i^t D_i^t} | N_r) \mathrm{Pr}(N_r) \Big] \\ &= \sum_{r=1}^{8} \mathrm{Pr}(N_r) \Big[\sum_{t=1}^{m} \mathrm{E}(G_{S_i^t D_i^t} | N_r) \Big] \\ &= \sum_{r=1}^{8} \mathrm{Pr}(N_r) \mu_r \end{aligned} \tag{3.2}$$

where μ_r denotes the genetic effect for breed group r. In the absence of imprinting, crossbred AB and BA are assumed to have the same genetic effect, which is not a simply average of the purebred effects due to heterosis. As a result, N_2 has no difference to N_5 in Equation 3.2 where the total number of breed groups can be reduced from eight to seven. The probability of N_r $\mathrm{Pr}(N_r)$ can also be interpreted as the proportions of the alleles originated from the breed group r. The genetic mean, which accounts for the breed composition of the individual, therefore is a function of the effects of breed groups and the proportions of the alleles originated from the corresponding breed groups. Table 3.2 shows the values of $\mathrm{Pr}(N_r)$ in some training populations as an example.

3.2.2 Additive Model

The following mixed linear model was used to estimate SNP effects assuming additive gene action:

$$y_i = \mu + \sum_{j=1}^{k} X_{ij} \alpha_j + e_i \tag{3.3}$$

where y_i is the phenotype of animal i, μ is the overall mean, X_{ij} is the copy number of a given allele of SNP j centered by the mean for this SNP over all individuals, α_j is the allele substitution effect for SNP j, and e_i is the residual effect for animal i. If breed composition was taken into account for animal i from an admixed population, μ would be replaced by the term $\sum_{r=1}^{7} \mathrm{Pr}(N_r)\mu_r$ from Equation 3.2.

Following BayesCπ method proposed by Habier et al. (2011), all parameters in the model were treated as random with informative and uninformative prior distributions. A flat prior

was used for μ or μ_r if breed composition was fitted in the model.

In order to concentrate the signal and reduce noise, only a proportion of SNP are fitted in the model whose effects are assumed to have a scaled multivariate t-distribution. In other words, conditional on σ^2_α, the common variance of random substitution effects for all SNP, α_j has a mixture prior of a normal distribution and a point mass at zero:

$$\alpha_j|\sigma^2_\alpha = \begin{cases} 0 & \text{with probability } \pi \\ \sim N(0, \sigma^2_\alpha) & \text{with probability } 1-\pi \end{cases} \tag{3.4}$$

The proportion π of SNP that have no effects on the trait is considered as an unknown with a uniform prior between 0 and 1:

$$\pi \sim \mathrm{U}(0,1) \tag{3.5}$$

A conjugate scaled inverse Chi-suqare distribution with degrees of freedom $\nu_\alpha = 4$ and scale parameter S^2_α is specified as a prior for σ^2_α:

$$\sigma^2_\alpha \sim S^2_\alpha \chi^{-2}_{\nu_\alpha} \tag{3.6}$$

The prior knowledge of S^2_α is obtained by using the fact that the expectation of a scaled inverse Chi-square variable is a function of S^2_α:

$$\mathrm{E}(\sigma^2) = \frac{S^2\nu}{\nu-2} \tag{3.7}$$

As linkage equilibrium between loci is assumed and given that the allele frequency is independent to the SNP effects, it can be shown (Fernando et al., 2008) that

$$\mathrm{E}(\sigma^2_\alpha) = \frac{V_A}{k(1-\pi_0)\mathrm{E}(2pq)} \tag{3.8}$$

where k is the total number of SNP, π_0 is the prior probability that a SNP has no effect, $p = 1-q$ is the allele frequency, so E(2pq) is the average heterozygosity, and V_A is the additive genetic variance for the trait explained by the SNP.

The residual e_i has a normal prior with the variance also from a scaled inverse Chi-square:

$$e_i \sim N(0, \sigma^2_e), \tag{3.9}$$

$$\sigma^2_e \sim S^2_e \chi^{-2}_{\nu_e} \tag{3.10}$$

where $\nu_e = 4$. The value of S_e^2 is obtained from Equation 3.7 with $\mathrm{E}(\sigma_e^2) = V_E$ where V_E is the residual variance that cannot be explained by the SNP.

The estimates of V_A and V_E can be obtained from an animal model using restricted maximum likelihood estimation (REML).

3.2.3 Dominance Model

The dominance model, as shown below, simultaneously fits the additive and dominance effects of SNP in the model.

$$y_i = \mu + \sum_{j=1}^{k}(X_{ij}a_j + W_{ij}d_j) + e_i \tag{3.11}$$

where y_i, μ, X_{ij} is as defined in the additive model, W_{ij} is the indicator variable for the heterozygous genotype of SNP j that is centered by the mean, a_j is the additive effect and d_j the dominance effect for SNP j, and e_j is the residual. In theory, the residual term in the dominance model only contains non-genetic effects, while that in the additive model also includes dominance deviations to the genotypic values. Thus, the additive model can be viewed as a reduced model of the dominance model.

The Bayesian hierarchical modeling used in the dominance model is comparable to that in the additive model. Conditional on π_a (the probability that a_j is zero) and σ_a^2 (the variance of a_j when it is nonzero), the prior for a_j is a mixture of normals as given in the additive model (Equation 3.4). Similarly, the prior for d_j is a mixture of normals given π_d and σ_d^2 with corresponding definitions. What differs from the prior specification for a_j is, in order to account for the directionality of dominance, the normal component of the prior for d_j has an unknown mean:

$$d_j|\mu_d, \sigma_d^2 = \begin{cases} 0 & \text{with probability } \pi_d \\ \sim N(\mu_d, \sigma_d^2) & \text{with probability } 1 - \pi_d \end{cases} \tag{3.12}$$

For convenience, the mean of the normal is assumed dependent to the variance. That is, the prior for μ_d is conditional on the sampled value of σ_d^2. The joint density of μ_d and σ_d^2 can then be written as:

$$f(\mu_d, \sigma_d^2) = f(\mu_d|\sigma_d^2)f(\sigma_d^2) \tag{3.13}$$

The prior for μ_d is a normal and therefore has the form of

$$\mu_d|\sigma_d^2 \sim N(\gamma, \sigma_d^2/k_0) \tag{3.14}$$

where γ is our prior belief about μ_d and k_0 is the "prior sample size", which expresses the strength of the prior belief in terms of the variations from the "data". The value of γ can be obtained by the information from a cross. Assuming independency between dominance effects and allele frequencies and ignoring selection, it can be shown from Equation 2.5 that the mean of dominance effects is a function of heterosis in crossbreds (H_{F_1}) and the disparity of allele frequencies in parental populations (y):

$$\begin{aligned} H_{F_1} &= k_d \mathrm{E}(dy^2) \\ &= k_d \mathrm{E}(d)\mathrm{E}(y^2) \end{aligned} \tag{3.15}$$

where k_d is the number of loci assumed nonzero dominance effects. Rearranging Equation 3.15 gives

$$\mathrm{E}(d) = \frac{H_{F_1}}{k(1-\pi_{d,0})\mathrm{E}(y^2)} \tag{3.16}$$

where $\pi_{d,0}$ is the prior proportion of loci that have nonzero dominance effects. Under the assumption that each QTL is associated with at least one SNP, $\pi_{d,0}$ should be at most 0.9 as 100 QTL versus 1,000 SNP were simulated. The calculation in Equation 3.11 gives the value of $\gamma = \mathrm{E}(d)$. The value of k_0 is set to 10 allowing the data to "dominate" the posterior of μ_d.

The variance components σ_a^2 and σ_d^2 are assumed to have independent scaled inverse Chi-square distributions. As shown in Equation 3.7, the specifications of the hyper parameters S_a^2 and S_d^2 require the knowledge of $\mathrm{E}(\sigma_a^2)$ and $\mathrm{E}(\sigma_d^2)$. In the additive model, the relationship between $\mathrm{E}(\sigma_\alpha^2)$ between V_A has been shown from Equation 3.8. The following describes how $\mathrm{E}(\sigma_a^2)$ or $\mathrm{E}(\sigma_d^2)$ is connected to V_A or V_D, where V_D is the dominance genetic variance.

Given independence between loci holds, Falconer and Mackay (1996) have shown that

$$V_D = \sum_{j=1}^{k} (2p_j q_j d_j)^2 \tag{3.17}$$

Assuming, again, independence between the effects and the allele frequencies, Equation 3.17

can be written as

$$\begin{aligned} V_D &= k(1-\pi_0)\mathrm{E}[(2pqd)^2] \\ &= k(1-\pi_0)\mathrm{E}[(2pq)^2]\mathrm{E}(d^2) \\ &= k(1-\pi_0)\mathrm{E}[(2pq)^2]\{(1+1/k_0)\sigma_d^2 + [\mathrm{E}(d)]^2\} \end{aligned} \tag{3.18}$$

Rearranging this and using $\gamma = \mathrm{E}(d)$ in Equation 3.18, we have

$$\sigma_d^2 = (\frac{V_D}{k(1-\pi_0)\mathrm{E}[(2pq)^2]} - \gamma^2)/(1+1/k_0) \tag{3.19}$$

Falconer and Mackay (1996) have also shown that

$$V_A = \sum_{j=1}^{k}(2p_jq_j\alpha_j^2) \tag{3.20}$$

Under the same assumptions as made in Equation 3.17 in addition to the assumption that additive effects have a mean zero and are independent to the dominance effects, Equation 3.20 becomes

$$\begin{aligned} V_A &= k(1-\pi_0)\mathrm{E}(2pq\alpha^2) \\ &= k(1-\pi_0)\mathrm{E}(2pq)\mathrm{E}(\alpha^2) \end{aligned} \tag{3.21}$$

where

$$\begin{aligned} \mathrm{E}(\alpha^2) &= \mathrm{E}\{[a+(1-2p)d]^2\} \\ &= \mathrm{E}(a^2) + 2\mathrm{E}[a(1-2p)d] + \mathrm{E}\{[(1-2p)d]^2\} \\ &= \mathrm{E}(a^2) + \mathrm{E}[(1-2p)^2]\mathrm{E}(d^2) \\ &= \sigma_a^2 + \mathrm{E}[(q-p)^2](\sigma_d^2+\gamma^2) \end{aligned} \tag{3.22}$$

Substituting it in 3.21 and turning the equation around, then,

$$\sigma_a^2 = \frac{V_A}{k(1-\pi_0)\mathrm{E}(2pq)} - \mathrm{E}[(1-2p)^2](\sigma_d^2+\gamma^2) \tag{3.23}$$

3.2.4 Breed-specific SNP Allele Model

As shown in Ibanez-Escriche et al. (2009) (with slightly different notation), the breed-specific SNP allele model (BSAM) fits SNP allele states in the model as below:

$$y_i = \mu + \sum_{j=1}^{k}(A_{ij}^r\alpha_j^r + A_{ij}^{r'}\alpha_j^{r'}) + e_i \tag{3.24}$$

where A_{ij}^{r} or $A_{ij}^{r'}$ with value (0, 1) is the SNP allele at locus j of breed origin r or r' that animal i received from its sire or dam, α_j^r or $\alpha_j^{r'}$ is the breed-specific substitution effect for allele A_{ij}^{r} or $A_{ij}^{r'}$. The other parameters are defined as in the additive or dominance model. In BSAM, the SNP allele effects have breed-specific variance $\sigma^2_{\alpha^r}$ and $\sigma^2_{\alpha^{r'}}$, and breed-specific π parameter π_{α^r} and $\pi_{\alpha^{r'}}$. However, the same prior used in the additive model is used for $\sigma^2_{\alpha^r}$ and $\sigma^2_{\alpha^{r'}}$.

The parental origin of alleles were known without error in the analysis. For the training individuals whose sires are all from one breed and dams all from another, such as crossbred AB and AC, knowing the parental origin of alleles is equivalent to knowing the breed origin. When the parents of the training individuals are from crossbreds or admixed populations, the breed origin of the alleles were unknown.

It can be argued, however, that because the estimated allele effects were only validated in breed A, which is the sire breed, therefore only breed-specific effects for the sire were under concern. Given sires are all homogeneous purebreds, the heterogeneity from dams can be ignored. We argue that fitting a common allele effect for the heterogeneous dams is adequate to help the model explore the allele effect specific to the sire breed. Thus, BSAM would be also applicable for crossbred A(BC) and admixed population A(MIX2) and A(MIX4).

Nevertheless, when the sires themselves are crossbreds or in admixture, BSAM would not work with breed origin unknown. Thus, BSAM was not applied to the training population (AB)(CD), $(AB)^2$ and A+B in this study.

3.3 Inference for Model Parameters

Markov Chain Monte Carlo (MCMC) sampling was used for the parameter inference. In particular, Gibbs sampling was used to sample parameters in turn each from their full conditional distributions. The full conditional distribution for each parameter that is fitted in the model is derived next. Since the implementation of Gibbs sampler in the additive model has been well described by Habier et al. (2011), here we focus on the algorithm for the dominance model (Equation 3.11).

3.3.1 Full Conditionals for Model Parameters

Assuming normality and homogeneous residual variances, the full conditional density of μ with a flat prior can be written in a matrix form as

$$\begin{aligned} f(\mu|\boldsymbol{y},\boldsymbol{a},\boldsymbol{d},\sigma_e^2) &\propto f(\boldsymbol{y}|\mu,\boldsymbol{a},\boldsymbol{d},\sigma_e^2)f(\mu) \\ &\propto (\sigma_e^2)^{-\frac{n}{2}}\exp\Big\{-\frac{[\mathbf{1}\mu-(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{a}-\boldsymbol{W}\boldsymbol{d})]'[\mathbf{1}\mu-(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{a}-\boldsymbol{W}\boldsymbol{d})]}{2\sigma_e^2}\Big\} \\ &\propto \exp\Big\{-\frac{n}{2\sigma_e^2}\Big[\mu-\frac{\mathbf{1}'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{a}-\boldsymbol{W}\boldsymbol{d})}{n}\Big]^2\Big\} \end{aligned} \tag{3.25}$$

Thus, the full conditional density of μ is a normal:

$$\mu|\boldsymbol{y},\boldsymbol{a},\boldsymbol{d},\sigma_e^2 \sim N\Big(\frac{\mathbf{1}'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{a}-\boldsymbol{W}\boldsymbol{d})}{n},\frac{\sigma_e^2}{n}\Big) \tag{3.26}$$

Given the conjugate prior (Equation 3.10), the full conditional of σ_e^2 is also a scaled inverse Chi-square,

$$\begin{aligned} f(\sigma_e^2|\boldsymbol{y},\mu,\boldsymbol{a},\boldsymbol{d}) &\propto f(\boldsymbol{y}|\mu,\boldsymbol{a},\boldsymbol{d},\sigma_e^2)f(\sigma_e^2) \\ &\propto (\sigma_e^2)^{-\frac{2+n+\nu_e}{2}}\exp\{-\frac{\boldsymbol{e}'\boldsymbol{e}+\nu_e S_e}{2\sigma_e^2}\} \end{aligned} \tag{3.27}$$

Thus, we have

$$\sigma_e^2|\boldsymbol{y},\mu,\boldsymbol{a},\boldsymbol{d} \sim \tilde{S}_e\chi_{\tilde{\nu}_e}^{-2} \tag{3.28}$$

where $\tilde{\nu}_e = n+\nu_e$ and $\tilde{S}_e = \frac{\boldsymbol{e}'\boldsymbol{e}+\nu_e S_e}{\tilde{\nu}_e}$ are the degrees of freedom and scale of the full conditional posterior.

Let β_j denote either a_j or d_j. The mixture prior for β_j allows it to be included in the model or be left out. Let $\delta_{j,\beta}$ denote a model inclusion indicator variable defined as,

$$\delta_{j,\beta} = \begin{cases} 1 & \text{then } \beta_j \sim \text{Normal} \\ 0 & \text{then } \beta_j = 0 \end{cases} \tag{3.29}$$

with a prior probability $1-\pi_\beta$ that $\delta_{j,\beta}=1$. The posterior probability that $\delta_{j,\beta}=1$ is then calculated by

$$\begin{aligned} Pr(\delta_{j,\beta}=1|\boldsymbol{y},\boldsymbol{\theta}_{else}) &= \frac{f(\boldsymbol{y}|\delta_{j,\beta}=1,\boldsymbol{\theta}_{else})Pr(\delta_{j,\beta}=1)}{\sum_{\delta_{j,\beta}} f(\boldsymbol{y}|\delta_{j,\beta},\boldsymbol{\theta}_{else})Pr(\delta_{j,\beta})} \\ &= \frac{f(\boldsymbol{y}|\delta_{j,\beta}=1,\boldsymbol{\theta}_{else})(1-\pi)}{f(\boldsymbol{y}|\delta_{j,\beta}=0,\boldsymbol{\theta}_{else})\pi + f(\boldsymbol{y}|\delta_{j,\beta}=1,\boldsymbol{\theta}_{else})(1-\pi)} \end{aligned} \tag{3.30}$$

where $\boldsymbol{\theta}_{else}$ denotes other parameters besides δ_j.

We sample $\delta_{j,\beta}$ and β_j jointly by first sampling $\delta_{j,\beta}$ from its marginal distribution and then sampling β_j conditional on $\delta_{j,\beta}$. Thus, β_j is integrated out from the likelihood in Equation 3.30. For $\delta_{j,d}$, we have that

$$\begin{aligned}
f(\boldsymbol{y}|\delta_{j,d}=1,\boldsymbol{\theta}_{else}) &= \int f(\boldsymbol{y}|\delta_{j,d}=1,\boldsymbol{\theta}_{else}) f(d_j)\,\mathrm{d}d_j \\
&= \int (2\pi)^{-\frac{n}{2}}(\sigma_e^2)^{-\frac{n}{2}}\exp\{-\frac{(\boldsymbol{v}-\boldsymbol{W}_j d_j)'(\boldsymbol{v}-\boldsymbol{W}_j d_j)}{2\sigma_e^2}\}(\sigma_d^2)^{-\frac{1}{2}}\exp\{-\frac{(d_j-\mu_d)^2}{2\sigma_d^2}\}\,\mathrm{d}d_j \\
&= (2\pi)^{-\frac{n}{2}}(\sigma_e^2)^{-\frac{n}{2}}(\sigma_d^2)^{-\frac{1}{2}}\int \exp\{-\frac{C_{j,d}[d_j-(\hat{d}_j+\lambda C_{j,d}^{-1}\mu_d)]^2}{2\sigma_e^2}\} \\
&\quad\cdot\exp\{-\frac{\boldsymbol{v}'\boldsymbol{v}+\lambda\mu_d^2-C_{j,d}(\hat{d}_j+\lambda C_{j,d}^{-1}\mu_d)^2}{2\sigma_e^2}\}\,\mathrm{d}d_j \\
&= (2\pi)^{-\frac{n}{2}}(\sigma_e^2)^{-\frac{n}{2}}(\sigma_d^2)^{-\frac{1}{2}}(\frac{C_{j,d}}{\sigma_e^2})^{-\frac{1}{2}}\exp\{-\frac{\boldsymbol{v}'\boldsymbol{v}+\lambda\mu_d^2-C_{j,d}(\hat{d}_j+\lambda C_{j,d}^{-1}\mu_d)^2}{2\sigma_e^2}\}
\end{aligned} \tag{3.31}$$

The log likelihood is then

$$\begin{aligned}
\log f(\boldsymbol{y}|\delta_{j,d}=1,\boldsymbol{\theta}_{else}) = -\frac{1}{2}\{&n\log 2\pi + n\log\sigma_e^2 + \boldsymbol{v}'\sigma_e^{-2}\boldsymbol{v} \\
&+\log\sigma_d^2+\log C_{j,d}\sigma_e^{-2}+\mu_d^2\sigma_d^{-2}-C_{j,d}\sigma_e^{-2}(\hat{d}_j+\lambda C_{j,d}^{-2}\mu_d)^2\}
\end{aligned} \tag{3.32}$$

Similarly, for $\delta_{j,a}$, we have that

$$\begin{aligned}
\log f(\boldsymbol{y}|\delta_{j,a}=1,\boldsymbol{\theta}_{else}) = -\frac{1}{2}\{&n\log 2\pi + n\log\sigma_e^2 + \boldsymbol{v'}\sigma_e^{-2}\boldsymbol{v} \\
&+\log\sigma_a^2+\log C_{j,a}\sigma_e^{-2}-C_{j,a}\sigma_e^{-2}(\hat{a}_j)^2\}
\end{aligned} \tag{3.33}$$

For $\delta_{j,\theta}=0$, the log likelihood is just

$$\log f(\boldsymbol{y}|\delta_{j,\beta}=0,\boldsymbol{\theta}_{else}) = -\frac{1}{2}\{n\log 2\pi + n\log\sigma_e^2 + \boldsymbol{v'}\sigma_e^{-2}\boldsymbol{v}\} \tag{3.34}$$

Given $\boldsymbol{\delta_\beta}$, the full conditional for π_β that has a uniform prior is a Beta distribution.

$$\pi_\beta \sim Beta(k-l+1, l+1) \tag{3.35}$$

where $l = \boldsymbol{\delta'_\beta}\boldsymbol{\delta_\beta}$ indicates the observed number of loci that have nonzero θ.

Given that $\delta_{j,a} = 1$, a_j has a normal prior centered at zero as in Equation 3.4. Otherwise it is zero. Let

$$\begin{aligned} \boldsymbol{u} &= \boldsymbol{y} - \mathbf{1}\mu - \boldsymbol{W}\boldsymbol{d} - \sum_{j' \neq j} \boldsymbol{X}_{j'} a_{j'} \\ &= \boldsymbol{X}_j a_j + \boldsymbol{e} \end{aligned} \tag{3.36}$$

and $\boldsymbol{a}_{-j}$ denote the other additive effects besides that for SNP j, then,

$$\begin{aligned} f(a_j|\boldsymbol{y}, \mu, \boldsymbol{a}_{-j}, \boldsymbol{d}, \sigma_a^2, \sigma_e^2) &\propto f(\boldsymbol{y}|\mu, \boldsymbol{a}_{-j}, \boldsymbol{d}, \sigma_e^2) f(a_j|\sigma_a^2) \\ &\propto (\sigma_e^2)^{-\frac{n}{2}} \exp\Big\{-\frac{(\boldsymbol{u} - \boldsymbol{X}_j a_j)'(\boldsymbol{u} - \boldsymbol{X}_j a_j)}{2\sigma_e^2}\Big\} (\sigma_a^2)^{-\frac{1}{2}} \exp\Big\{-\frac{a_j^2}{2\sigma_a^2}\Big\} \\ &\propto \exp\Big\{-\frac{1}{2\sigma_e^2}[\boldsymbol{u}\boldsymbol{u} - 2\boldsymbol{X}_j'\boldsymbol{u} a_j + (\boldsymbol{X}_j'\boldsymbol{X}_j + \frac{\sigma_e^2}{\sigma_a^2}) a_j^2]\Big\} \\ &\propto \exp\Big\{-\frac{1}{2\sigma_e^2}[C_{j,a}(a_j - \hat{a}_j)^2 + \boldsymbol{u'u} - C_{j,a}\hat{a}_j^2]\Big\} \\ &\propto \exp\Big\{-\frac{(a_j - \hat{a}_j)^2}{2\sigma_e^2 C_{j,a}^{-1}}\Big\} \end{aligned} \tag{3.37}$$

where $C_{j,a} = \boldsymbol{X}_j'\boldsymbol{X}_j + \frac{\sigma_e^2}{\sigma_a^2}$ and $\hat{a}_j = \frac{\boldsymbol{X}_j\boldsymbol{u}}{C_{j,a}}$ are, respectively, the coefficient of the mixed-model equation and the BLUP estimate for a_j. Thus, the full conditional distribution for a_j is also a normal:

$$a_j|\boldsymbol{y}, \mu, \boldsymbol{a}_{-j}, \boldsymbol{d}, \sigma_a^2, \sigma_e^2 \sim N(\hat{a}_j, \frac{\sigma_e^2}{C_{j,a}}) \tag{3.38}$$

The derivation for the dominance effect for SNP j is similar to that for a_j. Given $\delta_{j,d} = 1$, d_j, however, has a normal with a nonnull mean. Let

$$\begin{aligned} \boldsymbol{v} &= \boldsymbol{y} - \mathbf{1}\mu - \boldsymbol{X}\boldsymbol{a} - \sum_{j' \neq j} \boldsymbol{W}_{j'} d_{j'} \\ &= \boldsymbol{W}_j d_j + \boldsymbol{e} \end{aligned} \tag{3.39}$$

and $\boldsymbol{d}_{-j}$ denote the other dominance effects besides that for SNP j. Then,

$$\begin{aligned} f(d_j|\boldsymbol{y}, \mu, \boldsymbol{d}_{-j}, \boldsymbol{a}, \mu_d, \sigma_d^2, \sigma_e^2) &\propto f(\boldsymbol{y}|\mu, \boldsymbol{d}_{-j}, \boldsymbol{a}, \sigma_e^2) f(d_j|\mu_d, \sigma_d^2) \\ &\propto (\sigma_e^2)^{-\frac{n}{2}} \exp\{-\frac{(\boldsymbol{v} - \boldsymbol{W}_j d_j)'(\boldsymbol{v} - \boldsymbol{W}_j d_j)}{2\sigma_e^2}\} (\sigma_d^2)^{-\frac{1}{2}} \exp\{-\frac{(d_j - \mu_d)^2}{2\sigma_d^2}\} \\ &\propto \exp\{-\frac{\Delta}{2\sigma_e^2}\} \end{aligned} \tag{3.40}$$

where

$$\begin{aligned}
\Delta &= \boldsymbol{v'v} - 2\boldsymbol{W}'_j\boldsymbol{v}d_j + \boldsymbol{W}'_j\boldsymbol{W}_j d_j^2 + \lambda(d_j^2 - 2\mu_d d_j + \mu_d^2) \\
&= (\boldsymbol{W}'_j\boldsymbol{W}_j + \lambda)d_j^2 - 2(\boldsymbol{W}'_j\boldsymbol{v} + \lambda\mu_d)d_j + \boldsymbol{v'v} + \lambda\mu_d^2 \\
&\propto C_{j,d}d_j^2 - 2(C_{j,d}\hat{d}_j + \lambda\mu_d)d_j \\
&= C_{j,d}\Big\{d_j^2 - 2\Big(\hat{d}_j + \frac{\lambda}{C_{j,d}}\mu_d\Big)d_j + \Big(\hat{d}_j + \frac{\lambda}{C_{j,d}}\mu_d\Big)^2 - \Big(\hat{d}_j + \frac{\lambda}{C_{j,d}}\mu_d\Big)^2\Big\} \\
&= C_{j,d}\Big\{d_j - \Big(\hat{d}_j + \frac{\lambda}{C_{j,d}}\mu_d\Big)\Big\}^2 - C_{j,d}\Big(\hat{d}_j + \frac{\lambda}{C_{j,d}}\mu_d\Big) \\
&\propto C_{j,d}\Big\{d_j - \Big(\hat{d}_j + \frac{\lambda}{C_{j,d}}\mu_d\Big)\Big\}^2
\end{aligned} \tag{3.41}$$

where $\lambda = \frac{\sigma_e^2}{\sigma_d^2}$, $C_{j,d} = \boldsymbol{W}'_j\boldsymbol{W}_j + \lambda$ and $\hat{d}_j = \frac{\boldsymbol{W}_j\boldsymbol{v}}{C_{j,d}}$. Thus,

$$d_j|\boldsymbol{y}, \mu, \boldsymbol{d}_{-j}, \boldsymbol{a}, \mu_d, \sigma_d^2, \sigma_e^2 \sim N(\hat{d}_j + \frac{\lambda}{C_{j,d}}\mu_d, \frac{\sigma_e^2}{C_{j,d}}) \tag{3.42}$$

The variance variable for the additive or dominance effect is sampled from the full conditional that does not depend on the likelihood of the model.

$$\begin{aligned}
f(\sigma_a^2|\boldsymbol{y}, \mu, \boldsymbol{a}, \boldsymbol{d}, \sigma_e^2, S_a) &\propto f(\boldsymbol{y}|\mu, \boldsymbol{a}, \boldsymbol{d}, \sigma_e^2)f(\boldsymbol{a}|\sigma_a^2)f(\sigma_a^2|S_a) \\
&\propto f(\boldsymbol{a}|\sigma_a^2)f(\sigma_a^2|S_a) \\
&\propto (\sigma_a^2)^{-\frac{k}{2}}\exp\{-\frac{\boldsymbol{a'a}}{2\sigma_a^2}\} \cdot (\sigma_a^2)^{-\frac{\nu_a+2}{2}}\exp\{-\frac{\nu_a S_a}{2\sigma_a^2}\} \\
&\propto (\sigma_a^2)^{-\frac{k+\nu_e+2}{2}}\exp\{-\frac{\boldsymbol{a'a} + \nu_a S_a}{2\sigma_a^2}\}
\end{aligned} \tag{3.43}$$

Due to the conjugacy, the full conditional for σ_a^2 is also a scaled inverse Chi-square,

$$\sigma_a^2|\boldsymbol{y}, \mu, \boldsymbol{a}, \boldsymbol{d}, \sigma_e^2, S_a \sim \tilde{S}_a\chi_{\tilde{\nu}_a}^{-2} \tag{3.44}$$

where $\tilde{\nu}_a = k + \nu_a$ and $\tilde{S}_a = \frac{\boldsymbol{a'a} + \nu_a S_a}{\tilde{\nu}_a}$.

Similarly, the full conditional for σ_d^2 is,

$$\sigma_d^2|\boldsymbol{y}, \mu, \boldsymbol{a}, \boldsymbol{d}, \mu_d, \sigma_e^2, S_d \sim \tilde{S}_d\chi_{\tilde{\nu}_d}^{-2} \tag{3.45}$$

where $\tilde{\nu}_d = k + \nu_d$ and $\tilde{S}_d = \frac{(\boldsymbol{d} - \mathbf{1}\mu_d)'(\boldsymbol{d} - \mathbf{1}\mu_d) + \nu_d S_d}{\tilde{\nu}_d}$ given the sampled value of μ_d.

Due to the interdependency between μ_d and σ_d^2, the full conditional for μ_d depends on the sample of σ_d^2:

$$\begin{aligned} f(\mu_d|\boldsymbol{y},\mu,\boldsymbol{a},\boldsymbol{d},\sigma_d^2,\sigma_e^2,\gamma) &\propto f(\boldsymbol{d}|\mu_d,\sigma_d^2)f(\mu_d|\gamma,\sigma_d^2) \\ &\propto \exp\{-\frac{(\boldsymbol{d}-\mathbf{1}\mu_d)'(\boldsymbol{d}-\mathbf{1}\mu_d)}{2\sigma_d^2}\}\exp\{-\frac{(\mu_d-\gamma)^2}{2(\sigma_d^2/k_0)}\} \\ &\propto \exp\{-\frac{(\mu_d-\frac{\mathbf{1}'\boldsymbol{d}+k_0\gamma}{k+k_0})^2}{2\sigma_d^2/(k+k_0)}\} \end{aligned} \tag{3.46}$$

The full conditional distribution of μ_d given the sampled σ_d^2 is then

$$f(\mu_d|\boldsymbol{y},\mu,\boldsymbol{a},\boldsymbol{d},\sigma_d^2,\sigma_e^2,\gamma) \sim N\Big(\frac{\mathbf{1}'\boldsymbol{d}+k_0\gamma}{k+k_0},\frac{\sigma_d^2}{k+k_0}\Big) \tag{3.47}$$

3.3.2 Implementation and Software

The analyses described here were implemented by modifing *GenSel* (Fernando and Garrick, 2009) to allow dominance and allele specific effects. The Markov chain used for inference consisted of 11,000 samples with the first 1,000 discarded as a burn-in. A longer chain did not affect the prediction accuracy. Parameters were estimated from the mean of the resulting 10,000 posterior samples.

3.4 True and Genomic Estimated Breeding Values

For animal i from breed r, the true breeding value is given by,

$$TBV_i^r = \sum_{t=1}^{m} T_{it}\alpha_t^r \tag{3.48}$$

where T_{it} is the genotype and α_t^r is the true allele substitution effect for QTL t, and the genomic estimated breeding value is given by,

$$GEBV_i^r = \sum_{j=1}^{k} Z_{ij}\hat{\alpha}_j^r \tag{3.49}$$

where Z_{ij} is the genotype and $\hat{\alpha}_j^r$ is the estimated allele substitution effect for SNP j. The definition of α_t^r for a purebred animal with a breeding goal of maximizing the performance of the crossbred descendents is described next.

Suppose L_1 and L_2 are two alleles at locus L. Let p^S and q^S denote the frequencies of L_1 and L_2 in the sire breed and p^D and q^D denote the frequencies of L_1 and L_2 in the dam breed. The genotypic values (G) of genotypes L_1L_1, L_1L_2 and L_2L_2 are a, d and $-a$, respectively. The average effect of an L_1 allele from the sire is defined as the expected genotypic value of a crossbred offspring that received L_1 from the sire minus the crossbred population mean. Let S denote the allele that animal i inherited from its sire. From Table 3.3,

$$\begin{aligned} \alpha_1^S &= \mathrm{E}(G|S = L_1) - \mu \\ &= p^D a + q^D d - \mu \end{aligned} \tag{3.50}$$

Similarly, the average effect of L_2 allele from the sire is,

$$\begin{aligned} \alpha_2^S &= \mathrm{E}(G|S = L_2) - \mu \\ &= -q^D a + p^D d - \mu \end{aligned} \tag{3.51}$$

The difference between these average effects is the substitution effect for the sire:

$$\begin{aligned} \alpha^S &= \alpha_1^S - \alpha_2^S \\ &= a + (1 - 2p^D)d \end{aligned} \tag{3.52}$$

Similarly, the substitution effect for the dam is,

$$\alpha^D = a + (1 - 2p^S)d \tag{3.53}$$

As a result, the allele substitution effects for a purebred parent used for crossbreeding are breed-specific and defined in terms of the allele frequencies in the breed of the other parent.

In summary, for a purebred r, α_t^r in Equation 3.48 is defined as

$$\alpha_t^r = a_t + (1 - 2p_t^{r'})d_t \tag{3.54}$$

where r' is the breed of the other parent of the crossbreds. In BSAM, α_j^r is directly estimated for the prediction of $GEBV_i^r$ in Equation 3.49, while it is indirectly estimated from the dominance model by combining the estimates of a_j and d_j with the current value of $p_j^{r'}$ from breed r' in Equation 3.54. The additive model does not estimate α_j^r at all. In stead, it estimates a common α_j for SNP j alleles, which is not specific to any breed.

Table 3.1 The constitutions of training populations for genomic prediction and validation and the corresponding target crossbreds to be improved

Training Population	**Constitution**	**Target Crossbreds**
AB	AB	AB
A(BC)	A(BC)	A(BC)
(AB)(CD)	(AB)(CD)	(AB)(CD)
$(AB)^2$	(AB)(AB)	$(AB)^2$
A(MIX2)	A, AB, A(AB), A(A(AB)), A((AB)B)	A(MIX2)
A(MIX4)	A, AB, AC, AD, A(AB), A(CD), A(A(AB)), A((AB)B), A(C(CD)), A((CD)D), A(A(BC)) , A((AB)(AB)), A((CD)(CD)), A((AB)(CD))	A(MIX4)
A+B	A, B	AB
AC	AC	AB
A	A	AB
B	B	AB

AB is the cross of breed A and B; A(BC) is the three-way crossbreds; (AB)(CD) is the four-way crossbreds; $(AB)^2$ is the F2 crossbreds; MIX2 denotes the admixture of breed A, B and their heterogeneous crossbreds; MIX4 denotes the admixture of four breeds and their heterogeneous crossbreds; A(MIX2) and A(MIX4) are corresponding admixed populations; A+B is the combined population of breed A and B; AC is the cross of breed A and C; A or B is the purebred A or B.

Table 3.2 The proportions of the alleles ($\Pr(N_r)$) originated from different breed groups (N_r)

	N_1	$N_{2,5}$	N_3	N_4	N_6	N_7	N_8
Population	**A**	**AB**	**AC**	**AD**	**B**	**BC**	**BD**
A	1	0	0	0	0	0	0
B	0	0	0	0	1	0	0
AB	0	1	0	0	0	0	0
AC	0	0	1	0	0	0	0
A(BC)	0	0.5	0.5	0	0	0	0
(AB)(CD)	0	0	0.25	0.25	0	0.25	0.25
$(AB)^2$	0.25	0.5	0	0	0.25	0	0
A+B	1	0	0	0	1	0	0

A or B is the purebred A or B; AB or AC is the cross of breed A and B or breed A and C; A(BC) is the three-way crossbreds; (AB)(CD) is the four-way crossbreds; $(AB)^2$ is the F2 crossbreds; A+B is the combined population of breed A and B.

Table 3.3 Allele frequencies and genotypic values for locus L

		Alleles (allele frequency) from the dam breed		
		L_1 (p^D)	L_2 (q^D)	E($G|S$)
Alleles (allele frequency) from the sire breed	L_1 (p^S)	a	d	$p^D a + q^D d$
	L_2 (q^S)	d	-a	$p^D d - q^D a$
	E($G|D$)	$p^S a + q^S d$	$p^S d - q^S a$	

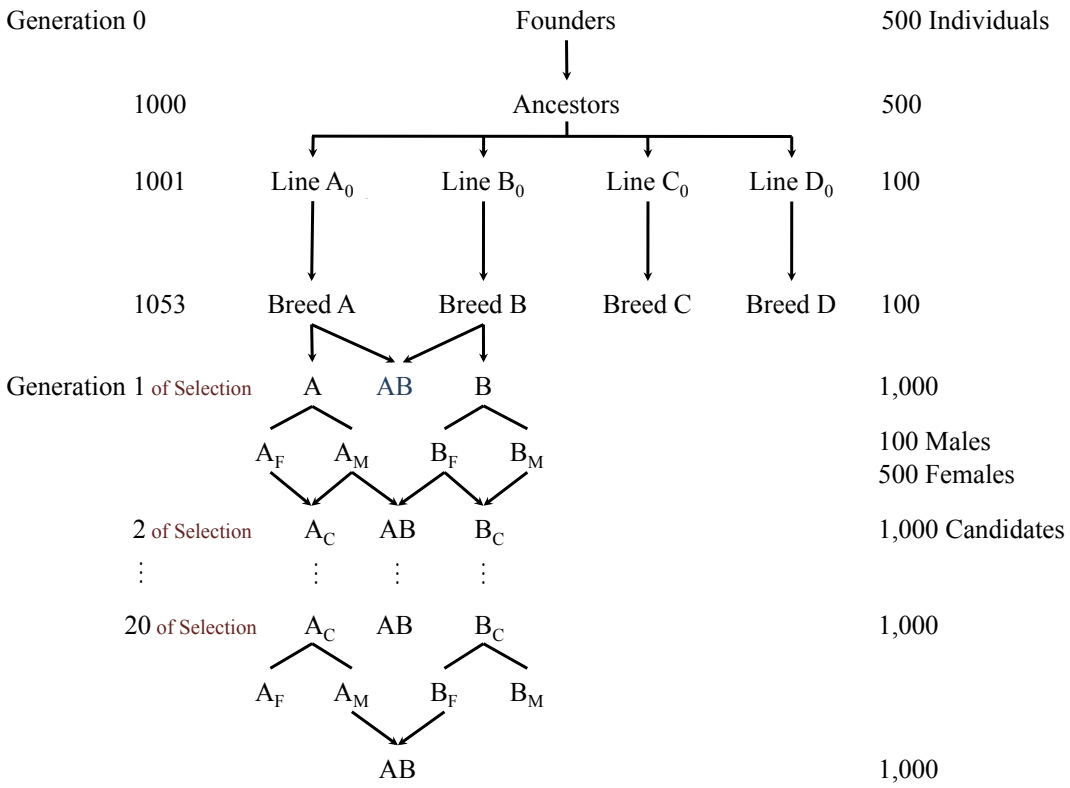

Figure 3.1 Schematic representation of the simulated population history and the two-way crossbreeding program that consisted of 20 generations of purebred selection for crossbred performance. Crossbred AB in blue is the training population; A_M or B_M is the selected breed A or B males; A_F or B_F is the selected breed A or B females; A_C or B_C is the breed A or B selection candidates. Lines without arrows connecting Y and X represent selecting X from Y; lines with an arrow pointing from Y to X represent reproducing X from Y.

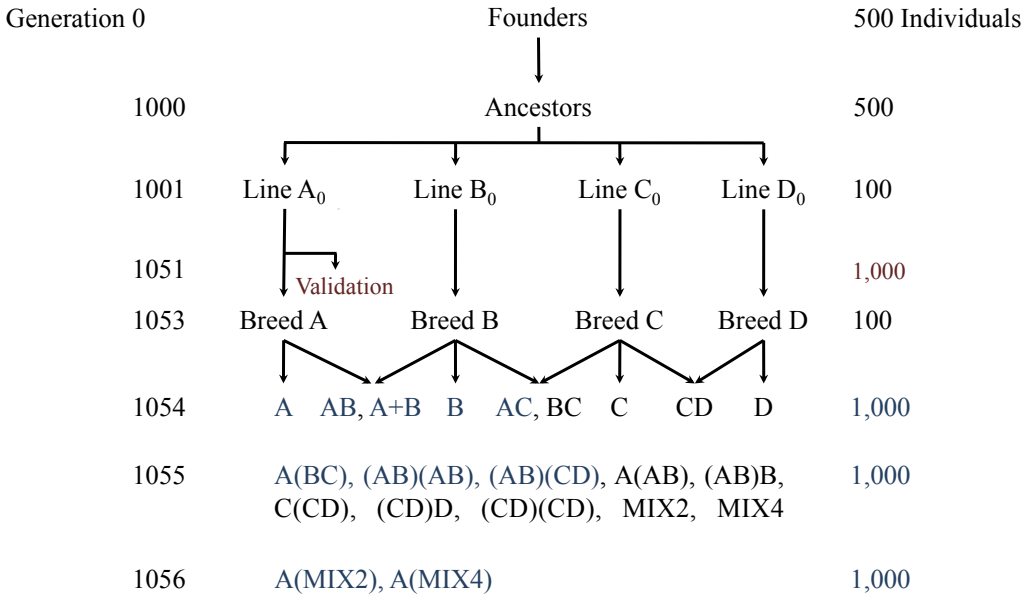

Figure 3.2 Schematic representation of the simulated population history and the different types of crossbred and admixed populations that were simulated for training (blue) and validation (red). A to D is the purebred A to D. AB is the cross of breed A and B; A(BC) is the three-way crossbreds; (AB)(CD) is the four-way crossbreds; $(AB)^2$ is the F2 crossbreds; MIX2 denotes the admixture of breed A, B and their heterogeneous crossbreds; MIX4 denotes the admixture of four breeds and their heterogeneous crossbreds; A(MIX2) and A(MIX4) are corresponding admixed populations; A+B is the combined population of breed A and B.

CHAPTER 4. RESULTS

4.1 Preliminary Simulations

4.1.1 Response to Selection

Figure 4.1 depicts the effects of the additive model, BSAM, and the dominance model on the cumulative response to 20 generations of GS in the crossbreeding program by summarizing over a total of 800 replicates of the preliminary simulations. The cumulative response is defined as the mean phenotypic progress in the crossbreds due to the selection in the purebreds standardized by the phenotypic standard deviation of the selection candidates in the first generation of selection. The dominance model consistently had a higher cumulative response than the additive model and BSAM over generations of selection, when the SNP effects were only estimated once in the crossbred AB that initiated the program. The superiority of the dominance over additive model was first observed in generation two with an advantage of 8.7%. This advantage grew fast as selection proceeded until in generation nine (29.5%) and by generation 20 it had slightly increased to 32.2%. Compared to the clear benefits observed from the dominance model, the advantage of BSAM over the additive model was limited, starting with an advantage of 1.0%, increasing to the maximum of 5.1% in generation nine and finally reducing to 2.2% by generation 20. It can be found in Figure 4.1 that the mean cumulative response from the dominance model had surpassed one standard deviation (colored shadows) of the response from the additive model since generation six and of that from BSAM since generation 12. The differences of the cumulative response by generation 20 between the dominance and any other models for GS are statistically significant at the type I error rate < 0.001, which were concluded from a set of hypothesis tests by fitting the value of the response in generation 20 with a model where the fixed effects of the models for GS were blocked in eight repetitions

of the simulation (APPENDIX A). However, the p-value for testing that difference between BSAM and the additive model was only 0.009. The variation of breed differences across repetitions due to genetic drift hardly altered the order of the model performance except that of BSAM and the additive model, though magnitudes of the differences between models varied considerably (Figure 4.2).

4.1.2 Accuracy of Selection

The accuracy of selection was calculated as the correlation between TBV from Equation 3.48 and GEBV from Equation 3.49 of the purebred selection candidates. The results shown in Figure 4.3 agree with the results of the response to selection. Without retraining, BSAM or the additive model had a greater loss in accuracy over generations than the dominance model, though all of them started out with very similar accuracies in generation 1 (0.866 for the additive model, 0.862 for BSAM, and 0.878 for the dominance model). The difference in accuracy between the dominance model and either BSAM or the additive model became wider until generation 5 where a maximum difference of 0.172 with BSAM or of 0.205 with the additive model was observed. The decline of accuracy toward zero for BSAM or the additive model had approximately the same shape. For the last few generations, the accuracies from any models were close to zero. The loss of the selection accuracy and the reduction of the genotypic variance shown in Figure 4.4 due to the selection in the purebreds jointly slowed down the growth of the cumulative response in the crossbreds. However, with continued selection, GS with the dominance model is expected to converge to a higher stable level than the other models (Figure 4.1).

4.1.3 Selected Parental Average and Heterosis in Crossbreds

From the definition of heterosis, the expected crossbred performance (CP) could be written as:

$$E(CP) = SPA + H, \tag{4.1}$$

where SPA denotes the selected (purebred) parental average and H the heterosis in crossbreds. Thus, the observed advantage of the dominance model may due to greater response in

SPA or H or in both. The contributions of SPA and H to cumulative response can be seen from Figure 4.5, where response using the one model (y-axis) is plotted against response using another (x-axis) for SPA (red squares) and H (blue dots). The dominance model led to substantially greater response in H than the alternatives and this advantage increased persistently as selection proceeded. Compared to the dominance model, although response in SPA was slightly higher for the additive model, in particular, after generation five of selection (but this phenomenon was not observed clearly for BSAM), the combination of these effects resulted in a better performance of crossbred descendents by using the dominance model, which agrees with Figures 4.1 and 4.2. Consider generation 20 for instance. The dominance model resulted in an about 0.3 phenotypic standard deviations (std) lower level of SPA than the additive model. This loss, however, was compensated by over 1.1 phenotypic std higher heterosis in crossbreds, which led to an advantage of over 0.8 phenotypic std for crossbred performance. On the other hand, even though the loss of SPA seems smaller for BSAM than the dominance model as compared to the additive model (Figure 4.5 (a, c)), the advantage in H for BSAM over the additive model was not large enough to show an advantage in crossbred performance as clear as in the dominance model.

4.1.4 Fixation of Over-dominant QTL

The benefit of fixing over-dominant QTL in the purebreds for the crossbred performance is maximized only when alternate alleles are fixed in the two parental breeds. This has been examined in Figure 4.6 where the average frequency of heterozygous genotypes for the over-dominant QTL in crossbreds is plotted against the generations of selection under different models. Starting at similar levels, the heterozygous genotypic frequency under the additive model reached the maximal value of 0.437 in generation five and gradually decreased to 0.415 in generation 20, while a consistent increase of heterozygous genotypic frequency was observed under either the dominance model or BSAM. However, the frequency from the dominance model stabilized to a substantially higher level (0.512) than that from BSAM (0.482).

In Figure 4.7, the change in allele frequency for a couple of over-dominant QTL by the additive and dominance models is plotted during the selection period. Figure 4.7 (a) depicts

a typical situation where the fixation of alternate alleles in the two breeds was more rapid in the dominance than in the additive model. In addition to the difference in the rate of fixation, a more unfavorable case was observed in Figure 4.7 (b) where the allele frequencies moved to one in both breeds with the additive model.

4.1.5 Accuracy of Genomic Prediction

The accuracies of genomic prediction shown in Table 4.1 are correlations between TBV and GEBV of breed A animals for crossbred performance averaging over 24 preliminary simulations, when a variety of populations were used for training. In this table, the accuracy reaches the highest when 1) breed A animals were the sires of all individuals in the training population, and 2) the population that used for training was also the target crossbreds to be improved through selection in breed A. These training populations are crossbred AB, A(BC), A(MIX2) and A(MIX4). The accuracy decreases when a breed A animal is the grandsire of a training individual but condition (2) still holds, such as the individual from crossbred (AB)(CD) and $(AB)^2$. The decrease of accuracy is even greater when condition (1) holds but the training population is not the target population. The crossbred AC or purebred A is one of those training populations. When both conditions were not met there was a substantial loss of accuracy. The accuracy from training on breed B only gave about half of the highest accuracy seen in the table. Training on the pooled population only performs better than training on crossbred AC and purebred B.

The advantages of the dominance model over the other models are also shown for the accuracy of genomic prediction (Table 4.1). For different training populations, the additional accuracy from the dominance model ranged from 1.1% to 8.1% compared to the additive model and it ranged from 2.8% to 23.7% compared to BSAM. The results of using BSAM in the training population (AB)(CD), $(AB)^2$ and A+B were not obtained because BSAM was not applied to these populations as explained in Chapter 2. The superiority of the dominance model over the additive model in populations A+B, AC and A, which ranged from 7.2 to 8.1% was substantially higher than that in the other populations (only 1.1 to 2.5%). BSAM was expected to give a better result than at least the additive model, when crossbreds were used

for training. However, the accuracy from BSAM was considerably lower than that from either the additive or dominance model, except when crossbred AB and A(BC) was used for training. Even the additive model outperforms BSAM with an advantage of 22.3% of additional accuracy in breed B and of up to 8.3% of additional accuracy in the other populations.

Table 4.2 depicts the prediction accuracy in admixed populations when breed composition was either explicitly considered or ignored in alternative models. An advantage of fitting breed composition explicitly was not detected for any of the models. It even slightly decreased the accuracy from 0.739 to 0.733 for the dominance model when the combined population A+B was used for training.

4.2 More Realistic Simulations

As described in Chapter 2, the parameters in these simulations including the size of dominance variance and heterosis were more realistic in order to examine if the advantages that were observed in the preliminary simulations still holds. Results were based on eight simulations.

4.2.1 Response to Selection

The results of the cumulative response to selection in the more realistic simulations are in Figure 4.8. This figure shows the superiority of the dominance model over the others was still present, although the magnitude of the differences were smaller relative to that observed in the preliminary simulations. Here, however, the difference in response between the additive model and BSAM was considerably larger that in the preliminary simulations. The superiority of the dominance model did not show until generation four and that of BSAM over the additive model was observed after generation 10. Through 20 generations of selection, crossbred performance had been improved by 3.89, 3.64 and 3.49 phenotypic std of the selection candidates in the first generation, respectively, for the dominance model, BSAM and the additive model. Thus, The cumulative response was only 11.5% higher for the dominance model than for the additive model and of only 6.9% higher than for BSAM at generation 20. Note that the corresponding values in the preliminary simulations were 32.2% and 29.2%. In contrast of the reduced advantage of the dominance model in the more realistic simulations, the advantage of BSAM over the additive

model was 4.3%, which is larger than the number of 2.2% in the preliminary simulations. Results of the hypothesis tests (APPENDIX A) suggested that the differences between pairwise models were all significant at the type I error rate less than 0.001.

4.2.2 Accuracy of Genomic Prediction

Table 4.3 depicts the prediction accuracies in different training populations using alternative models in the more realistic simulations. In these simulations, the additive variance was as much as four times higher than the dominance variance. Thus, except when training was AB, A(BC), (AB)(CD) or B, the accuracies in these simulations were substantially higher than in the preliminary simulations where the additive variance was about twice as large as the dominance variance. In population AB and B, the accuracy decreased as compared to that in the preliminary simulation. Further, the differences between models also became smaller. The largest observed superiority of the dominance model over the additive model dropped from 8.1% in the preliminary simulations to 2.0% in these simulations. Similarly, there was a drop from 23.7% to 11.2% for the comparison between the dominance model and BSAM. Overall, the dominance model was less favored in the more realistic simulations than in the preliminary simulations. As in the preliminary simulations, an advantage of fitting breed composition in alternative GS models was also not detected in the more realistic simulations (Results not shown).

Table 4.1 Correlations between TBV and GEBV of validation purebred A animals for crossbred performance obtained by model used for GS and training population in the preliminary simulations

Training Population	**Target Population**	**Additive Model**	**BSAM**	**Dominance Model**
AB	AB	0.806	0.797	0.819
A(BC)	A(BC)	0.805	0.798	0.823
(AB)(CD)	(AB)(CD)	0.722	-	0.740
$(AB)^2$	$(AB)^2$	0.761	-	0.773
A(MIX2)	A(MIX2)	0.809	0.750	0.828
A(MIX4)	A(MIX4)	0.800	0.760	0.819
A+B	AB	0.685	-	0.739
AC	AB	0.679	0.652	0.728
A	AB	0.704	0.650	0.761
B	AB	0.438	0.358	0.443

AB is the cross of breed A and B; A(BC) is the three-way crossbreds; (AB)(CD) is the four-way crossbreds; $(AB)^2$ is the F2 crossbreds; MIX2 denotes the admixture of breed A, B and their heterogeneous crossbreds; MIX4 denotes the admixture of four breeds and their heterogeneous crossbreds; A(MIX2) and A(MIX4) are corresponding admixed populations; A+B is the combined population of breed A and B; AC is the cross of breed A and C; A or B is the purebred A or B.

Table 4.2 Correlations between TBV and GEBV of validation purebred A animals for crossbred performance when admixed populations were used in training and breed composition was considered or ignored in alternative GS models

Training Population	**Model**	**Considered**	**Ignored**
A(MIX2)	Additive	0.809	0.809
	BSAM	0.749	0.750
	Dominance	0.830	0.828
A(MIX4)	Additive	0.802	0.800
	BSAM	0.760	0.760
	Dominance	0.817	0.819
A+B	Additive	0.682	0.685
	Dominance	0.733	0.739

MIX2 denotes the admixture of breed A, B and their heterogeneous crossbreds; MIX4 denotes the admixture of four breeds and their heterogeneous crossbreds; A(MIX2) and A(MIX4) are corresponding admixed populations; A+B is the combined population of breed A and B; AC is the cross of breed A and C; A or B is the purebred A or B.

Table 4.3 Correlations between TBV and GEBV of validation purebred A animals for crossbred performance obtained by model used for GS and training population in the more realistic simulations

Training Population	**Target Population**	**Additive Model**	**BSAM**	**Dominance Model**
AB	AB	0.808	0.796	0.812
A(BC)	A(BC)	0.794	0.776	0.800
(AB)(CD)	(AB)(CD)	0.729	-	0.736
$(AB)^2$	$(AB)^2$	0.793	-	0.806
A(MIX2)	A(MIX2)	0.830	0.788	0.836
A(MIX4)	A(MIX4)	0.830	0.780	0.834
A+B	AB	0.761	-	0.776
AC	AB	0.768	0.744	0.773
A	AB	0.793	0.722	0.794
B	AB	0.402	0.367	0.408

AB is the cross of breed A and B; A(BC) is the three-way crossbreds; (AB)(CD) is the four-way crossbreds; $(AB)^2$ is the F2 crossbreds; MIX2 denotes the admixture of breed A, B and their heterogeneous crossbreds; MIX4 denotes the admixture of four breeds and their heterogeneous crossbreds; A(MIX2) and A(MIX4) are corresponding admixed populations; A+B is the combined population of breed A and B; AC is the cross of breed A and C; A or B is the purebred A or B.

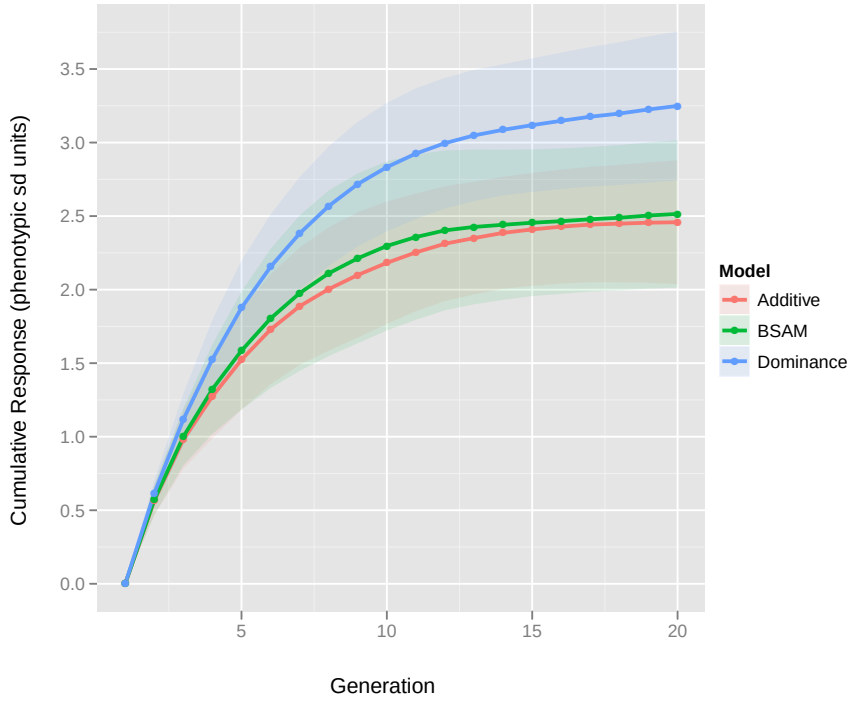

Figure 4.1 Cumulative response to selection standardized by phenotypic deviations over generations in the crossbreeding program obtained by different GS models, averaged across 800 replicates of the preliminary simulations. Shadows represent standard deviations.

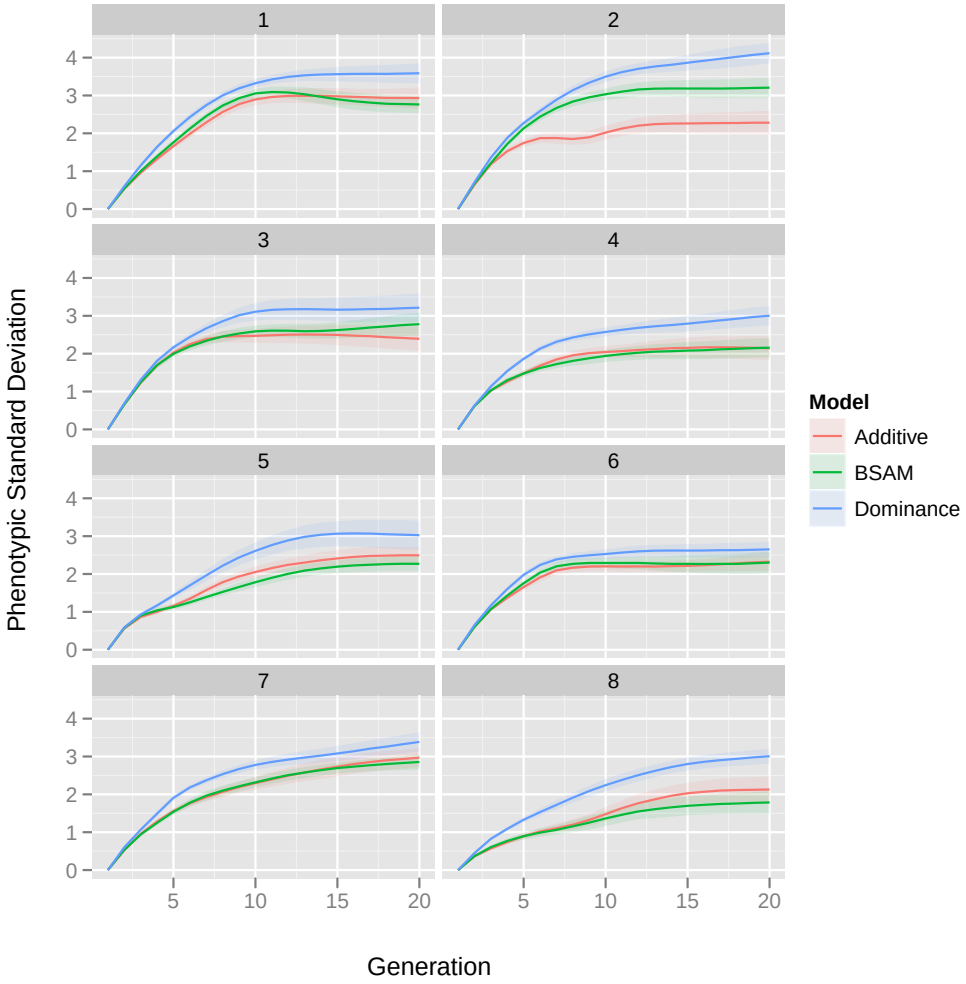

Figure 4.2 Cumulative response to selection standardized by phenotypic deviations over generations in the crossbreeding program obtained by different GS models, each averaged across 100 replicates, respectively, for the eight preliminary simulations.

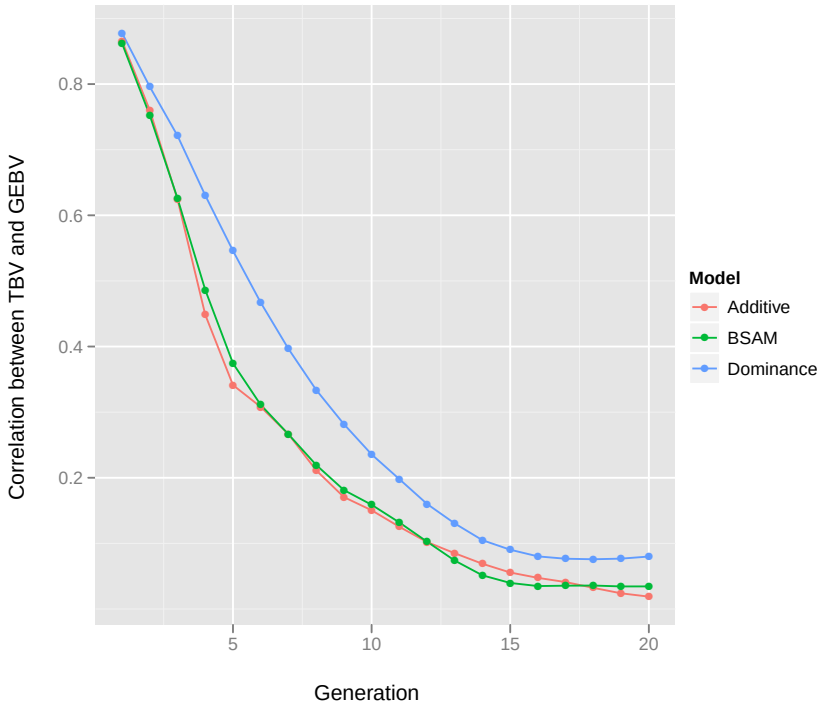

Figure 4.3 Correlation between TBV and GEBV of selection candidates over generations in the crossbreeding program obtained by different GS models, averaged across parental breeds and simulation replicates.

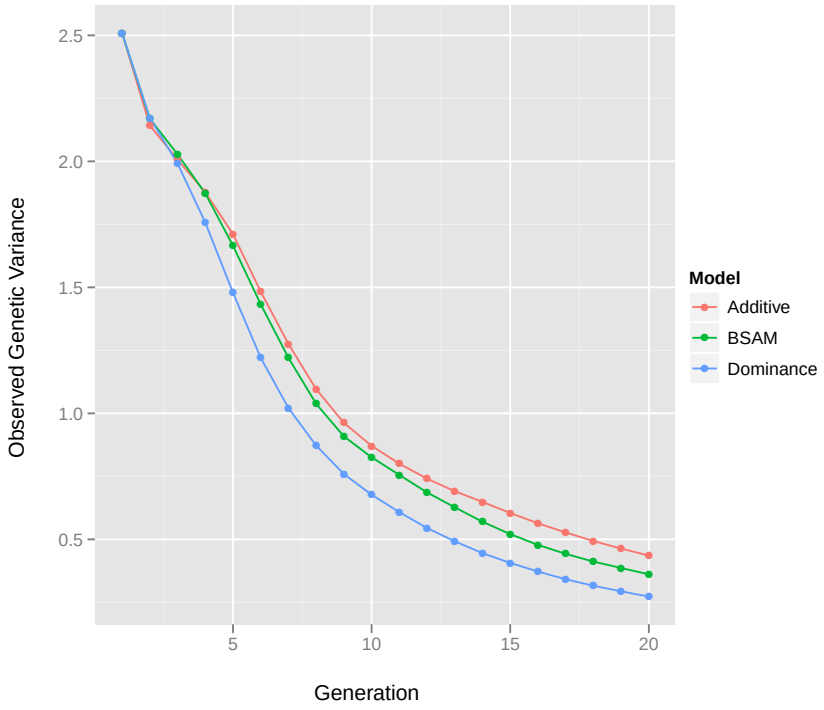

Figure 4.4 Observed total genetic variance in selection candidates over generations in the crossbreeding program under different GS models, averaged across parental breeds and simulation replicates.

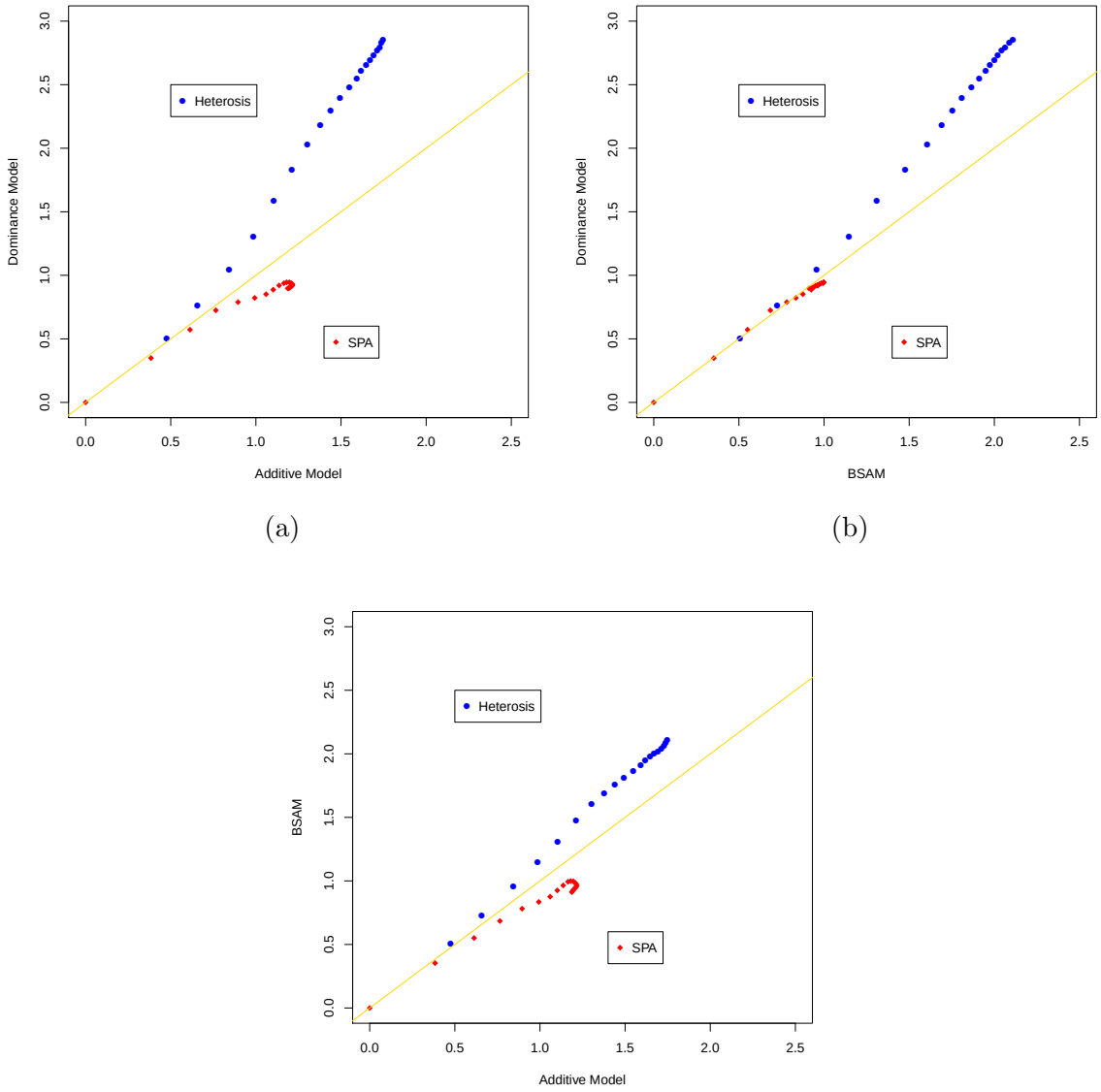

Figure 4.5 (a) Response using the dominance model (y-axis) against response using the additive model (x-axis), (b) response using the dominance model against response using BSAM, and (c) response using BSAM against the additive model, for selected (purebred) parental average (red squares) and heterosis in crossbreds (blue dots), averaged across preliminary simulation replicates. The solid line is y=x.

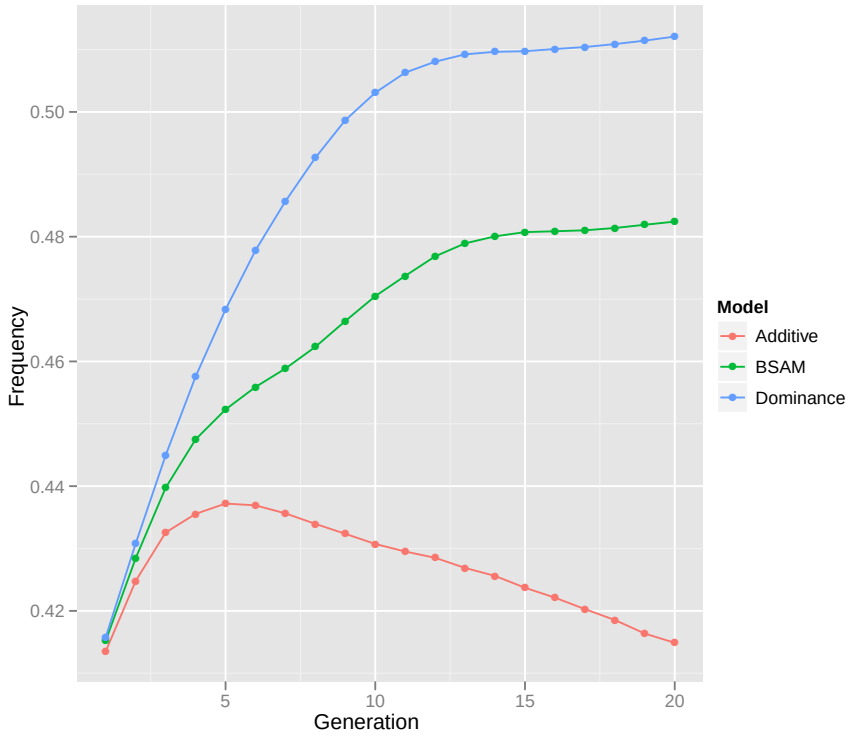

Figure 4.6 Change in heterozygous genotype frequency in crossbreds over generations in the crossbreeding program under different GS models, averaged across simulation replicates.

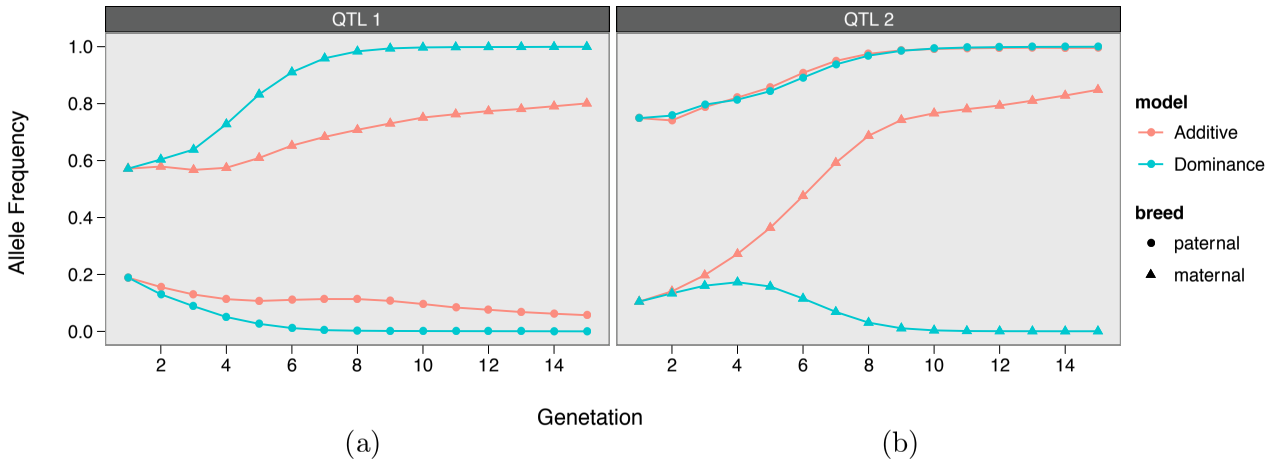

Figure 4.7 Change in allele frequencies of two over-dominant QTL with major dominance effects in both sire and dam breeds over generations of selection in a given simulation replicate. (a) shows alternate alleles approaching fixation in sire and dam breeds more rapidly with dominance than additive model; (b) shows the same allele approaching fixation in both parental breeds with the additive model, which is not desirable, and this did not happen with the dominance model.

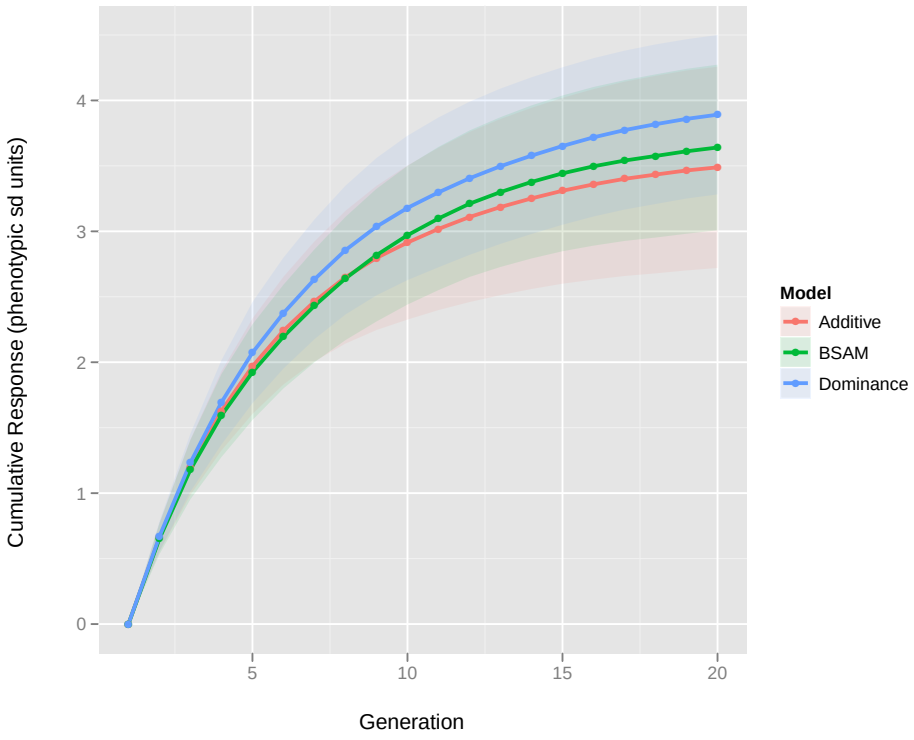

Figure 4.8 Cumulative response to selection standardized by phenotypic deviations over generations in the crossbreeding program obtained by different GS models, averaged across 800 replicates of the more realistic simulations. Shadows represent standard deviations.

CHAPTER 5. SUMMARY AND DISCUSSION

The primary objective of this study was to investigate the advantages of including dominance in the model used for genomic selection (GS) of purebred for crossbred performance. A secondary objective was to examine if breed composition can be ignored when the training population for GS is an admixture of breeds. In both preliminary and more realistic simulations, results show that including dominance effects in addition to additive effects in the model was advantageous for response to selection and prediction of GEBV. Breed composition of admixed populations was ignorable for all GS models as modeling breed composition did not increase the accuracy of genomic prediction.

The advantage of the dominance model is attributable to the use of breed-specific SNP allele substitution effects to predict purebred GEBV for crossbred performance. Previous studies have shown that in theory GS using BSAM can result in a substantial higher response than GS using the additive model (Dekkers, 2007; Kinghorn et al., 2010). However, perfect estimation of breed-specific allele substitution effects was assumed in their studies. When the number of observations is limited, the parameters in BSAM may be less accurately estimated than those of an the additive model because BSAM has twice the number of parameters as the additive model. Ibanez-Escriche et al. (2009) reported that under additive gene action the additive model can give a higher prediction accuracy than BSAM if the breeds are related and the training population size is small relative to the number of markers. This agrees with the results in this study where BSAM had slightly lower accuracies than the additive model. However, results from the 20 generations of selection showed that BSAM had a higher cumulative response than the additive model though the accuracy of selection was about the same in each generation.

The dominance model has about the same number of parameters as BSAM. With comparable model complexity, the dominance model, however, outperformed BSAM in both response

to selection and accuracy of prediction. This is because estimates of additive and dominance effects from the dominance model were combined with the observed allele frequencies in the other parental breed to calculate the breed-specific allele substitution effects. In BSAM, however, breed-specific substitution effects are estimated directly. Thus, the allele frequencies used in BSAM are actually estimates of the frequencies in the training population of the alleles inherited from the other parental breed. Note that the alleles inherited by the training population is a random sample of those from the parental population, and therefore its frequency will deviate from that of the parental population. Thus, the use of observed allele frequencies from the parental population to compute the breed-specific substitution effects favors the dominance model over BSAM.

Further, as selection progressed and allele frequencies changed due to selection, in each generation of selection, the observed allele frequencies from that generation were combined with the estimates of additive and dominance effects obtained during training. On the other hand, with the additive model and with BSAM, the substitution effects estimated during training were repeatedly used to compute GEBV of selection candidates ignoring the change in allele frequencies. It can be seen from Figure 4.3 that the drop in accuracy for the additive model and BSAM was greater than for the dominance model through the generations of selection. Thus, if a retaining is carried out when the accuracy drops by some pre-determined amount, use of the dominance model would require lower frequency of retraining than use of either BSAM or the additive model. This is appealing for traits that are hard or expensive to measure.

The dominance model is also advantageous when the purebred candidates are to be selected for performance in a crossbred target population that is different from the training population. Tables 4.1 and 4.3 show that prediction accuracy from the dominance model was higher than those from the additive model when population A, AC or A+B was used for training but the target crossbred was AB. This suggests that the dominance model can be used to select purebreds for performance in a range of target crossbreds as long as the allele frequencies that are used for calculating the substitution effects are from the other parental breed of the crossbred target population.

However, the advantages of the dominance model due to using the appropriate allele fre-

quencies may not hold if LD is different across the breeds. As a result, the SNP that track QTL well in one breed may fail in another. Thus, even though the QTL additive and dominance effects are independent of breed, the estimates of SNP effects may differ between breeds due to differences in LD, and with differences that are large enough, the dominance model may become inferior to BSAM, where breed-specific substitution effects are estimated that accommodate differences in LD in addition to differences in allele frequencies. However, the dominance model in this study can also be extended to include breed-specific additive and dominance effects to accommodate the LD specific to breed.

With a similar statistical model that includes dominance, Toro and Varona (2010) explored the advantages of GS in a purebred population and found using a dominancemodel gave a higher response than using an additive model in the first generation of selection, but this advantage was not seen in subsequent generations. Using a dominance model for GS would be more important, however, when purebreds are selected for crossbred performance, especially when heterosis is present as in this study. With the dominance model, the alternate alleles of over-dominant QTL approached to fixation at a more rapid pace with fewer errors (Figure 4.7). The progress towards fixation, however, was retarded by the decline of accuracy in the long-term. This explains why heterosis increased nonlinearly with generations of selection as shown in Figure 4.5. On the contrary, improvement in the purebreds was slower with the dominance model because fixing over-dominant QTL reduced heterogosity, which is inversely related to the purebred performance when there is directional dominance. Overall, the response to selection was dominated by the large amount of heterosis in the crossbreds. Therefore, by using the dominance model, the breeding goal of maximizing crossbred performance was more effectively fulfilled at some cost of improvement in the purebreds.

In addition to model complexity, another possible reason to explain the lower accuracy of BSAM relative to the additive model as shown in Ibanez-Escriche et al. (2009) is described below. Consider a locus that is segregating in breed A but fixed in breed B. Because the additive model regresses phenotypes only on the segregating alleles, the common substitution effect for this locus is actually the one specific to the breed A just as in BSAM. However, BSAM includes an additional substitution effect for breed B where the allele is fixed. The substitution effect

estimated from BSAM for the breed B allele will only add noise to the prediction of GEBV. Therefore, when several loci are nearly fixed in one of the parental breeds but segregating in the other, the additive model would show an advantage over BSAM.

The absence of epistasis and genotype by environment interactions were assumed in this study. Thus, as explained below, differences in accuracy of prediction observed in Table 4.1 can be explained as being due to the differences between breeds in allele frequency, in heterozygosity or in LD, or any combination of these.

When the dominance model is used for GS, the difference in allele frequencies do not contribute to any loss in accuracy. Thus, the large difference in accuracy between training in breed A and training in breed B for performance in crossbred AB using the dominance model (0.761 vs. 0.443) is primarily due to LD differences between breeds because the expected level of heterozygosity should be the same for both breeds as they were identically simulated. The same explanation holds for the difference observed between training in crossbred AB and AC (0.819 vs. 0.728). The LD is almost the same in population AB and A+B, therefore the difference observed between these two populations (0.819 vs. 0.739) must be due to the higher level of expected heterozygosity in AB.

When the additive model or BSAM is used for GS, the difference in allele frequencies becomes relevant. Comparing accuracy from the additive and dominance models when training was in breed A (0.704 vs. 0.761), the difference can be attributed to the substitution effects estimated in the additive model being based on allele frequencies in breed A rather than in breed B. On the other hand, comparing accuracy from the additive and dominance models when training was in breed B (0.438 vs. 0.443), the difference was negligible because the substitution effects estimated in the additive model were now based on allele frequencies in breed B, which are the appropriate allele frequencies.

Harris et al. (2008) found that the accuracy of GS of Holstein-Friesian with training in Jersey and vice versa was as low as -0.1 to 0.3 over traits. Toosi et al. (2010) showed by simulation that the accuracy of prediction decreased from 0.80 to 0.55 when a different breed was used for training. In this study, the drop of accuracy from training in breed A (0.704) to training in breed B (0.438) was about 0.26 in the additive model. Because the accuracy

of 0.438 that was obtained from training in breed B was for evaluating breed A animals for performance of crossbred AB offspring, the allele frequencies implicitly used in the calculation of substitution effects (Equation 3.54) when training in B were the appropriate ones. Thus, LD differences between breeds is the only cause for the low accuracy. In Harris et al. (2008), the accuracy, however, was for evaluating the performance within-breed, therefore the lower value of accuracy can be attributable to both LD and allele frequency differences between breeds.

Results from the more realistic simulations show reduced differences between alternative models as compared to those from the preliminary simulations because of the larger additive component. However, the superiority of the dominance model was still significant in generation 20 of selection and the advantage of the dominance model still held in accuracy of prediction.

One of the hypotheses of this study was that breed composition can be ignored even when dominance gene action is present provided that a dominance model is used for GS. However, even with an additive model, including breed composition in GS did not increase accuracy of prediction. To understand this, note that breed composition can be ignored when BSAM is used for GS because BSAM accounts for dominance gene action by fitting separate substitution effects for parental breeds. However, when the additive model is used for crossbreds or admixed populations, it can give even better results than BSAM (Ibanez-Escriche et al., 2009) as explained previously. This is fortunate as breed composition is often unknown in commercial beef cattle populations.

In conclusion, when dominance gene action is present, using a dominance model for GS would result in a greater response to selection in purebred animals for crossbred performance. The dominance model allows GS over generations or across breeds using one set of estimates of the additive and dominance effects with a higher accuracy than either BSAM or the additive model. Further, breed composition can be ignored in GS even when an admixed population is used for training.

APPENDIX A.

Scaling procedure for QTL additive and dominance effects

Let V_A and V_D denote the observed additive and dominance genetic variance of the trait. Assuming no genotype by genotype interactions among QTL that define the trait, the genetic variance components can be written as the sum of the variance explained by each QTL (Falconer and Mackay, 1996):

$$V_A = \sum_j 2p_j q_j \alpha_j \tag{A.1}$$

$$V_D = \sum_j (2p_j q_j d_j)^2 \tag{A.2}$$

where $p_j = 1 - q_j$ is the observed allele frequency for QTL j, d_j is the QTL dominance effect, and α_j is the QTL allele substitution effect defined as

$$\alpha_j = a_j + (q_j - p_j)d_j \tag{A.3}$$

where a_j is the QTL additive effect.

Let V_A^* and V_D^* denote the corresponding desired genetic variance components and a_j^*, d_j^*, α_j^* the corresponding scaled QTL effects. Let

$$s = \frac{V_D^*}{V_D} \tag{A.4}$$

From Equations A.2 and A.4, we have that

$$\sum_j (2p_j q_j d_j^*)^2 = s \sum_j (2p_j q_j d_j)^2 \tag{A.5}$$

Thus,

$$d_j^* = \sqrt{s} d_j \tag{A.6}$$

Similar to $\sqrt{s}$ the scalar for dominance effects, then we find a scalar t for additive effects such that

$$a_j^* = ta_j \tag{A.7}$$

and

$$\begin{aligned} V_A^* &= \sum_j 2p_j q_j \alpha_j^* \\ &= \sum_j 2p_j q_j (a_j^* + (q_j - p_j)d_j^*)^2 \end{aligned} \tag{A.8}$$

Substituting Equation A.7 in A.8 and rearranging this, we have that

$$t^2 \sum_j 2p_j q_j a_j^2 + t \sum_j (q_j - p_j) a_j d_j + \sum (q_j - p_j)^2 d^{*2} - V_A^* = 0 \tag{A.9}$$

This can be seen as a quadratic equation with variable t unknown. Thus, the scalar t for the additive effects can be obtained by solving this equation.

Hypothesis test for the GS model effect

We aim to test if any difference in cumulative response to GS observed at the 20 generation of selection between the additive model, BSAM, and the dominance model is statistically significant. As described in Chapter 2, data were collected from eight simulations each with 100 replicates resulting in a total of 800 observations. The following mixed linear model was used to fit the data:

$$y_{ij} = \mu + m_i + b_j + e_{ij} \tag{A.10}$$

where y_{ij} is the response from GS model i in simulation j, m_i is the fixed effect for GS model $i = \{1, 2, 3\}$, $b_j \sim N(0, \sigma_b^2)$ is the random blocking effect for simulation replicate $j = \{1, ..., 8\}$, and $e_{ij} \sim N(0, \sigma_e^2)$ is the residual. A set of t-tests was used for testing the null hypothesis that 1) $m_1 = m_2$, 2) $m_1 = m_3$, or 3) $m_2 = m_3$.

BIBLIOGRAPHY

Abasht, B. and Lamont, S. J. (2007). Genome-wide association analysis reveals cryptic alleles as an important factor in heterosis for fatness in chicken f2 population. *Anim Genet*, 38(5):491–498.

Bell, A. E. (1982). Selection for heterosis - results with laboratory and domestic animals. volume 6, pages 206–277. Proc. 2nd World Cong. Genet. Appl. Livest. Prod.

Bell, A. E., Moore, C. H., and Warren, D. C. (1950). Systems of breeding designed to give maximum heterosis in chickens. *Poultry Science*, 29:749.

Bennewitz, J. and Meuwissen, T. H. E. (2010). The distribution of qtl additive and dominance effects in porcine f2 crosses. *J Anim Breed Genet*, 127(3):171–179.

Berger, E. (1976). Heterosis and the maintenance of enzyme polymorphism. *Am. Nat.*, 110(823-839).

Bijma, P. and Arendonk, J. A. M. v. (1998). Maximizing genetic gain for the sire line of a crossbreeding scheme utilizing both purebred and crossbred information. *Animal Science*, 66:529–542.

Bijma, P., Woolliams, J. A., and van Arendonk, J. A. M. (2001). Genetic gain of pure line selection and combined crossbred purebred selection with constrained inbreeding. *Animal Science*, 72:225–232.

Birchler, J. A., Yao, H., Chudalayandi, S., Vaiman, D., and Veitia, R. A. (2010). Heterosis. *Plant Cell*, 22(7):2105–2112.

Bowman, J. C. (1959). Selection for heterosis. *Anim. Breed. Abstr.*, 27:261–273.

Boysen, T. J., Tetens, J., and Thaller, G. (2010). Detection of a quantitative trait locus for ham weight with polar overdominance near the ortholog of the callipyge locus in an experimental pig f2 population. *J Anim Sci*, 88(10):3167–3172.

Bruce, A. B. (1910). The mendelian theory of heredity and the augmentation of vigor. *Science*, 32(827):627–628.

Caballero, A. and Keightley, P. D. (1994). A pleiotropic nonadditive model of variation in quantitative traits. *Genetics*, 138(3):883–900.

Calhoon, R. E. and Bohren, B. B. (1974). Genetic gains from reciprocal recurrent and within-line selection for egg production in the fowl. *Theor Appl Genet*, 44:364–372.

Calus, M. P. L., Meuwissen, T. H. E., de Roos, A. P. W., and Veerkamp, R. F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, 178(1):553–561.

Charlesworth, D. and Willis, J. H. (2009). The genetics of inbreeding depression. *Nat Rev Genet*, 10(11):783–796.

Comings, D. E. and MacMurray, J. P. (2000). Molecular heterosis: a review. *Mol Genet Metab*, 71(1-2):19–31.

Comstock, R. E., Robinson, H. F., and Harvey, P. H. (1949). A breeding procedure designed to make maximum use of both general ans specific combining ability. *Agronomy Journal*, 41:360–367.

Crnokrak, P. and Roff, D. A. (1995). Dominance variance: associations with selection and fitness. *Heredity*, 75:530–540.

Crow, J. F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*, page 79. Harper and Row, New York, US.

Culbertson, M. S., Mabry, J. W., Misztal, I., Gengler, N., Bertrand, J. K., and Varona, L. (1998). Estimation of dominance variance in purebred yorkshire swine. *J Anim Sci*, 76(2):448–451.

Cundiff, L. V. and Gregory, K. E. (1999). What is systematic crossbreeding? *Proc. NCBA Cattleman's College, Charlotte, NC*.

Daetwyler, H. D., Villanueva, B., Bijma, P., and Woolliams, J. A. (2007). Inbreeding in genome-wide selection. *J Anim Breed Genet*, 124(6):369–376.

Dagnachew, B., Thaller, G., Lien, S., and Adnoy, T. (2011). Casein snp in norwegian goats: additive and dominance effects on milk composition and quality. *Genet Sel Evol*, 43(1):31.

Davenport, C. B. (1908). Degeneration, albinism and inbreeding. *Science*, 28(718):454–455.

Dekkers, J. C. M. (2007). Marker-assisted selection for commercial crossbred performance. *J Anim Sci*, 85(9):2104–2114.

Dekkers, J. C. M. and Chakraborty, R. (2004). Optimizing purebred selection for crossbred performance using qtl with different degrees of dominance. *Genet Sel Evol*, 36(3):297–324.

Duvick, D. (1999). Heterosis: feeding people and protecting natural resources. In Coors, J. and Pandey, S., editors, *The Genetics and Exploitation of Heterosis in Crops*, pages 19–29. American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

East, E. M. (1908). *Inbreeding in corn*, pages 419–428. Reports of the Connecticut Agricultural Experiment Station for Years 1907-1908.

East, E. M. (1936). Heterosis. *Genetics*, 21(4):375–397.

Estelle, J., Gil, F., Vazquez, J. M., Latorre, R., Ramirez, G., Barragan, M. C., Folch, J. M., Noguera, J. L., Toro, M. A., and Perez-Enciso, M. (2008). A quantitative trait locus genome scan for porcine muscle fiber traits reveals overdominance and epistasis. *J Anim Sci*, 86(12):3290–3299.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Longmans Green, Harlow, Essex, UK, 4 edition.

Fernando, R., Habier, D., Stricker, C., Dekkers, J., and Totir, L. (2008). Genomic selection. *Acta Agriculturae Scandinavica, Section A - Animal Science*, 57(4):182–195.

Fernando, R. L. and Garrick, D. J. (2009). *GenSel - user manual*.

Flint-Garcia, S. A., Thornsberry, J. M., and Buckler, E. S. t. (2003). Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol*, 54:357–374.

Frascaroli, E., Cane, M. A., Landi, P., Pea, G., Gianfranceschi, L., Villa, M., Morgante, M., and Pe, M. E. (2007). Classical genetic and quantitative trait loci analyses of heterosis in a maize hybrid between two elite inbred lines. *Genetics*, 176(1):625–644.

Gengler, N., I, M., and K, B. J. (1997). Relationship between estimates of heterosis and dominance variance for post-weaning gain in us limousin cattle. *J Anim Sci*, Suppl. 1(a):149.

Goddard, M. E. and Hayes, B. J. (2007). Genomic selection. *J Anim Breed Genet*, 124(6):323–330.

Habier, D., Fernando, R., Kizilkaya, K., and Garrick, D. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12(1):186.

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

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., and Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in german holstein cattle. *Genet Sel Evol*, 42:5.

Harris, B., Johnson, D., and Spelman, R. (2008). Genomic selection in New Zealand and the implications for national genetic evaluation. Niagara Falls, Ont. Proc. Interbull Meeting.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: progress and challenges. *J Dairy Sci*, 92(2):433–443.

Hill, W. (1982). Dominance and epistasis as components of heterosis. *J Anim Breed Genet*, 99:161–168.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6(2):95–108.

Hoeschele, I. (1991). Additive and nonadditive genetic variance in female fertility of holsteins. *J Dairy Sci*, 74(5):1743–1752.

Ibanez-Escriche, N., Fernando, R. L., Toosi, A., and Dekkers, J. C. M. (2009). Genomic selection of purebreds for crossbred performance. *Genet Sel Evol*, 41:12.

Ishikawa, A. (2009). Mapping an overdominant quantitative trait locus for heterosis of body weight in mice. *J Hered*, 100(4):501–504.

Johnson, R. K. (1980). *Heterosis and Breed Effects in Swine*. NC Reg. Pub 262.

Jones, D. F. (1917). Dominance of linked factors as a means of accounting for heterosis. *Genetics*, 2(5):466–479.

Kacser, H. and Burns, J. A. (1981). The molecular basis of dominance. *Genetics*, 97:639–666.

Karlin, S. (1984). Theoretical aspects of genetic map functions in recombination processes. In Chakravarti, A., editor, *Human Population Genetics: The Pittsburgh Symposium*, pages 209–228, Van Nostrand Reinhold, New York, NY.

Kennedy, B. W., Quinton, M., and van Arendonk, J. A. (1992). Estimation of effects of single genes on quantitative traits. *J Anim Sci*, 70(7):2000–2012.

Kinghorn, B., Hickey, J., and van der Werf (2010). Reciprocal recurrent genomic selection (rrgs) for total genetic merit in crossbred individuals. Proceedings of 9th WCGALP.

Kosba, M. A. (1978). Heterosis and phenotypic correlations for shank length, body weight and egg production traits in the alexandria strains and their crosses with fayoumi chickens. *Beitr Trop Landwirtsch Veterinarmed*, 16(2):187–198.

Lee, S. H., van der Werf, J. H. J., Hayes, B. J., Goddard, M. E., and Visscher, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome snp data. *PLoS Genet*, 4(10):e1000231.

Li, L., Lu, K., Chen, Z., Mu, T., Hu, Z., and Li, X. (2008). Dominance, overdominance and epistasis condition the heterosis in two heterotic rice hybrids. *Genetics*, 180(3):1725–1742.

Li, Z. K., Luo, L. J., Mei, H. W., Wang, D. L., Shu, Q. Y., Tabien, R., Zhong, D. B., Ying, C. S., Stansel, J. W., Khush, G. S., and Paterson, A. H. (2001). Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. i. biomass and grain yield. *Genetics*, 158(4):1737–1753.

Lippman, Z. B. and Zamir, D. (2007). Heterosis: revisiting the magic. *Trends Genet*, 23(2):60–66.

Lo, L. L., Fernando, R. L., Cantet, R. J. C., and Grossman, M. (1995). Theory for modelling means and covariances in a two-breed population with dominance inheritance. *Theor Appl Genet*, 90:49–62.

Lo, L. L., Fernando, R. L., and Grossman, M. (1993). Covariance between relatives in multi-breed populations: Additive model. *Theor. Appl. Genet.*, 87:423–430.

Lo, L. L., Fernando, R. L., and Grossman, M. (1997). Genetic evaluation by blup in two-breed terminal crossbreeding systems under dominance. *J Anim Sci*, 75(11):2877–2884.

Lopez-Villalobos, N., Garrick, D. J., Holmes, C. W., Blair, H. T., and Spelman, R. J. (2000). Profitabilities of some mating systems for dairy herds in New Zealand. *J Dairy Sci*, 83(1):144–153.

Luo, L. J., Li, Z. K., Mei, H. W., Shu, Q. Y., Tabien, R., Zhong, D. B., Ying, C. S., Stansel, J. W., Khush, G. S., and Paterson, A. H. (2001). Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. ii. grain yield components. *Genetics*, 158(4):1755–1771.

Lutaaya, E., Misztal, I., Mabry, J. W., Short, T., Timm, H. H., and Holzbauer, R. (2001). Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. *J Anim Sci*, 79(12):3002–3007.

Meffert, L. M., Hicks, S. K., and Regan, J. L. (2002). Nonadditive genetic effects in animal behavior. *Am Nat*, 160 Suppl 6:S198–213.

Merks, J. and de Vries, A. W. (2002). New sources of information in pig breeding. Number 03-01, Montpellier. F. Minivielle, ed. INRA, Paris, France. Proc. 7th World Congr. Genet. Appl. Livest. Prod.

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

Meuwissen, T. H. E., Karlsen, A., Lien, S., Olsaker, I., and Goddard, M. E. (2002). Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics*, 161(1):373–379.

Misztal, I., Lawlor, T. J., and Gengler, N. (1997). Relationships among estimates of inbreeding depression, dominance and additive variance for linear traits in holsteins. *Genet Sel Evol*, 29:319–326.

Moav, R. (1973). *Agricultural Genetics, Selected Topics*, pages 319–352. Wiley, New York, US.

Mujibi, F., Nkrumah, J., Durunna, O., Stothard, P., Mah, J., Wang, Z., Basarab, J., Plastow, G., Crews DH, J., and Moore, S. (2011). Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle. *J Anim Sci*.

Nitter, G. (1978). Breed utilization for meat production in sheep. *Animal Breeding Abstracts*, 46:131–143.

Pirchner, F. and Mergl, R. (1977). Overdominance as cause for heterosis in poultry. *Journal of Animal Breeding and Genetics*, 94:151–158.

Piyasatian, N., Fernando, R. L., and Dekkers, J. C. M. (2007). Genomic selection for marker-assisted improvement in line crosses. *Theor Appl Genet*, 115(5):665–674.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–909.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations. *Am J Hum Genet*, 67(1):170–181.

Rabinowitz, D. (1997). A transmission disequilibrium test for quantitative trait loci. *Hum Hered*, 47(6):342–350.

Sheridan, A. K. (1981). Crossbreeding and heterosis. *Anim. Breed. Abstr*, 49:131–144.

Shull, G. H. (1908). The composition of field of maize. *Am. Breed. Assn. Rep.*, 4:296–301.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet*, 52(3):506–516.

Toosi, A., Fernando, R. L., and Dekkers, J. C. M. (2010). Genomic selection in admixed and crossbred populations. *J Anim Sci*, 88(1):32–46.

Toro, M. A. and Varona, L. (2010). A note on mate allocation for dominance handling in genomic selection. *Genet Sel Evol*, 42:33.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J Dairy Sci*, 91(11):4414–4423.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., and Schenkel, F. S. (2009). Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci*, 92(1):16–24.

Wallace, B. (1968). *Topics in Population Genetics*. Norton, New York.

Wei, M. and van der Steen, H. A. M. (1991). Comparison of reciprocal recurrent selection with pure-line selection systems in animal breeding (a review). *Anim. Breed. Abstr.*, 59:281–298.

Wei, M. and van der Werf, J. H. (1993). Animal model estimation of additive and dominance variances in egg production traits of poultry. *J Anim Sci*, 71(1):57–65.

Wei, M. and van der Werf, J. H. (1995). Genetic correlation and heritabilities for purebred and crossbred performance in poultry egg production traits. *J Anim Sci*, 73(8):2220–2226.

Wei, M. and van der Werf, J. H. J. (1994). Maximizing genetic response in crossbreds using both purebred and crossbred information. *Animal Production*, 59(401-413).

Wellmann, R. and Bennewitz, J. (2010). Considering dominance in genomic selection. Proceedings of 9th WCGALP.

Xiao, J., Li, J., Yuan, L., and Tanksley, S. D. (1995). Dominance is the major genetic basis of heterosis in rice as revealed by qtl analysis using molecular markers. *Genetics*, 140(2):745–754.

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics*, 163(2):789–801.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*, 38(2):203–208.

Zeng, J., Toosi, A., Fernando, R. L., Dekkers, J. C., and Garrick, D. (2011). Genomic selection of purebred animals for crossbred performance under dominance. San Diego, CA. Plant and Animal Genomes XIX Conference.