Genomic selection for crossbred performance

Hadi Esfandyari

Thesis committee

Main supervisor
Dr A. C. Sorensen
Senior Researcher, Center for Quantitative Genetics and Genomics
Aarhus University, Tjele, Denmark

Promotor
Prof. Dr J.A.M van Arendonk
Professor of Animal Breeding and Genetics
Wageningen University, The Netherlands

Co-promotor
Dr P. Bijma
Assistant professor, Animal Breeding and Genomics Center
Wageningen University, The Netherlands

Other members (assessment committee)
Prof. Dr F.E. van Eeuwijk, Wageningen University
Dr A. Legarra, National Institute of Agricultural Research, Toulouse, France
Prof. Dr J. Jensen, Aarhus University, Tjele, Denmark
Dr E.F. Knol, Topigs Norsvin, Beuningen

This research was conducted under the joint auspices of the Graduate School of Science and Technology (GSST), Aarhus University and the Graduate School of Wageningen Institute of Animal Sciences (WIAS), Wageningen University and is part of the Erasmus Joint Doctorate Program ''EGS-ABG''.

Genomic selection for crossbred performance

Hadi Esfandyari

Thesis

submitted in fulfillment of the requirements for the joint degree of doctor between

Aarhus University

by the authority of the Head of Graduate School of Science and Technology

and

Wageningen University

by the authority of the Rector Magnificus, Prof. Dr A.P.J. Mol,

in the presence of the

Thesis Committee appointed by the Academic Board at Wageningen University and

the Head of Graduate School of Science and Technology at Aarhus University

to be defended in public

on Friday February 12, 2016

at 8.30 a.m. in the Aula, Wageningen University

Esfandyari, H.
Genomic selection for crossbred performance
182 pages

Joint PhD thesis, Aarhus University, Denmark and Wageningen University, the Netherlands (2016)
With summaries in English and Danish
ISBN: 978-94-6257-652-0

Abstract

Esfandyari, H. (2016). Genomic selection for crossbred performance

Joint PhD thesis, Aarhus University, Denmark and Wageningen University, the Netherlands

Crossbreeding programs are used intensively in livestock production systems. The aim of selective-breeding programs in many of these systems is to maximize crossbred performance (CP), where selection is carried out within pure-lines using data from purebred animals. However, selection based on performance of purebred parents may not maximize performance of their crossbred descendants due to the genetic and environmental differences between purebred and crossbred animals. Genomic selection (GS) can be used to select purebreds for CP and has some advantages, such as it does not require pedigree information on crossbreds and can make accommodating non-additive gene action easier. The overall objective of this PhD project was to assess the possibilities of using dominance in genomic crossbreeding programs. Dominance is important in crossbreeding programs as it is the likely genetic basis of heterosis. It is also expected to be one of the factors causing the genetic correlations between crossbred and purebred performance to be smaller than one. Using stochastic simulations, response to selection in a two-way crossbreeding system was investigated. Under the hypothesis that performance of crossbred animals differs from that of purebred animals due to dominance, it was found that a dominance model can be used for GS of purebred individuals for CP, without using crossbred data. Furthermore, results showed that, if the correlation of linkage disequilibrium phase between pure lines is high, accuracy of selection can be increased by combining the two pure lines into a single reference population to estimate marker effects. In addition, response to selection of crossbreds with either a purebred or crossbred training population under a dominance model was compared. It was found that response to selection in crossbreeding programs can be increased by training on crossbred genotypes and phenotypes. Moreover, if the reference population is sufficiently large and both pure lines are not very closely related, tracing the line origin of alleles in crossbreds improved the accuracy of genomic prediction. Finally, real data of purebred Landrace and Yorkshire pigs were analysed to compare the predictive ability of genomic prediction models with either additive, or both additive and dominance effects, when the validation criterion was CP. The results showed some gains in prediction accuracy for CP by including dominance and combining both pure lines into a single reference population for training. In conclusion, GS can be used for efficient selection of purebreds for CP by addressing

the factors that cause the genetic correlations between crossbreds and purebreds to be lower than one.

To my grandmother, deeply missed

Contents

1

General introduction

1.1 Crossbreeding

Animal genetic improvement programs involve two main methodologies for increasing the productivity of farm animals; selection of best animals within a breed or population and using the best breeds or breed combination through crossbreeding systems. 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 heterosis, known as hybrid vigor, which has been generally found to occur in swine (Johnson, 1980; Toelle and Robison, 1983; Mclaren et al., 1987), poultry (Bell et al., 1950; Kosba, 1978), sheep (Fahmy, 1970; Fogarty, 1981; Farid, 1989), beef (Gregory et al., 1965; Gregory et al., 1992), and dairy cattle (Ahlbornbreier and Hohenboken, 1991; Lopez-Villalobos et al., 2000; Penasa et al., 2010). Even more than heterosis, crossbreeding is sought for breed complementarity, which is to combine different desirable characteristics from pure lines or breeds (Cundiff et al., 1986). 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 descendants of commercial crossbreds.

1.2 Non-additive genetic effects and heterosis

Despite the rediscovery of heterosis about a century ago and the suggestion of various genetic models to explain this phenomenon, little consensus has yet been reached about the genetic basis of heterosis (Xiao et al., 1995; Birchler et al., 2006; Lippman and Zamir, 2007). The most prominent genetic hypotheses to explain heterosis are the ‘dominance’ and ‘overdominance’ hypotheses. The dominance hypothesis attributes heterosis to canceling of deleterious or inferior recessive alleles contributed by one parent, by beneficial or superior dominant alleles contributed by the other parent in the heterozygous genotypes at different loci (Bruce, 1910; Jones, 1917). 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 (Falconer and Mackay, 1996).

The overdominance hypothesis attributes heterosis to the superior fitness of heterozygous genotypes over homozygous genotypes at a single locus (East, 1908; Shull, 1908). 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). 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).

Apart from the two main theories which have been proposed to explain the genetic basis of heterosis, epistasis is also considered to be associated with heterosis (Schnell and Cockerham, 1992; Li et al., 2001). Epistasis refers to interaction between alleles of two or more different loci. In summary, all genetic hypotheses suggest that the contribution of many genes is responsible for the more vigorous phenotypes of hybrids over parental lines. This also implies that positive and negative effects of various loci might compensate each other, which makes it difficult to support one hypothesis over the other (Melchinger, 1999).

Besides the genetic models, 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% (Stuber, 1994), while heterosis in animal crosses is about 0–10% for growth traits and 5–25% for fertility traits (Kosba, 1978; Johnson, 1980; Koch et al., 1985; Gregory et al., 1992).

Falconer and Mackay (1996) comprehensively formulated how the dominance and the difference in allele frequency between parental populations jointly affect the level of heterosis in a cross. In summary, for a single bi-allelic locus that has effect d at the heterozygous genotype, given Hardy-Weinberg equilibrium holds in parental populations and the sires are randomly mated to the dams, the amount of heterosis at this locus, expressed as the difference between the crossbred and the average parental means, is $\mathrm{H} = \mathrm{d}\Delta\mathrm{q}^2$ where $\Delta\mathrm{q}$ is difference in allele frequency between sires and dams. In the absence of epistatic interaction between loci, total heterosis is the additive combination of the heterosis effects of the loci that jointly affect the trait, $\mathrm{H} = \sum \mathrm{d}\Delta\mathrm{q}^2$.

It can be concluded from the above equation that the difference in allele 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 allele 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; Estelle et al., 2008; Li et al., 2008), though it may be important to some traits (Meffert et al., 2002; Abasht and Lamont, 2007).

1.3 Importance of non-additive effects in animal breeding

Genetic evaluations in livestock breeding programs are generally based on additive genetic models, e.g. sire or animal models. Total genetic values of animals may also contain non-additive components (Falconer and Mackay, 1996). Although non-additive genetic effects are not directly transmitted from parents to offspring, knowledge about these effects can be beneficial. In particular, dominance as a non-additive effect, is of theoretical and practical importance, because it is heavily used in crosses of animal breeds. In addition, inclusion of dominance effects in models to predict genomic breeding values could increase prediction accuracy and decrease the bias of estimated breeding values (Toro and Varona, 2010; Su et al., 2012). Furthermore, a model that includes additive and non-additive genetic effects could be beneficial for exploiting specific combining ability. Breeders should continue to select for additive merit but can also improve non-additive merit by considering interactions in mating programs (Van Raden, 2006). Sun et al. (2013) showed that mating programs that include dominance effects can increase expected progeny value for milk yield compared with mating programs that only include additive genetic effects. Dominance effects could also be included in mating programs to estimate inbreeding losses more precisely (Toro and Varona, 2010).

Several studies have estimated non-additive variances in livestock using traditional pedigree information (Hoeschele, 1991; Fuerst and Solkner, 1994; Misztal et al., 1997; Culbertson et al., 1998; Palucci et al., 2007) and reported a small but significant non-additive variance. However, it is difficult to estimate non-additive variance because it is often, at least partially, confounded with other effects such as common environment or maternal effects. Also, there is a lack of informative pedigrees, typically with large full-sib families, which are needed for accurate estimates of dominance effects (Misztal et al., 1998). In view of this, it is not surprising that most genetic evaluation systems use an additive model and ignore non-additive effects, especially considering that their aim is to estimate breeding values or additive genetic values. In addition, Hill et al. (2008) argued that even if gene effects are not additive, most of the genetic variance is still expected

to be additive variance. However, the recent advent of dense SNP panels, has ignited interest in the prediction of non-additive genetic effects (Su et al., 2012; Ertl et al., 2014; Lopes et al., 2014; Moghaddar et al., 2014; Sun et al., 2014). In fact, the availability of SNP genotypes represent a new opportunity to estimate non-additive effects at individual loci and to estimate non-additive variances.

1.4 Conventional methods for selection of purebreds for crossbred performance

The aim of selective-breeding programs in many of livestock production systems is to maximize crossbred performance (CP), where selection is carried out within pure-lines using data from purebred animals (Wei and Steen, 1991). However, traits that are evaluated in purebred populations may be genetically different from traits at the commercial production level because the genetic correlations between crossbred and purebred performance (r_{pc}) are usually less than one (Wei and Vanderwerf, 1994; Dekkers, 2007). Evidence for r_{pc} values less than one has been observed in livestock species (Lutaaya et al., 2001; Zumbach et al., 2007). Deviations of r_{pc} from 1 are often caused by genotype by environment (G×E) interactions and non-additive (particularly dominance) genetic effects (Wei et al., 1991). Thus, when r_{pc} is substantially lower than 1, the conventional strategy that relies on selection of purebreds or pure lines on their own performance (PLS) is not effective to improve the CP. Several methods have been proposed as alternatives to pure line selection to obtain greater response in crossbred populations. These methods can be classified into three groups: reciprocal recurrent selection (RRS), combined crossbred and purebred selection (CCPS) and genomic selection (GS).

Reciprocal recurrent selection (RRS), originally proposed by Comstock et al. (1949) is a cyclical breeding procedure designed for the genetic improvement of quantitative traits and has been applied to a variety of animal and plant species (i.e., poultry, swine, maize, etc.). In this procedure, nucleus individuals are selected based on the hybrid performance of their sibs or descendants. Even though, RRS can more efficiently exploit non-additive genetic variance than PLS (Comstock et al., 1949; Bell et al., 1950), the practical value of RRS, however, was not as encouraging as expected in most of the experiments (Calhoon and Bohren, 1974; Wei and van der Steen, 1991).

CCPS aims to maximize the genetic response by using information on both purebred and crossbred performance (Wei and Steen, 1991; Lo et al., 1993)). 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 CP as genetically different traits and use selection index theory to estimate the purebred breeding values for CP (Wei and Vanderwerf, 1994; Bijma and van Arendonk, 1998). Alternatively, genetic evaluations of purebreds for CP 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 CCPS increases the probability of co-selection of family members (Bijma et al., 2001; Dekkers, 2007). In addition, to implement CCPS requires routine collections of crossbred phenotypes and pedigree that can link crossbred descendants 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.

1.5 Genomic selection in crossbreeding schemes

Genomic selection proposed by Meuwissen et al. (2001) is an extension of marker-assisted selection (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 a training population, they can be applied to predict the breeding values of genotyped animals at an early stage without own phenotypic records. 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 accurately the 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 predicted SNP effects can be used over a few generations with limited loss in prediction accuracy (Habier et al., 2007; Sonesson and Meuwissen, 2009). With such advantages, recent studies have shown encouraging results of GS in the selection of purebreds

(Meuwissen et al., 2001; Muir, 2007; Hayes et al., 2009; VanRaden et al., 2009; Habier et al., 2011).

Recent studies have shown that GS is also an appealing method to select purebreds for CP, particularly when the crossbreds are used for training (Dekkers, 2007; Piyasatian et al., 2007; Ibanez-Escriche et al., 2009; Kinghorn et al., 2010; Toosi et al., 2010; Zeng et al., 2013). As compared to alternative methods that use covariance theory, such as combined crossbred and purebred selection proposed by (Wei and 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 (Daetwyler et al., 2007; Dekkers, 2007), and it does not require a systematic collection of pedigree that connects crossbreds to purebreds. Dekkers (2007) demonstrated that MAS or GS with marker effects derived from the commercial crossbred population led to substantially higher crossbred response and a lower rate of inbreeding compared to CCPS and PLS when the estimation of marker effects was accurate.

For the implementation of GS in crossbreeding programs, several studies have focused on crossbred data for the prediction of marker effects and also an additive model has been used (Dekkers, 2007; Ibanez-Escriche et al., 2009; Toosi et al., 2010). Given that the SNP effects in a crossbred population originate from 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 of purebreds for CP. 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 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 is 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.

It has been argued that dominance is the likely genetic basis of heterosis (Falconer and Mackay, 1996; Charlesworth and Willis, 2009). Therefore explicitly including dominance in the GS model may be beneficial for selection of purebreds for CP. A model that explicitly includes dominance effects (the dominance model)

provides estimates of both additive and dominance effects and therefore enables the computation of allele substitution effects using appropriate allele frequencies. Once estimates of SNP effects are obtained from training, they can be successively applied over generations with updated allele frequencies to develop prediction equations specific to that generation. Compared to the BSAM model that breed origin of SNP alleles must be known or inferred, such knowledge is not needed for the dominance model. Zeng et al. (2013) compared additive and dominance models for GS of purebred animals for CP and came to the conclusion that, when dominance is the sole driver of heterosis, using a dominance model for GS results in greater cumulative response to selection of purebred animals for CP than either BSAM or the additive model. However, based on their simulation study, the extent of this additional response to selection depended on the size of dominance effects at the QTL and the power of detection of dominance effects through SNP genotypes.

1.6 Aim and outline of this thesis

The aim of this PhD project was to assess the possibilities of using non-additive genetic effects in the selection of animals to use in cross breeding programs.

If improvement is to be continued in a breeding programme, or if there is to be the opportunity to redirect the programme to improve different traits or respond to environmental or production constraints, genetic variability and in particular additive genetic variance (V_A) has to be present. Genetic variation is lost as a result of sampling or genetic drift, due to finite population size, and as a result of selection. Chapter 2 of this thesis explores the effect of presence of non-additive effects in genetic models and assess their importance in medium to long term selection experiments. It was investigated how the genetic variance and genetic gain are affected by the presence of non-additive genetic effects, using BLUP-EBV and phenotypes as selection criteria.

The next two chapters of this thesis deal with GS of purebreds for CP. In both chapters, conclusions were based on simulated data. In the previous studies on the selection of purebred animals for CP (Ibanez-Escriche et al., 2009; Zeng et al., 2013) crossbred data have been used to estimate marker effects, which requires collecting genotypes and phenotypes on crossbred animals. This can substantially increase the required financial investment of the breeding program, since crossbred animals are usually not individually identified and individual performance is not recorded. It is interesting to evaluate the potential benefit of GS within purebred lines when the objective is to improve performance of crossbred animals,

by using marker effects that are estimated from pure line data. In other words, additive and dominance effects of alleles can be estimated from pure line data, and subsequently breeding values for CP can be estimated by using the appropriate allele frequencies. Thus, in chapter 3, the aim was to investigate the benefits of GS of purebred animals for CP based on purebred information, compared to traditional selection for purebred performance. The effect of the correlation of LD phase between the two pure breeds on the consequences of combining two pure lines to a single reference population to estimate marker effects was also investigated.

Previous studies on the implementation of GS in crossbreeding programs focused either on crossbred (Ibanez-Escriche et al., 2009; Zeng et al., 2013) or purebred (Esfandyari et al., 2015) data for prediction of marker effects, without explicitly comparing responses to selection obtained with both methods. Therefore, in chapter 4 the aim was to compare response to selection in crossbreds by simulating a two-way crossbreeding program with either a purebred or crossbred training population under a dominance model. In addition, the benefit of GS of purebreds for CP using a crossbred training population was compared when breed origin of alleles was either accounted for or not in the calculation of breeding values.

In chapter 5, to confirm the findings of the simulation study in chapter 3, the aim was to compare the predictive ability of genomic prediction models with either additive, or both additive and dominance effects, when the validation criterion was CP. For this purpose the real genomic data of purebred Landrace and Yorkshire pigs were analysed. Finally, in chapter 6, I discuss the relevance of my findings and place them in a broader context. I reflect on the advantages and shortcomings of GS in crossbreeding schemes and discuss future perspectives.

References

Abasht, B., and S. J. Lamont. 2007. Genome-wide association analysis reveals cryptic alleles as an important factor in heterosis for fatness in chicken F-2 population. Anim Genet 38: 491-498.

Ahlbornbreier, G., and W. D. Hohenboken. 1991. Additive and nonadditive genetic-effects on milk-production in dairy-cattle - evidence for major individual heterosis. J Dairy Sci 74: 592-602.

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

Bijma, P., and J. A. M. van Arendonk. 1998. Maximizing genetic gain for the sire line of a crossbreeding scheme utilizing both purebred and crossbred information. Anim Sci 66: 529-542.

Birchler, J. A., H. Yao, and S. Chudalayandi. 2006. Unraveling the genetic basis of hybrid vigor. Proceedings of the National Academy of Sciences of the United States of America 103: 12957-12958.

Boysen, T. J., J. Tetens, and G. Thaller. 2010. Detection of a quantitative trait locus for ham weight with polar overdominance near the ortholog of the callipyge locus in an experimental pig F-2 population. J Anim Sci 88: 3167-3172.

Bruce, A. B. 1910. The Mendelian theory of heredity and the augmentation of vigor. Science 32 627–628.

Charlesworth, D., and J. H. Willis. 2009. Fundamental concepts in genetics: The genetics of inbreeding depression. Nat Rev Genet 10: 783-796.

Comstock, R. E., H. F. Robinson, and P. H. Harvey. 1949. A breeding procedure designed to make maximum use of both general and specific combining ability. Agron J 41: 360-367.

Crow, J. F., and M. Kimura. 1970. An introduction to population genetics theory.

Culbertson, M. S. et al. 1998. Estimation of dominance variance in purebred Yorkshire swine. J Anim Sci 76: 448-451.

Cundiff, L. V., K. E. Gregory, R. M. Koch, and G. E. Dickerson. 1986. Genetic diversity among cattle breeds and its use to increase beef production efficiency in a temperate environment. In: 3rd World Congress on Genetics Applied to Livestock Production, Lincoln, USA

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

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

East, E. M. 1908. Inbreeding in corn, Reports of the Connecticut Agricultural Experiment Station for Years 1907-1908, USA.

East, E. M. 1936. Heterosis. Genetics 21: 375-397.

Ertl, J. et al. 2014. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genetics Selection Evolution 46.

Esfandyari, H., A. C. Sorensen, and P. Bijma. 2015. Maximizing crossbred performance through purebred genomic selection. Genetics Selection Evolution 47.

Estelle, J. et al. 2008. A quantitative trait locus genome scan for porcine muscle fiber traits reveals overdominance and epistasis. J Anim Sci 86: 3290-3299.

Fahmy, M. H. 1970. Homogametic heterosis in crossbreeding experiments with Sheep. Can J Anim Sci 50: 377-&.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. 4 ed. Pearson

Farid, A. 1989. Direct, maternal and heterosis effects for slaughter and carcass characteristics in 3 breeds of fat tailed sheep. Livest Prod Sci 23: 137-162.

Fogarty, N. M. 1981. Heterosis and genetic-parameters for reproduction in sheep. J Aust I Agr Sci 47: 219-220.

Fuerst, C., and J. Solkner. 1994. Additive and non-additive genetic variances for milk-yield, fertility, and lifetime performance traits of dairy-cattle. J Dairy Sci 77: 1114-1125.

Gregory, K. E., L. V. Cundiff, and R. M. Koch. 1992. Breed Effects and Heterosis in Advanced Generations of Composite Populations for Reproduction and Maternal Traits of Beef-Cattle. J Anim Sci 70: 656-672.

Gregory, K. E. et al. 1965. Heterosis in preweaning traits of beef cattle. J Anim Sci 24: 21-25.

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

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. Bmc Bioinformatics 12.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges (vol 92, pg 433, 2009). J Dairy Sci 92: 1313-1313.

Hill, W. G., M. E. Goddard, and P. M. Visscher. 2008. Data and theory point to mainly additive genetic variance for complex traits. Plos Genet 4.

Hoeschele, I. 1991. Additive and non-additive genetic variance in female fertility of Holsteins. J Dairy Sci 74: 1743-1752.

Ibanez-Escriche, N., R. L. Fernando, A. Toosi, and J. C. M. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. Genetics Selection Evolution 41.

Ishikawa, A. 2009. Mapping an Overdominant Quantitative Trait Locus for Heterosis of Body Weight in Mice. J Hered 100: 501-504.

Johnson, R. K. 1980. Heterosis and breed effects in swine. North Central Regional Pub.

Jones, D. F. 1917. Dominance of linked factors as a means of accounting for heterosis. . In: Proc. Natl. Acad. Sci. , USA p310–312.

Kinghorn, B. P., J. M. Hickey, and J. H. J. van der Werf. 2010. Reciprocal recurrent genomic selection for total genetic merit in crossbred individuals. In: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production: 1-6 August 2010; Leipzig. Paper 36; 2010.

Koch, R. M., G. E. Dickerson, L. V. Cundiff, and K. E. Gregory. 1985. Heterosis retained in advanced generations of crosses among Angus and Hereford cattle. J Anim Sci 60: 1117-1132.

Kosba, M. A. 1978. Heterosis and phenotypic correlations for shank length, body-weight and egg-production traits in Alexandria strains and their crosses with Fayoumi chickens. Beitr Trop Landwirt 16: 187-198.

Li, L. Z. et al. 2008. Dominance, overdominance and epistasis condition the heterosis in two heterotic rice hybrids. Genetics 180: 1725-1742.

Li, Z. K. et al. 2001. Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. I. Biomass and grain yield. Genetics 158: 1737-1753.

Lippman, Z. B., and D. Zamir. 2007. Heterosis: revisiting the magic. Trends Genet 23: 60-66.

Lo, L. L., R. L. Fernando, and M. Grossman. 1993. Covariance between relatives in multibreed populations - Additive-model. Theoretical and Applied Genetics 87: 423-430.

Lopes, M. S., J. W. M. Bastiaansen, L. Janss, H. Bovenhuis, and E. F. Knol. 2014. Using SNP markers to estimate additive, dominance and imprinting genetic variance. In: 10th World Congress on Genetics Applied to Livestock Production, Vancouver, BC, Canada

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

Luo, L. J. et al. 2001. Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. II. Grain yield components. Genetics 158: 1755-1771.

Lutaaya, E. et al. 2001. Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. J Anim Sci 79: 3002-3007.

Mclaren, D. G., D. S. Buchanan, and R. K. Johnson. 1987. Individual heterosis and breed effects for postweaning performance and carcass traits in 4 breeds of swine. J Anim Sci 64: 83-98.

Meffert, L. M., S. K. Hicks, and J. L. Regan. 2002. Nonadditive genetic effects in animal behavior. Am Nat 160: S198-S213.

Melchinger, A. E. 1999. Genetic diversity and heterosis, American Society of Agronomy, Inc and Crop Science Society of America, Inc, USA.

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

Misztal, I., T. J. Lawlor, and N. Gengler. 1997. Relationships among estimates of inbreeding depression, dominance and additive variance for linear traits in Holsteins. Genetics Selection Evolution 29: 319-326.

Misztal, I. et al. 1998. Studies on the value of incorporating the effect of dominance in genetic evaluations of dairy cattle, beef cattle and swine. Biotechnol Agron Soc Environ 2: 227-233.

Moghaddar, N., A. A. Swan, and J. H. J. van der Werf. 2014. Comparing genomic prediction accuracy from purebred, crossbred and combined purebred and crossbred reference populations in sheep. Genetics Selection Evolution 46.

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

Palucci, V., L. R. Schaeffer, F. Miglior, and V. Osborne. 2007. Non-additive genetic effects for fertility traits in Canadian Holstein cattle (Open Access publication). Genetics Selection Evolution 39: 181-193.

Penasa, M. et al. 2010. Heterosis effects in a black and white dairy cattle population under different production environments. Livest Sci 131: 52-57.

Piyasatian, N., R. L. Fernando, and J. C. M. Dekkers. 2007. Genomic selection for marker-assisted improvement in line crosses. Theoretical and Applied Genetics 115: 665-674.

Schnell, F. W., and C. C. Cockerham. 1992. Multiplicative vs arbitrary gene-action in heterosis. Genetics 131: 461-469.

Shull, G. H. 1908. The composition of field of maize, Am. Breed. Assn. Rep., USA.

Sonesson, A. K., and T. H. E. Meuwissen. 2009. Testing strategies for genomic selection in aquaculture breeding programs. Genetics Selection Evolution 41.

Stuber, C. W. 1994. Heterosis in plant breeding, in plant breeding reviews. In: U. Purdue University (ed.) Plant Breeding Reviews No. 12. John Wiley & Sons, Inc., Oxford, UK.

Su, G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund. 2012. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PloS one 7: e45293.

Sun, C., P. M. VanRaden, J. B. Cole, and J. R. O'Connell. 2014. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. PloS one 9.

Sun, C., P. M. VanRaden, J. R. O'Connell, K. A. Weigel, and D. Gianola. 2013. Mating programs including genomic relationships and dominance effects. J Dairy Sci 96: 8014-8023.

Toelle, V. D., and O. W. Robison. 1983. Breed prenatal, breed postnatal and heterosis effects for post-weaning traits in swine. J Anim Sci 57: 313-319.

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

Toro, M. A., and L. Varona. 2010. A note on mate allocation for dominance handling in genomic selection. Genetics, selection, evolution : GSE 42: 33.

Van Raden, P. M. 2006. Predicting genetic interactions within and across breeds. In: 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, MG, Brazil

VanRaden, P. M. et al. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92: 16-24.

Wei, M., and H. Steen, van der.,. 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 J. H. J. Vanderwerf. 1994. Maximizing genetic response in crossbreds using both purebred and crossbred information. Anim Prod 59: 401-413.

Wei, M., J. H. J. Vanderwerf, and E. W. Brascamp. 1991. Relationship between purebred and crossbred parameters .2. genetic correlation between purebred and crossbred performance under the model with 2 Loci. J Anim Breed Genet 108: 262-269.

Xiao, J. H., J. M. Li, L. P. Yuan, and S. D. Tanksley. 1995. Dominance is the major genetic-basis of heterosis in Rice as revealed by QTL analysis using molecular markers. Genetics 140: 745-754.

Zeng, J., A. Toosi, R. L. Fernando, J. C. Dekkers, and D. J. Garrick. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genetics, selection, evolution : GSE 45: 11.

Zumbach, B. et al. 2007. Genetic correlations between two strains of Durocs and crossbreds from differing production environments for slaughter traits. J Anim Sci 85: 901-908.

2

Medium to long-term effects of selection in finite locus models with non-additive effects

Hadi Esfandyari[1,2*], Mark Henryon[3], Peer Berg[4], Jorn Rind Thomasen[1,5], Piter Bijma[2], Anders Christian Sørensen[1]

[1] Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Denmark
[2] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands
[3] Seges, Danish Pig Research Centre, Copenhagen, 1609, Denmark
[4] Nordic Genetic Resource Center, Ås, Norway
[5] VikingGenetics, Ebeltoftvej 16, 8860 Assentoft, Denmark

To be submitted

Abstract

Background: Under the finite-locus model in the absence of mutation, the additive genetic variation is expected to decrease when directional selection is acting on a population, according to quantitative-genetic theory. However, some theoretical studies of selection suggest that the level of additive variance can be sustained or even increased when non-additive genetic effects are present. We tested the hypothesis that finite-locus models with both additive and non-additive genetic effects maintain more additive genetic variance (V_A) and realize larger medium-to-long term genetic gains than models with only additive effects when the trait under selection is subject to truncation selection.

Methods: Four genetic models that included additive, dominance, and additive-by-additive epistatic effects were simulated. The simulated genome for individuals consisted of 25 chromosomes, each with a length of 1M. One hundred bi-allelic QTL, four on each chromosome, were considered. In each generation, 100 sires and 100 dams were mated, producing five progeny per mating. The population was selected for a single trait (h2=0.1) for 100 discrete generations with selection on phenotype or BLUP-EBV.

Results: V_A decreased with directional truncation selection even in presence of non-additive genetic effects. Non-additive effects influenced long-term response to selection and among genetic models additive gene action had highest response to selection. In addition, in all genetic models, BLUP-EBV resulted in a greater fixation of favourable and unfavourable alleles and higher response than phenotypic selection.

Conclusions: In the schemes we simulated, the presence of non-additive genetic effects had little effect in changes of additive variance and $\mathbf{V_A}$ decreased by directional selection.

2.1 Introduction

Dominance and epistatic effects have been detected in many experimental populations (Carlborg and Haley, 2004) and these allelic interactions can influence the amount of additive genetic variance (V_A) in populations (Goodnight, 1987, Cheverud and Routman, 1995, Hansen and Wagner, 2001, Caballero and Toro, 2002, Barton and Turelli, 2004). It has been argued that epistatic variance may be "converted" into additive variance by genetic drift when a population passes through a population bottleneck (Goodnight, 1995, Cheverud and Routman, 1996, Cheverud et al., 1999, Lopez-Fanjul et al., 2002, Barton and Turelli, 2004). However, this argument is not restricted to genetic drift. Changes in V_A occur with variations in the genetic background, and any process that changes allele frequencies, including selection, can change V_A (Hansen and Wagner, 2001). Gene interactions may also affect response to selection through a build-up of linkage disequilibrium associated with favourable gene combinations, since parents transmit not only half of the additive effects to offspring, but also a quarter of the pairwise epistatic effects (A×A) and smaller fractions of higher-order interactions (Lynch and Walsh, 1998). This suggests that some of the linkage disequilibrium built by epistatic selection can be converted into response to selection (Griffing, 1960).

The model most commonly used for genetic evaluation is the infinitesimal model. It assumes large numbers of genes affecting traits with each gene having a small additive effect (Fisher, 1918). In this model, selection does not change the allele frequency significantly at any individual locus, nor does it change the genetic variance except for non-permanent changes caused by gametic phase disequilibrium (the Bulmer effect; Bulmer (1971)). The assumptions of the infinitesimal model are incorrect. In practice, there are individual genes, sometimes with large effects, and many genes showing dominance and epistasis (Mackay, 2001b, a). An alternative to the infinitesimal model is a finite-locus model, which can accommodate non-additive inheritance. Under the finite-locus model, the additive genetic variation is expected to decrease when directional selection is acting on a population, according to quantitative-genetic theory (Crow and Kimura, 1970, Falconer and Mackay, 1996). However, some theoretical studies of selection suggest that the level of additive variance can be sustained or even increased when non-additive genetic effects are present, in a manner similar to the action of genetic drift (Fuerst et al., 1997, Carter et al., 2005). Experimental evidence for this phenomenon was found by Martinez et al. (2000), when they selected mice for body fat, and by Sorensen and Hill (1982), who selected D. melanogaster for abdominal bristle number. Furthermore, Carlborg et al. (2006) showed that epistatic interactions between four loci mediated a considerably higher response to

selection for growth in chicken than predicted by a single-locus model. In a simulation study, Hallander and Waldmann (2007) investigated changes in V_A in the presence of non-additive effects in a trait subjected to directional selection. They showed that by including dominance and epistatic effects, V_A was increased during the initial generations of selection. Fuerst et al. (1997) showed similar results by using a two-locus genetic model to simulate a trait with different levels of additive and non-additive genetic effects. In these simulation studies V_A increased by including non-additive effects, but it seems considering 2 to 4 loci with equal additive and dominance effects across all loci with initial frequency of 0.5 is not a realistic model of the underlying genes. In fact, finite-locus models used in these studies to test quantitative theory are too restrictive. Indeed, it is possible to fit less restrictive models with many genes each having a unique effect, thus allowing a range from genes of large to zero effect. These genes could display non-additive effects such as dominance or epistasis. In theory, such a model seems to agree more closely with what we know about the genetics of quantitative traits than the simple models (Goddard, 2001).

Previous studies for investigating effect of selection on additive genetic variance in presence of non-additive effects have focused on phenotypic selection (Fuerst et al., 1997, Carter et al., 2005, Hallander and Waldmann, 2007). An alternative to selection based only on the phenotypic record of the individual is selection based on best linear unbiased predictor (BLUP) of breeding value (Henderson, 1975), which uses records on all relatives, in addition to the individual's own record, in genetic evaluation. However, for a trait under selection, there is no theory to quantify the differences between phenotypic selection or BLUP-EBV selection when the trait is controlled both by additive and non-additive genetic effects. This work explores the effect of presence of non-additive effects in genetic models and to assess their importance in medium to long term selection experiments. For this objective we examine how the genetic variance and genetic gain are affected by the presence of non-additive genetic effects, using BLUP-EBV and phenotypes as selection criteria.

2.2 Methods

2.2.1 Procedure

We used stochastic simulation to estimate genetic gain and monitor changes in genetic-variance components generated by four genetic models, with two different selection criteria and two distributions of QTL effects. The models were applied to a population undergoing directional truncation selection for a single trait over 100 discrete generations. During each generation, the genetic gain and changes in

additive, dominance and epistasis variances were calculated. The simulations were carried out using a modified version of ADAM (Pedersen et al., 2009). For each scenario, 100 replicates were performed.

2.2.2 Genetic models

The four genetic models, two selection criteria, and two distributions of QTL effects are presented in Table 2.1. The first genetic model (A) assumed that the trait is controlled by additive-gene actions. In the second model (A/D), the trait was controlled by additive and dominance effects. In the third model (A/AA), additive and additive × additive epistatic effects were included. In the final model, a full-genetic model (A/D/AA) consisted of additive, dominance, and additive × additive epistatic effects.

Table 2.1 Values of initial ratio of non-additive variances on additive variance, QTL effects distribution and selection criteria in simulated scenarios. P stands for phenotypic selection.

Tested Parameters	**Values**	
Distribution of QTL effects	Mixture, Gamma	
Selection criteria	P, BLUP	
Genetic Model	V_D/V_A	V_{AA}/V_A
A	0	0
A/D	1/2	0
A/AA	0	1/4
A/D/AA	1/2	1/4

2.2.3 Selection criteria

Truncation selection was applied in each generation and the criterion for truncation selection was either the phenotypic observation of the individual or EBVs obtained from standard BLUP evaluations of phenotypic records and pedigree information.

2.2.4 Simulated genome and distributions of QTL effects

The simulated genome consisted of 25 chromosomes. One hundred QTL, 4 on each chromosome, were considered. All chromosomes had a length of 1M and the QTL were assumed to be positioned randomly on each chromosome following a uniform distribution. All QTL were bi-allelic and initial frequencies of the alleles followed a U-shaped distribution as suggested by Crow and Kimura (Crow and Kimura, 1970). The U-shaped distribution of gene frequencies explains why selection response does not decline in the first few generations, because selection increases the frequency of rare favorable alleles and, hence, increases the genetic

variance due to these loci, which compensates for the loss of variance caused by selection for common favorable alleles (Goddard, 2001).

Two distributions were fitted for QTL effects (a_i); either a gamma distribution (0.4, 1.66) or a mixture of a double exponential distribution and a normal distribution, i.e. $a_i \sim 0.95\,.\,L(0, u^2) + 0.05.\,N(0, (5u)^2)$

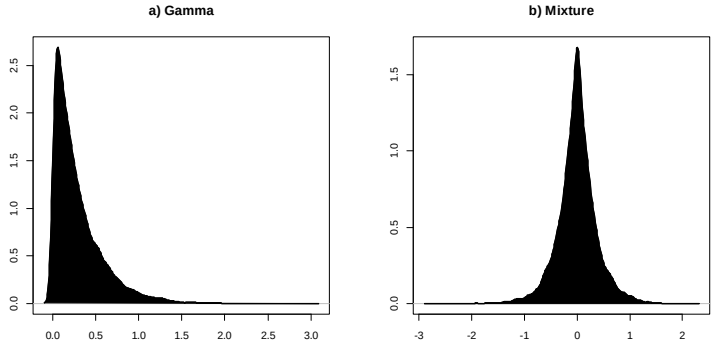


Figure 2.1 Distribution of QTL effects. (a) Gamma distribution, (b) Mixture of a double exponential distribution and a normal distribution.

(Figure 2.1). With appropriate u > 0. These distributions resulted in many QTL with small effects and few QTL with large effect (Bennewitz and Meuwissen, 2010). Dominance effects (d_i) were generated between alleles at the same locus and interactions between loci (aa_{ij}) only occurred for each pair of neighboring loci, and a locus had only interaction with one other locus. Dominance degrees (h_i) were normally distributed with mean $\mu_h = 0.5$, variance V_h=1, and they were independent of the additive effects. Dominance effects were then calculated as $d_i = |h_i.\,a_i|$, so they became dependent to the additive effects (Wellmann and Bennewitz, 2011). First-order (additive-by-additive) epistatic degrees (kij) also followed a normal distribution (0,1), and the epistatic effects were calculated as $aa_{ij} = |k_i|.\,a_i.\,a_j$, where aa_{ij} is the epistatic effect of the two adjacent loci, k_i is the epistatic degree, a_i and a_j are the additive effects at the first and second locus.

The total genotypic value of an individual was obtained by summing the genotypic contribution of each locus pair (Table 2.2). The genetic variance components depend on gene frequencies and values of different genetic effects. Following Fuerst et al. (1997), they were computed as:

$$V_A = \sum_{l}^{np} \{2p_1 q_1[a_1 + (q_1 - p_1)d_1 + (p_2 - q_2)aa_{12}]^2 + 2p_2 q_2[a_2 + (q_2 - p_2)d_2 + (p_1 - q_1)aa_{12}]^2\}_l \quad (1)$$

$$V_D = \sum_{l}^{np} \{4\, p_1^2 q_1^2 d_1^2 + 4p_2^2 q_2^2 d_2^2\}_l \quad (2)$$

$$V_{AA} = \sum_{l}^{np} \{4\, p_1 q_1 p_2 q_2 aa_{12}^2\}_l \quad (3)$$

Where V_A is the additive variance (variance of breeding values), np is the total number of pairs of loci, V_D is the dominance variance (variance of dominance deviations), V_{AA} is the additive-by-additive variance, a_1 and a_2 are the additive effect at loci 1 and 2, d_1 and d_2 are the dominance effects, and aa_{12} is the additive-by-additive effect at the pair (1 and 2). The gene frequencies of alleles A, B, a and b are p_1, p_2, q_1 and q_2.

Table 2.2 Genetic effects of genotypes for a pair of loci based on Fuerst et al (1997)

Genotype at locus 1	Genotype at locus 2		
	BB	**Bb**	**bb**
AA	$a_1+a_2+aa_{12}$	a_1+d_2	$a_1-a_2-aa_{12}$
Aa	d_1+a_2	d_1+d_2	d_1-a_2
aa	$-a_1+a_2-aa_{12}$	$-a_1+d_2$	$-a_1-a_2+aa_{12}$

2.2.5 Population structure

One hundred sires and 100 dams were in the base population (generation 0) of each scenario and were mated randomly to produce the first generation of offspring. Offspring produced by the base population were selected in generation 1 and the first generation of offspring from selected parents was produced in generation 2. Offspring produced in generation 100 were the result of 99 generations of selection. One hundred sires and 100 dams were selected in each generation, each sire was mated with one dam, and 5 full-sib offspring were produced per mating, resulting in 500 offspring in each generation.

2.2.6 Trait

The trait under selection had a narrow-sense heritability of 0.10 in the base population. The environmental values were sampled from a normal distribution for each individual with mean zero and variance $V_E = V_P - (V_A + V_D + V_{AA})$.

2.3 Results

2.3.1 Additive variance

In order to compare trends of V_A across scenarios, additive genetic variance in each scenario was scaled by dividing all values to initial additive variance of each scenario. Figure 2.2 shows changes of V_A over generations in four genetic models. There was a systematic pattern in the loss of V_A when comparing the four genetic models, and differences between genetic models were relatively small. In the A/AA model, the loss of V_A was faster than in the other genetic models. Initially, the decline in V_A was smallest in the A-model. In the long term, however, the amount of retained V_A was highest in full genetic model (A/D/AA). There were no differences in changes of V_A over generations between the two distributions fitted for QTL effects. However, two distinct selection criteria showed differences in reduction of V_A over time, as the BLUP-EBV selection criteria accelerated the reduction in V_A.

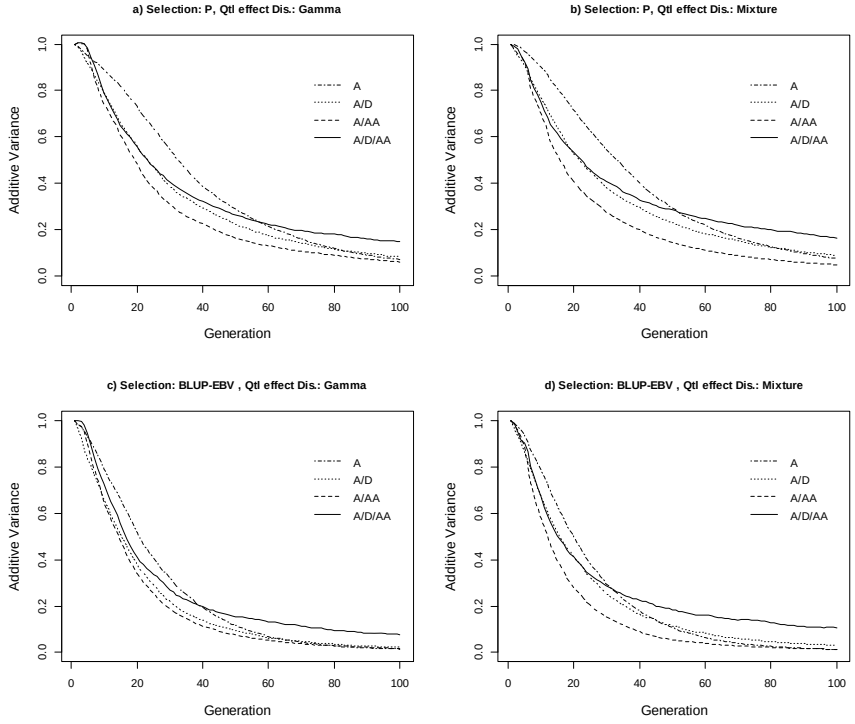


Figure 2.2 Changes in additive genetic variance (V_A) in four genetic models separated by selection criteria and QTL effect distribution. The plotted additive variance for each scenario are means from 100 replicates, standardized by the initial additive variance of each scenario. (a) Selection method: Phenotypic and QTL effect distribution: Gamma, (b) Selection method: Phenotypic and QTL effect distribution: Mixture, (c) Selection method: BLUP-EBV and QTL

effect distribution: Gamma, (d) Selection method: BLUP-EBV and QTL effect distribution: Mixture.
P stands for phenotypic selection and *Mixture* stands for a mixture of normal and double exponential distribution.

2.3.2 Dominance variance

In the two genetic models, A/D/AA and A/D, where dominance effects were included, V_D increased or was constant in the initial generations. Afterwards V_D decreased in the A/D model. Similar to V_A, changes in V_D followed the same trend for both distributions of QTL effects. Changes in V_D over time differed for the two selection criteria. When selection was based on BLUP-EBV, the reduction of V_D over time was faster than with phenotypic selection (Figure 2.3c and 2.3d). A striking result was that phenotypic selection hardly affected V_D in the A/D/AA model over the entire period of 100 generations.

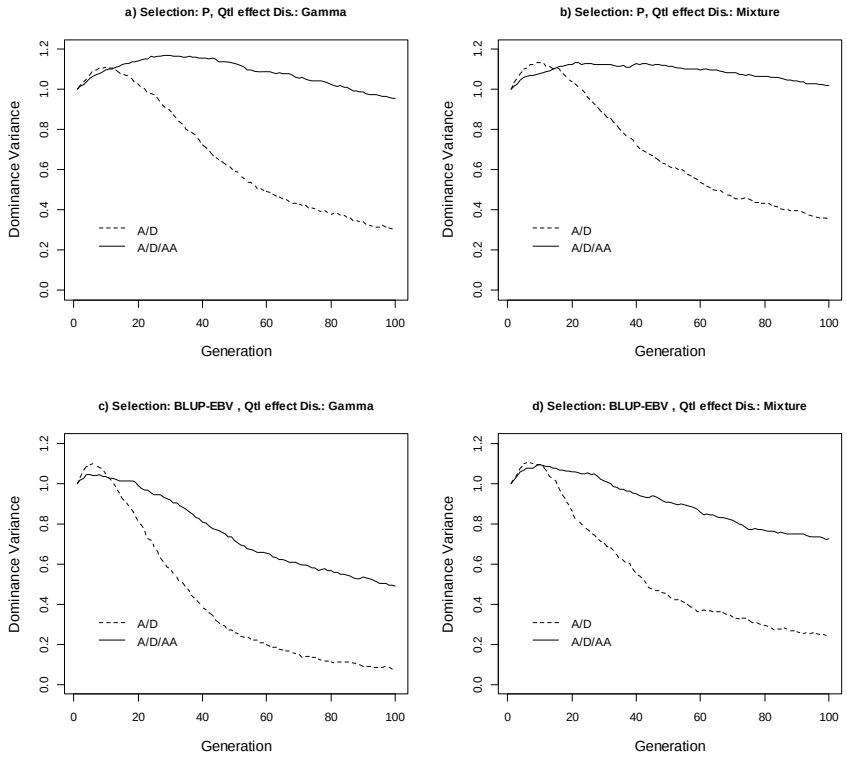


Figure 2.3 Changes in dominance genetic variance (V_D) in two genetic models (A/D/AA and A/D) separated by selection method and QTL effect distribution. The plotted dominance variance for each scenario are means from 100 replicates, standardized by the initial dominance variance of each scenario. (a) Selection method: Phenotypic and QTL effect distribution: Gamma, (b) Selection method: Phenotypic and QTL effect distribution: Mixture,

(c) Selection method: BLUP-EBV and QTL effect distribution: Gamma, (d) Selection method: BLUP-EBV and QTL effect distribution: Mixture.
P stands for phenotypic selection and *Mixture* stands for a mixture of normal and double exponential distribution.

2.3.3 Additive-by-additive genetic variance

The epistatic variance decreased in a similar way over the 100 generations of selection in both genetic models, both selection criteria, and both distributions of QTL-effects (Figure 2.4). The decrease was fastest for selection on BLUP-EBV, where V_{AA} approached its lowest level (~0) at generation 60. For phenotypic selection, longer time was needed for V_{AA} to vanish completely (Figure 2.4a and 2.4b).

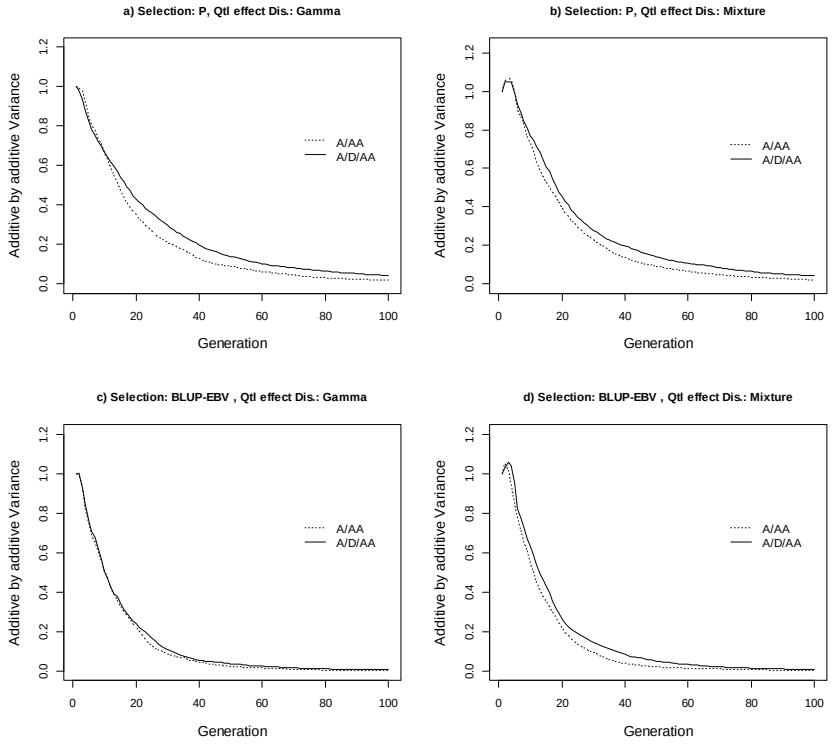


Figure 2.4 Changes in additive-by-additive genetic variance (V_{AA}) in two genetic models (A/D/AA and A/AA) separated by selection method and QTL effect distribution. The plotted additive by additive variance for each scenario are means from 100 replicates, standardized by the initial additive by additive variance of each scenario. (a) Selection method: Phenotypic and QTL effect distribution: Gamma, (b) Selection method: Phenotypic and QTL effect distribution: Mixture, (c) Selection method: BLUP-EBV and QTL effect distribution: Gamma, (d) Selection method: BLUP-EBV and QTL effect distribution: Mixture. P stands for phenotypic selection and Mixture stands for a mixture of normal and double exponential

distribution.

2.3.4 Response to selection

The mean observed genotypic value of individuals in each generation, expressed in initial additive genetic standard deviations and as a deviation from the initial mean, is plotted in Figure 2.5. As selection proceeds, the difference between genetic models became larger. In generation 100, the additive model had the highest cumulative response to selection, and the full genetic model (A/D/AA) had the lowest cumulative response in long term. Similar to results for the genetic variance components, there were no major differences in genetic gain between both distributions of QTL effects. When selection was based on BLUP-EBV, for all genetic models response plateaued earlier than for phenotypic selection.

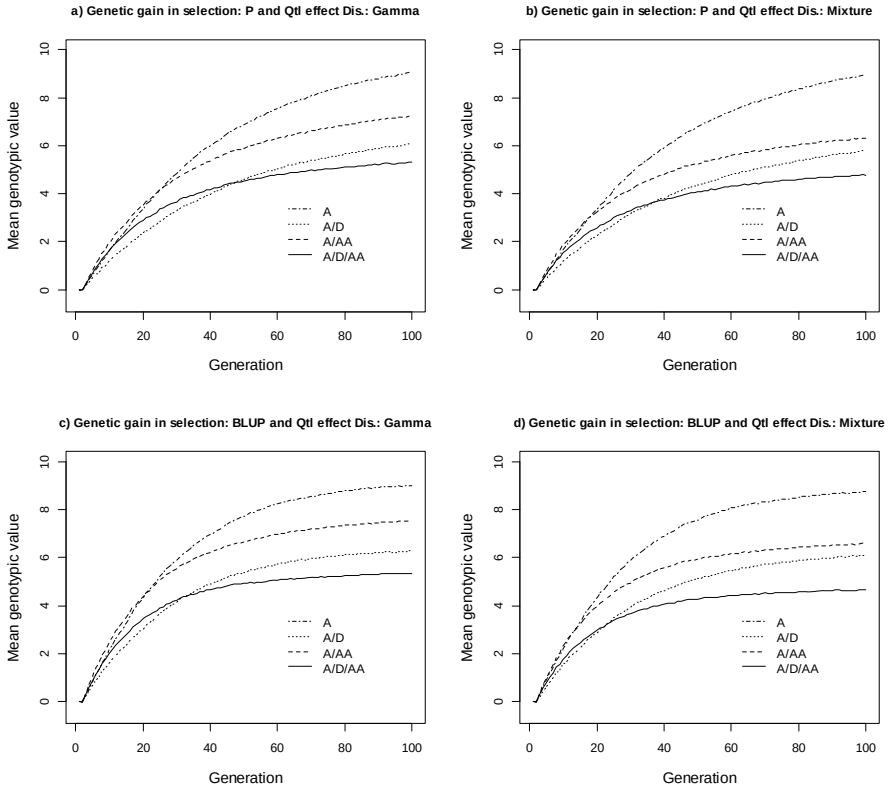


Figure 2.5 Response to selection over 100 generations of selection separated by selection method and QTL effect distribution. The plotted mean genotypic value for each scenario are means from 100 replicates, standardized by initial additive genetic standard deviations of each scenario. (a) Genetic gain in selection method: Phenotypic and QTL effect distribution: Gamma, (b) Genetic gain in selection method: Phenotypic and QTL effect distribution:

Mixture, (c) Genetic gain in selection method: BLUP-EBV and QTL effect distribution: Gamma, (d) Genetic gain in selection method: BLUP-EBV and QTL effect distribution: Mixture.
P stands for phenotypic selection and *Mixture* stands for a mixture of normal and double exponential distribution.

2.3.5 Fixation and loss of favorable QTL

The additive genetic model (A) had the highest percentage of fixation of favorable alleles at generation 100 when QTL-effects were gamma distributed (Figure 2.6; results for the mixture distribution were similar and are not shown). This result agrees with the highest response found for the additive model in Figure 2.5. The percentage of fixed favorable alleles decreased when more non-additive effects were added to the model. In all genetic models, the percentage of favorable alleles that became fixed was higher than percentage of favorable alleles that were lost. Of the total fraction of fixed alleles, however, the additive model had a greater proportion of loci fixed for the favorable allele. BLUP-EBV generated higher levels of fixation and loss of favorable alleles than phenotypic selection, which agrees with the earlier plateau of response seen with BLUP selection in Figure 2.5. In addition, the ratio of favorable fixed over total allele fixation was higher in BLUP-EBV than phenotypic selection.

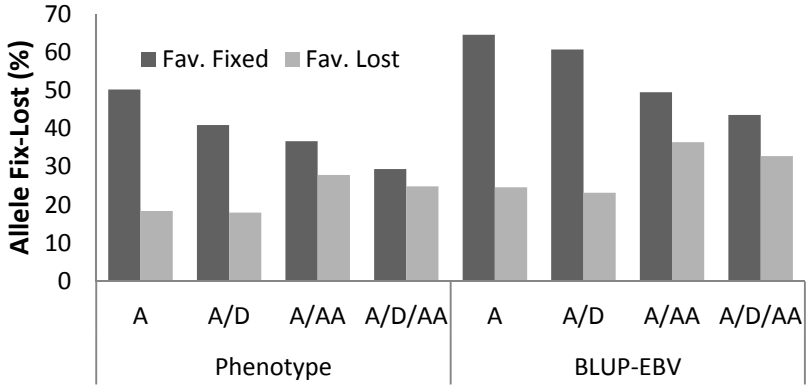


Figure 2.6 Percentage of allele fixation and lost at generation 100 in four genetic models separated by selection method when QTL effect distribution was gamma.
Fav. Fix: Percentage of QTL in which frequency of favorable allele is 1 in generation 100,
Fav. Lost: Percentage of QTL in which frequency of favorable allele is 0 in generation 100.

2.4 Discussion

Our findings did not support the hypothesis that finite-locus models with both additive and non-additive genetic effects maintain more V_A and realize larger medium-to-long term genetic gains than models with only additive effects when the trait under selection is subject to truncation selection. We used four genetic models to simulate a population undergoing directional truncation selection. In all four models, V_A decreased by directional selection, also in the presence of non-additive effects, but the rate at which variation decreased varied among genetic models and selection criteria.

2.4.1 Changes in variance

In our finite locus model, V_A decreased by selection, which is in agreement with results from other studies (Villanueva and Kennedy, 1990, Fuerst et al., 1997). Directional selection changes the mean of a trait and it can also change the variance. First, directional selection decreases V_A due to the generation of negative gametic phase disequilibrium (Bulmer, 1971). In the intermediate term, regardless of selection criteria and QTL distribution, genetic models with epistatic terms (A/AA and A/D/AA), showed faster reduction in V_A compared to the purely additive model (Figure 2.2). One explanation for this could be that the double homozygote genotype is more favoured by selection and that reduces V_A by inducing negative linkage disequilibrium among selected genes. If the selected genes are linked, the decay of linkage disequilibrium is delayed, and the reduction of V_A is enhanced (Nomura, 2005). However, it has been shown in several studies as well as experimental results that epistatic variance and to some extent dominance variance might convert to additive genetic variance (Fuerst et al., 1997, Hallander and Waldmann, 2007). In addition, any changes in V_A might depend on the ratio of VAA to V_A. If VAA is smaller than V_A, as in our case, then epistatic values for a pair of loci might be small and this will cause more decrease in V_A by selection in comparison of purely additive gene action (Mueller and James, 1983).

Second, a more important cause for changes in V_A is due to changes in allele frequency. If all alleles are at intermediate frequencies, it can be expected that variance will decline monotonically with time (assuming additivity), whereas if some are at low frequency, an initial increase in variance might be observed. If the distribution of frequencies is U shaped, as in this study, then the increase in variance due to alleles at low frequency might be expected to outweigh the decrease from those at high frequency (Hill and Bünger, 2010). However, analyses undertaken by Hill and Rasbash (1986) for finite populations indicate that the pattern of response and change in V_A is somewhat robust to the gene frequency

distribution and V_A decreases eventually as favourable alleles are moved to fixation.

Third, in finite populations, V_A also declines due to drift and, in the absence of selection, this decline can be predicted as $V_{At} = V_{A0}(1 - 1/2Ne)^t$ when assuming additive-gene action (Robertson, 1960), where V_{At} is additive variance at generation t, V_{A0} is intial additive variance and Ne is effective population size. We plotted V_A predicted by this formula versus observed V_A for the additive model (A) to see how they would differ (Figure 2.7). The large difference between predicted and observed V_A in Figure 2.7 is due to the fixation of favourable alleles by selection, as the formula based on effective population size (or inbreeding) assumes changes the allele frequency due to drift only. The lines for predicted V_A in Figure 2.7 show that, as expected, selection based on BLUP-EBV generates more fixation due to drift than phenotypic selection and as pointed out earlier fraction of favorable fixed over total fixation was higher in BLUP-EBV.

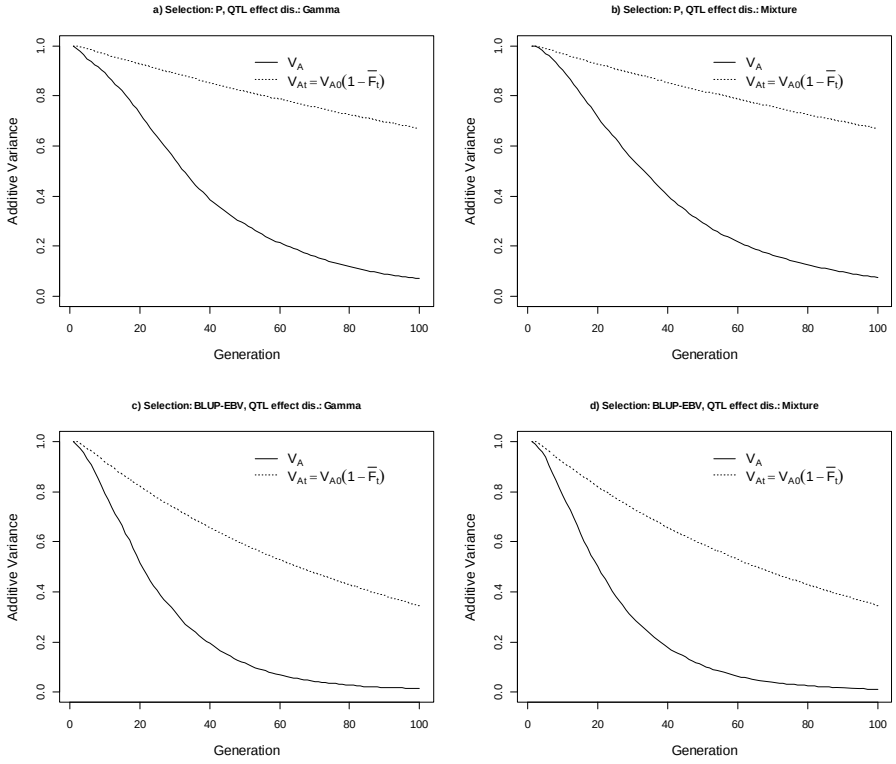


Figure 2.7 Comparison of changes in observed V_A in pure additive genetic model with V_A based on inbreeding calculated as $V_{At} = V_{A0}(1 - \bar{F}_t)$.
(a) Selection method: Phenotypic and QTL effect distribution: Gamma, (b) Selection method:

Phenotypic and QTL effect distribution: Mixture, (c) Selection method: BLUP-EBV and QTL effect distribution: Mixture, (d) Selection method: BLUP-EBV and QTL effect distribution: Mixture.
P stands for phenotypic selection and *Mixture* stands for a mixture of normal and double exponential distribution.

In contrast to V_A, which decreased by selection, V_D was preserved especially by phenotypic selection over generations. This maintenance of dominance variation can be explained by the occurrence of overdominance at some loci. In the A/D/AA genetic model, around 75 and 50% of loci showed overdominance for the mixture and gamma distribution, respectively. It has been shown that overdominance results in stabilizing selection, maintaining heterozygosity in the population rather than driving one allele to fixation.

2.4.2 Response to selection

Our results demonstrate that non-additive effects may affect response to selection in long term. Comparing genetic models, A and A/AA models had higher response in long term than the models having a dominance component (Figure 2.5). The reason for the greater response is probably due to the constellation of the genotypic values, where one double homozygote pair of loci has higher value. Hansen and Wagner (2001) [6]argued that non-additive effect such as directional epistasis will affect the response to selection due to systematic changes in the effects of alleles as their genetic background changes. On the other hand, if the epistatic interactions are random and non-directional as in this simulation, these effects will tend to cancel out or add random noise. However, over many generations, the dynamics of gene effect reinforcement and competition can become very complex, and may lead to substantial departures from simple additive response to selection (Carter et al., 2005). One deduction of these results could be that estimates of classical epistatic variance components are of little value in predicting response in short term, as these estimation do not distinguish between directional and non-directional forms of functional epistasis. Griffing (1960) showed when directional epistasis is present, gametic-phase disequilibrium increases the response to directional selection, with the response augmented by $\frac{S\sigma^2_{AA}}{2\sigma^2_p}$ where S is intensity of selection σ^2_{AA} is additive by additive variance and σ^2_p is phenotypic variance. This increase in rate of response has been termed the "Griffing effect''. Thus, in the presence of directional epistasis, disequilibrium is on the one hand expected to increase the rate of response, while it is also expected to decrease the rate of response by decreasing additive genetic variance (the Bulmer

effect). Based on a small simulation study, Mueller and James (1983) concluded that if epistatic variance is small relative to additive variance and the proportion of pairs showing epistasis is also small, the Bulmer effect dominates the Griffing effect, and disequilibrium reduces the response to selection (Walsh and Lynch, 2010).

We did not observe difference between two distributions fitted for QTL effects. Assuming that effects of mutant genes follow a gamma distribution but their frequencies are independent of their effects (i.e., a neutral model), Hill and Rasbash (1986) examined the influence of number and effects of mutant genes on response to selection and variance in response among replicates and found that the shape of the distribution of effects of mutant genes on the quantitative trait is not usually important, which is in agreement with our findings.

In the short term, selection response depends on additive effects and heritability. In long term, models including dominance (A/D and A/D/AA) had lower amount of response. This can be due to the rather frequent overdominance in this simulation, so loci with positive overdominance get stuck at intermediate frequencies. In fact, the presence of dominance might have some effect on the cumulated response to selection. Gill (1965b) in a simulation study showed that dominance (i.e. positive dominance effect) reduces selection advance. In fact, in gaining selection response, negative dominance effects are better than positive effects, and positive additive-by-additive effects are better than negative effects. (Fuerst et al., 1997).

To compare our results with the findings of Hallander and Waldmann (2007), we simulated their genetic model. When the initial allele frequency and additive effects were same across all loci, as in their analysis, we obtained similar results and including non-additive effects increased V_A in the initial generations. In our study, we found a different trend for V_A when including non-additive effects, as we had more loci, a U shaped initial frequency, and a different distribution of additive and non-additive effects.

In conclusion, in the schemes we simulated, additive genetic variance decreased by directional truncation selection, also in presence of non-additive genetic effects. The distribution of QTL effects underlying the trait and the presence of non-additive genetic effects had relatively small effects on the changes in additive variance. Response was relatively robust to non-additive genetic effects in short term, but dominance decreased long-term response to selection.

2.5 Conclusion

In conclusion, in the schemes we simulated, additive genetic variance decreased by directional truncation selection, also in presence of non-additive genetic effects. The distribution of QTL effects underlying the trait and the presence of non-additive genetic effects had relatively small effects on the changes in additive variance. Response was relatively robust to non-additive genetic effects in short term, but dominance decreased long-term response to selection.

References

Barton, N. H. and M. Turelli. 2004. Effects of genetic drift on variance components under a general model of epistasis. Evolution 58(10):2111-2132.

Bennewitz, J. and T. H. E. Meuwissen. 2010. The distribution of QTL additive and dominance effects in porcine F2 crosses. J Anim Breed Genet 127(3):171-179.

Bulmer, M. G. 1971. The effect of selection of genetic variability. The American Naturalist 105(943):201-211.

Caballero, A. and M. A. Toro. 2002. Analysis of genetic diversity for the management of conserved subdivided populations. Conserv Genet 3(3):289-299.

Carlborg, O. and C. S. Haley. 2004. Epistasis: too often neglected in complex trait studies? Nat Rev Genet 5(8):618-U614.

Carlborg, O., L. Jacobsson, P. Ahgren, P. Siegel, and L. Andersson. 2006. Epistasis and the release of genetic variation during long-term selection. Nat Genet 38(4):418-420.

Carter, A. J. R., J. Hermisson, and T. F. Hansen. 2005. The role of epistatic gene interactions in the response to selection and the evolution of evolvability. Theor Popul Biol 68(3):179-196.

Cheverud, J. M. and E. J. Routman. 1995. Epistasis and its contribution to genetic variance-components. Genetics 139(3):1455-1461.

Cheverud, J. M. and E. J. Routman. 1996. Epistasis as a source of increased additive genetic variance at population bottlenecks. Evolution 50(3):1042-1051.

Cheverud, J. M., T. T. Vaughn, L. S. Pletscher, K. King-Ellison, J. Bailiff, E. Adams, C. Erickson, and A. Bonislawski. 1999. Epistasis and the evolution of additive genetic variance in populations that pass through a bottleneck. Evolution 53(4):1009-1018.

Crow, J. F. and M. Kimura. 1970. An Introduction to Population Genetics Theory.

Falconer, D. S. and T. F. C. Mackay. 1996. Introduction To Quantitative Genetics. 4 ed. Longman.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52:399-433

Fuerst, C., J. W. James, J. Solkner, and A. Essl. 1997. Impact of dominance and epistasis on the genetic make-up of simulated populations under selection: A

model development. Journal of Animal Breeding and Genetics-Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie 114(3):163-175.

Gill, J. L. 1965a. Effects of finite size on selection advance in simulated genetic populations. Aust J Biol Sci 18(3):599-617.

Gill, J. L. 1965b. Selection and linkage in simulated genetic populations. Aust J Biol Sci 18(6):1171-1187.

Goddard, M. E. 2001. The validity of genetic models underlying quantitative traits. Livest Prod Sci 72(1-2):117-127.

Goodnight, C. J. 1987. On the effect of founder events on epistatic genetic variance. Evolution 41(1):80-91.

Goodnight, C. J. 1995. Epistasis and the increase in additive genetic variance - implications for phase-1 of Wrights Shifting-Balance process. Evolution 49(3):502-511.

Griffing, B. 1960. Theoretical consequences of truncation selection based on the individual phenotype. Aust J Biol Sci 13:307–343.

Hallander, J. and P. Waldmann. 2007. The effect of non-additive genetic interactions on selection in multi-locus genetic models. Heredity 98(6):349-359.

Hansen, T. F. and G. P. Wagner. 2001. Modeling genetic architecture: A multilinear theory of gene interaction. Theor Popul Biol 59(1):61-86.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31(2):423-447.

Hill, W. G. and L. Bünger. 2010. Inferences on the genetics of quantitative traits from long-term selection in laboratory and domestic animals. Pages 169-210 in Plant Breeding Reviews. John Wiley & Sons, Inc.

Hill, W. G. and J. Rasbash. 1986. Models of long-term artificial selection in finite population with recurrent mutation. Genetical Research 48(2):125-131.

Lopez-Fanjul, C., A. Fernandez, and M. A. Toro. 2002. The effect of epistasis on the excess of the additive and nonadditive variances after population bottlenecks. Evolution 56(5):865-876.

Lynch, M. and B. Walsh. 1998. Genetics and Analysis of Quantitative Traits.

Mackay, T. F. C. 2001a. The genetic architecture of quantitative traits. Annu Rev Genet 35:303-339.

Mackay, T. F. C. 2001b. Quantitative trait loci in Drosophila. Nat Rev Genet 2(1):11-20.

Martinez, V., L. Bunger, and W. G. Hill. 2000. Analysis of response to 20 generations of selection for body composition in mice: fit to infinitesimal model assumptions. Genet Sel Evol 32(1):3-21.

Mueller, J. P. and J. W. James. 1983. Effect on linkage disequilibrium of selection for a quantitative character with epistasis. Theor Appl Genet 65(1):25-30.

Nomura, T. 2005. Joint effect of selection, linkage and partial inbreeding on additive genetic variance in an infinite population. Biometrical J 47(4):527-540.

Pedersen, L. D., A. C. Sorensen, M. Henryon, S. Ansari-Mahyari, and P. Berg. 2009. ADAM: A computer program to simulate selective breeding schemes for animals. Livest Sci 121(2-3):343-344.

Sorensen, D. A. and W. G. Hill. 1982. Effect of short-term directional selection on genetic-variability - experiments with Drosophila-Melanogaster. Heredity 48(Feb):27-33.

Villanueva, B. and B. W. Kennedy. 1990. Effect of Selection on genetic-parameters of correlated traits. Theor Appl Genet 80(6):746-752.

Walsh, B. and M. Lynch. 2010. Short-term Changes in the variance. in evolution and selection of quantitative traits: I. Foundations. Vol. 2.

Wellmann, R. and J. Bennewitz. 2011. The contribution of dominance to the understanding of quantitative genetic variation. Genet Res 93(2):139-154.

3

Maximizing crossbred performance through purebred genomic selection

Hadi Esfandyari[1,2], Anders Christian Sørensen[1], Piter Bijma[2]

[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Denmark
[2]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands

GSE (2015) 47:16

Abstract

Background: In livestock production, many animals are crossbred, with two distinct advantages: heterosis and breed complementarity. Genomic selection (GS) can be used to select purebred parental lines for crossbred performance (CP). Dominance being the likely genetic basis of heterosis, explicitly including dominance in the GS model may be an advantage to select purebreds for CP. Estimated breeding values for CP can be calculated from additive and dominance effects of alleles that are estimated using pure line data. The objective of this simulation study was to investigate the benefits of applying GS to select purebred animals for CP, based on purebred phenotypic and genotypic information. A second objective was to compare the use of two separate pure line reference populations to that of a single reference population that combines both pure lines. These objectives were investigated under two conditions, i.e. either a low or a high correlation of linkage disequilibrium (LD) phase between the pure lines.

Results: The results demonstrate that the gain in CP was higher when parental lines were selected for CP, rather than purebred performance, both with a low and a high correlation of LD phase. For a low correlation of LD phase between the pure lines, the use of two separate reference populations yielded a higher gain in CP than use of a single reference population that combines both pure lines. However, for a high correlation of LD phase, marker effects that were estimated using a single combined reference population increased the gain in CP.

Conclusions: Under the hypothesis that performance of crossbred animals differs from that of purebred animals due to dominance, a dominance model can be used for GS of purebred individuals for CP, without using crossbred data. Furthermore, if the correlation of LD phase between pure lines is high, accuracy of selection can be increased by combining the two pure lines into a single reference population to estimate marker effects.

3.1 Introduction

One of the main limitations of many livestock breeding programs is that selection is carried out in purebred nucleus lines or breeds that are housed in high-health environments, whereas the goal of selection is to improve crossbred performance (CP) under field conditions. Due to genetic differences between purebred and crossbred animals and to environmental differences between nucleus and field conditions, performance of purebred parents can be a poor predictor of the performance of their crossbred descendants (Dekkers, 2007). Several methods have been proposed as alternatives to pure line selection to obtain greater response in crossbred populations. These methods can be classified into three groups: reciprocal recurrent selection, combined crossbred and purebred selection (CCPS) and genomic selection (GS).

Numerous studies have provided encouraging results regarding the application of GS in purebred populations (Meuwissen et al., 2001, Hayes et al., 2009). However, in livestock production systems, many animals are crossbred, with two distinct advantages i.e. heterosis and breed complementarity. Different GS models have been proposed and used to select purebred animals for CP (Dekkers, 2007, Ibanez-Escriche et al., 2009). Dekkers (2007) demonstrated that marker-assisted selection or GS with marker effects derived at the commercial crossbred level can lead to substantially higher gain in CP and a lower rate of inbreeding compared to CCPS when marker effects were estimated accurately.

If one accepts that GS is an appropriate tool to select animals for CP, then another issue to solve is: should marker effects be estimated from purebred or crossbred animals? Using simulated data on training populations that consisted of crossed or mixed breeds, Toosi et al. (2010) reported that the accuracy of GS was lower than when using purebred data for training, but not substantially lower. However, the GS model used in Toosi et al. (2010) assumed that single nucleotide polymorphism (SNP) allele effects were the same in all breeds. In crossbred populations, effects of SNPs may be breed-specific because the extent of linkage disequilibrium (LD) between SNPs and quantitative trait loci (QTL) can differ between breeds. SNP effects may also differ due to dominance and epistasis. Moreover, the LD may not be restricted to markers that are tightly linked to the QTL. Both these problems have been addressed by using a model with breed-specific effects of SNP alleles (BSAM) (Dekkers, 2007) and 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 was beneficial only when the parental breeds were distantly related and the number of SNPs was small relative to the size of the training population (Ibanez-Escriche et al., 2009).

In most studies, additive gene action or perfect knowledge of allele substitution effects or both are assumed (Ibanez-Escriche et al., 2009, Toosi et al., 2010). It has been argued that dominance is the likely genetic basis of heterosis (Falconer and Mackay, 1996), therefore explicitly including dominance in the GS model may be an advantage to select purebred animals for CP. With dominance, allele substitution effects and individual breeding values depend on allele frequency and, thus, change over time, which alters the ranking of individuals. This problem can be overcome by applying a dominance model, which provides estimates of both additive and dominance effects and, therefore, enables the computation of allele substitution effects using appropriate allele frequencies. Once SNP effects are estimated for the training population, they can be successively applied over generations with updated allele frequencies to develop prediction equations specific to a given generation (Zeng et al., 2013). Zeng et al. (2013) compared additive and dominance models for GS of purebred animals for CP and came to the conclusion that, when dominance is the sole driver of heterosis, using a dominance model for GS is expected to result in greater cumulative response to selection of purebred animals for CP than either BSAM or the additive model. The extent of this additional response to selection depended on the size of dominance effects at the QTL and the power of detection of dominance effects through SNP genotypes. The results of Zeng et al. (2013) suggested that in the presence of dominant gene action, compared with BSAM and additive models, GS with a dominance model is better at maximizing CP through purebred selection, especially when no retraining is carried out at each generation.

Previous studies on the selection of purebred animals for CP (Ibanez-Escriche et al., 2009, Toosi et al., 2010, Zeng et al., 2013) focused on crossbred data to estimate marker effects, which requires collecting genotypes and phenotypes on crossbred animals. This can substantially increase the required financial investment of the breeding program, since crossbred animals are usually not individually identified and individual performance is not recorded. It is interesting to evaluate the potential benefit of GS within purebred lines when the objective is to improve performance of crossbred animals, by using marker effects that are estimated from pure line data. In other words, additive and dominance effects of alleles can be estimated from pure line data, and subsequently breeding values for CP can be estimated by using the appropriate allele frequencies. Thus, our objective was to investigate the benefits of GS of purebred animals for CP based on purebred information, compared to traditional selection for purebred performance. A second objective was to compare the use of two separate pure line reference populations with that of a single reference population that combined the pure lines. These

objectives were investigated under two conditions, i.e. either a low or a high correlation of LD between the pure lines.

3.2 Methods

3.2.1 Population structure

Using the QMSim software (Sargolzaei and Schenkel, 2009), a historical population was simulated forward in time. Subsequent generations, GS, and evaluation were simulated using a script developed in R version 2.15.2 (R Development Core Team, 2014) (Table 3.1 and Figure 3.1). In the first simulation step, 1000 discrete generations with a constant population size of 2000 were simulated, followed by 1000 generations with a gradual decrease in population size from 2000 to 100 in order to create initial LD. The number of individuals of each sex remained the same in this step and the mating system was based on random union of gametes that were randomly sampled from the male and female gamete pools. Therefore, only two evolutionary forces were considered in this step: mutation and drift. To simulate the two recent purebred populations (referred to as breeds A and B, hereafter), two random samples of 50 animals were drawn from the last generation of the historical population and each animal was randomly mated for another 100 generations (step 2).

In the next simulation step (step 3), in order to enlarge population size for breeds A and B, eight generations were simulated with ten offspring per dam. The mating within each breed was again based on random union of gametes and no selection was considered in this step. Within each breed, all animals in generation 8 of this step were considered as training population for the estimation of marker effects.

In the next step (step 4), for each breed, 100 males and 200 females were sampled randomly from the last generation of step 3 and mated randomly to produce 1000 purebred animals (A0 and B0). In the subsequent generations (step 5), a two-way crossbreeding program with five generations of selection was simulated, as illustrated in Figure 3.1. The goal was to improve CP through selection in the two parental breeds (breeds A and B acted as sire and dam breeds, respectively). The selection criterion in the purebred population was either the rank of the individual's genomic estimated breeding value (GEBV) for purebred performance (GEBVP), or its GEBV for crossbred performance (GEBVC). SNP effects for the prediction of GEBV for each breed were estimated only once, using the purebred reference population of generation 8 of step 3 (these are the parents of generations A0 and B0). These estimates of SNP effects were then repeatedly applied to predict either GEBVP or GEBVC in the following five generations of

Table 3.1 Parameters of the simulation process

Population structure	
Step 1: Historical generations (HG)	
Number of generations(size) - phase 1	1000 (2000)
Number of generations(size) - phase 2	1000 (gradual decrease)
Selection and mating	Random
Step 2: Breed formation (BF)	
Number of founder males from HG	50
Number of founder females from HG	50
Number of generations	100
Number of offspring per dam	5
Selection and mating	Random
Step 3*: Expanded generations (EG)	
Number of founder males from BF	100
Number of founder females from BF	100
Number of generations	8
Number of offspring per dam	10
Selection and mating	Random
Step 4: Purebred A0 and B0	
Number of founder males/females from EG breed A	100/200
Number of founder males/females from EG breed B	100/200
Number of offspring per dam	5
Mating system	Random
Selection and mating	Random
Step 5: Purebred A and B	
Number of males/females from A0	100/200
Number of males/females from B0	100/200
Number of offspring per dam	5
Selection	GEBV
Mating system	Random
Heritability of the trait	0.3
Phenotypic variance	1
Genome	
Number of chromosomes	1
Number of SNPs	1000
SNP distribution	Random
Number of QTL	100
QTL distribution	Random
MAF of SNPs	0.05
MAF of QTL	0.05
Additive allelic effects for SNPs	Neutral
Additive allelic effects for QTL	Gamma
Rate of recurrent mutation	2.5×10^{-4}

*All of the individuals from the last generation of step 3 (Generation 8) was the training set.

selection of the pure breeds. In generation 1 through 5, 300 animals (the top 100 males and top 200 females) were selected from the 1000 available candidates in each parental breed, based on their GEBV. Thus, the selected proportions were 20% (100 out of 500) in males and 40% in females (200 out of 500). The selected animals were randomly mated within each breed to produce 1000 purebred replacement animals for the next generation. Meanwhile, the 100 selected males of breed A were randomly mated to the 200 selected females of breed B to produce 1000 crossbred progeny (step 5). The phenotypic mean of crossbred animals was computed for each generation of selection (AB_1 to AB_5) to evaluate the cumulative response to selection.

3.2.2 Genome and trait phenotypes

A genome consisting of one chromosome of 1 Morgan with 100 segregating QTL and 1000 markers was simulated (Table 3.1). Both QTL and markers were randomly distributed over the chromosome. To reach the required number of segregating loci after 2000 generations, about two to three times as many bi-allelic loci were simulated with starting allele frequencies sampled from a uniform distribution and a recurrent mutation rate of 2.5 × 10-4. To build the SNP panel, 1000 SNPs were randomly drawn from segregating SNPs that had a minor allele frequency (MAF) of at least 0.05, in the last historical generation. The additive effect (a) of a QTL was defined as half the difference in genotypic value between alternate homozygotes and the dominance effect (d) as the deviation of the value of the heterozygote from the mean of the two homozygotes (Falconer and Mackay, 1996). A gamma distribution with shape and scale parameters of 0.4 and 1.66, respectively, was used to generate the unsigned value of the additive effect for each QTL. This provided an L-shaped distribution of QTL effects. With equal probability, one of the two alleles was chosen to be positive or negative. Previous studies have not shown a consistent relationship between additive and dominance effects of QTL (Bennewitz and Meuwissen, 2010). Similar to Wellmann and Bennewitz (Wellmann and Bennewitz, 2011, 2012), we simulated relative dominance degrees h_i that were normally distributed, $N(0.5, 0.1)$, and independent of the additive effects. Next, absolute dominance effects were $d_i = h_i.|a_i|$ where $|a_i|$ is the absolute value of the additive effect. Thus, additive and dominance effects were dependent. Additive and dominance effects were scaled in each replicate of each scenario such that additive and dominance variances were equal to 0.3 and 0.1, respectively, in the last historical generation. This was done to ensure that each scenario had the same genetic variance, such that this could not contribute to differences among scenarios. After scaling, 10 to 15% of QTL showed overdominance. Trait

phenotypes were simulated by adding a standard normal residual effect to the genotypic value of each animal. The variance of the residual effects was chosen such that broad-sense heritability H^2 of the trait was equal to 0.4. As a result, phenotypic variance (σ_p^2) was 1, narrow-sense heritability h^2 was equal to 0.3 and dominance variance was $0.1\sigma_p^2$.

3.2.3 Estimation of marker effects

The Bayesian LASSO proposed by Park and Casella (2008) and developed by de los Campos et al. (2009) was used to estimate marker effects. The difference between Bayesian LASSO and the Bayesian approaches developed by Meuwissen et al. (2001) (BayesA and BayesB) stems from the specification of the a priori variance of the marker-specific regression coefficient. We used the BLR "Bayesian linear regression" R package developed by Perez et al. (2010). The following model was used to estimate the genetic effect associated with each marker:

$y_i = \mu + \sum X_{ij} a_j + \sum Z_{ij} d_j + e_i,$

where y_i is the phenotypic value of individual i in the training data, μ is the overall mean, X_{ij} is the copy number of a given allele of marker j, coded 0, 1 and 2 for aa, aA and AA, respectively, a_j is the random unknown additive effect for marker j, Z_{ij} is the indicator variable for heterozygosity of individual i at marker j, with $Z_{ij} = 0$ when individual i is homozygous at marker j (aa or AA) and $Z_{ij} = 1$ if individual i is heterozygous at marker j (aA), d_j is the random unknown dominance effect for SNP j, and e_i is the residual effect for animal i and Σ denotes summation over all marker loci j.

The prior distribution of the residual variance was a scaled inverse χ^2 such that $\sigma_e^2 \sim \chi^{-2}(df_e, S_e)$. The degrees of freedom (df_e) and the scale parameter (S_e) for residual variance were set at 3.5 and 3, respectively. The conditional prior distribution of the marker effects was a Gaussian distribution with prior variance specific to each marker: $a_j \sim N(0, \sigma^2 \epsilon \tau_j^2)$ for j=1,…, m, with τ_j^2 following an exponential prior distribution defined by $\tau_j^2 \sim \exp(\lambda^2)$. The regularisation parameter λ^2 followed a Gamma distribution, as suggested in Park and Casella (2008). In addition, an inverted Chi-square distribution was used for the variance of dominance effects: $\sigma_d^2 \sim \chi^{-2}(df_d, S_d)$ with df_d = 3 and $S_d = 0.0005$. The parameters of the prior distributions were computed according to the guidelines of the BLR package (de los Campos et al., 2009, Perez et al., 2010). The BLR method used an MCMC algorithm to generate 10 000 samples, with the first 1500 samples discarded as burn-in.

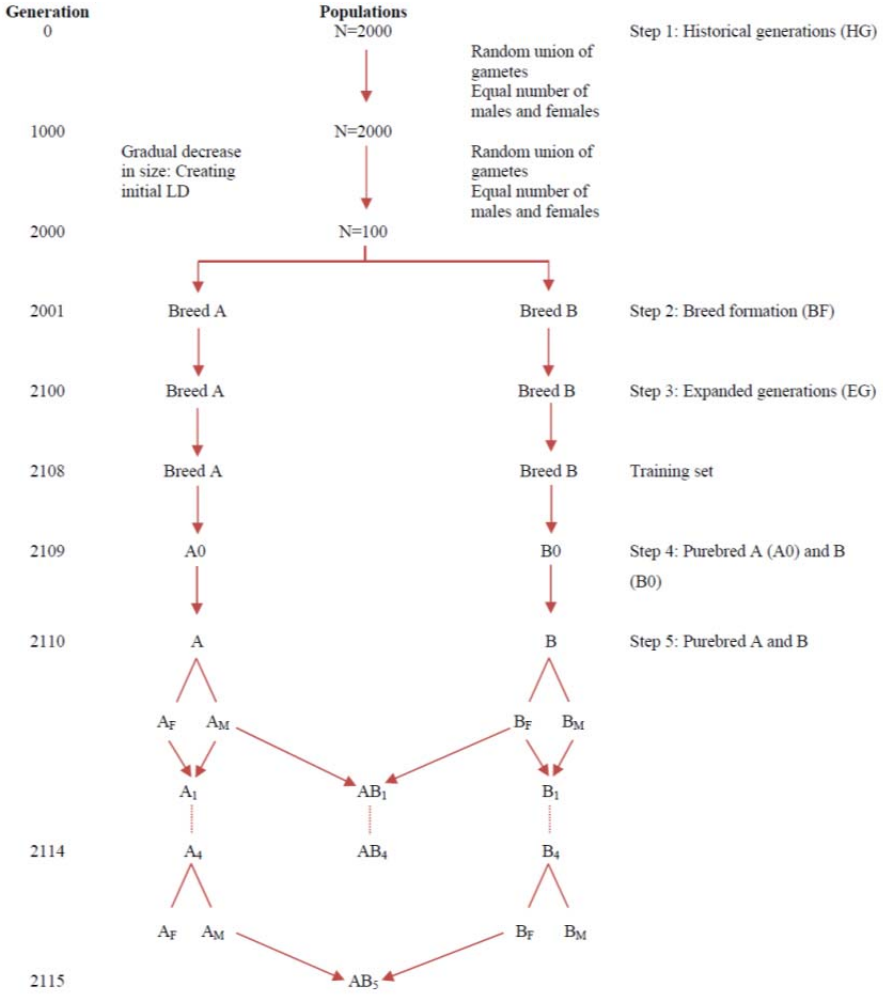


Figure 3.1 Schematic representation of the simulation steps. The crossbreeding program started in step 5 and consisted of five generations of purebred selection for crossbred performance; a random sample of individuals from the last generation of step 3 (Generation 2108) constitutes the training population; A_M and B_M represent the males selected from breeds A and B, respectively; A_F and B_F represent the females selected breeds A and B, respectively; lines with arrows denote reproduction, while lines without arrows denote selection.

3.2.4 True and genomic estimated breeding values

Two types of true breeding values (TBV) were calculated, i.e. TBV for purebred performance (TBVP) and TBV for crossbred performance (TBVC). The TBV were calculated as the expected genotypic value of the offspring of a parent carrying a

certain QTL-genotype, when this parent is mated at random to its own line (TBVP) or to the other pure line (TBVC). Thus, for animal i from breed r, the TBV for purebred performance was calculated as:

$$\mathrm{TBVP}_{ir} = \sum_{j=1}^{100}[(x_{ij})(p_{jr}a_j + q_{jr}d_j)]$$
$$+[(y_{ij})(0.5p_{jr}a_j + 0.5q_{jr}d_j + 0.5p_{jr}d_j - 0.5q_{jr}a_j)]$$
$$+[(z_{ij})(-q_{jr}a_j + p_{jr}d_j)], \qquad (1)$$

where x_{ij}, y_{ij} and z_{ij} are indicator functions of the genotype of the j^{th} QTL of the i^{th} individual, with $x_{ij} = 1$ when the genotype is AA and otherwise 0, $y_{ij} = 1$ when the genotype is Aa or aA and otherwise 0, and $z_{ij} = 1$ when the genotype is aa and otherwise 0. Moreover, p_{jr} and q_{jr} are the allelic frequencies (A and a) for the j^{th} QTL in breed r, and a_j and d_j are true additive and dominance effects of the j^{th} QTL. For example, for an AA parent at locus j, a fraction p_{jr} of its offspring will have genotype AA, while a fraction q_{jr} of its offspring will have genotype Aa. Hence, for locus j, the breeding value of this parent equals $(p_{jr}a_j + q_{jr}d_j)$, which is the first term in Equation 1.

For crossbred offspring, the expected genotype frequencies of the offspring of a parent depend on the allele frequency in the other pure line (denoted r' here). Thus, for animal i from breed r, the TBV for CP was calculated using Equation 1, however p_{jr} and q_{jr} were replaced by $p_{jr'}$ and $q_{jr'}$, where $p_{jr'}$ and $q_{jr'}$ are the allele frequencies (A and a) for the j^{th} QTL in breed r'. We also calculated the correlation ($r_{tbvp,tbvc}$) between TBVP and TBVC, which is known as the purebred-crossbred genetic correlation, denoted as r_{pc} by Wei and Vanderwerf (1994).

Genomic estimated breeding values were calculated in the same way, but using SNP genotypes rather than QTL genotypes, and estimated effects rather than true effects. Thus, from the estimates of additive ($\hat{a}$) and dominance effects ($\hat{d}$), the GEBVP (for purebred performance) for animal i from breed r was calculated as:

$$\mathrm{GEBVP}_{ir} = \sum_{j=1}^{1000}[(x_{ij})(p_{jr}\hat{a}_j + q_{jr}\hat{d}_j)]$$
$$+[(y_{ij})(0.5p_{jr}\hat{a}_j + 0.5q_{jr}\hat{d}_j + 0.5p_{jr}\hat{d}_j - 0.5q_{jr}\hat{a}_j)]$$
$$+[(z_{ij})(-q_{jr}\hat{a}_j + p_{jr}\hat{d}_j)]. \qquad (2)$$

For the calculation of GEBVC (for crossbred performance), SNP frequencies in the other breed were used i.e. p_{jr} and q_{jr} in Equation 2 were replaced by $p_{jr'}$ and

$q_{jŕ}$ where $p_{jŕ}$ and $q_{jŕ}$ are the allele frequencies (A and a) for the j^{th} marker in breed r'. SNP frequencies in the other breed were calculated based on marker genotypes of all selection candidates in that breed.

3.2.5 Accuracies of additive and dominance effects

In order to evaluate the accuracy of estimated additive and dominance effects separately, both true and estimated breeding values of an individual were partitioned into components of additive and dominance effects. For example, according to Equation 1, the TBV of an individual i is a function of additive effects, dominance effects and allele frequencies, and can be written as $\mathrm{TBV_i} = \sum \mathrm{TBV_{Add}} + \sum \mathrm{TBV_{Dom}}$, where $\sum \mathrm{TBV_{Add}}$ is the component of the TBV of animal i that is due to additive effects, and $\sum \mathrm{TBV_{Dom}}$ is the component of the TBV of animal i that is due to dominance effects. Equations 3 and 4 show the calculation of the TBV due to additive and dominance effects for animal i respectively:

$$\mathrm{TBV_{Add}} = \sum_{j=1}^{100} [(x_{ij})(p_{jr}a_j)] + [(y_{ij})(0.5p_{jr}a_j - 0.5q_{jr}a_j)] + [(z_{ij})(-q_{jr}a_j)] \quad (3)$$

and

$$\mathrm{TBV_{Dom}} = \sum_{j=1}^{100} [(x_{ij})(q_{jr}d_j)] + [(y_{ij})(0.5q_{jr}d_j + 0.5p_{jr}d_j)] + [(z_{ij})(p_{jr}d_j)] \quad (4)$$

Symbols used in Equations 3 and 4 are the same as in Equation 1. Similarly, the GEBV of an individual i was calculated as $\mathrm{GEBV_i} = \sum \mathrm{GEBV_{Add}} + \sum \mathrm{GEBV_{Dom}}$, where $\sum \mathrm{GEBV_{Add}}$ and $\sum \mathrm{GEBV_{Dom}}$ are the components of the estimated breeding value of animal i due to estimated additive and dominance effects, respectively. GEBV due to additive and dominance effects were calculated in the same way as in Equations 3 and 4, but using SNP genotypes rather than QTL genotypes, and estimated effects rather than true effects. After partitioning the breeding value of each individual, the accuracy of estimated additive effects was calculated as the correlation between the TBV due to additive effects ($\mathrm{TBV_{Add}}$) and the GEBV due to additive effects ($\mathrm{GEBV_{Add}}$). Similarly, the accuracy of estimated dominance effects was calculated as the correlation between the TBV due to dominance effects ($\mathrm{TBV_{Dom}}$) and the GEBV due to dominance effects ($\mathrm{GEBV_{Dom}}$).

3.2.6 Scenarios

Response to selection in CP was examined in five scenarios (Table 3.2). Simulated scenarios differed in structure of the training population and also in the criterion of selection. In all scenarios, breed A acted as the sire breed and breed B acted as the dam breed. In the reference scenario, both pure lines were selected for purebred performance, and both pure lines had their own reference population. In all other scenarios, breed A was selected for CP. Selection in breed B was for purebred performance in scenarios 1 and 3, and for CP in scenarios 2 and 4. In scenarios 1 and 3, both populations had their own reference population, while the reference population was combined in scenarios 2 and 4. In order to increase resolution between scenarios, we used the same population simulated from step 1 to step 3 (Figure 3.1) for a given replicate of each scenario. Each scenario was replicated 30 times.

We compared our scenarios under two conditions, i.e. low and high correlation of LD phase between the two breeds. In order to increase the correlation of LD phase between the two breeds, we increased LD in the common ancestral population by decreasing effective population size. Sved et al. (2008) showed that, if two populations diverge from a common ancestral population, their correlation of LD phase is approximately equal to $r_0^2(1-c)^{2T}$, where r_0^2 is LD in the common ancestral population, c is the recombination rate between markers, and T is the time since breed divergence in generations.

Table 3.2 Simulated scenarios

Scenarios	Selection criterion		Training population structure
	Breed A	Breed B	
Reference scenario	GEBVP	GEBVP	Separate
Scenario 1	GEBVC	GEBVP	Separate
Scenario 2	GEBVC	GEBVC	Separate
Scenario 3	GEBVC	GEBVP	Common
Scenario 4	GEBVC	GEBVC	Common

GEBVP: selection in purebred breeds A and B is based on genomic estimated breeding value for purebred performance; GEBVC: selection in purebred breeds A and B is based on genomic estimated breeding value for crossbred performance; separate training means that each breed had its own training set; common stands for the combination of animals from breeds A and B to estimate marker effects.

3.2.7 LD and correlation of LD phase

To evaluate the extent and magnitude of LD in the training populations and its impact on accuracy, LD was measured by r^2 (Hill, 1973). Only markers with a MAF greater than 0.1 were considered in this analysis, because the power of detection of LD between two loci is minimal when at least one of the loci has an extreme allele frequency (Goddard et al., 2000). To determine the decay of LD with increasing distance between SNPs, the average r^2 within each breed was expressed as a function of distance between SNPs. SNP pairs were grouped by their pairwise distance into intervals of 1 cM, starting from 0 up to 100 cM. The average r^2 for SNP pairs in each interval was estimated as the mean of all r^2 within that interval.

To estimate persistence of LD phase, only segregating SNPs (MAF > 0) in both breeds were included in the analysis. Persistence of LD phase was estimated following Badke et al. (2012) as:

$$R_{AB} = \frac{\sum_{(i,j)\in p}(r_{ij(A)} - \bar{r}_A)(r_{ij(B)} - \bar{r}_B)}{sd(A)sd(B)},$$

where $R_{A,B}$ is the correlation between $r_{ij(A)}$ in breed A and $r_{ij(B)}$ in breed B, sd(A) and sd(B) are the standard deviations of $r_{ij(A)}$and $r_{ij(B)}$, respectively, and $\bar{r}_A$ and $\bar{r}_B$ are the average r_{ij} across all SNPs i and j within interval p for breeds A and B, respectively. Correlation of LD between the two lines was estimated for intervals of 1 cM (from 0 to 50 cM). SNPs with a pairwise distance greater than 50 cM were excluded since estimates of average r^2 at greater distances are close to 0, which would result in the correlation of LD phase to be close to 0 as well.

3.3 Results

3.3.1 Distribution of marker allele frequencies

Figure 3.2 shows the distribution of marker allele frequencies for the last generation of the historical population. Since the initial allele frequencies were sampled from a uniform distribution, a kind of uniform distribution was expected with some fluctuations after 2000 generations of random mating, under a balance between mutation and random genetic drift due to finite population size. Although, a U-shaped distribution is typically observed with sequence data (Daetwyler et al., 2014), allele frequencies on SNP chips tend to be uniform (Ramos et al., 2009).

3.3.2 Linkage disequilibrium

To estimate LD, we used SNP genotypes of animals in the training set of both breeds. An average r^2 of 0.43 and 0.42 for adjacent SNPs was found for breeds A and B, respectively. These average r^2 between adjacent SNPs are similar to those reported by Badke et al. (2012) for four US pig breeds that ranged from 0.36 to

0.46 for animals genotyped using the Illumina PorcineSNP60 (number of markers M = 62 163). Another study by Du et al. (2007) that investigated the range and extent of LD in six commercial pig lines (two terminal sire lines and four maternal lines) for 4500 autosomal SNPs , reported an average r2 of 0.2 and 0.07 for all pairs of SNPs that were approximately 1 and 5 cM apart, respectively, whereas we found average r2 of 0.29 and 0.08 at those distances. Figure 3.3 displays an overview of the decline of r2 over distance in both breeds. As expected, in both breeds the most tightly linked SNP pairs had the highest average r^2, and the observed average r2 decreased rapidly as the map distance increased.

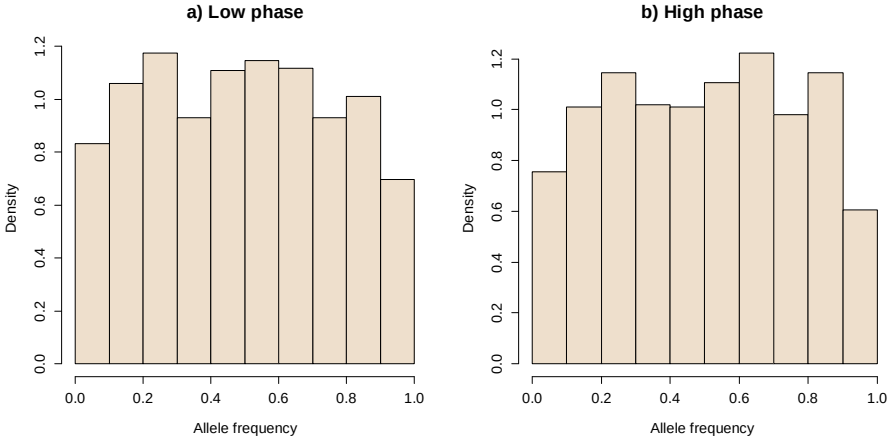


Figure 3.2 Distribution of allele frequencies in the last generation of the historical population for a low (a) and a high correlation of LD phase (b). The bounds are 0.01 and 0.99. The plots are the result of one replicate.

3.3.3 Persistence of LD phase

Persistence of LD phase among breeds can be used to infer on the history of a species and relatedness of breeds within that species, as well as on the reliability of across-population prediction of genome-wide association studies (GWAS) and GEVB (de Roos et al., 2008). Figure 3.4 shows the persistence of LD phase between adjacent SNPs, measured by the correlation of r between the two breeds. A greater correlation implies that the SNP-SNP (and most probably the SNP-QTL) LD is more consistent between the two breeds. As distance in time between subpopulations increases, there is a greater chance for recombination to break down the LD that was present in the ancestral population and for drift to create new LD within each subpopulation. Both mechanisms decrease the correlation of LD phase between the two breeds (Hill and Robertson, 1968, Goddard et al., 2006). For SNPs with a

pairwise distance of 1 cM, persistence of LD phase between breeds A and B was estimated at 0.2 and 0.7 for cases with a low and high correlation of LD phase, respectively. Persistence of LD phase has been reported for Duroc, Landrace, Yorkshire pig breeds. For SNPs with a pairwise distance less than 50 kb, Badke et al. (2012) reported a correlation of LD of 0.85 between Landrace and Yorkshire breeds and of 0.82 between Duroc and Landrace and between Duroc and Yorkshire breeds. Assuming 1 cM is approximately 1 Mb, we found correlations of LD equal to 0.38 and 0.87 for SNPs with a pairwise distance less than 50 kb for cases with low and high correlations of LD phase between two breeds, respectively. The correlation of LD phase between pig breeds in different studies ranged from 0.80 to 0.92 for SNPs with a pairwise distance less than 10 kb. In a study on the extent and persistence of LD phase in Holstein-Friesian, Jersey, and Angus cattle, de Roos et al. (2008) reported a correlation of LD phase that ranged from 0.7 to 0.97 between two breeds for SNPs with a pairwise distance less than 10 kb and a decline of this correlation as the distance between SNPs or divergence between breeds increased. In our study, as distance between SNPs increased, the correlation of LD phase between the two breeds decreased (0.5 at an average pairwise SNP distance of 1 cM). It has been reported that, while correlation of LD phase is similar for pig breeds and dairy cattle at short distance ranges (< 10 kb), pig breeds generally show greater correlations of LD phase than dairy cattle at larger SNP distances (Badke et al., 2012).

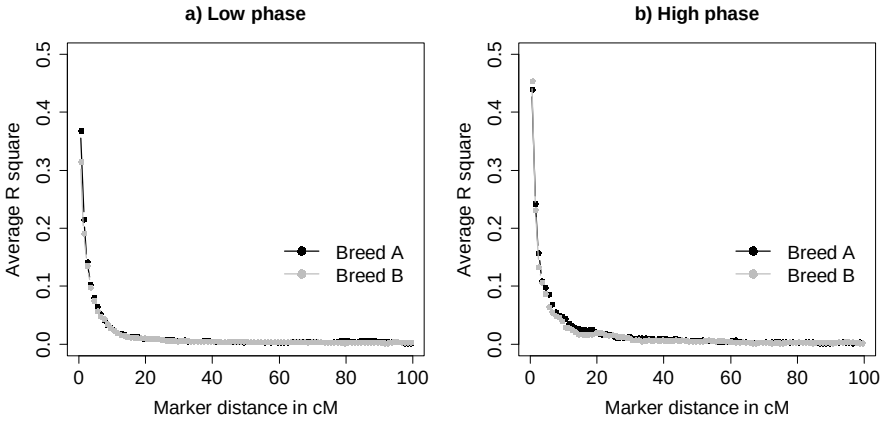


Figure 3.3 Decay of average r^2 over distance for a low (a) and a high correlation of LD phase (b). Average r^2 between SNPs in breed A and breed B at various distances in base pairs ranging from 1 to 100 cM. The plots are the result of one replicate.

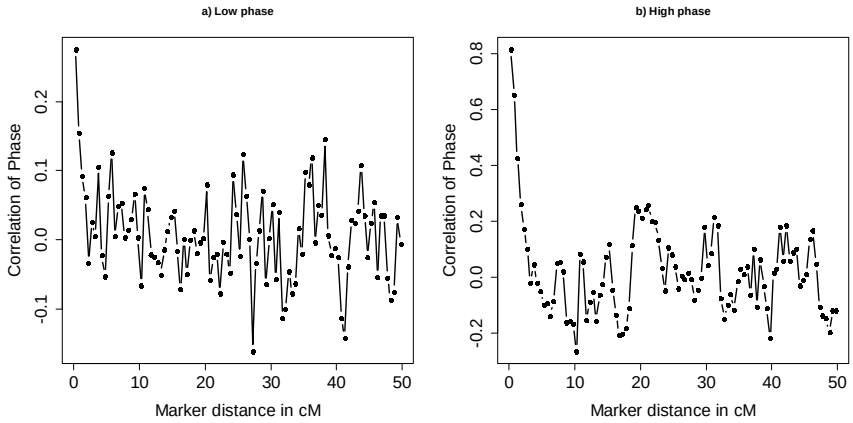


Figure 3.4 Correlation of gametic phase compared across two breeds over distance for a low (a) and a high correlation of LD phase (b). Correlation of LD phase between the two breeds for SNP pairs grouped by distance in intervals of 1 cM and covering 0 to 50 cM across the genome. The plots are the result of one replicate.

3.3.4 Response to selection in crossbred animals

The purebred-crossbred genetic correlation, i.e. the correlation between TBVP and TBVC ($r_{tbvp,tbvc}$), was equal to 0.66 and 0.70 on average for low and high correlations of LD phase, respectively. Figure 3.5 shows the mean values of phenotypes for crossbred animals in five generations under the five simulated scenarios with either a low (r = 0.2 in 1cM) or a high correlation of LD phase (r = 0.7 in 1 cM) between the two breeds. When the correlation of LD phase was low between the two breeds, the ranking of scenarios in terms of mean phenotype of crossbred animals shows that breeding for CP led to higher gains in crossbred animals. By generation 5, scenario 2, in which both breeds were selected for CP, had a higher mean phenotype in the crossbred offspring than other scenarios. Scenario 1 also resulted in higher gain than the reference scenario since, in this scenario, one of the breeds was selected for CP. In the reference scenario, in which both breeds were selected for purebred performance, response to selection was lower than for the other scenarios. Graph a in Figure 3.5 shows that, when each breed had a separate training set to estimate marker effects (scenarios 1 and 2), the performance of their crossbred offspring improved compared to that with the alternative scenarios for which a common reference was used to estimate marker effects (scenarios 3 and 4). For example, although in scenarios 1 and 3 one of the breeds (breed A) was selected for CP and because in scenario 1 each breed had its own training set, the response for scenario 1 was greater than for scenario 3.

In addition, when the correlation of LD phase was high between the two breeds, selection for CP improved the response in crossbred animals and the use of a combined reference population of the two breeds improved response even more. For scenarios 3 and 4, response in crossbred animals was greater than for the other scenarios, since these scenarios used a common training set to estimate marker effects.

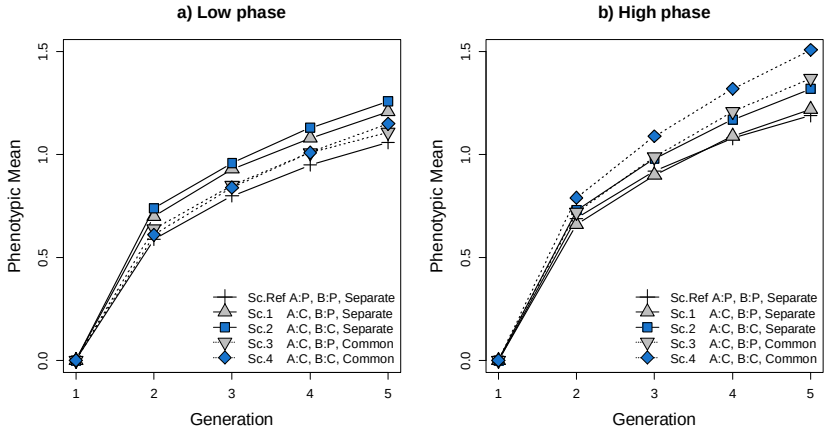


Figure 3.5 Mean phenotype of crossbred individuals. (**a**) Results for a low correlation of LD phase between breeds A and B (r = 0.2 for markers 1 cM apart) (**b**) Results for a high correlation of LD phase between breeds A and B (r = 0.7 for markers 1 cM apart). The plotted responses are means from 30 replicates. **Sc. Ref**: Selection criteria in both breed A and B was for purebred performance (P) and both breeds had **Separate** training sets. **Sc.1**: Selection criteria in breed A was for crossbred performance (C) and selection criteria in breed B was for purebred performance and both breeds had separate training sets. **Sc.2**: Selection criteria in both breed A and B was for crossbred performance and both breeds had separate training sets. **Sc.3**: Selection criteria in breed A was for crossbred performance and selection criteria in breed B was for purebred performance and both breeds had a **Common** training sets. **Sc.4**: Selection criteria in both breed A and B was for crossbred performance and both breeds had a common training set. Standard error of phenotypic means for simulated scenarios in generation 5 ranged from 0.03 to 0.04.

3.3.5 Heterosis in crossbred animals

Based on the definition of heterosis, expected CP can be written as CP = BA + H, where BA denotes the breed average of pure lines and H the heterosis present in the crossbred animals. Thus, the observed advantage of selection for CP in some scenarios may be due to greater response in BA or in H, or in both. Heterosis was calculated at each generation of the crossbred population (Figure 3.6) and Table 3.3 shows BA values for each scenario. Since heterosis was simulated due to

dominance, total heterosis was simply the sum of heterosis at each locus, $H = \sum d_l(p_{A,l} - p_{B,l})^2$, where d_l is the dominance effect at QTL l, $p_{A,l}$ is the allele frequency at QTL l in breed A, and $p_{B,l}$ is the allele frequency at QTL l in breed B (Falconer and Mackay, 1996). For both low and high correlations of LD phase, the amount of heterosis in the reference scenario was constant over generations but in other scenarios in which at least one breed was selected for CP, the amount of heterosis increased in each generation, which indicates that selection for CP resulted in greater heterosis and finally in improved performance of crossbred animals. Since heterosis depends on the difference in allele frequencies between the two breeds, these results suggest that selection for CP moves allele frequencies in the two breeds in opposite directions and causes divergence in allele frequencies between both breeds.

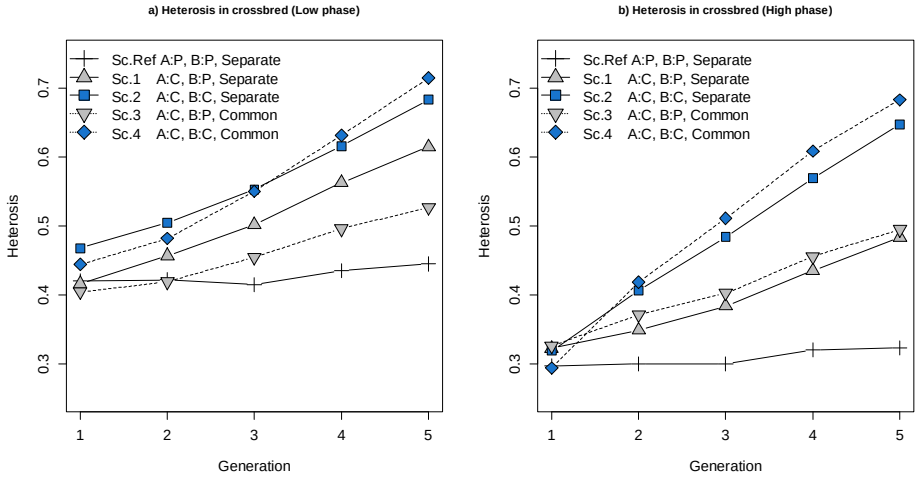


Figure 3.6 Heterosis in crossbred individuals. (**a**) Results for a low correlation of LD phase between breeds A and B (*r* = 0.2 for markers 1 cM apart) (**b**) Results for a high correlation of LD phase between breeds A and B (*r* = 0.7 for markers 1 cM apart).The plotted heterosis values are means from 30 replicates. **Sc. Ref**: Selection criteria in both breed A and B was for purebred performance (P) and both breeds had **Separate** training sets. **Sc.1**: Selection criteria in breed A was for crossbred performance (C) and selection criteria in breed B was for purebred performance and both breeds had separate training sets. **Sc.2**: Selection criteria in both breed A and B was for crossbred performance and both breeds had separate training sets. **Sc.3**: Selection criteria in breed A was for crossbred performance and selection criteria in breed B was for purebred performance and both breeds had a **Common** training set. **Sc.4**: Selection criteria in both breed A and B was for crossbred performance and both breeds had a common training set.

Table 3.3 Mean phenotypic average of breeds A and B in simulated scenarios

	Low correlation of LD phase					High correlation of LD phase				
G	Sc.Ref	Sc.1	Sc.2	Sc.3	Sc.4	Sc.Ref	Sc.1	Sc.2	Sc.3	Sc.4
1	1.33	1.25	1.33	1.21	1.37	1.12	1.19	1.04	1.04	0.93
2	1.97	1.88	1.94	1.79	1.96	1.81	1.84	1.68	1.71	1.60
3	2.02	2.04	2.11	1.96	2.12	2.04	2.03	1.86	1.90	1.80
4	2.32	2.14	2.21	2.07	2.20	2.18	2.17	1.97	2.12	1.95
5	2.40	2.21	2.28	2.15	2.26	2.29	2.24	2.03	2.23	2.05

G = generation; Sc. Ref = reference scenario; Sc. 1 = scenario 1; Sc. 2 = scenario 2; Sc. 3 = scenario 3; Sc. 4 = scenario 4

3.3.6 Accuracy of selection

Prediction accuracy, i.e. correlation between the breeding values predicted by GS and the TBV obtained from simulation, ranged from 0.69 to 0.86 in the validation population (generation 1) across the different scenarios analysed (Figure 3.7). It should be noted that accuracies in Figure 3.7 always refer to the selection criterion. In other words, when selection is for purebred performance, accuracy is the correlation between TBVP and GEBVP, i.e. ($r_{tbvp,gebvp}$). Conversely, when selection is for CP, accuracy is the correlation between TBVC and GEBVC, i.e. ($r_{tbvc,gebvc}$). Hence, this comparison shows that selection for crossbred performance would be more difficult than selection for purebred performance.

For a low correlation of LD phase, Figures 3.7a and 3.7b show that accuracy of selection for breed A was greater in the reference scenario (in which breed A was selected for purebred performance) than in the other scenarios (in which breed A was selected for CP. Accuracy of selection in breed B (Figure 3.7b) was also greater when selection in this breed was for purebred performance (reference scenario and scenarios 1 and 3) than when selection was for CP (scenarios 2 and 4). Thus, predicting GEBVC based on purebred data is more difficult than predicting GEBVP on such data.

For a high correlation of LD phase (Figure 3.7c and 3.7d), accuracies ranged from 0.78 to 0.88 in the first generation, which suggests that when the correlation of LD phase between breeds is high, there is a smaller difference in accuracy between purebred and crossbred selection ($r_{tbvp,gebvp} \sim r_{tbvc,gebvc}$). Finally, for both low and high correlations of LD phase, prediction accuracy declined over generations in all scenarios.

3.3.7 Accuracies of additive and dominance effects

The accuracies reported above are correlations between TBV and GEBV and include both additive and dominance components of the breeding values per se. In order

to compare the accuracy of estimates of additive and dominance effects separately, both true and estimated breeding values of an individual were partitioned into components due to additive and dominance effects to its comprising components. Table 3.4 includes accuracies of estimated breeding values, as well as accuracies of the additive and dominance components of estimated breeding values for low and high correlations of LD phase between the two breeds. It should be noted that accuracies of estimated breeding values in Table 3.4 always refer to the selection criterion. In other words, when selection in a breed is for purebred performance, accuracy is the correlation between TBVP and GEBVP. Conversely, when selection in a breed is for CP, accuracy is the correlation between

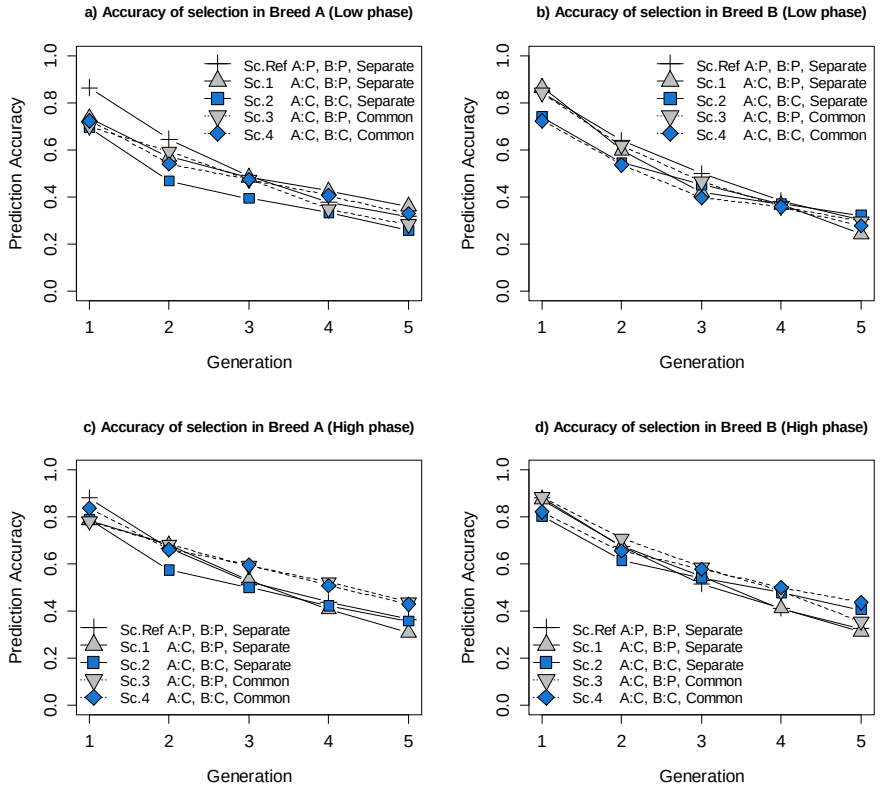


Figure 3.7 Accuracy of selection in breeds A and B in five scenarios. (**a**) and (**b**) Results for a low correlation of LD phase between breeds A and B (r = 0.2 for markers 1 cM apart) (**c**) and (**d**) Results for a high correlation of LD phase between breeds A and B (r = 0.7 for markers 1 cM apart). The plotted accuracies are means from 30 replicates. **Sc. Ref**:

Selection criteria in both breed A and B was for purebred performance (P) and both breeds had **Separate** training sets. **Sc.1**: Selection criteria in breed A was for crossbred performance (C) and selection criteria in breed B was for purebred performance and both breeds had separate training sets. **Sc.2**: Selection criteria in both breed A and B was for crossbred performance and both breeds had separate training sets. **Sc.3**: Selection criteria in breed A was for crossbred performance and selection criteria in breed B was for purebred performance and both breeds had a **Common** training set. **Sc.4**: Selection criteria in both breed A and B was for crossbred performance and both breeds had a common training set. It should be noted that accuracies in this Figure are correlations between the selection criterion and the EBV of interest. Thus, when selection is for purebred performance, accuracy is the correlation between GEBVP and TBVP, while when selection is for crossbred performance, accuracy is the correlation between GEBVC and TBVC.

TBVC and GEBVC. Generally, in all scenarios, accuracies of estimated breeding values due to additive effects were greater than accuracies of estimated breeding values due to dominance effects. These differences in accuracies were clearer for scenarios in which selection within a breed was for CP (e.g. breed B in scenarios 2 and 4 in Table 3.4). However, when selection in a breed was for purebred performance, accuracies of estimated breeding values due to additive and dominance effects were not very different (e.g. breed B in the reference scenario and scenarios 1 and 3). In summary, for both selection criteria, accuracies of estimated breeding values were as high as accuracies due to additive effects. However, when selection within a breed was for CP, accuracies due to dominance effects were higher than accuracies due to dominance effects for selection on purebred performance. The same trend was observed with a high correlation of LD phase between the two breeds [See Additional file 1].

3.3.8 Response to selection in purebred animals

Figure 3.8 shows the response to selection in both purebred populations of breeds A and B over five generations. For a low correlation of LD phase between breeds A and B (Figures 3.8a and 3.8b), response to selection in both breeds in the reference scenario was higher than the other scenarios, since selection in this scenario was for purebred performance. In the other scenarios, response to selection was lower for breed A than in the reference scenario, since in these scenarios the selection criterion was CP (Figure 3.8a). Figure 3.8b shows that response to selection for breed B in scenarios 3 and 4, which used a common reference population, was lower than in the other scenarios.

Table 3.4 Partitioning accuracies of breeding values due to additive and dominance effects for a low correlation of LD phase

		Ref scenario			**Scenario 1**			**Scenario 2**			**Scenario 3**			**Scenario 4**		
Breed A	G	BV	A	Dom	BV	Add	Dom	BV	Add	Dom	BV	Add	Dom	BV	Add	Dom
	1	0.86	0.81	0.53	0.73	0.80	0.22	0.69	0.80	0.15	0.70	0.78	0.26	0.72	0.76	0.31
	2	0.64	0.69	0.56	0.57	0.65	0.20	0.46	0.69	0.19	0.59	0.69	0.27	0.54	0.65	0.22
	3	0.48	0.63	0.57	0.48	0.50	0.23	0.39	0.63	0.20	0.47	0.61	0.21	0.47	0.61	0.22
	4	0.37	0.59	0.60	0.42	0.52	0.24	0.33	0.57	0.21	0.34	0.54	0.18	0.40	0.58	0.24
	5	0.31	0.56	0.61	0.36	0.47	0.23	0.25	0.52	0.20	0.28	0.48	0.20	0.32	0.52	0.26
		Ref scenario			Scenario 1			Scenario 2			Scenario 3			Scenario 4		
Breed B	G	BV	Add	Dom	BV	Add	Dom	BV	Add	Dom	BV	Add	Dom	BV	Add	Dom
	1	0.85	0.77	0.47	0.87	0.81	0.56	0.74	0.81	0.13	0.88	0.85	0.60	0.72	0.82	0.19
	2	0.64	0.65	0.43	0.60	0.64	0.55	0.55	0.68	0.16	0.71	0.76	0.59	0.54	0.69	0.18
	3	0.50	0.58	0.49	0.42	0.59	0.55	0.45	0.59	0.18	0.59	0.70	0.63	0.40	0.62	0.16
	4	0.38	0.58	0.53	0.37	0.56	0.54	0.37	0.54	0.19	0.49	0.65	0.68	0.36	0.56	0.15
	5	0.30	0.55	0.56	0.24	0.54	0.58	0.32	0.49	0.18	0.35	0.60	0.68	0.28	0.48	0.14

Reference scenario = selection criteria in both breeds A and B were for purebred performance (P) and both breeds had each a separate training set; scenario 1 = selection criteria in breed A were for crossbred performance (C) and selection criteria in breed B were for purebred performance and both breeds had each a separate training set; scenario 2 = selection criteria in both breeds A and B were for crossbred performance and both breeds had each a separate training set; scenario 3 = selection criteria in breed A were for crossbred performance and selection criteria in breed B were for purebred performance and both breeds had a common training set; scenario 4 = selection criteria in both breeds A and B were for crossbred performance and both breeds had a common training set.

For a high correlation of LD phase between breeds A and B, response to selection for breed A was lower in scenario 2 than in the other scenarios (Figure 3.8c). Figure 3.8c also shows that for a high correlation of LD phase between breeds, the use of a common reference population to estimate marker effects improved the performance of purebred animals, i.e. scenario 3 performed better than scenario 1, and scenario 4 performed better than scenario 2.

In conclusion, for both low and high correlations of LD phase, selection for CP generated a loss in response to selection in purebred animals.

3.4 Discussion

The purpose of this study was to evaluate the potential benefit of GS within purebred lines, when the objective is to improve performance of crossbred populations at the commercial level and phenotypic information is collected only on purebred animals. We compared response to selection in crossbred animals in five scenarios, where individuals were selected either on GEBVP or GEBVC, and marker effects were estimated either from two separate purebred reference populations or a combined purebred reference population. In a two-way crossbreeding system, we found that selection for GEBVC increased response in crossbred animals compared to selection for GEBVP. We also investigated the effect of the correlation of LD phase between the two pure breeds on the consequences of combining both reference populations. The results revealed that, for a high correlation of LD phase, combining both populations into a single reference population increased response to selection in crossbred animals.

3.4.1 Persistence of LD phase

The value of SNPs effect estimated for populations other than the reference population depends on the persistence of LD phase between the reference population and the other population (Dekkers and Hospital, 2002). For example, a SNP that was identified as being in LD with the QTL in one breed may not be in LD with the QTL in another breed. The level of LD is more likely to be different between two populations when these populations have diverged for many generations and the effective population size becomes small, and when distance between the SNP and the QTL is large, since these factors will either break down LD in the ancestral population or create new LD within the subpopulation (Hill and Robertson, 1968, Hayes et al., 2009).

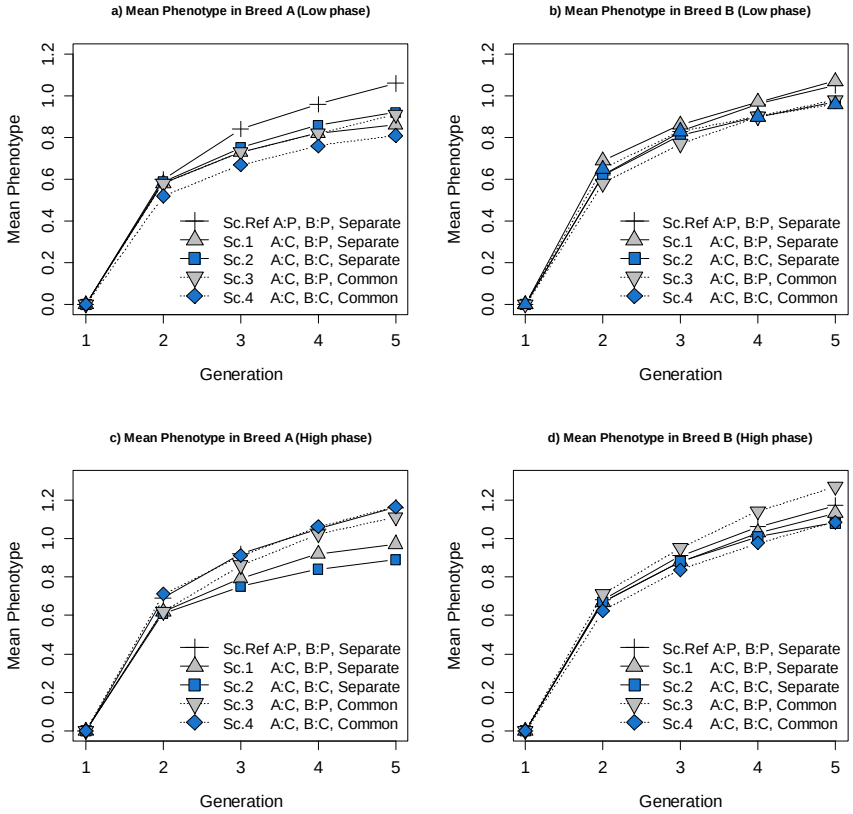


Figure 3.8 Mean phenotype of purebred individuals. (**a**) and (**b**) Results for a low correlation of LD phase between breeds A and B (r = 0.2 for markers 1 cM apart) (**c**) and (**d**) Results for a high correlation of LD phase between breeds A and B (r = 0.7 for markers 1 cM apart). The plotted responses are means from 30 replicates. **Sc. Ref**: Selection criteria in both breed A and B was for purebred performance (P) and both breeds had **Separate** training sets. **Sc.1**: Selection criteria in breed A was for crossbred performance (C) and selection criteria in breed B was for purebred performance and both breeds had separate training sets. **Sc.2**: Selection criteria in both breed A and B was for crossbred performance and both breeds had separate training sets. **Sc.3**: Selection criteria in breed A was for crossbred performance and selection criteria in breed B was for purebred performance and both breeds had a **Common** training set. **Sc.4**: Selection criteria in both breed A and B was for crossbred performance and both breeds had a common training set.

For a low correlation of LD phase, combining data from both breeds to estimate marker effects (scenarios 3 and 4) had no effect on the accuracy of GS. It has been reported that using multiple breeds to predict GEBV can be effective to increase the size of the reference population and in turn increase accuracy of selection (Pryce et al., 2011). However, the benefit of combining reference populations

depends on the size of the reference population, since there is a diminishing return relationship between size and accuracy of reference populations. Hence, if the reference population is small, combining populations may help when the correlation of LD phase is sufficiently high but will have a limited benefit or may even be detrimental when the reference population is large.

For a high correlation of LD phase, combining animals from the two breeds in the training set improved the accuracy of selection in scenarios 3 and 4. These results are consistent with those of Ibanez-Escriche et al. (2009) and de Roos et al. (2009), who concluded that across-population evaluations were preferred to within-population evaluations when the populations were closely related, marker density was high, or the number of animals with phenotypic records was small.

3.4.2 Non-additive effects and response to selection

It has been argued that dominance is the likely genetic basis of heterosis (Falconer and Mackay, 1996), therefore explicitly including dominance in the GS model may be an advantage when selecting purebred animals for CP, i.e. it may increase heterosis. In this study, we assumed dominance variance to be one third of the additive genetic variance. This ratio resulted in 10 to 15% of loci showing overdominance. When overdominance is present, crossbred performance is maximized if alternate alleles are fixed in the two purebred populations. In fact with overdominance, allele substitution effects may have opposite signs in the parental breeds, depending on allele frequencies in the two breeds. In this case, the two parental breeds are expected to be fixed for alternate alleles of overdominant QTL, which increases the frequency of favourable heterozygotes in crossbred progeny and can explain the benefit of selection based on GEBVC. However, it should be noted that existence of overdominance is not the only driver of divergence in allele frequencies in parental breeds. It has been shown that partial dominance can play a role in influencing changes in allele frequencies and have favourable effects on heterosis, especially when the number of QTL that affect the trait is large (Kinghorn et al., 2011).

3.4.3 Genotype-by-environment and genotype-by-genetic interactions

In our simulation, we assumed that the additive and dominance effects of the QTL alleles were similar in both breeds. For some QTL, which have been traced to known mutations, the alleles do act reasonably similarly in different breeds and populations (Spelman et al., 2002). However, this assumption is violated when there are QTL-by-environment interactions or QTL-by-genetic background

interactions (epistasis). With substantial QTL-by-environment interactions or epistasis, it will be less advantageous to combine populations in a training set, because marker effects will differ between populations (de Roos et al., 2009). In addition, with genotype-by-environment (G×E) interaction and epistasis, the main complication is that the dominance model does not fully explain the incomplete genetic correlation between crossbred and purebred individuals (r_{pc}). In fact, an incomplete genetic correlation between purebred and crossbred performance can be due to both non-additive effects (dominance and epistasis), and G×E interaction. In our simulation, the correlation between TBVP and TBVC ($r_{tbvp,tbvc}$) was 0.66 and 0.7 on average for low and high correlations of LD phase between two breeds, respectively, which was purely due to dominance and differences in allele frequencies between the two purebred lines.

In this study, we focused on using purebred data to improve CP. In fact, selection at the purebred level reduces the need for the crossbred testing that is required for CCPS, thereby saving important test resources and enabling the short generation intervals of purebred selection. However, Dekkers and Chakraborty (2004) discussed the benefit of GS for improving CP and suggested that it may be limited if marker effects are estimated from purebred nucleus data since the resulting EBV are strictly relevant to the studied population and environment only and may not help much to improve selection for CP if substantial G×E and genotype-by-genetic (G×G) background interactions are present. In this study, we considered the G×G due to dominance and not that due to differences in the physical environment. In principle, one could use a dominance model and multi-trait analysis to partition the purebred-crossbred genetic correlation into a component due to dominance and a remaining component due to G×E and epistasis. However, accurate partitioning would require a small standard error of the estimated purebred-crossbred genetic correlation, and thus very large datasets (Bijma and Bastiaansen, 2014).

In this study, directional dominance was simulated since dominance coefficients (h_i) were normally distributed with a positive mean, $N(0.5, 0.1)$. Consequently, dominance effects (d_i) were on average greater than 0 ($d > 0$). However, in the statistical model used to estimate the genetic effects associated with each marker, dominance effects were considered as random unknown effects with a mean of 0. The simulation of dominance effects that are on average greater than 0 has two consequences. First, the overall average trait value may increase. This will be accounted for by the fixed effects component of the model (Xb). Second, directional dominance leads to inbreeding depression. Thus, animals with different inbreeding levels will have systematically different trait phenotypes. This probably

means that our model could be improved by including a regression on inbreeding coefficients. However, we think this effect is probably limited since we simulated only five discrete generations of data with random mating among selected animals. Thus, the range of inbreeding coefficients may not have been sufficiently large to affect the results.

3.5 Conclusion

Under the hypothesis that crossbred animals differ from purebred animals because of dominance, GS can be applied to select purebred individuals for CP without collecting crossbred phenotypic or genotypic data, by using a dominance model. We found that in a two-way crossbreeding system, response to selection in crossbred individuals was higher when selection was for GEBV for CP, although data were collected on purebred individuals. Furthermore, if the correlation of LD phase between two breeds is high, there can be an added benefit in terms of accuracy of GEBV if animals from both breeds are combined into a single reference population to estimate marker effects.

3.6 Authors' contributions

HE wrote the simulation scripts, carried out the analyses and drafted the manuscript. HE, ACS and PB conceived and designed the study. PB helped in interpreting the output and edited the drafted manuscript. All authors read and approved the final manuscript.

3.7 Acknowledgements

HE has benefited from a joint grant from the European Commission and Aarhus University within the framework of the Erasmus-Mundus joint doctorate 'EGS-ABG'. Comments from the reviewers greatly helped to improve the quality of the work.

3.8 Appendix

Partitioning accuracies of breeding values due to additive and dominance effects for a high correlation of LD phase.

Partitioning accuracies of breeding values due to additive and dominance effects for a high correlation of LD phase.

		Ref scenario			**Scenario 1**			**Scenario 2**			**Scenario 3**			**Scenario 4**		
Breed A	G	BV	Add	Dom	BV	Add	Dom	BV	Add	Dom	BV	Add	Dom	BV	Add	Dom
	1	0.88	0.82	0.50	0.73	0.80	0.22	0.78	0.83	0.45	0.78	0.84	0.17	0.84	0.85	0.45
	2	0.67	0.71	0.53	0.57	0.65	0.20	0.68	0.75	0.46	0.68	0.75	0.35	0.66	0.72	0.49
	3	0.52	0.66	0.61	0.48	0.50	0.23	0.53	0.67	0.45	0.59	0.70	0.36	0.60	0.70	0.51
	4	0.44	0.63	0.65	0.42	0.52	0.24	0.41	0.62	0.44	0.52	0.66	0.39	0.51	0.65	0.52
	5	0.36	0.62	0.66	0.36	0.47	0.23	0.31	0.57	0.48	0.44	0.64	0.36	0.43	0.60	0.51
		Ref scenario			Scenario 1			Scenario 2			Scenario 3			Scenario 4		
Breed B	G	BV	Add	Dom	BV	Add	Dom	BV	Add	Dom	BV	Add	Dom	BV	Add	Dom
	1	0.88	0.84	0.51	0.87	0.82	0.47	0.80	0.83	0.37	0.88	0.85	0.60	0.82	0.83	0.49
	2	0.67	0.71	0.55	0.68	0.70	0.51	0.62	0.73	0.42	0.71	0.76	0.59	0.65	0.75	0.41
	3	0.51	0.68	0.59	0.55	0.66	0.58	0.54	0.70	0.44	0.59	0.70	0.63	0.58	0.70	0.41
	4	0.41	0.64	0.63	0.41	0.60	0.64	0.48	0.64	0.41	0.49	0.65	0.68	0.50	0.65	0.39
	5	0.33	0.63	0.65	0.31	0.54	0.63	0.41	0.59	0.38	0.35	0.60	0.68	0.44	0.62	0.38

BV: Accuracy of breeding values that is correlation between the selection criterion and the EBV of interest. Thus, when selection is for purebred performance, accuracy is the correlation between GEBVP and TBVP, while when selection is for crossbred performance, accuracy is the correlation between GEBVC and TBVC. Add: Accuracy of breeding values due to additive effects. Dom: Accuracy of breeding values due to dominance effects. G: generation.

Reference scenario = selection criteria in both breeds A and B were for purebred performance (P) and both breeds had each a separate training set; scenario 1 = selection criteria in breed A were for crossbred performance (C) and selection criteria in breed B were for purebred performance and both breeds had each a separate training set; scenario 2 = selection criteria in both breeds A and B were for crossbred performance and both breeds had each a separate training set; scenario 3 = selection criteria in breed A were for crossbred performance and selection criteria in breed B were for purebred performance and both breeds had a common training set; scenario 4 = selection criteria in both breeds A and B were for crossbred performance and both breeds had a common training set.

References

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, and J. P. Steibel. 2012. Estimation of linkage disequilibrium in four US pig breeds. Bmc Genomics 13.

Bennewitz, J. and T. H. E. Meuwissen. 2010. The distribution of QTL additive and dominance effects in porcine F2 crosses. J Anim Breed Genet 127(3):171-179.

Bijma, P. and W. M. Bastiaansen. 2014. Standard error of the genetic correlation: how much data do we need to estimate a purebred-crossbred genetic correlation? Genetics Selection Evolution 46:79.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brondum, X. P. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46(8):858-865.

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

de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183:1545-1553.

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179(3):1503-1512.

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

Dekkers, J. C. M. and R. Chakraborty. 2004. Optimizing purebred selection for crossbred performance using QTL with different degrees of dominance. Genetics Selection Evolution 36(3):297-324.

Dekkers, J. C. M. and F. Hospital. 2002. The use of molecular genetics in the improvement of agricultural populations. Nat Rev Genet 3(1):22-32.

Du, F. X., A. C. Clutter, and M. M. Lohuis. 2007. Characterizing linkage disequilibrium in pig populations. Int J Biol Sci 3(3):166-178.

Falconer, D. S. and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. Vol. 4. 4 ed. Pearson

Goddard, K. A. B., P. J. Hopkins, J. M. Hall, and J. S. Witte. 2000. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66(1):216-234.

Goddard, M. E., B. Hayes, H. McPartlan, and A. J. Chamberlain. 2006. Can the same genetic markers be used in multiple breeds? Pages 22-16. Instituto Prociência, Minas Gerais.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges (vol 92, pg 433, 2009). J Dairy Sci 92(3):1313-1313.

Hill, W. G. 1973. Linkage Disequilibrium among Neutral Genes in Finite Populations. Genetics 74(Jun):S115-S115.

Hill, W. G. and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38(6):226-231.

Ibanez-Escriche, N., R. L. Fernando, A. Toosi, and J. C. M. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. Genetics Selection Evolution 41.

Kinghorn, B. P., J. M. Hickey, and J. H. J. van der Werf. 2010. Reciprocal Recurrent Genomic Selection for Total Genetic Merit in Crossbred Individuals. in Proc. Proceedings of the 9th World Congress on Genetics Applied to Livestock Production: 1-6 August 2010; Leipzig. Paper 36; 2010.

Kinghorn, B. P., J. M. Hickey., and J. H. J. v. d. Werf. 2011. Long-range phasing and use of crosbred data in genomic selection. in Proc. 7th European Symposium on Poultry Genetics, Edinburgh, Scotland.

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

Park, T. and G. Casella. 2008. The Bayesian Lasso. J Am Stat Assoc 103(482):681-686.

Perez, P., G. de los Campos, J. Crossa, and D. Gianola. 2010. Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R. Plant Genome-Us 3(2):106-116.

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Solkner, M. E. Goddard, and B. J. Hayes. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. J Dairy Sci 94(5):2625-2630.

R Development Core Team. 2014. R: A language and environment for statistical computing. Vienna, Austria.

Ramos, A. M., R. P. M. A. Crooijmans, N. A. Affara, A. J. Amaral, A. L. Archibald, J. E. Beever, C. Bendixen, C. Churcher, R. Clark, P. Dehais, M. S. Hansen, J. Hedegaard, Z. L. Hu, H. H. Kerstens, A. S. Law, H. J. Megens, D. Milan, D. J. Nonneman, G. A. Rohrer, M. F. Rothschild, T. P. L. Smith, R. D. Schnabel, C. P. Van Tassell, J. F. Taylor, R. T. Wiedmann, L. B. Schook, and M. A. M. Groenen. 2009. Design of a high density snp genotyping assay in the pig using snps identified and characterized by next generation sequencing technology. PloS one 4(8).

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

Spelman, R. J., C. A. Ford, P. McElhinney, G. C. Gregory, and R. G. Snell. 2002. Characterization of the DGAT1 gene in the New Zealand dairy population. J Dairy Sci 85(12):3514-3517.

Sved, J. A., A. F. McRae, and P. M. Visscher. 2008. Divergence between human populations estimated from linkage disequilibrium. Am J Hum Genet 83(6):737-743.

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

Wei, M. and J. H. J. Vanderwerf. 1994. Maximizing Genetic Response in Crossbreds Using Both Purebred and Crossbred Information. Anim Prod 59:401-413.

Wellmann, R. and J. Bennewitz. 2011. The contribution of dominance to the understanding of quantitative genetic variation. Genet Res 93(2):139-154.

Wellmann, R. and J. Bennewitz. 2012. Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genet Res 94(1):21-37.

Zeng, J., A. Toosi, R. L. Fernando, J. C. Dekkers, and D. J. Garrick. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genetics, selection, evolution : GSE 45(1):11.

4

A crossbred reference population can improve the response to genomic selection for crossbred performance

Hadi Esfandyari[1,2], Anders Christian Sørensen[1], Piter Bijma[2]

[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Denmark
[2]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands

GSE (2015) 47:76

Abstract

Background: Breeding goals in a crossbreeding system should be defined at the commercial crossbred level. However, selection is often performed to improve purebred performance. A genomic selection (GS) model that includes dominance effects can be used to select purebreds for crossbred performance. Optimization of the GS model raises the question of whether marker effects should be estimated from data on the pure lines or crossbreds. Therefore, the first objective of this study was to compare response to selection of crossbreds by simulating a two-way crossbreeding program with either a purebred or a crossbred training population. We assumed a trait of interest that was controlled by loci with additive and dominance effects. Animals were selected on estimated breeding values for crossbred performance. There was no genotype by environment interaction. Linkage phase and strength of linkage disequilibrium between quantitative trait loci (QTL) and single nucleotide polymorphisms (SNPs) can differ between breeds, which causes apparent effects of SNPs to be line-dependent. Thus, our second objective was to compare response to GS based on crossbred phenotypes when the line origin of alleles was taken into account or not in the estimation of breeding values.

Results: Training on crossbred animals yielded a larger response to selection in crossbred offspring compared to training on both pure lines separately or on both pure lines combined into a single reference population. Response to selection in crossbreds was larger if both phenotypes and genotypes were collected on crossbreds than if phenotypes were only recorded on crossbreds and genotypes on their parents. If both parental lines were distantly related, tracing the line origin of alleles improved genomic prediction, whereas if both parental lines were closely related and the reference population was small, it was better to ignore the line origin of alleles.

Conclusions: Response to selection in crossbreeding programs can be increased by training on crossbred genotypes and phenotypes. Moreover, if the reference population is sufficiently large and both pure lines are not very closely related, tracing the line origin of alleles in crossbreds improves genomic prediction.

4.1 Introduction

Breeding goals in a crossbreeding system should be defined at the commercial crossbred level. However, selection is often optimized to improve animals within pure lines or breeds (Hartmann, 1992). Performance of purebred parents can be a poor predictor of the performance of their crossbred descendants in the presence of non-additive gene action or genotype by environment (G×E) interaction. A number of methods have been proposed as alternatives to pure line selection to obtain greater response to selection in crossbreds. These methods can be classified into three groups: reciprocal recurrent selection (Comstock et al., 1949), combined crossbred and purebred selection (CCPS) (Wei and Steen, 1991, Lo et al., 1993, Bijma and van Arendonk, 1998) and genomic selection (GS) (Dekkers and Chakraborty, 2004, Dekkers, 2007).

Recent studies have shown that GS can be applied to select purebreds for crossbred performance (CP), (Dekkers, 2007, Ibanez-Escriche et al., 2009, Kinghorn et al., 2010, Zeng et al., 2013). Compared to alternative methods such as CCPS, GS can lead to substantially greater response to selection (Dekkers, 2007, Piyasatian et al., 2007), lower the rate of inbreeding (Daetwyler et al., 2007, Dekkers, 2007), and does not require systematic collection of pedigree information between crossbreds and purebreds. Moreover, measuring the phenotypes of crossbred animals at each generation of GS may not be necessary, because in theory, predicted SNP effects can be used over a few generations with limited loss in prediction accuracy (Habier et al., 2007, Sonesson and Meuwissen, 2009).

For traits with significant non-additive variance, explicitly including dominance in the GS model may increase response to selection of purebreds for CP. Esfandyari et al. (2015) investigated the benefits of GS of purebreds for CP, based on purebred information under two conditions, i.e. a low or high correlation of linkage disequilibrium (LD) phase between the two pure lines. They concluded that a dominance model can be used to increase CP, without using crossbred data. Furthermore, they showed that, if the correlation of LD phase between both pure lines is high, accuracy of selection can be increased by combining both pure lines into a single reference population to predict marker effects.

Accepting that GS is an appropriate tool to select animals for CP raises another question i.e. should marker effects be predicted from pure line or crossbred data. On the one hand, if training is carried out on pure lines for traits with significant non-additive variance and therefore potential heterosis, the purebred performance is not a good predictor of crossbred performance. On the other hand, if training is done on crossbreds, it is necessary to record genotype and phenotype data on crossbreds, which can substantially increase the required investment in the

breeding program, since crossbred animals are usually not individually identified and individual performances are not recorded. Furthermore, SNP effects in crossbred animals may be specific to the parental line origin, because the extent of LD between SNPs and QTL can differ between the pure lines. Moreover, LD may not be restricted to markers that are tightly linked to the QTL (Dekkers, 2007). Nevertheless, training on crossbred data for GS accounts for genetic differences between purebred and crossbred animals and potential genotype by environment effects, and we expect that it can be beneficial to improve crossbred performance.

Previous studies on the implementation of GS in crossbreeding programs focused either on crossbred (Ibanez-Escriche et al., 2009, Zeng et al., 2013) or purebred (Esfandyari et al., 2015) data for prediction of marker effects, without explicitly comparing responses to selection obtained with both methods. For example, Zeng et al. (2013) compared additive and dominance models for GS of purebred animals for CP by training only on crossbred animals. Therefore, the first objective of our study was to compare response to selection of crossbreds by simulating a two-way crossbreeding program with either a purebred or crossbred training population under a dominance model. In addition, in the dominance model previously proposed by Zeng et al. (2013) for the application of GS in crossbreeding programs, alternate heterozygotes (based on breed origin) were assumed to have the same effect. Thus, the second objective of our study was to compare the benefits of GS of purebreds for CP using a crossbred training population when breed origin of alleles was either accounted for or not in the calculation of breeding values. In other words, this study includes models in which alternate heterozygotes can have different effects.

4.2 Methods

4.2.1 Scenarios

Response to selection in crossbreds was examined in six different scenarios (Table 4.1). For all scenarios, breed A acted as sire breed and breed B acted as dam breed. Scenarios differed in the structure of the training population. In Scenario 1, both lines A and B had their own purebred training population (separate). In Scenario 2, animals from both breeds A and B were combined into a single purebred training population (combined). In Scenario 3 and 4, crossbred animals had phenotypes but no genotypes, thus the phenotypes of crossbred animals were linked to the genotypes of the purebred animals to predict marker effects, assuming that pedigree information for both purebred and crossbred animals was available. The difference between Scenarios 3 and 4 was that, for Scenario 3, alleles of heterozygous individuals were not traced back to the purebred line of origin, and

thus alternate heterozygotes (i.e. genotype Aa or aA) were considered as identical, whereas for Scenario 4, they could be distinguished. For Scenarios 5 and 6, the training population consisted of crossbred animals with both phenotypes and genotypes and, as for Scenarios 3 and 4, alternate heterozygotes were considered as identical in Scenario 5 but could be distinguished in Scenario 6. In the six scenarios presented in Table 4.1, breeds A and B shared a common ancestor 300 generations back, which means that each breed had 300 generations of independent breeding.

Table 4.1 Simulated scenarios.

Scenarios	Training population structure
Scenario 1	PB Separate (A and B)
Scenario 2	PB Combined (A+B)
Scenario 3	Crossbred (P1)
Scenario 4	Crossbred (P2)
Scenario 5	Crossbred (PG1)
Scenario 6	Crossbred (PG2)

"Separate" means that training was done separately for each pure line; "Combined" means that training was done on a combination of purebred lines A and B; "Crossbred (P1)" means that training was done on crossbred animals with phenotypes and genotype probabilities and it was assumed that the alternate heterozygotes were identical in crossbred animals. "Crossbred (P2)" means that training was done on crossbred animals with phenotypes and genotype probabilities and it was assumed that the alternate heterozygotes could be distinguished in crossbred animals. "Crossbred (PG1)" means that training was done on crossbred animals with phenotypes and genotypes and it was assumed that the alternate heterozygotes were identical in crossbred animals. "Crossbred (PG2)" means that training was done on crossbred animals with phenotypes and genotypes and it was assumed that the alternate heterozygotes could be distinguished in crossbred animals.

In order to evaluate the impact of relatedness between both pure lines (measured as the number of generations since they separated) and of the size of crossbred training population on response to selection, additional scenarios were simulated for Scenarios 5 and 6 only in which: (1) the number of generations to the most recent common ancestor between breeds A and B varied as follows 1, 50, 100, 200 or 400 generations and (2) the size of the training population varied with 500, 2000 or 8000 randomly selected individuals. All simulated scenarios were replicated 50 times.

4.2.2 Population structure

The QMSim software (Sargolzaei and Schenkel, 2009) was used to simulate a historical population of 2000 generations with a constant size of 2000 individuals for 1000 generations, followed by a gradual decrease in population size from 2000

to 1000 to create initial LD (Figure 4.1). The number of individuals of each sex was equal and mating was performed by randomly drawing the parents of an animal from the animals of the previous generation (step 1). To simulate the two purebred recent populations (referred to as breeds A and B, hereafter), two random samples of 100 animals were drawn from the last generation of the historical population and, within each sample, animals were randomly mated for another 300 generations (step 2); 300 generations of random mating for breed formation may seem unrealistic but this was done to simulate two distantly related breeds. In step 3, in order to expand breeds A and B, eight generations were simulated with five

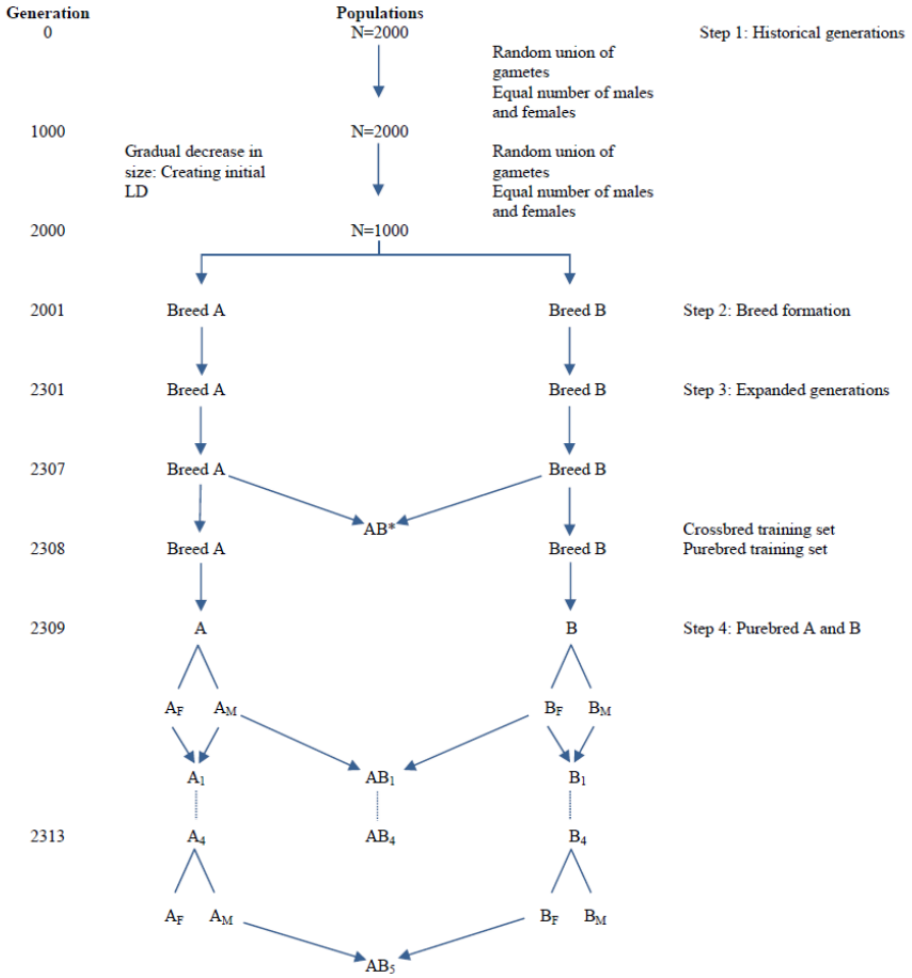


Figure 4.1 Schematic representation of the simulation steps.
The crossbreeding program started in step 4 and consisted of five generations of purebred

selection for crossbred performance; the random sample of individuals from the last generation of step 3 (Generation 2308) composed the purebred training set, and crossbred animals (AB*) in generation 2307 composed the crossbred training set; AM and BM represent the selected males of breeds A and B, AF and BF the selected females of breeds A and B; lines with arrows denote reproduction, while lines without arrows denote selection; the size of the reference population for scenarios with purebred training was 1000 within each pure breed, and 2000 for the scenarios with crossbred training; thus all scenarios had a total reference population size of 2000.

offspring per dam. Random mating within each breed was also assumed and no selection was considered in this step.

Since we considered two types of training populations; crossbred and purebred, 400 males and 400 females were selected randomly from generation 7 of step 3 and were randomly mated to produce crossbred offspring, of which 2000 randomly selected animals served as the crossbred training population. Within each pure breed, 1000 randomly selected animals from generation 8 of step 3 were used as the purebred training population for prediction of additive and dominance effects. In subsequent generations (step 4), a two-way crossbreeding program with five generations of selection was simulated. There was no updating of predicted marker effects. The goal was to improve CP through selection in both parental breeds and the selection criterion in both purebred lines was based on genomic estimated breeding values for crossbred performance (GEBVC). The phenotypic mean of crossbreds was computed for each generation of selection (AB1 to AB5) to evaluate the realized cumulative response to selection.

4.2.3 Genome and trait phenotypes

A genome consisting of one chromosome of 100 cM with 100 segregating QTL and 1000 SNPs was simulated (Table 4.2). This small genome size was chosen to limit computing time. In addition, since our objective was to compare CP between simulated scenarios, the absolute level of response to selection and accuracy were not of primary interest. Both QTL and SNPs were randomly distributed over the chromosome. To obtain the required number of segregating loci after 2000 generations, about two to three times as many bi-allelic loci were simulated by sampling initial allele frequencies from a uniform distribution and applying a recurrent mutation rate of 2.5 × 10-5. Mutation rates of loci were determined on the basis of the number of polymorphic loci in generation 2000 of the preliminary analysis that were necessary to obtain 1000 polymorphic SNPs and 100 QTL. SNPs and QTL were distinct loci and were randomly drawn from segregating loci, with a minor allele frequency (MAF) higher than 0.05, in generation 2000. It should be

noted that this MAF criterion refers to the common ancestral population 300 generations back and thus, lower MAF can occur in the reference population.

The additive effect (a) of a QTL, defined as half the difference in genotypic value between alternate homozygotes, was sampled from a gamma distribution (0.4, 1.66). Dominance effects (d) were defined as the deviation of the genotypic value of the heterozygote from the mean of the genotypic values of the two homozygotes. Similar to Wellmann and Bennewitz (Wellmann and Bennewitz, 2011, 2012), first, dominance degrees at the i^{th} QTL (h_i) were sampled from a normal distribution, $N(0.5, 0.1)$, and then dominance effects were calculated as $d_i = h_i . |a_i|$, where $|a_i|$ is the absolute value of the additive effect for each QTL. Thus, the absolute magnitudes of additive and dominance effects were not independent, i.e. loci with large additive effects were also more likely to have large dominance effects. Moreover, since the average h was greater than zero, the average dominance effect was greater than zero, indicating directional dominance. The additive and dominance effects were scaled for each replicate of each scenario to reach additive and dominance variances of 0.3 and 0.1, respectively. After scaling, about 12% of the loci showed overdominance. Furthermore, additive and dominance effects of QTL alleles were assumed to be the same in both breeds. In other words, G×E interactions and epistasis were not simulated. The phenotypes of the trait were simulated by adding a standard normal residual effect to the genotypic value of each animal. The variance of the residual effects was chosen such that broad sense heritability H^2 of the trait was equal to 0.4. As a result, phenotypic variance (σ_p^2), narrow sense heritability h^2 and dominance variance were equal to 1, 0.3 and 0.1, respectively.

Table 4.2 Parameters of the simulated genome

Number of chromosomes	1
Number of markers	1000
Marker distribution	Random
Number of QTL	100
QTL distribution	Random
Initial MAF for markers	0.05
Initial MAF for QTL	0.05
Additive allelic effects for markers	Neutral
Additive allelic effects for QTL	Gamma (0.4,1.66)
Dominance degree for QTL (h_i)	$N(0.5,0.1)$
Dominance allelic effects for QTL	$d_i = h_i . \|a_i\|$
Rate of recurrent mutation	2.5 × 10−5

4.2.4 Prediction of marker effects

The Bayesian ridge regression was used to predict marker effects. We used the BGLR "Bayesian general linear regression" R package developed by Perez and de los Campos (2014) and its built-in default rules to set values of hyper-parameters. The following two models were used to predict the genetic effects associated with each marker:

(a) The first model was used for Scenarios 1, 2, 3 and 5, for which alternate heterozygotes (Aa and aA) could not be distinguished. The model used to predict genotypic values was as follows:

$$y_i = \sum X_{AA_{ij}} g_{1j} + \sum X_{Aa_{ij}} g_{2j} + \sum X_{aa_{ij}} g_{3j} + e_i,$$

where y_i is the phenotypic value of individual i in the training data. For Scenarios 1, 2 and 5, $X_{..ij}$ is an indicator variable of the genotype of individual i at SNP j, with $X_{AA_{ij}} = 1$ when individual i is AA and $X_{AA_{ij}} = 0$ otherwise. Similarly, $X_{Aa_{ij}} = 1$ when individual i is Aa and $X_{Aa_{ij}} = 0$ otherwise, and with $X_{aa_{ij}} = 1$ when individual i is aa and $X_{aa_{ij}} = 0$ otherwise. g_{1j}, g_{2j} and g_{3j} are the random unknown genotype effects for marker j, and e_i is the residual effect for animal i. The Σ denotes summation over all SNPs j.

For Scenario 3, animals in the training population had phenotypes but no genotypes. Therefore, in this scenario, $X_{..ij}$ were genotype probabilities based on the genotypes of parents, rather than indicator variables. To calculate the three genotype probabilities P(AA), P(Aa), and P(aa) for a bi-allelic SNP with two alleles, A and a, for animal i, we considered the genotyped sire and dam of the animal. For any genotyped parent, the probability to transmit allele A is P(A) = 1 for the homozygous state (AA), P(A) = 0.5 for the heterozygous state (Aa and aA), and P(A) = 0 for the alternative homozygous state (aa). The probability to transmit allele a is P(a) = 1-P(A). Thus, based on the genotypes of the parents, the values of X were equal to 0, 0.25, 0.5 or 1. For example, if both the sire and dam of animal i were heterozygous (Aa or aA), then the probabilities of observing genotypes AA, Aa and aa in the offspring were equal to 0.25, 0.5 and 0.25, respectively.

(b) The second model was used for Scenarios 4 and 6, for which alternate heterozygotes Aa and aA could be distinguished, and was as follows:

$$y_i = \sum X_{AA_{ij}} g_{1j} + \sum X_{Aa_{ij}} g_{2j} + \sum X_{aA_{ij}} g_{3j} + \sum X_{aa_{ij}} g_{4j} + e_i.$$

For Scenario 6, X-elements are the same as for Scenarios 1, 2 and 5. However, since, in this model, alternate heterozygotes Aa and aA could be distinguished in

crossbreds, $X_{Aa_{ij}} = 1$ when individual i is Aa and $X_{Aa_{ij}} = 0$ otherwise, while $X_{aA_{ij}} = 1$ when individual i is aA and $X_{aA_{ij}} = 0$ otherwise.

For Scenario 4, animals used for training had phenotypes but no genotypes and thus, X_{ij} were genotype probabilities based on the genotypes of parents, rather than indicator variables. Since in this model, alternate heterozygotes Aa and aA could be distinguished in crossbreds, four genotype probabilities P(AA), P(Aa), P(aA), and P(aa) were considered. For example, if the sire and dam of animal i were both heterozygous (Aa) at marker j, the probabilities of observing any of the genotypes AA, Aa, aA and aa in a crossbred offspring were all equal to 0.25.

4.2.5 True and genomic estimated breeding values

The true breeding value for crossbred performance (TBVC) for each animal was calculated as the expected genotypic value in the offspring of a parent carrying a certain QTL-genotype, when this parent was randomly mated to an individual of the other pure line. For crossbred offspring, the expected genotype frequencies of the offspring of a parent depend on the allele frequencies in the other pure line (denoted ŕ here). Thus, for animal i from breed r, the true breeding value for CP was calculated as:

$$\begin{aligned} TBVC_{ir} = & \sum_{j=1}^{100} [(X_{AA_{ij}})(p_{jŕ}a_j + q_{jŕ}d_j)] \\ & + [(X_{Aa\&aA_{ij}})(0.5p_{jŕ}a_j + 0.5q_{jŕ}d_j + 0.5p_{jŕ}d_j - 0.5q_{jŕ}a_j)] \\ & + [(X_{aa_{ij}})(-q_{jŕ}a_j + p_{jŕ}d_j)], \end{aligned} \quad (1)$$

where $X_{AA_{ij}}, X_{Aa\&aA_{ij}}$ and $X_{aa_{ij}}$ are indicator variables of the genotype at the j^{th} QTL of the i^{th} purebred individual. Thus, $X_{AA_{ij}} = 1$ when the genotype is AA and zero otherwise, $X_{Aa\&aA_{ij}} = 1$ when the genotype is Aa or aA and 0 otherwise and $X_{aa_{ij}} = 1$ when the genotype is aa and 0 otherwise. Moreover, $p_{jŕ}$ and $q_{jŕ}$ are the allelic frequencies (A and a) for the j^{th} QTL in breed ŕ and a_j and d_j are true additive and dominance effects at the j^{th} QTL. For example, for a parent with genotype AA at locus j, a fraction $p_{jŕ}$ of its offspring will have genotype AA, while a fraction $q_{jŕ}$ of its offspring will have genotype Aa. Hence, for locus j, the breeding value of this parent equals $(p_{jŕ}a_j + q_{jŕ}d_j)$, which is the first term of Equation 1. Equations 1 and 2 are simply the expected crossbred progeny averages for an animal with a certain genotype. These could also have been calculated from Fisher's average effect (Falconer and Mackay, 1996) for CP, which would yield identical results.

Genomic estimated breeding values were calculated in the same way except that SNP genotypes rather than QTL genotypes, and predicted genotypic effects were used. Thus, for Scenarios 1, 2, 3 and 5, genomic predicted breeding values for crossbred performance (GEBVC) for animal i from breed r was calculated as:

$$\begin{aligned}\mathrm{GEBVC_{ir}} = &\sum_{j=1}^{1000}[(X_{AA_{ij}})(p_{j\acute{r}}\hat{g}_{1j} + q_{j\acute{r}}\hat{g}_{2j})]\\ &+[(X_{Aa\&aA_{ij}})(0.5p_{j\acute{r}}\hat{g}_{1j}+0.5q_{j\acute{r}}\hat{g}_{2j} + 0.5p_{j\acute{r}}\hat{g}_{2j}+0.5q_{j\acute{r}}\hat{g}_{3j})]\\ &+[(X_{aa_{ij}})(q_{j\acute{r}}\hat{g}_{3j} + p_{j\acute{r}}\hat{g}_{2j})], \qquad (2)\end{aligned}$$

where, $\hat{g}_{1j}$, $\hat{g}_{2j}$ and $\hat{g}_{3j}$ are predicted genotypic effects for SNP genotypes AA, Aa and aA, and aa, respectively.

In Scenarios 4 and 6, for which alternate heterozygotes Aa and aA could be distinguished, GEBVC for animal i from breed r was calculated as:

$$\begin{aligned}\mathrm{GEBVC_{ir}} = &\sum_{j=1}^{1000}[(X_{AA_{ij}})(p_{j\acute{r}}\hat{g}_{1j} + q_{j\acute{r}}\hat{g}_{2j})]\\ &+[(X_{Aa\&aA_{ij}})(0.5p_{j\acute{r}}\hat{g}_{1j}+0.5q_{j\acute{r}}\hat{g}_{2j} + 0.5p_{j\acute{r}}\hat{g}_{3j}+0.5q_{j\acute{r}}\hat{g}_{4j})]\\ &+[(X_{aa_{ij}})(q_{j\acute{r}}\hat{g}_{4j} + p_{j\acute{r}}\hat{g}_{3j})], \qquad (3)\end{aligned}$$

where $\hat{g}_{1j}, \hat{g}_{2j}, \hat{g}_{3j}$ and $\hat{g}_{4j}$ are predicted genotypic values of AA, Aa, aA and aa genotypes at the j^{th} marker, respectively.

4.2.6 Correlation of LD phase between breeds A and B

Correlation of LD phase between breeds A and B was estimated to evaluate the degree of relatedness between the two simulated breeds. To estimate persistence of LD phase between two lines, only the segregating SNPs with a MAF greater than 0 in both breeds were included in the analysis. Persistence of LD phase was estimated following Badke et al. (2012) as follows:

$$R_{AB} = \frac{\sum(r_{ij(A)}-\bar{r}_A)(r_{ij(B)}-\bar{r}_B)}{sd(A)sd(B)},$$

where $R_{A,B}$ is the correlation of phase between $r_{ij(A)}$ in breed A and $r_{ij(B)}$ in breed B, r_{ij} is the correlation coefficient as a measure of LD for each pair of SNPs, sd(A) and sd(B) are the standard deviations of $r_{ij(A)}$ and $r_{ij(B)}$, respectively, and $\bar{r}_A$ and $\bar{r}_B$

are the average r_{ij} across all SNPs i and j within interval p for breeds A and B, respectively.

4.3 Results

4.3.1 Purebred-crossbred genetic correlation

The genetic correlation between TBVP and TBVC, which is known as the purebred-crossbred genetic correlation ([r_{pc}, Wei and Vanderwerf (1994)) was on average equal to 0.78±0.02. Since G×E interaction was not included in the simulations, the deviation of r_{pc} from 1 was purely due to dominance effects and differences in allele frequencies between the two purebred lines.

4.3.2 Response to selection in crossbreds

The increase in phenotypic mean of crossbred animals was measured over four generations of selection in the six simulated scenarios for which breeds A and B had diverged 300 generations back (Figure 4.2). Ranking of scenarios in terms of phenotypic mean of crossbreds showed that training on crossbreds (Scenarios 3, 4, 5 and 6) resulted in greater response to selection than training on the pure lines separately (Scenario 1) or on the pure lines combined (Scenario 2), although selection was based on GEBVC in all cases and no G×E interaction was included. Response to selection was greater when training was on crossbred animals for

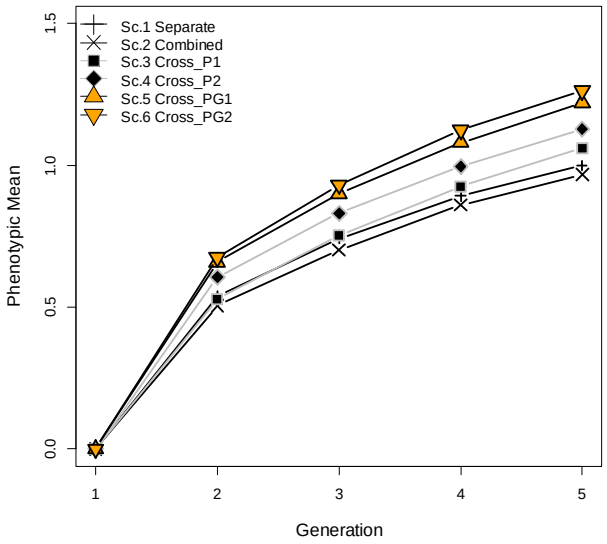


Figure 4.2 Phenotypic mean of crossbred animals.

Scenario 1: separate training in both breeds A and B. Scenario 2: training on a combined set of animals from both breeds A and B. Scenario 3: training on crossbred animals with phenotypes and genotype probabilities and it was assumed that the alternate heterozygotes Aa and aA were identical in crossbred animals. Scenario 4: training on crossbred animals with phenotypes and genotype probabilities and it was assumed that the alternate heterozygotes could be distinguished in crossbred animals. Scenario 5: training on crossbred animals with phenotypes and genotypes and it was assumed that the alternate heterozygotes were identical in crossbred animals. Scenario 6: training on crossbred animals with phenotypes and genotypes and it was assumed that the alternate heterozygotes could be distinguished in crossbred animals; standard errors of phenotypic means ranged from 0.02 to 0.03.

which both phenotypes and genotypes were available (Scenarios 5 and 6) than when training was on crossbreds for which only phenotypes were available and genotype probabilities based on their parents' genotypes were used (Scenarios 3 and 4). In addition, when alternate heterozygotes Aa and aA could be distinguished in crossbred animals (Scenario 6), response to selection was greater than when they could not be distinguished (Scenario 5). Similarly, response to selection was greater when training was on genotype probabilities of crossbred animals for which alternate heterozygotes Aa and aA could be distinguished (Scenario 4) than when they were pooled together (Scenario 3). The phenotypic mean of crossbreds increased when breeds A and B had separate training populations (Scenario 1) compared to when a common training population consisting of animals from both breeds A and B was used (Scenario 2).

Finally, the difference in the amount of response to selection in the first generation compared to that in the subsequent generations is due to marker effects being estimated in the base generation only and to using these estimates to calculate the GEBVC in all subsequent generations. Thus, there was no retraining in each generation and GEBVC accuracy decreased over generations of selection, which caused a decline in genetic gain.

4.3.3 Response to selection in purebreds

CP can be written as CP = BA + H, where BA denotes the breed average of pure lines and H the heterosis present in crossbreds (Falconer and Mackay, 1996). Thus, the observed superiority of some scenarios may be due to a greater response in BA or in H, or in both. The cumulative response to selection averaged over breeds A and B for four generations of selection is in Table 4.3. Contrary to what was observed for response to selection for CP, response to selection within pure lines was greater when training was on pure lines although selection was based on GEBVC in all scenarios. Response to selection was greatest for Scenario 1 and

smallest for Scenarios 3 and 4 with training on crossbred animals and using their genotype probabilities. However, when training was on phenotypes and genotypes of crossbreds (Scenarios 5 and 6), response to selection within pure lines was almost comparable to that for scenarios with training on pure lines. Similar to response for CP, response to selection within pure lines was greater when the alternate heterozygotes Aa and aA could be distinguished, i.e. Scenario 4 performed better than Scenario 3 and Scenario 6 performed better than Scenario 5.

Table 4.3 Mean phenotypic average of pure lines.

	G1	G2	G3	G4	G5
Scenario 1	0.00	0.55	0.72	0.85	0.93
Scenario 2	0.00	0.50	0.67	0.78	0.86
Scenario 3	0.00	0.35	0.48	0.57	0.62
Scenario 4	0.00	0.42	0.54	0.63	0.70
Scenario 5	0.00	0.49	0.64	0.75	0.83
Scenario 6	0.00	0.49	0.66	0.77	0.87

Scenario 1: separate training in both breeds A and B; Scenario 2: training on a combined set of animals from both breeds A and B; Scenario 3: training on crossbred animals with phenotypes and genotype probabilities and it was assumed that the alternate heterozygotes Aa and aA were identical in crossbred animals; Scenario 4: training on crossbred animals with phenotypes and genotype probabilities and it was assumed that the alternate heterozygotes could be distinguished in crossbred animals. Scenario 5: training on crossbred animals with phenotypes and genotypes and it was assumed that the alternate heterozygotes were identical in crossbred animals. Scenario 6: training on crossbred animals with phenotypes and genotypes and it was assumed that the alternate heterozygotes could be distinguished in crossbred animals; standard errors of phenotypic means for simulated scenarios in generation 5 ranged from 0.03-0.04.

4.3.4 Heterosis in crossbreds

Heterosis refers to the superior performance of crossbred animals compared to the average performance of its purebred parents. Figure 4.3 shows the amount of heterosis over generations for the simulated scenarios. Total heterosis was calculated as the sum of heterosis at each locus based on $H = \sum d_l(p_{A,l} - p_{B,l})^2$, where d_l is the dominance effect at QTL l, $p_{A,l}$ is the allele frequency at QTL l in breed A, and $p_{B,l}$ is the allele frequency at QTL l in breed B (Falconer and Mackay, 1996). In all scenarios, the amount of heterosis increased over generations, however, the rate of increase differed among scenarios. The amount of heterosis in Scenarios 1 and 2 increased a little from generation 1 to 5, whereas it increased much more sharply in the other scenarios in which training was on crossbreds. Since heterosis depends on the differences in allele frequencies between two breeds, this increase suggests that training on crossbreds together with selection

for CP result in diverging allele frequencies between the two breeds. This could be caused by allele frequencies moving in different directions in both breeds or by selection acting on different loci in the two breeds.

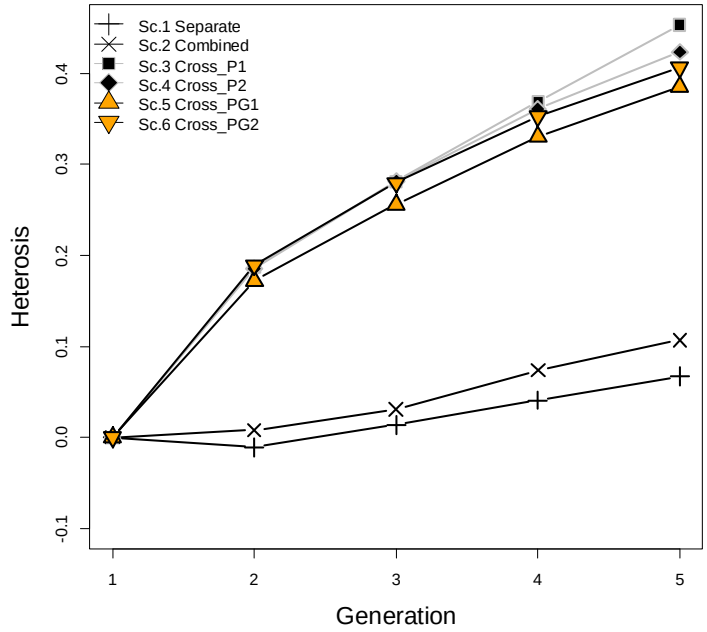


Figure 4.3 Heterosis in crossbreds.
Scenario 1: separate training in both breeds A and B; Scenario 2: training on a combined set of animals from both breeds A and B; Scenario 3: training on crossbred animals with phenotypes and genotype probabilities and it was assumed that the alternate heterozygotes Aa and aA were identical in crossbred animals; Scenario 4: training on crossbred animals with phenotypes and genotype probabilities and it was assumed that the alternate heterozygotes could be distinguished in crossbred animals. Scenario 5: training on crossbred animals with phenotypes and genotypes and it was assumed that the alternate heterozygotes were identical in crossbred animals. Scenario 6: training on crossbred animals with phenotypes and genotypes and it was assumed that the alternate heterozygotes could be distinguished in crossbred animals.

4.3.5 Correlation of LD phase between breeds A and B

We estimated the correlation of LD phase between breeds A and B for the scenarios in which time of breed divergence (1, 50, 100, 200, 300 and 400 generations back) varied. In these scenarios, the correlation of LD phase for SNPs with a pairwise distance of 1 cM decreased as the number of generations since separation increased, i.e. correlations of 0.39, 0.22, 0.11, 0.05, 0.0 and -0.04 were

obtained for scenarios including 1, 50, 100, 200, 300 and 400 generations since divergence, respectively.

4.3.6 Effect of being able to distinguish between alternate heterozygotes

Table 4.4 shows the effect of being able to distinguish between alternate heterozygotes Aa and aA by comparing Scenarios 5 and 6, for different times since breeds A and B diverged. Time since divergence affected the relative ranking of Scenarios 5 and 6: when the two breeds were closely related (i.e. 1, 50 and 100 generations of separation), response to selection for CP was greater for Scenario 5 than for Scenario 6, when time since divergence increased to 200 generations, response to selection was almost the same for both scenarios and when time since divergence increased to 300 and 400 generations, response to selection was greater for Scenario 6 than for Scenario 5. Thus, these results showed that being able to distinguish between alternate heterozygotes Aa and aA (Scenario 6) increases response to selection when breeds have diverged a long time ago.

Table 4.4 Mean phenotype of crossbreds in generation five without or with distinguishing between both heterozygotes (Scenario 6 vs Scenario 5), for different times since divergence of the pure lines

Scenarios	Time since divergence				
	1	50	100	200	400
Scenario 5	1.21	1.32	1.33	1.20	0.94
Scenario 6	1.15	1.28	1.30	1.19	0.99

Scenario 5: training on crossbred animals with phenotypes and genotypes and it was assumed that the alternate heterozygotes were identical in crossbred animals. Scenario 6: training on crossbred animals with phenotypes and genotypes and it was assumed that the alternate heterozygotes could be distinguished in crossbred animals; standard errors of phenotypic means ranged from 0.02 to 0.03; note that the mean phenotype of crossbreds cannot be compared for different times since divergence, as they are results of distinct simulations.

4.3.7 Effect of the training population size on the response to selection

Figure 4.4 shows the cumulative response to selection in Scenarios 5 and 6 for varying sizes of the training population. To evaluate the impact of training population size on the relative ranking of Scenarios 5 and 6, 200 generations of divergence between breeds A and B were considered, since response to selection for these two scenarios was almost the same for this time since divergence and a training population size of 2000. As expected, response to selection with both scenarios increased as the size of the training population increased. However, the

relative ranking of Scenarios 5 and 6 changed as the size of training population increased. If the size of the training population was 500, response to selection was greater for Scenario 5 than for Scenario 6, but with a 4- and 16-fold increase, response to selection was greater for Scenario 6 than for Scenario 5. Thus, these results showed that being able to distinguish between alternate heterozygotes Aa and aA was beneficial when the training population was large.

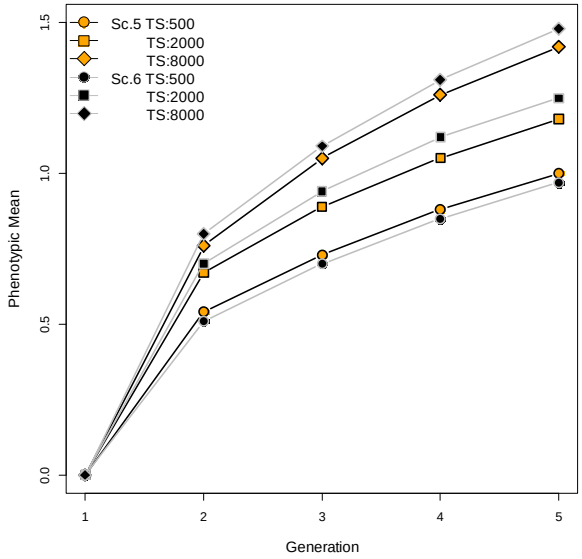


Figure 4.4 Cumulative response to selection for varying sizes of training populations.
Training on crossbred animals with phenotypes and genotypes. Scenario 5: Alternate heterozygotes Aa and aA were assumed identical in crossbred animals; Scenario 6: Alternate heterozygotes could be distinguished in crossbred animals; TS stands for training population size of 500, 2000 and 8000; standard errors of phenotypic means ranged from 0.02 to 0.03.

4.4 Discussion

We investigated response to selection in crossbred performance in a two-way crossbreeding system of two related breeds for five generations. To estimate SNP effects, training was either on pure lines or crossbred animals, animals were selected on GEBVC, and there was no G×E interaction. Thus, the deviation of the purebred-crossbred genetic correlation (r_{pc}) from 1 originated purely from dominance effects and differences in allele frequencies between the two purebred lines. We also investigated the effect of being able to distinguish between alternate heterozygotes Aa and aA in crossbred animals.

4.4.1 Training on crossbred animals vs pure lines

A general finding of our study was that training on crossbred animals led to greater phenotypic response in crossbred animals compared to training on purebred lines. To identify the potential reasons for the superiority of training on crossbred animals, we partitioned the EBV of animals in the pure lines into components due to additive and dominance effects (see Esfandyari et al. (2015) for partitioning of breeding values). We found that, by training on crossbred animals, we could predict dominance effects and consequently breeding values of the animals in pure lines more accurately. Accuracy of EBV due to dominance effects when training was on crossbred animals was on average equal to 0.24, whereas when training was on pure lines, it was equal to 0.16 [see Additional file 1]. This indicates that a higher prediction accuracy of dominance effects by training on crossbred animals is associated with a higher level of heterozygosity in the crossbred animals. Observed heterozygosity in the crossbred training population was 0.49 on average, which was higher than that found for the pure lines, i.e. 0.33 and 0.34 on average for breeds A and B, respectively. Logically, dominance effects can be predicted more accurately when the level of heterozygosity is higher.

In this study, we did not simulate environmental differences for purebred and crossbred animals. However, in practice, environments in which purebreds and crossbreds are kept are often different. Thus, selection of purebreds to improve crossbred performance in a commercial environment involves not only the r_{pc} caused by non-additive genetic effects, but also a possible G×E interaction (Dekkers and Chakraborty, 2004). For instance, for a r_{pc} of 0.8 due to dominance effects, it might be possible to reach the maximum accuracy (i.e. 1) by using an infinite amount of information on purebred animals under a dominance model. However, for a r_{pc} of 0.8 only due to G×E interactions, the maximum achievable accuracy by using purebred information is 0.8. Thus, the mechanism that results in r_{pc} less than 1 has an impact on response to selection under a dominance model. Nonetheless, by using crossbred data, it might be possible to reach maximum accuracy as well. Thus, a loss in genetic gain should be expected in the presence of G×E interactions compared to no G×E interactions. In other words, if a r_{pc} less than 1 was partly due to G×E interactions, training on crossbred animals would be even more beneficial than the results show in this study.

Although training on crossbred animals led to greater response to selection in crossbreds, it requires the collection of data at the crossbred level. Since commercial crossbred animals are usually not individually identified and individual performances are not recorded, it might be difficult and expensive to collect phenotype and genotype data on crossbred individuals, whereas most breeding

programs have routine phenotyping and genotyping of nucleus animals in the pure lines. If genotyping but not phenotyping of crossbred animals is a limiting factor, one could do training on crossbred animals with phenotypes and use genotype probabilities based on the genotypes of their purebred parents (Scenarios 3 and 4 in our simulation). With this strategy, it is possible to gain some of the benefits of crossbred training without genotyping crossbred animals (see Figure 4.2). However, this strategy does require pedigree identification of crossbreds.

4.4.2 Distinguishing between heterozygotes in crossbred animals

Our results showed that being able to distinguish between alternate heterozygotes Aa and aA in crossbred animals and to predict two distinct genetic values for these genotypes will lead to greater response to selection in crossbreds when the two purebred lines are distantly related. The reasons for this superiority are both differences in SNP and QTL frequencies between the two lines as well as differences in the amount and extent of LD between SNPs and QTL between the lines. Any difference in QTL and SNP frequencies and in LD between the pure lines can result in the two alternate heterozygotes at a SNP having different probabilities for a heterozygous QTL in the crossbreds. These differences suggest that one should distinguish between the two alternate heterozygotes in the crossbred when a dominance model is used for crossbred training.

Due to the genetic differences among the pure lines, we expected that being able to distinguish between alternate heterozygotes when training on crossbreds would always perform better. However, we found that this superiority was associated with two other factors; time since divergence of the two lines and number of records used in the training. The results showed that being able to distinguish between alternate heterozygotes was favourable only for distantly related lines (Table 4.4). In fact, in distantly related lines, the chance that recombination breaks down the shared ancestral haplotypes (and even reverse the LD phase) across the populations is greater. Hence, reverse LD phase between SNPs and QTL between the two lines for distantly related breeds can cause the two alternate heterozygotes at a SNP to have different QTL alleles in the crossbreds. Apparently, by predicting two genetic effects for alternate heterozygote genotypes, this difference in LD phase was captured and resulted in greater response to selection when pure lines were distantly related.

In our simulations, the number of records used in the training population also contributed to the observed differences in response for Scenarios 5 and 6. We found that with a small number of records used in the training data, response to selection was greater in Scenario 5 than in Scenario 6 (Figure 4.4). This is probably

due to the difference in number of effects that need to be predicted in the two scenarios. For Scenario 6, where alternate heterozygotes could be distinguished, four genotypic effects had to be predicted, whereas for Scenario 5 only three genotypic effects had to be predicted. Hence, because the number of effects to be predicted in Scenario 6 was greater, it was at a disadvantage over Scenario 5 with a small number of records. However, this disadvantage disappears as the training population size increases. In other words, as the number of records used for training increases, more information becomes available to predict the effects of SNPs and, given the sufficient number of records for training, differences in SNP effects between lines render Scenario 6 more advantageous. This result agrees with those of Ibanez-Escriche et al. (2009), who showed that breed-specific allele substitution effects (BSAM) will have an advantage over across-breed allele substition effects, provided sufficient information is available for estimating the additional breed-specific effects.

Finally, it should be mentioned that a prerequisite for distinguishing between alternate heterozygotes in our study and for the implementation of BSAM in Ibanez-Escriche et al. (2009) is that the purebred origin of SNP alleles in crossbreds is known, which may not be easily obtained for real data. Nevertheless, given the very high SNP density, it may be possible to trace alleles to ancestors accurately (Meuwissen and Goddard, 2007). In a recent study, Bastiaansen et al. (2014) suggested a method to determine breed origin of alleles in crossbreds using long-range phasing without the need for tracking pedigree relationships of crossbreds. Based on this method, it is not even necessary to have close relationship between the crossbred and genotyped purebred animals since long-range phasing will work even with distant purebred relatives of the crossbreds (Bastiaansen et al., 2014). Hence, tools are available to distinguish between alternate heterozygotes, and also to take advantage of the associated benefits in practical situations.

4.4.3 Simulation model

For reasons of computation time, simulation studies usually use a genome size which is smaller than that of most livestock species (Meuwissen et al., 2001). In our simulations, we used a genome with one chromosome 100 cM long. By assuming a phenotypic variance of 1, QTL on this chromosome result in an additive variance of 0.3. However, in real livestock genomes (e.g., a genome of 30 M for cattle), QTL on a chromosome of this length would cause an additive variance of only ~0.01. One consequence is that the sizes of the QTL effects in our simulation are substantially larger than those of real QTL, which means that the effects of simulated QTL were predicted more accurately than what may be possible with a real dataset.

Daetwyler et al. (2008) and Goddard (2009) predicted that the accuracy of genomic selection depends on the parameter $\rho^2 = \frac{Th^2}{ML}$, where h^2 is the heritability of the trait, T is the number of records in the training data, M is the effective number of loci per Morgan ($2Ne$), and L is the genome size in Morgan. This relationship predicts that accuracy will be the same for all cases where ρ^2 is the same. So, under optimal conditions, a genome of 30 chromosomes of 1 M each requires 30 times as many training records to achieve the same accuracy as a genome with 1 chromosome 1 M long.

We did not check whether the number and effect of QTL or the density of SNPs affected the relative ranking of Scenarios 5 and 6. However, most probably by increasing the genome size and keeping all other parameters constant (i.e., SNP density, training population size and values of variance components), Scenario 6 would be at a disadvantage over Scenario 5 due to the greater number of effects that need to be predicted. This suggests that the benefit of being able to distinguish between alternate heterozygotes is expected to decrease as the genetic architecture becomes more polygenic. In addition, SNP density may affect the difference between Scenarios 5 and 6 as well. As SNP density increases, the model will include SNPs that are closer to the QTL. In a finite population, SNP alleles that are closer to the QTL will more accurately reflect the state of the QTL alleles (Ibanez-Escriche et al., 2009). Thus, as the SNP density increases, the need for distinguishing between alternate heterozygotes may be reduced.

Besides the small genome size that may cause overestimation of the accuracy of GEBV in our simulation, additive effects were sampled from a gamma distribution, which results in some QTL with a large effect that may account for a substantial part of the additive variance. Hence, genomic breeding values may be predicted more accurately than for a purely polygenic trait. In addition, in real populations, QTL effects may be line-dependent due to epistatic interactions, which may be negligible if selection is performed within a population but not if effects are estimated simultaneously for several populations. In fact, presence of epistatic interactions among genes may cause the lack of consistency across breeds. In this case, the effect of a particular QTL depends on the allelic frequency of genes it interacts with (Carlborg et al., 2003). Since these frequencies can differ among breeds it results in breed-specific effects. Thus, combining animals from two breeds into a single training population may not be advantageous in the presence of substantial epistasis.

In this study, generation interval was the same for all scenarios with purebred or crossbred training. In other words, randomly selected sires of breed A in

generation 2307 were mated to the dams of this breed to produce purebred offspring and also to the dams of breed B to produce crossbred offspring. Training was on randomly selected individuals from these offspring. Thus, scenarios with crossbred training did not require an additional generation compared to purebred training to create the training population.

Finally, the difference in the amount of response to selection in the first generation compared to that in the subsequent generations is due to marker effects being estimated in the base generation only and to using these estimates to calculate the GEBVC in all subsequent generations. Thus, there was no retraining in each generation and GEBVC accuracy decreased over generations of selection, which caused a decline in genetic gain.

In our simulations, regardless of whether training was on pure lines or crossbreds, a dominance model based on own performance of the animals was used to estimate the GEBVC for the selection candidates. However, an alternative approach would be to carry out training on pure lines based on the yield deviations of their crossbred progeny and to use an additive model to estimate breeding values. In other words, training can be done on purebred animals with genotypes and the mean phenotypes of their crossbred progeny can be used as response variable. We compared performance of Scenario 1 to such a scenario (referred to as additive scenario, hereafter) where training was on purebred animals, mean performance of crossbred progeny was used as response variable and an additive model was used to estimate GEBV. The size of the reference population for the additive scenario was 1000 within each pure line and each of the animals in the training set had 10 crossbred progeny. Result showed that using crossbred progeny information yielded a greater response to selection than using the animals' own records although in both cases, training was on pure lines. [See Additional file 2]. Scenario 1 with a dominance model resulted in a smaller breed average response compared to the additive scenario which resulted in a smaller overall crossbred response. The superiority of the additive scenario is due to the increased accuracy of selection in pure lines by using crossbred progeny information. In fact, for Scenario 1, own performance of the animals in the training set was used as response variable, whereas for the additive scenario more information was available by using the mean performance of 10 crossbred progeny. Zeng et al. (2013) compared additive and dominance models for GS in purebreds for CP and came to the conclusion that, when dominance is the sole driver of heterosis, using a dominance model for GS is expected to result in greater cumulative response to selection of purebred animals for CP than an additive model.

4.5 Conclusion

Genomic selection can be very valuable in crossbreeding programs since it allows efficient selection for crossbred performance. To reach greater response to selection when crossing two distantly related lines, it is better to do training on crossbred animals rather than on pure lines to predict genetic effects. In addition, being able to distinguish between alternate heterozygotes in the crossbred training set by taking into account the breed origin of alleles increases response to selection, except when breeds are closely related and the reference population is small. Finally, our results showed that response to selection in crossbreds was greater when both phenotypes and genotypes were collected on crossbreds, compared to having only phenotypes on the crossbreds and genotypes on their parents.

4.6 Authors' contributions

HE, ACS and PB conceived the study. HE carried out the simulation and drafted the manuscript. PB helped in the interpretation of results. ACS and PB edited the drafted manuscript. All authors read and approved the final manuscript.

4.7 Acknowledgements

HE has benefited from a joint grant from the European Commission and Aarhus University within the framework of the Erasmus-Mundus joint doctorate 'EGS-ABG'.

4.8 Appendix A

Partitioning accuracies of breeding values due to additive and dominance effects for Scenario 1 and Scenario 5.

Partitioning accuracies of breeding values due to additive and dominance effects

		Scenario 1			Scenario 5		
Breed A	G	BV	Add	Dom	BV	Add	Dom
	1	0.65	0.69	0.19	0.80	0.57	0.36
	2	0.49	0.56	0.18	0.61	0.40	0.22
	3	0.38	0.50	0.20	0.50	0.37	0.19
	4	0.30	0.47	0.19	0.43	0.34	0.19
	5	0.26	0.44	0.19	0.39	0.32	0.24
	Mean	0.42	0.53	0.19	0.54	0.40	0.24
		Scenario 1			Scenario 5		
Breed B	G	BV	Add	Dom	BV	Add	Dom
	1	0.64	0.71	0.15	0.79	0.61	0.36
	2	0.45	0.58	0.13	0.60	0.44	0.22
	3	0.35	0.50	0.11	0.49	0.38	0.19
	4	0.28	0.45	0.12	0.40	0.34	0.22
	5	0.26	0.43	0.15	0.32	0.30	0.22
	Mean	0.40	0.53	0.13	0.52	0.41	0.24

G: generation

BV: Accuracy of breeding values that is correlation between TBVC and GEBVC Add: Accuracy of breeding values due to additive effects. Dom: Accuracy of breeding values due to dominance effects.

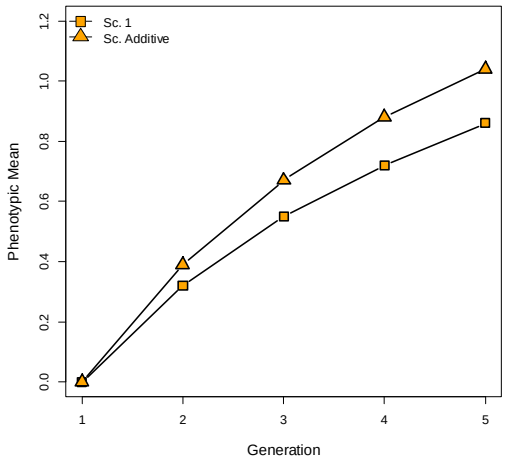
Scenario 1: Separate training in both breed A and B.

Scenario 5: Training on crossbred animals with phenotypes and genotypes. Two types of heterozygotes were assumed the same in crossbred animals.

4.9 Appendix B

Comparison of Scenario 1 with an additive scenario.

We compared the performance of Scenario 1 to an additive scenario (Scenario Additive) where training was on purebred animals, mean performance of crossbred progeny was used as response variable and an additive model was used to estimate genomic estimated breeding values. The size of the reference population for the additive scenario was 1000 within each pure line and each of the animals in the training set had 10 crossbred progeny.



References

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, and J. P. Steibel. 2012. Estimation of linkage disequilibrium in four US pig breeds. Bmc Genomics 13.

Bastiaansen, J. W. M., H. Bovenhuis, M. S. Lopes, F. F. Silva, H. J. Megens, and M. P. L. Calus. 2014. SNP Effects Depend on Genetic and Environmental Context. in Proc. 10th World Congress on Genetics Applied to Livestock Production, Vancouver, BC, Canada.

Bijma, P. and J. A. M. van Arendonk. 1998. Maximizing genetic gain for the sire line of a crossbreeding scheme utilizing both purebred and crossbred information. Anim Sci 66:529-542.

Carlborg, O., S. Kerje, K. Schutz, L. Jacobsson, P. Jensen, and L. Andersson. 2003. A global search reveals epistatic interaction between QTL for early growth in the chicken. Genome Res 13(3):413-421.

Comstock, R. E., H. F. Robinson, and P. H. Harvey. 1949. A Breeding Procedure Designed to Make Maximum Use of Both General and Specific Combining Ability. Agron J 41(8):360-367.

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

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PloS one 3(10).

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

Dekkers, J. C. M. and R. Chakraborty. 2004. Optimizing purebred selection for crossbred performance using QTL with different degrees of dominance. Genetics Selection Evolution 36(3):297-324.

Esfandyari, H., A. C. Sorensen, and P. Bijma. 2015. Maximizing crossbred performance through purebred genomic selection. Genetics Selection Evolution 47.

Falconer, D. S. and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. Vol. 4. 4 ed. Pearson

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

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

Hartmann, W. 1992. Evaluation of the Potentials of New Scientific Developments for Commercial Poultry Breeding. World Poultry Sci J 48(1):17-27.

Ibanez-Escriche, N., R. L. Fernando, A. Toosi, and J. C. M. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. Genetics Selection Evolution 41.

Kinghorn, B. P., J. M. Hickey, and J. H. J. van der Werf. 2010. Reciprocal Recurrent Genomic Selection for Total Genetic Merit in Crossbred Individuals. in Proc. Proceedings of the 9th World Congress on Genetics Applied to Livestock Production: 1-6 August 2010; Leipzig. Paper 36; 2010.

Lo, L. L., R. L. Fernando, and M. Grossman. 1993. Covariance between Relatives in Multibreed Populations - Additive-Model. Theoretical and Applied Genetics 87(4):423-430.

Meuwissen, T. H. E. and M. E. Goddard. 2007. Multipoint identity-by-descent prediction using dense markers to map quantitative trait loci and estimate effective population size. Genetics 176(4):2551-2560.

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

Perez, P. and G. de los Campos. 2014. Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics 198(2):483-U463.

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

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

Sonesson, A. K. and T. H. E. Meuwissen. 2009. Testing strategies for genomic selection in aquaculture breeding programs. Genetics Selection Evolution 41.

Wei, M. and H. Steen, van der.,. 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 J. H. J. Vanderwerf. 1994. Maximizing Genetic Response in Crossbreds Using Both Purebred and Crossbred Information. Anim Prod 59:401-413.

Wellmann, R. and J. Bennewitz. 2011. The contribution of dominance to the understanding of quantitative genetic variation. Genet Res 93(2):139-154.

Wellmann, R. and J. Bennewitz. 2012. Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genet Res 94(1):21-37.

Zeng, J., A. Toosi, R. L. Fernando, J. C. Dekkers, and D. J. Garrick. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genetics, selection, evolution : GSE 45(1):11.

5

Genomic prediction of crossbred performance based on purebred Landrace and Yorkshire data using a dominance model

Hadi Esfandyari[1,2*], Piter Bijma[2], Mark Henryon[3], Ole Fredslund Christensen[1], Anders Christian Sørensen[1]

[1] Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Denmark
[2] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands
[3] Seges, Danish Pig Research Centre, Copenhagen, 1609, Denmark

Submitted

Abstract

Background: In pig breeding, selection is usually carried out in purebred populations, although the final goal is to improve crossbred performance. Genomic selection can be used to select purebred parental lines for crossbred performance. Dominance is the likely genetic basis of heterosis and explicitly including dominance in the genomic selection model may be an advantage to select purebreds for crossbred performance. Thus, the first objective of this study was to compare the predictive ability of genomic prediction models with either additive, or both additive and dominance effects, when the validation criterion is crossbred performance. The second objective was to compare the use of two separate pure line reference populations to that of a single reference population that combines both pure lines.

Methods: The data used concerned pigs from two pure lines (Landrace and Yorkshire) and their reciprocal crosses, and the trait of interest was litter size in the first parity. Training was carried out on (i) the separate pure-bred sows of Landrace (2085) and Yorkshire (2145) and (ii) combined pure lines (4230) that were genotyped for 38k SNPs. Prediction accuracy was measured as the correlation between genomic-estimated breeding values of boars in pure lines and mean corrected crossbred-progeny performance, divided by the average accuracy of mean-progeny performance. Next to a model with additive effects only (MA), we evaluated a model with both additive and dominance effects (MAD). Two types of genomic estimated breeding values were computed; GEBV for purebred performance (GEBV) based on either MA or the MAD, and GEBV for crossbred performance (GEBV-C) based on MAD. GEBV-C was calculated based on SNP allele frequencies of genotyped animals in the opposite breed.

Results: Compared to MA, MAD improved prediction accuracy in both breeds. Within MAD, GEBV-C improved prediction accuracy compared to GEBV. Prediction accuracy for Landrace boars was 0.11 based on MA and 0.13 and 0.14 for GEBV and GEBVC based on MAD, respectively. The corresponding values for Yorkshire boars were 0.32, 0.34 and 0.36. Combining animals from both breeds into a single reference population yielded 12 to 46% higher accuracies than training separately in both pure lines. In conclusion, the use of a dominance model increased the accuracy of genomic predictions of crossbred performance that were based on purebred data.

5.1 Introduction

The effect of dominance, a non-additive genetic effect, has traditionally been ignored in genetic evaluation of livestock populations. There are three reasons for this. First, there is a lack of informative pedigrees, typically with large full-sib families, which are needed for accurate estimates of dominance effects (Misztal et al., 1998). Second, litter effects are often confounded with family effects, particularly in prolific species, such as chicken and pigs. Third, prediction of dominant effects involves complex computations that are often cumbersome (Misztal et al., 1998, Mrode and Thompson, 2005). The recent advent of dense SNP panels, however has ignited interest in the prediction of non-additive genetic effects (Su et al., 2012, Lopes et al., 2014, Moghaddar et al., 2014, Sun et al., 2014, Wittenburg et al., 2015). The availability of SNP genotypes increases the potential to estimate dominance effects, because it enables us to determine which animals are heterozygotes for each SNP locus and to predict the genotypic value of future matings (Toro and Varona, 2010). So, dense SNP panels provide the technology required to exploit dominance effects in genetic evaluations.

In some livestock production systems, including pigs, crossbreds are used in commercial production to utilize heterosis and complementary effects. The aim of selective-breeding programs in many of these systems is to maximize crossbred performance, where selection is carried out within pure-lines using data from purebred animals (Wei and Steen, 1991). However, traits that are evaluated in purebred populations may be genetically different from traits at the commercial production level, because the genetic correlations between crossbred and purebred performance (r_{pc}) are usually less than one (Wei and Vanderwerf, 1994, Dekkers, 2007). Evidence for r_{pc} values less than one has been observed in livestock species (Lutaaya et al., 2001, Zumbach et al., 2007). They are often caused by genotype by environment (G×E) interactions and non-additive (particularly dominance) genetic effects (Wei et al., 1991).

One of the problems in the implementation of genomic selection schemes in crossbreeding programs is whether to predict marker effects from pure line or crossbred data. When non-additive gene action or G×E exists, the performance of purebred parents is likely to be a poor predictor of the performance of their crossbred descendants. As a result, training on crossbred data has been suggested (Dekkers, 2007, Zeng et al., 2013, Esfandyari et al., 2015a). It is expected that training on crossbred data accounts for genetic differences between purebred and crossbred animals and for G×E. However, in practice, crossbred information is often not available, since both performance records and genotypes can be difficult or expensive to obtain on crossbred animals. An alternative would be training on pure

lines using a dominance model, and this would offer a solution if some part of deviation of r_{pc} from one is due to dominance (Esfandyari et al., 2015b).

Improvement in prediction accuracies by including dominance in the genomic evaluation models has been reported, but most studies were using purebred genomic selection models (Su et al., 2012, Ertl et al., 2014, Sun et al., 2014). It is expected that including dominance in genomic selection models for crossbred performance would provide further improvement in prediction accuracies, as dominance is a genetic basis for heterosis. Furthermore, dominance is expected to be one of the factors contributing to the deviation of r_{pc} from unity. So, we hypothesized that including dominance effects in genomic prediction models increases the prediction accuracy of purebred animals that are selected for crossbred performance. We tested this hypothesis in two ways. First, we compared the predictive ability of genomic prediction models with either additive, or both additive and dominance effects, when the validation criterion was crossbred performance. Second, we compared the use of two separate pure-line reference populations to a single reference population that combines both pure lines.

5.2 Methods

The data used concerned pigs from two pure lines (Landrace and Yorkshire) and their reciprocal crosses, and the trait of interest was litter size in the first parity (Figure 5.1). The data were supplied by the Danish Pig Research Centre (Copenhagen, Denmark).

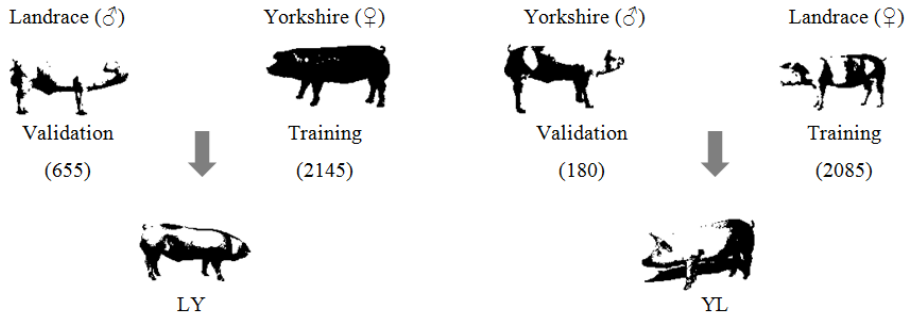


Figure 5.1 Schematic representation of the mating design. Boars from Landrace were mated to the sows of Yorkshire (and other way around) to produce crossbred progeny. Training in both breeds was on sows and validation was on boars.

5.2.1 Purebred data

Litter sizes of 489,523 Landrace and 316,127 Yorkshire sows were used to calculate corrected phenotypic values of litter size for each breed separately (see details below). Corrected phenotypic values of litter size at birth (LSc), instead of original observations, were used as response variables for genomic prediction and to estimate additive and dominance genetic variances. The reason for using LSc as response variable was to reduce noise by removing non-genetic effects, which could be estimated much more accurately using a large dataset including all contemporaries and relatives, rather than using only genotyped animals. The contemporary group effects were estimated using a traditional pedigree-based linear model including herd–year–season, month at farrowing and regressions on hybrid indicator (0 = pure litter and 1 = hybrid litter), age at first farrowing, AI (0 = natural mating and 1 = AI), service sire and animal additive genetic effects as well as random residuals. The LSc was defined as original observations of litter size adjusted for all non-genetic effects.

A total of 2740 Landrace pigs (2085 sows and 655 boars) and 2325 Yorkshire pigs (2145 sows and 180 boars) were genotyped using the Illumina PorcineSNP60 BeadChip (Illumina, San Diego, CA). Edits on the genotype data comprised removing SNPs with a call rate below 90%, a minor allele frequency (MAF) below 1% and SNPs that deviated strongly from Hardy Weinberg equilibrium ($P < 10\text{-}7$). SNPs with more than 2% missing genotypes were filtered out. For the SNPs with less than 2% missing genotypes, the most common genotype of each SNP was defined and missing genotypes were allocated the population common genotype. Animals with more than 10% of missing SNP genotypes were also removed. Missing genotypes of the remaining animals were allocated the common genotype of the population. After editing, there were 34216 and 35135 SNP markers available for 2085 Landrace and 2145 Yorkshire sows, respectively. More details about the data can be found in (Guo et al., 2015).

5.2.2 Crossbred data

There were 8303 sows in the crossbred dataset. The crossbred animals were the first generation of reciprocal crosses of Landrace and Yorkshire. The crossbred animals were 5575 Landrace×Yorkshire (sire - dam) and 2030 Yorkshire×Landrace (sire - dam) and were born between 2009 and 2012. The pedigrees for both purebreds and crossbreds were available and all the crossbreds could be traced back to their purebred parents. Similar to purebreds of Landrace and Yorkshire, the corrected phenotypic values for crossbred animals were calculated by using a traditional animal model with a pedigree-based relationship matrix. The model

included herd-year-group, month at farrowing and regressions on age at first farrowing and as well as animal additive effects and random residuals (Nielsen et al., 2014).

5.2.3 Training and validation dataset

The purebred genotyped animals were split into training and validation datasets to evaluate the accuracy of genomic prediction for crossbred performance (Figure 5.1). The training dataset for Landrace consisted of 2085 sows with genotypes and phenotypes pre-corrected for fixed effects. The Validation dataset for this breed was 655 boars, which had 5575 Landrace-Yorkshire (LY) crossbred offspring. The response variable for the Landrace boars in the validation data set was the mean LSc of their LY crossbred progeny. Out of 655 boars in the validation dataset, 32 boars also had daughters (No.=320) in the training data set of Landrace. For Yorkshire, training was on 2145 genotyped sows. Similar to Landrace, validation in this breed was on the 180 genotyped boars that had 1928 daughters in the Yorkshire-Landrace (YL) crossbred dataset, and there was no direct relationship between sows of training population and YL crossbred animals. The relationship between sows in the training set and boars in the validation set was minimal as only 3 boars out of 180 had daughters (No.=30) in the training dataset of Yorkshire. For across breed genomic predictions, genotyped sows from both breeds were combined together to make a single training population with the size of 4230.

5.2.4 Linear models for genomic prediction

Estimation of marker effects

Two models for genomic prediction were evaluated. The first model included only additive effects (MA). The model for estimation of the additive effect associated with each marker was

$$y_i = \mu + \sum X_{ij} a_j + e_i, \qquad \text{(MA)}$$

where y_i is the phenotypic value of individual i in the training data, μ is the overall mean, X_{ij} is the copy number of a given allele of marker j, coded 0, 1 and 2 for aa, aA and AA, respectively, a_j is the random unknown additive effect for marker j and e_i is the residual effect for animal i and Σ denotes summation over all marker loci j.

The second model (MAD) included both additive and dominance effects and the following model was used to estimate the genetic effects associated with each marker:

$$y_i = \mu + \sum X_{ij} a_j + \sum Z_{ij} d_j + e_i, \qquad \text{(MAD)}$$

The definitions of the elements are analogous to model MA. In addition, Z_{ij} is the indicator variable for heterozygosity of individual i at marker j, with $Z_{ij} = 0$ when individual i is homozygous at marker j (aa or AA) and $Z_{ij} = 1$ if individual i is heterozygous at marker j (aA) and d_j is the random unknown dominance effect for SNP j.

The BayesC method proposed by Habier et al. (2011) was used to estimate marker effects. We used the BGLR "Bayesian general linear regression" R package developed by Perez and de los Campos (2014) and its built-in default rules to set values of hyper-parameters. A total of 100 000 iterations of the sampler were run, with the first 10 000 iterations discarded as burn-in samples. The number of total iterations and the number of 'burn in' iterations of the chain was calculated using the raftery.diag function of the R package Coda (Plummer et al., 2006). Convergence of the resulting posterior distributions was assessed by the Geweke diagnostic using the Coda package (Plummer et al., 2006).

Genomic estimated breeding values

The GEBV were calculated as the expected genotypic value of the offspring of a boar. From the estimates of additive effects (â), the genomic estimated breeding value based on model MA, ($GEBV_{MA}$) for purebred boar i from breed r was calculated as (Falconer and Mackay, 1996)

$$GEBV_{MA} = \sum_{j=1}^{s} [(S_{ij}^1)(p_{jr}\hat{a}_j)] + [(S_{ij}^2)(0.5p_{jr}\hat{a}_j - 0.5q_{jr}\hat{a}_j)] + [(S_{ij}^3)(-q_{jr}\hat{a}_j)] \quad (1)$$

where S_{ij}^1, S_{ij}^2 and S_{ij}^3 are indicator variables of the genotype of the j^{th} SNP of the i^{th} individual, with $S_{ij}^1 = 1$ when the genotype is AA and 0 otherwise, $S_{ij}^2 = 1$ when the genotype is Aa or aA and 0 otherwise, and $S_{ij}^3 = 1$ when the genotype is aa and 0 otherwise. Moreover, p_{jr} and q_{jr} are the allele frequencies (A and a) for the j^{th} SNP in breed r, $\hat{a}_j$ is estimated additive effect of the j^{th} SNP and s is the total number of SNPs. The formula (1) can be reduced to the usual formula $GEBV_{MA} = \sum_{j=1}^{s} X_{ij}\,\hat{a}_j$, but the reason for presenting it in this way is for similarity with the formula given below for GEBV when including dominance. It should be noted that the reduced formula and formula (1) are the same up to a constant i.e., correlation of GEBV based on two formulas is one while a simple linear regression between them would result in coefficient of 0.5.

With the MAD, two types of GEBV were calculated; genomic estimated breeding value for purebred performance (GEBV) and genomic estimated breeding value for crossbred performance (GEBV-C). The GEBVs were calculated as the expected genotypic values of the offspring of a boar carrying a certain SNP-genotype, when this parent is mated at random to its own line (GEBV) or to the other pure line (GEBV-C). Thus, from the estimates of both additive (â) and dominance effects ($\hat{d}$), the GEBV from model MAD was calculated as:

$$\begin{aligned}\mathrm{GEBV} = \sum_{j=1}^{s}[(S_{ij}^{1})(p_{jr}\hat{a}_j + q_{jr}\hat{d}_j)] + [(S_{ij}^{2})(0.5p_{jr}\hat{a}_j + 0.5q_{jr}\hat{d}_j \\ + 0.5p_{jr}\hat{d}_j - 0.5q_{jr}\hat{a}_j)] + [(S_{ij}^{3})(-q_{jr}\hat{a}_j + p_{jr}\hat{d}_j)]\end{aligned} \quad (2)$$

The definition of the elements are analogous to $\mathrm{GEBV_{MA}}$.In addition, $\hat{d}_j$ is the estimated dominance effect of the j^{th} SNP.

For crossbred offspring, the expected genotype frequencies of the offspring of a parent depend on the allele frequency in the other pure line (denoted ŕ here). Thus, for animal i from breed r, the GEBV-C was calculated using Equation 2, where p_{jr} and q_{jr} were replaced by $p_{jŕ}$ and $q_{jŕ}$, where $p_{jŕ}$ and $q_{jŕ}$are the allele frequencies (A and a) for the j^{th} SNP in breed r′. SNP allele frequencies in the other breed were calculated based on marker genotypes of genotyped sows in that breed. As an example, to predict GEBV-C for a Landrace boar, we used Equation 2 with SNP allele frequencies calculated from the all genotyped sows in Yorkshire. We also calculated the correlation between GEBV and GEBV-C from MAD, which is an indication of the purebred-crossbred genetic correlation when there is no G×E interaction, denoted as r_{pc} by Wei and Vanderwerf (1994).

5.2.5 Variance components

In addition to the additive variance computed from a pedigree based animal model, we estimated genomic additive and dominance variances for the animals in the training set. A mixed linear model for the individuals breeding values (u) and dominance deviations (v) is as follows: $y = \mu + Z_1u + Z_2v + e$,

where y is a vector of phenotypic values, μ is the overall mean, $Z_1$and Z_2 are design matrices relating animals to their breeding values and dominance deviations, u is a vector of breeding values, v is a vector of dominance deviations of animals, and e is a vector of residuals. $V(u) = G\sigma_A^2$, G is the genomic relationship matrix which was calculated based on the approach of VanRaden (2008): $G = \frac{W_a W_a{}'}{2\sum_{k=1}^{m} p_k q_k}$, where matrix W has dimensions of the number of individuals (n) by the number of loci

(m), with elements that are equal to $2 - 2p_k$ and $-2p_k$ for opposite homozygous and $1 - 2p_k$ for heterozygous genotypes, p_k is the minor allele frequency of locus k, and $q_k = 1 - p_k$. The covariance matrix of dominance effects is $V(v) = D\sigma_D^2$ where D is the genomic dominance relationship matrix and σ_D^2 is the dominance variance. Matrix D was calculated as $D = \frac{W_d W_d{}'}{4\sum_{k=1}^{m} p_k^2 q_k^2}$, where W_d has dimensions of the number of individuals (n) by the number of loci (m), with elements that are equal to $-2q_k^2$ for genotype AA, $2p_k\ q_k$for genotype Aa, and $-2p_k^2$ for genotype aa.The estimation of additive and dominance variances using these parameterizations that matches with classical quantitative genetics theory (Falconer and Mackay, 1996) were carried out applying the average information restricted maximum likelihood algorithm (Gilmour et al., 1995) implemented in the GVCBLUP package (Wang et al., 2014).

5.2.6 Model validation

Goodness of fit for each model was evaluated by deviance information criterion (DIC) value based on the training dataset. The superiority of MAD over MA was tested by a likelihood ratio test.

The predictive ability of the model (with respect to accuracy and unbiasedness) was evaluated by comparing GEBV of the boars in the validation dataset and mean corrected phenotype of their crossbred offspring. Unbiasedness of genomic predictions was assessed by regressing mean corrected phenotypes of crossbreds on the predicted breeding values of the boars in both breeds. A necessary condition for unbiased predictions is that the regression coefficient does not deviate significantly from one.

Predictions based on MA and MAD were both evaluated. Prediction accuracy was measured as the correlation between genomic-estimated breeding values of boars in pure lines and mean corrected crossbred-progeny performance. This correlation was divided by the average accuracy of mean-progeny performance i.e., the mean of $\sqrt{\frac{n}{n+k}}$ where n is number of daughters for each boar and $k = (4 - h^2)/h^2$ (Mrode and Thompson, 2005). Here, the heritability h^2 was the narrow-sense heritability estimated from the pedigree based linear model.

5.3 Results

5.3.1 Prediction of breeding values

MAD had better predictive ability than MA in both breeds (Table 5.1). Including dominance in genomic prediction improved prediction accuracy of GEBV by 18 and 22% in Landrace and Yorkshire, respectively. Within MAD, prediction of crossbred performance based on GEBV-C was more accurate than based on GEBV in both breeds.

Enlarging the training dataset by combining animals from both breeds into a single training population improved prediction accuracy for both models in both breeds (Table 5.1). This improvement in prediction accuracy was more evident for Landrace as this breed had lower prediction accuracy by separate training. For instance, in MA, combined training caused an improvement of 46% for Landrace, while this improvement in prediction accuracy was 21% in Yorkshire. Regardless of training on pure lines separately or jointly, the results indicated that including dominance effects in a prediction model improved accuracy of genomic predictions.

5.3.2 Model goodness of fit

Measures of goodness of fit are given in Table 5.2. In both breeds, MAD fitted the data better than MA with additive effects only. MAD had lower DIC than MA in both breeds. Measures of goodness of fit based on likelihood ratio test also showed superiority of MAD over MA in fitting the data. However, this superiority was not statistically significant.

5.3.3 Bias of genomic prediction

The coefficients of regressing corrected phenotypes of crossbreds on the predicted breeding values of the boars in both breeds show that, for Landrace, the variance of the predicted values was overestimated, i.e. most of regression coefficients were smaller than 1.0 (Table 5.3). When training was on the combined dataset, regression coefficients were closer to one, suggesting that joining two breeds to a single reference population improved the unbiasedness of genomic predictions especially for the MA model.

Table 5.1 Prediction accuracy for boars of Landrace and Yorkshire under two genomic model

	Purebred			Combined		
	MA	MAD		MA	MAD	
	GEBV	GEBV	GEBV-C	GEBV	GEBV	GEBV-C
Landrace	0.114(0.03)	0.135(0.03)	0.144(0.03)	0.167(0.03)	0.179(0.03)	0.207(0.03)
Yorkshire	0.320(0.06)	0.339(0.06)	0.358(0.06)	0.391(0.06)	0.402(0.06)	0.426(0.06)

Purebred: training in both pure lines was separately. **Combined**: genotyped sows from both pure lines were combined together to make a single training population. **GEBV**: genomic estimated breeding value for purebred performance. **GEBV-C**: genomic estimated breeding value for crossbred performance. For both models validation criterion was crossbred performance.

Table 5.2 DIC, χ^2 value and the corresponding P-value of likelihood ratio.

	MA	MAD	χ^2 values	P-value
	DIC	DIC		
Landrace	11230.35	11227.60	2.17	0.14
Yorkshire	11131.54	11121.42	2.18	0.13

5.3.4 Estimation of variance components

Estimates of additive genetic variance and heritability obtained with the pedigree model differed from those obtained with the genomic models (Table 5.4). Pedigree based heritability in Landrace was higher than in Yorkshire. Genomic additive heritability was similar in Landrace and Yorkshire. Dominance genetic variance computed from genomic information accounted for 15% and 18% of additive genetic variance in Landrace and Yorkshire, respectively.

5.4 Discussion

We tested whether the predictive ability of genomic prediction models that included dominance effects is increased when the validation criterion is crossbred performance. We supported this premise by showing some gains in prediction accuracy for both Landrace and Yorkshire breeds. We also found that combining animals into a single reference population improved prediction accuracy for both breeds. Therefore, a dominance model can be used to increase accuracy of genomic predictions for crossbred performance.

5.4.1 Comparison of models

We found that by including dominance in our genomic models we could predict crossbred performance more accurately than with a purely additive model. In fact, accuracies of genomic predicted breeding value using MAD were higher than that using the additive genetic model. In addition, the models including dominance effects slightly improved unbiasedness of genomic prediction.

The improvement in genomic prediction by including dominance effects in the genetic evaluation models has been reported widely in different livestock species for purebred performance. Su et al. (2012) analyzed daily gain in Danish Duroc pigs using models with or without non-additive genetic effects. Their results showed that non-additive genetic effects are important sources of genetic variation for daily gain in pigs and genomic prediction models including non-additive genetic effects improved accuracy of genomic predicted breeding value. Sun et al. (2014) investigated the role of dominance in the Holstein and Jersey breeds for yield and non-yield traits and found that for yield traits, including additive and dominance effects fitted the data better than including only additive effects; average correlations between estimated genetic effects and phenotypes showed that prediction accuracy increased when both effects rather than just additive effects were included in the model. Moghaddar et al. (2014) used data on purebred Merino sheep to predict breeding values of purebred rams and found that fitting

Table 5.3 Regression coefficients (± standard errors) of corrected litter size of crossbreds on genomic estimated breeding value for the boars in the validation dataset

	Purebred			Combined		
	MA	MAD		MA	MAD	
	GEBV	GEBV	GEBV-C	GEBV	GEBV	GEBV-C
Landrace	0.44±0.11	0.60 ±0.14	0.73±0.17	0.71±0.13	0.87±0.16	1.26±0.21
Yorkshire	0.69±0.09	1.14 ±0.20	1.36±0.28	0.94±0.18	1.24±0.24	1.60±0.27

Purebred: training in both pure lines was separately. **Combined**: genotyped sows from both pure lines were combined together to make a single training population. **GEBV**: genomic estimated breeding value for purebred performance. **GEBV-C**: genomic estimated breeding value for crossbred performance. For both models validation criterion was crossbred performance.

Table 5.4 Estimates of additive genetic variance (σ_a^2), dominance variance (σ_d^2), and the proportions of these variances (h_a^2 and h_d^2) to phenotypic variance

	Landrace		Yorkshire	
Parameters	**Pedigree**	**Genomic**	**Pedigree**	**Genomic**
$\boldsymbol{\sigma_a^2}$	1.29 (0.03)	0.78 (0.13)	1.00 (0.03)	0.66 (0.12)
$\boldsymbol{\sigma_d^2}$	-	0.12(0.07)	-	0.12 (0.06)
$\boldsymbol{h_a^2}$	0.10 (0.002)	0.05 (0.02)	0.08 (0.003)	0.05 (0.02)
$\boldsymbol{h_d^2}$	-	0.007 (0.01)	-	0.01 (0.01)

both additive and dominance effects of marker genotypes provide either similar or higher genomic breeding value accuracy depending on the degree of dominance variance. To our knowledge, no study on real data has compared additive and dominance models for crossbred performance. However, in a simulation study, Zeng et al. (2013) compared additive and dominance models for genomic selection in purebreds for crossbred performance and came to the conclusion that, in the presence of dominant gene action, relative to the additive model, genomic selection with the dominance model is superior to maximize crossbred performance through purebred selection.

The additive model is the most simple and practical model for the estimation of breeding values in pure lines for crossbred performance both computationally and theoretically. However, in crossbreeding schemes this model may not be very efficient if the trait of interest is affected by non-additive effects or when the genetic correlation between the purebred and the crossbred performance is lower than 1 ($r_{pc} < 1$). In addition, it has been shown (Dekkers, 1999) that for a two-way cross, the allele substitution effects for QTL or markers in one parental breed depend on the allele frequencies in the other parental breed. Thus, in the computation of substitution effects, failure to use the appropriate allele frequencies may result in a loss of response to selection. This is one of the drawbacks of the additive model that in case of training on pure lines, the genomic estimated breeding value of an animal would be the same for purebred and crossbred performance and cannot maximize the genetic improvement in crossbreds. With presence of dominance, allele substitution effects and individual breeding values depend on allele frequencies. A dominance model provides estimates of both additive and dominance effects and therefore, enables the computation of allele substitution effects using appropriate allele frequencies.

Within the MAD, GEBV-C showed higher prediction accuracy for crossbred performance than GEBV in both breeds. GEBV is an estimated breeding value for purebred performance while GEBV-C is an estimated breeding value for crossbred performance. GEBV can be used as a selection criterion for genetic improvement within a pure line, while GEBV-C is a selection criterion to improve crossbred performance. We ranked the top 50 boars of Landrace based on both GEBV and GEBV-C, and found that 42 boars were common in the two rankings. The corresponding value was 43 for Yorkshire boars. These re-rankings indicate that ranking of boars would be different for purebred and crossbred performance and different selection criteria should be used depending on the breeding goal. Superiority of prediction accuracy based on GEBV-C over GEBV in our results is in agreement with findings of Esfandyari et al. (2015b) who showed with simulation

data that under a dominance model the response to selection in crossbred individuals was higher when selection was based on GEBV for crossbred performance, although data were collected on purebred individuals.
For the calculation of GEBV-C for the boars of purebreds we used SNP allele frequencies calculated from the genotyped sows in the opposite breed. However, the more accurate approach would be using the SNP allele frequencies calculated from the selection candidates in the opposite breed. For instance, in calculating GEBV-C for boars of Landrace, SNP allele frequencies could be calculated from 2450 Yorkshire sows that were mated to the boars of Landrace to produce crossbred progeny. However, as these sows were not genotyped, we used SNP allele frequencies calculated from the sows of the relevant generation that is an estimation of SNP allele frequencies for the selection candidates.

Prediction accuracy for crossbred performance in Yorkshire boars was higher than Landrace boars across two models, even though both breeds had almost the same size of training population and heritability. The pedigree based prediction accuracy was also higher for the Yorkshire boars compared to the Landrace boars (results not shown). We did not find a clear reason for these differences. However it seems the difference in environmental variance and also difference in the genetic level between two breeds might be an explanation for the observed differences in prediction accuracies [personal communication to B. Nielsen, Danish Pig Research Centre].

5.4.2 Gain of combined reference population for genomic prediction

Combining animals from pure lines into a single training set improved prediction accuracy for both Landrace and Yorkshire across all models. In fact, joining two or more populations from the same or different breeds into a common reference population is often argued to be an obvious way to increase the accuracies of GEBV (de Roos et al., 2009, Lund et al., 2014). However, the increase in accuracy of GEBV, when combining populations into a single reference, will depend on how closely related the populations are and how aligned the Linkage disequilibrium (LD) information used for genomic predictions is. According to Daetwyler et al. (2012), across breed accuracy depends on the LD between markers and QTL or persistence of LD phase. In a simulation study Esfandyari et al. (2015b) showed that in a two-way crossbreeding scheme if the correlation of LD phase between two breeds is high, there can be an added benefit in terms of accuracy of GEBV if animals from both breeds are combined into a single reference population to estimate marker effects. Persistance of LD phase has been reported in some pig breeds and knowledge about the persistence of LD phase between breeds would allow to find

out whether joining populations may help or not. Badke et al. (2012), evaluated correlation of phase among four US pig breeds and reported correlation of phase of 0.87 for Duroc-Yorkshire and 0.92 for Landrace-Yorkshire, for markers with a pairwise distance <10kb. For the same distance, Wang et al. (2013) found a persistence of phase of 0.61 for Duroc-Landrace, 0.57 for Duroc-Yorkshire and 0.66 for Landrace-Yorkshire. Wang et al. (2013) results concerning genome-wide LD confirmed the mixture history of Landrace and Yorkshire, which is also implied by the higher level of persistence of phase between Landrace and Yorkshire. This may explain the improvement in prediction accuracy by combining animals from Landrace and Yorkshire breeds to a single training population in our study.

In addition, the benefit of joining reference populations depends on the size of the reference population, as there is a diminishing return relationship between reference population size and accuracy. Hence, when the reference population is small, joining may help when the correlation of LD phase is sufficiently high, whereas, joining will have limited benefit or may even be detrimental when reference populations are large enough or correlation of LD phase is low. Moghaddar et al. (2014) compared the accuracy of genomic prediction in Australian sheep breeds by using data from purebreds, crossbreds or a combination of those in a reference population. The results of their study showed zero to small negative effects on genomic prediction accuracy when data from distant breeds were included in the reference population.. A number of studies have compared the predictive ability of genomic models trained in a joint reference population by combining populations of the same breed or populations of different breeds (for review see Lund et al. (2014))

5.4.3 Additive and dominance genetic variances of litter size

In our study to estimate additive and dominance variances, following Vitezica et al. (2013), we used breeding (or classical) model rather than the genotypic model. The breeding model partitions a genotypic value into breeding value and dominance deviation. Therefore, estimated variances are variances due to the individual additive value (breeding values) and dominance deviations and are comparable with pedigree based estimates. However, the genotypic model, partitions genetic variance into additive and dominance in such a way that estimated variances are not directly comparable to pedigree based estimates i.e., the additive variance is the variance of additive effects. The difference between two models has been discussed in (Vitezica et al., 2013).

Estimates of additive variance obtained from pedigree were different than those from genomic information in both breeds. The differences observed were

most likely because animals used for estimates of genomic variances were a small fraction of all animals. Based on the present data, the estimated dominance variance in proportion to additive variance was 15% and 18% for Landrace and Yorkshire, respectively. In pigs, significant contributions of dominance genetic variance have been reported. Lopes et al. (2014) by using genotypic model reported this ratio to be 13% and 21% for number of teats and back fat in Landrace. They, however, mentioned that by using breeding model these values decreased to 0.08% and 0.16% for number of teats and back fat, respectively. Vitezica et al. (2013) have argued that the genotypic model overestimates the dominance genetic variance and, consequently, underestimates additive genetic variance. In the study of Su et al. (2012) dominance variance accounted for 26% of the additive variance for daily gain in Danish Duroc pigs. However, as they have used the genotypic model, reported variance for dominance is overestimated. Based on pedigree estimates, Culbertson et al. (1998) reported that the ratio of dominance to additive variances for different traits in pigs ranged from 11 to 78%. These results indicate dominance genetic variation is important for complex traits.

5.4.4 Purebred-crossbred genetic correlation

In this study, the correlation between GEBV and GEBV-C from MAD model was 0.90 and 0.93 for Landrace and Yorkshire, respectively. The deviation of these correlations from one in both breeds is due to dominance effects only, and in the situation where there is no G × E interaction these correlations are an indicator of the purebred-crossbred genetic correlations (r_{pc}). Theoretically, assuming that the dominance model is true, i.e. there is no G×E interaction, having an infinite amount of information on purebred animals, then using the dominance model on purebred data would provide the maximum accuracy (i.e. unity) of prediction for crossbred performance. On the other hand, when $r_{pc}$is smaller than one solely due to G×E interaction, the maximum achievable accuracy by using purebred information will be r_{pc}. So, due to deviation of r_{pc} from unity it has been suggested that when aim of selection within a pure line is to improve crossbred performance, crossbred animals should be used for training (Dekkers and Chakraborty, 2004, Zeng et al., 2013, Esfandyari et al., 2015a). In fact, training on crossbred data for genomic selection accounts for genetic differences between purebred and crossbred animals and for genotype by environment effects. Esfandyari et al. (2015a) compared crossbred response in a two-way crossbreeding program by training either on purebred or on crossbred animals under dominance model in the absence of G × E interaction. According to their results, using data of crossbreds yielded a substantial improvement in crossbred performance and the explanation was that

by training on crossbred animals, they could predict dominance effects and consequently breeding values of the animals in pure lines more accurately. van Grevenhof and van der Werf (2015) investigated the benefit of including crossbred information in the reference population of a crossbreeding program using genomic selection. Based on a deterministic simulation they concluded that using crossbred rather than purebred individuals in a reference population for genomic selection can provide substantial advantages, but only when correlations between purebred and crossbred performances are not high. However, in their study the reason for $r_{pc} < 1$ was not clear. In fact, knowing the mechanism for the deviation of r_{pc} from one would be helpful to determine whether crossbred info should be used in genomic selection schemes when the aim of selection within pure lines is to improve crossbred performance. In principle, one could use a dominance model and multi-trait analysis to partition the r_{pc} into a component due to dominance and a remaining component due to G × E and epistasis. However, accurate partitioning would require a small standard error of the estimated purebred-crossbred genetic correlation, and thus very large datasets (Bijma and Bastiaansen, 2014).

5.5. Conclusions

Compared to additive model, the use of a dominance model increased the prediction accuracy of purebred animals for crossbred performance. This is probably due to the fact that using dominance model on purebred data can accounts for some part of $r_{pc} < 1$ in crossbreeding programs. Furthermore, we found that combining animals from both breeds into a single reference population improved prediction accuracy in both breeds.

5.6 Acknowledgements

HE has benefited from a joint grant from the European Commission and Aarhus University within the framework of the Erasmus-Mundus joint doctorate ‘EGS-ABG’.

References

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, and J. P. Steibel. 2012. Estimation of linkage disequilibrium in four US pig breeds. Bmc Genomics 13.

Bijma, P. and W. M. Bastiaansen. 2014. Standard error of the genetic correlation: how much data do we need to estimate a purebred-crossbred genetic correlation? Genetics Selection Evolution 46:79.

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

Daetwyler, H. D., K. E. Kemper, J. H. J. van der Werf, and B. J. Hayes. 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. J Anim Sci 90(10):3375-3384.

de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183:1545-1553.

Dekkers, J. C. M. 1999. Breeding values for identified quantitative trait loci under selection. Genetics Selection Evolution 31(5-6):421-436.

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

Dekkers, J. C. M. and R. Chakraborty. 2004. Optimizing purebred selection for crossbred performance using QTL with different degrees of dominance. Genetics Selection Evolution 36(3):297-324.

Ertl, J., A. Legarra, Z. G. Vitezica, L. Varona, C. Edel, R. Emmerling, and K. U. Gotz. 2014. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genetics Selection Evolution 46.

Esfandyari, H., A. C. Sorensen, and P. Bijma. 2015a. A crossbred reference population can improve the response to genomic selection for crossbred performance. Genetics, selection, evolution : GSE In press.

Esfandyari, H., A. C. Sorensen, and P. Bijma. 2015b. Maximizing crossbred performance through purebred genomic selection. Genetics Selection Evolution 47.

Falconer, D. S. and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. Vol. 4. 4 ed. Pearson

Gilmour, A. R., R. Thompson, and B. R. Cullis. 1995. Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51(4):1440-1450.

Guo, X., O. F. Christensen, T. Ostersen, Y. Wang, M. S. Lund, and G. Su. 2015. Improving genetic evaluation of litter size and piglet mortality for both genotyped and nongenotyped individuals using a single-step method1. J. Anim. Sci. 93(2):503-512.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. Bmc Bioinformatics 12.

Lopes, M. S., J. W. M. Bastiaansen, L. Janss, H. Bovenhuis, and E. F. Knol. 2014. Using SNP Markers to estimate additive, dominance and imprinting genetic

Variance. in Proc. 10th World Congress on Genetics Applied to Livestock Production, Vancouver, BC, Canada.

Lund, M. S., G. Su, L. Janss, B. Guldbrandtsen, and R. F. Brondurn. 2014. Invited review: Genomic evaluation of cattle in a multi-breed context. Livest Sci 166:101-110.

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

Misztal, I., L. Varona, M. Culbertson, J. K. Betrand, J. Mabry, T. J. Lawlor, Van.,, C. P. Tassel, and N. Gengler. 1998. Studies on the value of incorporating the effect of dominance in genetic evaluations of dairy cattle, beef cattle and swine. Biotechnol Agron Soc Environ 2:227-233.

Moghaddar, N., A. A. Swan, and J. H. J. van der Werf. 2014. Comparing genomic prediction accuracy from purebred, crossbred and combined purebred and crossbred reference populations in sheep. Genetics Selection Evolution 46.

Mrode, R. A. and R. Thompson. 2005. Linear Models for the prediction of animal breeding values. CABI Publishing.

Nielsen, B., I. Velander, T. Ostersen, M. Henryon, and O. F. Christensen. 2014. Nurse capacity in crossbred sows and genetic correlation to purebred fertility. in Proc. 10th World Congress on Genetics Applied to Livestock Production, Vancouver, BC, Canada.

Perez, P. and G. de los Campos. 2014. Genome-Wide regression and prediction with the BGLR statistical package. Genetics 198(2):483-U463.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: Convergence diagnosis and output analysis for MCMC. Vol. 6. R.

Su, G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund. 2012. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PloS one 7(9):e45293.

Sun, C., P. M. VanRaden, J. B. Cole, and J. R. O'Connell. 2014. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. PloS one 9(8).

Toro, M. A. and L. Varona. 2010. A note on mate allocation for dominance handling in genomic selection. Genetics, selection, evolution : GSE 42:33.

van Grevenhof, I. E. M. and J. H. J. van der Werf. 2015. Design of reference populations for genomic selection in crossbreeding programs. Genetics Selection Evolution 47.

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

Vitezica, Z. G., L. Varona, and A. Legarra. 2013. On the Additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195(4).

Wang, C. K., D. Prakapenka, S. W. Wang, S. Pulugurta, H. B. Runesha, and Y. Da. 2014. GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects. Bmc Bioinformatics 15.

Wang, L., P. Sorensen, L. Janss, T. Ostersen, and D. Edwards. 2013. Genome-wide and local pattern of linkage disequilibrium and persistence of phase for 3 Danish pig breeds. Bmc Genet 14.

Wei, M. and H. Steen, van der.,. 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 J. H. J. Vanderwerf. 1994. Maximizing Genetic Response in Crossbreds Using Both Purebred and Crossbred Information. Anim Prod 59:401-413.

Wei, M., J. H. J. Vanderwerf, and E. W. Brascamp. 1991. Relationship between purebred and crossbred parameters .2. Genetic correlation between purebred and crossbred performance under the model with 2 loci. J Anim Breed Genet 108(4):262-269.

Wittenburg, D., N. Melzer, and N. Reinsch. 2015. Genomic additive and dominance variance of milk performance traits. J Anim Breed Genet 132(1):3-8.

Zeng, J., A. Toosi, R. L. Fernando, J. C. Dekkers, and D. J. Garrick. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genetics, selection, evolution : GSE 45(1):11.

Zumbach, B., I. Misztal, S. Tsuruta, J. Holl, W. Herring, and T. Long. 2007. Genetic correlations between two strains of Durocs and crossbreds from differing production environments for slaughter traits. J Anim Sci 85(4):901-908.

6

General discussion

6.1 Introduction

One of the main limitations of many cross breeding programs is that selection is in purebred nucleus lines or breeds that are housed in high-health environments but the goal of selection is to improve crossbred performance (CP) under field conditions. Due to genetic differences between purebreds and crossbreds and environmental differences between nucleus and field conditions, performance of purebred parents can be a poor predictor of performance of their crossbred descendants (Dekkers, 2007). Furthermore, some important traits such as disease resistance cannot be measured in nucleus lines. In order to deal with these problems, it has been proposed to select purebred relatives based on CP using combined crossbred and purebred selection or CCPS (Wei and Steen, 1991, Lo et al., 1993, Lo et al., 1995, Lo et al., 1997). This approach can increase response to selection for CP relative to the classical method of selection on purebred performance (Bijma and van Arendonk, 1998). It has, however, not been extensively implemented in livestock due mainly to the difficulty and cost of routine collection of phenotypic and pedigree data from crossbreds in the field (Dekkers, 2007). In addition, using CCPS increases the rate of inbreeding (Bijma et al., 2001) and makes it difficult to accommodate non-additive gene action (Lo et al., 1997). As an alternative to CCPS, Dekkers (2007) proposed to select purebreds for commercial CP using genomic selection (GS). GS of purebreds for CP involves estimating effects of SNPs on CP, using phenotypes and SNP genotypes evaluated on crossbreds, and applying the resulting estimates to SNP genotypes obtained on purebreds (Dekkers, 2007). GS for CP has advantages over CCPS such as it does not require pedigree information on crossbreds, it reduces the rate of inbreeding (Daetwyler et al., 2007), and makes accommodating non-additive gene action easier (Dekkers, 2007).

This thesis primarily focused on implementation of dominance as a non-additive genetic effect in genomic crossbreeding programs. In chapter 3, the potential benefit of GS within purebred lines, when the objective is to improve performance of crossbred populations at the commercial level was evaluated. Both phenotypic and genotypic information was collected on purebred animals only. EBV for CP were obtained based on estimated dominance effects and the allele frequency in the other line. In a two-way crossbreeding system, it was found that selection for genomic estimated breeding value for crossbred performance (GEBVC) increased response in crossbred animals compared to selection for genomic estimated breeding value for purebred performance (GEBVP). The effect of the correlation of linkage disequilibrium (LD) phase between the two pure breeds on the

consequences of combining both reference populations was also investigated. The results revealed that, for a high correlation of LD phase, combining both populations into a single reference population increased response to selection in crossbred animals. In chapter 4, response to selection of crossbreds by simulating a two-way crossbreeding program with either a purebred or crossbred training population under a dominance model was compared. It was confirmed that, to reach greater response to selection when crossing two distantly related lines, it is better to do training on crossbred animals rather than on pure lines to predict genetic effects. In addition, being able to distinguish between alternate heterozygotes in the crossbred training set by taking into account the breed origin of alleles increased response to selection, except when breeds were closely related and the reference population was small. Finally in chapter 5, to confirm the findings of the simulation study in chapter 3, real genomic data of purebred Landrace and Yorkshire pigs were analyzed. It was tested whether the predictive ability of genomic prediction models for CP could be improved by including dominance. Training was on pure lines and we also compared the use of two separate pure-line reference populations to a single reference population that combines both pure lines. The results showed some gains in prediction accuracy for CP by including dominance and combining both pure lines into a single reference population for training.

Some topics have already been addressed in the discussion sections of the relevant chapters and will not be repeated here. Thus, this general discussion will concentrate on three main topics, i) Genomic models in crossbreeding, ii) Design of a reference population for GS in crossbreeding schemes and iii) Genomic selection and pig breeding. Finally some other relevant topics will be discussed briefly.

6.2 Genomic prediction models in crossbreeding schemes

Several genomic models have been suggested for the prediction of breeding values of the individuals in the purebred lines for CP in genomic crossbreeding programs. These models are namely the standard additive model, across-breed effects of SNP genotypes model (ASGM), breed-specific effects of SNP alleles model (BSAM) and the dominance model. Additive and dominance models can be used by training either on crossbreds or purebreds, however, ASGM and BSAM can be used for crossbred training only.

The additive model is the most simple and practical model for the estimation of breeding values in pure lines for CP both computationally and theoretically. However, in crossbreeding schemes this model may not be very efficient if the trait

of interest is affected by non-additive effects or when the genetic correlation between the purebred and the crossbred performance (r_{pc}) is lower than 1 ($r_{pc} < 1$). For example, accuracy of EBV of an animal in a pure line for a trait with heritability of 0.3 based on own phenotype for purebred performance would be 0.54, while for under an $r_{pc} = 0.5$, the accuracy of EBV for CP would be only 0.27 (i.e. $\sqrt{0.3} \times 0.5$). In addition, it has been shown (Dekkers, 1999) that for a two-way cross, the allele substitution effects for QTL or markers in one parental breed depend on the allele frequencies in the other parental breed when non-additive effects are present. Thus, in the computation of substitution effects, failure to use the appropriate allele frequencies may result in a loss of response to selection. This is one of the drawbacks of the additive model that in case of training on pure lines, the genomic estimated breeding value of an animal would be the same for purebred and CP and cannot maximize the genetic improvement in crossbreds.

Furthermore, using the additive model with crossbred training, a single substitution effect is estimated for each SNP, assuming it is the same for both parental breeds. Selection based on GEBV derived using such allele substitution effects is expected to fix the favourable allele in both breeds and thus reduce heterozygosity in crossbreds. Exceptions to this could occur due to genetic drift or the marker and QTL being in LD with opposite phases in the two parental breeds. When the two breeds have opposite LD phases, and a common nonzero substitution effect is estimated for a SNP in the additive model, the allele frequencies of associated QTL will move in opposite directions in the two breeds.

In some studies, additive gene action or perfect knowledge of allele substitution effects or both are assumed (Ibanez-Escriche et al., 2009, Toosi et al., 2010). However, it has been argued that dominance is the likely genetic basis of heterosis (Falconer and Mackay, 1996). Therefore explicitly including dominance in the GS model may be an advantage to select purebred animals for CP. With dominance, allele substitution effects and individual breeding values depend on allele frequency and, thus, change over time, which alters the ranking of individuals. This problem can be overcome by applying a dominance model, which provides estimates of both additive and dominance effects and, therefore, enables the computation of allele substitution effects using appropriate allele frequencies. In chapter 3 we used such a model for estimation of GEBV of animals in pure lines for CP while training was on pure lines. Under a dominance model we calculated allele substitution effects based on allele frequency of the other breed and could improve the response to selection compared to a model where allele frequencies were used from the line itself. However, in our dominance model, we used SNP

allele frequencies among selection candidates of the opposite breed for the calculation of GEBV. The more appropriate approach would be using the SNP allele frequencies of selected mates. However, allele frequencies of selected mates cannot be observed prior to computation of the substitution effects that are needed for selection. Nonetheless, one can argue that differences in allele frequencies between selected mates and selection candidates should be relatively small in most cases. In fact, for highly polygenic traits, the change due to selection will be small, and the main change in allele frequency may result from sampling (i.e., drift). Following Falconer and Mackay (1996), the change of allele frequency resulting from sampling is random in the sense that its direction is unpredictable but its magnitude can be predicted. The change of allele frequency, (Δq), resulting from sampling can be stated in terms of its variance as $\sigma^2_{\Delta q} = \frac{pq}{2N}$ where p and q are allele frequencies and N is the number of sampled individuals (selected as mates). Figure 1 shows variance and standard deviation of Δq at different values of allele frequency (p) due to sampling of 100 individuals. If we consider a locus win an equal allele frequencies, (p=q=0.5), variance of change in allele frequency in that locus due to sampling of e.g., 100 (number of selected females in our simulations) would be 0.0012 with standard deviation of ~ 0.03 that is negligible.

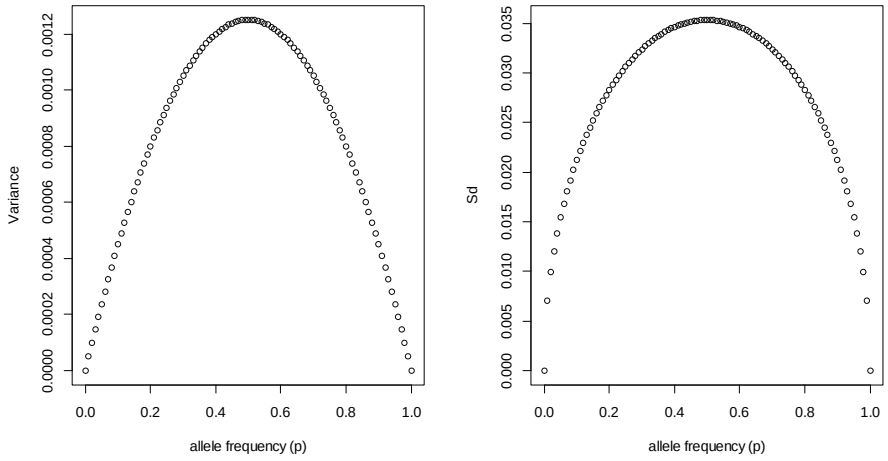


Figure 6.1 Variance and standard deviation of change in allele frequency, (Δq), due to sampling of 100 individuals.

Presence of overdominance in the genetic architecture of the trait of interest would further clarify the advantage of a dominance model over an additive model.

In our simulations in chapter 3 and 4, we assumed dominance variance to be one third of the additive genetic variance. This ratio resulted in 10 to 15% of loci showing overdominance. When overdominance is present, the allele substitution effect (i.e., $\alpha = a + (1 - 2p)d$) may have opposite signs in the parental breeds, depending on allele frequencies p in the two breeds (Falconer and Mackay, 1996). In this case, the two parental breeds are expected to be fixed for alternate alleles of over-dominant QTL, which increases the frequency of favourable heterozygotes in crossbred progeny. Note that under the additive model, fixation of the favourable allele in both breeds would result in lower heterozygosity in the crossbreds. So, it is expected that the dominance model results in substantially greater heterosis than the additive model. However, the purebred gain may be lower with the dominance model than with the additive model, because the unfavourable allele will be moved towards fixation in one parental breed at some loci.

Under the additive model, SNP allele effects are assumed the same in all breeds. However, in crossbred populations, effects of SNPs may be breed specific because the extent of LD between SNPs and QTL can differ between breeds. Moreover, the LD may not be restricted to markers that are tightly linked to the QTL. Both these problems were addressed by using a model with breed-specific effects of SNP alleles (BSAM) by Ibanez-Escriche et al. (2009). In this model breed-specific substitution effects for each allele in a SNP are estimated based on the breed origin of the allele, and it is assumed that breed origin of alleles are known without error. The estimated effects and the SNP genotypes of purebred candidates for selection, then are used to predict their breeding values for CP (Ibanez-Escriche et al., 2009).

In our simulations in chapter 4, it was assumed that the additive and dominance effects of QTL are the same for both breeds, which may not be the case with real populations. However, even when additive and dominance effects are consistent between breeds, allele substitution effects will be breed-specific if allele frequencies differ between breeds. In such a case, the estimates of breed-specific allele substitution effects in the dominance model are expected to be more accurate than those in BSAM. The first reason is that the estimates of additive and dominance effects from the dominance model are combined with the observed allele frequencies in the opposite parental breed to calculate the breed-specific allele substitution effects. In BSAM, however, breed-specific allele substitution effects are estimated directly. Thus, the allele frequencies used implicitly in BSAM are based on the frequencies in the training population of the alleles inherited from the opposite parental breed. Note that the alleles inherited by the training

population are a random sample of those from the parental population, and therefore their frequencies may deviate from those of the parental population. Thus, the use of observed allele frequencies from the parental population to compute breed-specific allele substitution effects favours the dominance model over BSAM. Second, in the dominance model, as selection progresses and allele frequencies change due to selection and drift, the observed allele frequencies in each generation are combined with the estimates of additive and dominance effects obtained in training to compute the current values of the breed-specific allele substitution effects. However, with the additive model and with BSAM, the allele substitution effects estimated in training are repeatedly used to compute GEBV of selection candidates, ignoring changes in allele frequencies. Therefore, the use of the dominance model is expected to require less frequent retraining than use of BSAM or the additive model. This is appealing for traits that are difficult or expensive to measure (Zeng et al., 2013).

In chapter 4, we used a dominance model that is a modified version of dominance and BSAM, and would be called breed-specific dominance model (BSDM). This model has some advantages over BSAM and the dominance model. Compared to the dominance model, in which alternate heterozygotes (based on breed origin) are assumed to have the same effect, in the BSDM model alternate heterozygotes can have different effects. This is relevant because for two breeds involved in crossbreeding it is very possible that SNP and QTL frequencies might be different between the two lines especially if two lines involved in crossbreeding are distantly related. In addition, there might be differences in the amount and extent of LD between SNPs and QTL between the lines. Any difference in QTL and SNP frequencies and in LD between the pure lines can result in the two alternate heterozygotes at a SNP having different probabilities for a heterozygous QTL in the crossbreds. These differences suggest that one should distinguish between the two alternate heterozygotes in the crossbred when a dominance model is used for crossbred training. Our results in chapter 4 showed that being able to distinguish between alternate heterozygotes, Aa and aA, while training on crossbreds and predicting two distinct genetic values for these genotypes can lead to greater response to selection in crossbreds compared to a standard dominance model. However, similar to BSAM, implementation of BSDM requires knowing the breed origins of SNP alleles which is not required for the dominance model. However, if this phasing can be done then BSDM can also account for imprinting, which may contribute to heterosis and is expected to contribute to the genetic architecture and evolution of complex traits (Cheng et al., 2013, Lawson et al., 2013).

To determine the breed origin of alleles in crossbreds, pedigree-based phasing methods may not be suitable, because usually in real breeding programs the pedigree of crossbred animals are not known and several generations may separate the genotyped purebred and crossbred animals (Bastiaansen et al., 2014). LD based phasing methods also may not be suitable because haplotypes within a LD block are often common between several breeds. However, long range phasing (Kong et al., 2008) overcomes both the issues of lacking pedigree and common haplotypes between breeds. In a recent study, Bastiaansen et al. (2014) suggested a method to determine the breed origin of alleles in crossbreds using long-range phasing without the need for tracking pedigree relationships of crossbreds. Based on this method, it is not even necessary to have close relationships between the crossbred and genotyped purebred animals since long-range phasing will work even with distant purebred relatives of the crossbreds (Bastiaansen et al., 2014).

In summary, all proposed models for GS of purebreds for CP have their own pros and cons and the decision to use a model mainly depends on the availability of required information (purebreds or crossbreds) and also on the factors that cause $r_{pc} < 1$. Table 6.1 summarizes the benefits of the crossbred and purebred training and also suggested genomic models for GS of purebreds for CP. For the models presented in the table, it is assumed that estimated additive ($\hat{a}$) and dominance ($\hat{d}$) effects rather than allele substitution effects are used for calculation of GEBVC (see footnote of Table 6.1). In general, training on crossbreds animals and using BSDM can be more beneficial in crossbreeding programs as this approach can account for almost all factors that cause $r_{pc} < 1$. However, in practice implementation of such a model may not be trivial as it requires large number of crossbred animals with both phenotype and genotype for accurate prediction of GEBVC. In addition, phasing of the genotypes may not be easy, particularly in the absence of informative pedigree.

6.3 Design of a reference population for genomic selection in crossbreeding schemes

In crossbreeding schemes, the ultimate goal is to improve the performance of the crossbred offspring of the pure breeding lines. GS uses marker genotypes and phenotypes in a reference population to predict breeding values of selection candidates that have been genotyped (Meuwissen et al., 2001). Similar to pure breeding, GS could benefit crossbreeding programs since it allows using information at an early age. Accepting that GS is an appropriate tool to select animals for CP raises a question i.e., should marker effects be predicted from pure

line or crossbred data. This question is relevant as the effectiveness of GS in crossbreeding schemes will depend among others on the composition of the reference population used for genomic predictions.

Table 6.1 Potential benefits of training on purebred or crossbreds and genomic models depending on nature of $\mathbf{r_{pc}}$

	Training		Prediction model			
Factors contributing to $\mathbf{r_{pc}}$ <1	PB	CB	Additive model	BSAM	Dominance model	BSDM
Dominance	×	×	×†	×†	×	×
Epistasis	×	×	×†	×†		
Imprinting		×				×
G×E		×	×*	×	× *	×

* Additive and dominance model would account for G×E if training is on crossbreds.
† If allele substitution effects are calculated, additive and BSAM models may account for dominance and epistasis as allele substitution effects capture a part of dominant and higher-order interactions across genes and alleles.

Estimation of marker effects for genomic prediction based on purebred data is appealing since large amounts of phenotypic as well as genotypic information on PB animals are usually already available. However, purebred training and estimation of SNP effects based on purebred individuals might be relevant for the genetic improvement within purebred lines but might not be efficient when the aim of selection within pure lines is to improve CP. One of the reasons would be that estimated SNP effects might be different in purebreds and crossbreds. In addition, for traits with significant non-additive variance and therefore potential heterosis and in situations where r_{pc} is lower than one, purebred performance is not a good predictor of CP. Furthermore, training on purebreds cannot maximize the performance of crossbred animals especially in presence of G×E. Nonetheless, in chapter 3, under the hypothesis that crossbred and purebred animals differ from each other due to dominance, we used GS to select purebred individuals for CP without collecting crossbred phenotypic or genotypic data. In a two-way crossbreeding system, we found that selection for GEBVC increased response in crossbred animals compared to selection for GEBVP.

While the use of crossbred phenotypes has been limited in applied breeding programs because tracing pedigree relationships in a crossbred production environment is not trivial, it has recently regained attention because genomic relations are a solution for the cumbersome pedigree tracing process. Dekkers (2007) proposed to use marker information that was estimated based on the

performance of commercial CB animals. He found a significant increase in the rates of genetic gain compared to using only PB phenotypic information, or combined PB and CB information, whereas the rate of inbreeding decreased. Despite of these advantages, for training on crossbreds, it is necessary to collect genotypic and phenotypic data at crossbred level, which can substantially increase the required investment in the breeding program, since crossbred animals are usually not individually identified and individual performances are not recorded. Nevertheless, as mentioned earlier training on crossbred data for GS accounts for genetic differences between purebred and crossbred animals and potential genotype by environment effects. In chapter 4, we investigated response to selection in CP in a two-way crossbreeding system of two distantly related breeds. To estimate SNP effects, training was either on pure lines or crossbreds and animals were selected on GEBVC. Our results showed that to reach greater response to selection when crossing two distantly related lines, it is better to do training on crossbred animals rather than on pure lines. However, in that simulation, the deviation of r_{pc} from one was due to dominance and differences in allele frequencies between the two lines. It can be expected that training on crossbreds is much more efficient if $r_{pc} < 1$ is due to G×E on top of the non-additive effects.

An alternative approach in designing a reference population in crossbreeding scheme would be to carry out training on pure lines based on the yield deviations of their crossbred progeny. In other words, training can be done on purebred animals with genotypes and the mean phenotypes of their crossbred progeny can be used as response variable. This strategy is appealing as large number of purebred genotypes may be available in pure lines. However, in case of a limited number of phenotyped progeny for each reference animal, progeny-based phenotypic records will be less accurate than own performance records. In fact, the accuracy of breeding values depends on the sources of information included in each phenotypic record. If this phenotypic record is based on progeny phenotypes, its accuracy is approximately equals to $\sqrt{0.25Nh^2/(1+0.25(N-1)h^2}$ where N is the number of phenotyped progeny and h^2 is the trait heritability. The more phenotyped progeny, the higher value of the progeny-based accuracy. The accuracy of a single phenotypic measurement of an animal itself is equal to the square root of the heritability ($\sqrt{h^2}$). Hence, particularly for traits of high h^2 an own performance would be superior.

As mentioned earlier it might be difficult and expensive to collect phenotype and genotype data on crossbred individuals, whereas most breeding programs have routine phenotyping and genotyping of the nucleus animals in the pure lines.

In such circumstances, if genotyping but not phenotyping of crossbred animals is a limiting factor, one could do training on crossbred animals with phenotypes and use genotype probabilities based on the genotypes of their purebred parents (Scenarios 3 and 4 in chapter 4). With this strategy, it is possible to gain some of the benefits of crossbred training without genotyping crossbred animals. However, this strategy does require pedigree identification of crossbreds.

Beyond the training on crossbred or purebreds, proper optimization of a reference population in crossbreeding programs should also be considered. As mentioned before, the greatest advantage of GS is the potential to predict GEBVs with high accuracy over several generations without repeated phenotyping, which results in lower costs and shorter generation intervals. This approach requires LD between marker loci and QTL, otherwise the accuracy is expected to decline fast in the generations following the estimation of marker effects. In addition to LD between markers and QTL, Habier et al. (2007) showed that the accuracy of GEBV for selection candidates depends also on the additive genetic relationship between individuals. In other words, accuracy of GS depends on distance between reference population and selection candidates and the accuracy decreases as the distance between selection candidates and reference population increases (Habier et al., 2007). Many other studies also have shown that accuracy of GEBV depends heavily on family relationships between the reference and test populations (Habier et al., 2010, Daetwyler et al., 2012, Wientjes et al., 2013). So, in crossbreeding programs, optimization of reference population should be considered because due to the pyramidal structure of crossbreeding programs, e.g. in chicken and poultry, there is generation lags from pedigree pure line animals to end-product crossbred animals, which in case of training on crossbreds can make a considerable distance between selection candidates and reference population.

One possibility for optimizing the reference population for GS is to consider relationships within the reference population and between the reference population and selection candidates. Closely related animals partly explain the same part of the genetic variation and therefore, they may also partly have similar phenotypes. When constructing a reference population, the goal is to capture in it as much of the usable genetic variation present in the whole population as possible. To do so, the animals in reference population should be distantly related to each other, but at the same time at least somehow be related to the potential selection candidates (Pszczola, 2013).

Training on a combination of crossbred and purebred animals may offer a solution for minimizing relationship among animals within reference population, while decreasing the distance between reference population and selection

candidates. However, in combining the CB and PB to a single reference population the correlation of phase between crossbreds (bottom of the pyramid) and purebreds (top of the pyramid) should be assessed. Veroneze et al. (2014), evaluated the persistence of LD and LD decay of pure and crossbred pig lines which were representing the crossbreeding structure of pig production using 60K SNP panel. They found a high correlation of phase between crossbred and their parental lines, suggesting that the available porcine single nucleotide polymorphism (SNP) chip should be dense enough to include markers that have the same LD phase with QTL across crossbred and parental pure lines. In chicken, Fu et al. (2015) characterized the consistency of LD and differences in LD between crossbred and their purebred populations using 60K SNP panel. They also found that correlations of phase were high (0.83 to 0.94) between these populations for closely spaced SNPs (0 to 10 kb). Both in pigs and chicken the 60K SNP panel seems to be sufficient to provide consistent LD between causative variants and markers across purebreds and crossbreds. So, increasing marker density which has been suggested to increase the accuracy of GS of multiple populations may not be relevant here. In dairy cattle, increasing marker density (700K) in two different breeds, Holstein and Jersey, (which have less genetic relatedness than purebreds and crossbreds) did not improve prediction accuracy (Erbe et al., 2012). The lack of improvement in prediction accuracy implies that other factors such as epistasis rather than LD phase may contribute to the accuracy of multiple populations predictions.

6.4 Purebred-crossbred genetic correlation

Although GS may have solved the problem of tracing pedigree by using crossbred information, still collecting CB information might be difficult, expensive and time consuming. The genetic correlation between the purebred and the crossbred trait (r_{pc}) is the key parameter that determines the need for crossbred information. van Grevenhof and van der Werf (2015) found that the effect of replacing PB with CB animals in the reference population was highly positive, but only when the correlation between PB and CB performance was low ($r_{pc} < 0.7$) and the breeding objective emphasis was mainly focused on improving CB performance. For example, with an r_{pc} of 0.7 and a breeding objective of CP performance, they found that accuracy increased from 0.52 to 0.55 by using a CB instead of a PB reference population. However, with an r_{pc} of 0.9, the accuracy decreased slightly from 0.64 to 0.62 for a CB compared to a PB reference population. So, apparently knowledge about r_{pc} is critical in assessing the composition of the reference population in crossbreeding programs.

The r_{pc} has been studied in some livestock species for some traits. In general it has been shown that production traits tend to have high values of r_{pc} (0.66 – 0.96), whereas reproduction traits tend to have low to moderate genetic correlations (0.21 – 0.52) (Lutaaya et al., 2001, Zumbach et al., 2007, Nielsen et al., 2014). Therefore, if the aim of crossbreeding program is to improve production traits of crossbreds, GS based on purebred training and selection based on purebred records seems to be an appropriate method for these types of trait. However, for reproduction traits, purebred training might be less relevant for use to improve crossbred reproduction traits because of low genetic correlations between purebreds and crossbreds. Therefore, the estimated r_{pc} can be used as an indicator for crossbred or purebred training in GS schemes in crossbreeding programs.

It has been shown that the deviation of r_{pc} from one can be both due to non-additive effects and G×E interaction (Wei et al., 1991). Even though the estimated r_{pc} can be used as an indicator for crossbred or purebred training in GS schemes, for an efficient design of a reference population the components of this correlation may be taken into account. In other words, knowledge about the mechanism underlying the deviation of r_{pc} from one and the possibility of partitioning this correlation into its components can further help in designing a reference population. For instance, if r_{pc} lower than one is due to dominance effects, then training on purebreds animals under a dominance model would be efficient to reach the maximum accuracy (i.e. 1) conditional on using an infinite amount of information on purebred animals. However, for an $r_{pc} < 1$ only due to G×E interaction, the maximum achievable accuracy by using purebred information is r_{pc}. Thus, a loss in genetic gain should be expected by training on purebreds in the presence of G×E interactions. Nonetheless, by using crossbred data, it might be possible to reach the maximum accuracy. Therefore, the mechanism that results in r_{pc} less than 1 has an impact on the optimal design of the reference population and on response to selection.

The experimental way to partition r_{pc} to its potential components such as dominance and G×E interaction is to have an appropriate design. In such a design, two different environments shall be considered (Nucleus and Commercial). In the nucleus environment sires are mated to the dams of their own line producing purebred offspring, and also to the dams of the other line, producing crossbred offspring. As progenies in both case are in the nucleus environment, the genetic correlation between purebreds and crossbreds offspring in this environment is an indicator of r_{pc} due to dominance effects (ignoring epistasis) ($r_{pc(D)}$). Similarly, if sires have purebred offspring both in the nucleus environment and in the

commercial environment, the genetic correlation between offspring in two distinct environments will be an indicator of r_{pc} due to G×E ($r_{pc(G\times E)}$). Theoretically, $r_{pc} = r_{pc(D)} \times r_{pc(G\times E)}$ So, based on the assumption that r_{pc} is known, estimation of either $r_{pc(D)}$ or $r_{pc(G\times E)}$ following designs mentioned above can give knowledge of the components of r_{pc}.

Even though theoretically such partitioning is possible, in practice there will be some relevant issues. Assuming that r_{pc} is known, if one gets an estimate of $r_{pc(D)}$~ 0.9 with an small standard error (± 0.05), the confidence interval for such an estimate would be 0.8 - 1. For this confidence interval, the equation, $r_{pc} = r_{pc(D)} \times r_{pc(G\times E),}$ indicates that r_{pc} is fully due to G×E (i.e., $r_{pc} = 1 \times r_{pc(G\times E)}$). Hence, despite having an accurate estimate of $r_{pc(D)}$, partitioning of r_{pc} is not very informative. Furthermore, getting such an standard error (0.05) requires very large datasets. Bijma and Bastiaansen (2014b) presented an equation to predict the standard error (SE) of additive genetic correlation between traits recorded on distinct individuals for nested full-half sib schemes with common-litter effects. They showed that the SE of the estimate of the purebred-crossbred genetic correlation is determined by the true value of r_{pc}, the number of sire families, and the reliabilities of sire EBV. In the following, the equation of Bijma and Bastiaansen (2014a) is used to get an indication of the sample size required to get an SE of 0.05. Consider a trait that has true purebred crossbred genetic correlation of 0.9, and its heritability is 0.3 for both purebreds and crossbreds. Each sire is mated to 10 dams of its own line and also to the same number of dams of the other line and the number of offspring per dam is 8 (e.g., in pigs) for both purebreds and crossbreds. In such a design more than 100 half-sib families are needed to get an SE of 0.05. In addition, it was assumed that there are no common litter effects for both purebreds and crossbred trait. If such effects exists then the required number of families will increase. In conclusion, accurate partitioning would require a small standard error of the estimated purebred-crossbred genetic correlation, and thus very large datasets.

In summary, breeders have to choose whether to do training on pure lines or on crossbreds. To answer this question, it seems necessary to know how the genetic and environmental components affect the genetic correlation. On general, if r_{pc} lower than one is due to non-additive effects, training on pure lines by using a model that accounts for non-additive effects such as dominance model in our studies should be an appropriate approach. On the other hand, for r_{pc} lower than one due to G×E, training on crossbreds will be more efficient. However, as r_{pc} lower than one can be due to both mechanisms, it seems the CCPS model

presented by Wei and Vanderwerf (1994) for genetic evaluation using information from both purebred and crossbred animals should be reconsidered in GS schemes. However, in practice it is rare (yet) that all relevant animals for GS of purebreds for CP to be genotyped. Thus, the so-called single-step methods (Legarra et al., 2009, Aguilar et al., 2010, Christensen and Lund, 2010) would provide a coherent approach for genomic crossbreeding programs. These methods incorporate marker genotypes into a traditional animal model by using a combined relationship matrix that extends the marker-based relationship matrix of VanRaden (2008) to non-genotyped animals, and they have been shown to perform well for genomic evaluation of dairy cattle, pigs and chickens. Recently, Christensen et al. (2014) developed a single-step method for genomic evaluation of both purebred and CPs for a two-breed crossbreeding system. In summary, the method incorporates marker genotypes into the Wei and van der Werf model for genetic evaluation using both purebred and crossbred information. Extending the model to incorporate genomic information requires the construction of two combined breed specific partial relationship matrices. In fact, partial relationship matrices based on pedigree in CCPS are replaced by combined partial relationship matrices. The assumption of the model is that the marker genotypes of crossbreds can be phased such that the breed of origin of alleles is known. The model can be implemented using a software package for multivariate mixed models (e.g., DMU, WOMBAT, ASReml). In a study on real dataset, Xiang et al. (2015b) applied this single-step method to analyse data for total number of piglets born in Danish Landrace, Yorkshire and two-way crossbred pigs. The results confirmed that including genomic information, especially crossbred genomic information, improved reliabilities of purebred boars for their CP, and also improved the predictive ability for crossbred animals and reduced the bias of prediction (Xiang et al., 2015b). So, apparently the new single-step BLUP method is an applicable tool in the genetic evaluation for CP in purebred animals.

6.5 Three-Way crossbreeding

Throughout this thesis the aim was to improve CP under a two way crossbreeding system using GS methodology. GS may be applied also for a three-way or four-way crossbreeding systems. However, there are some difficulties involved in 3-way or 4-way crossbreeding systems. The main problem would be defining a proper training population to predict breeding values for CP within pure lines. For example, in a 3-way crossbreeding, by training on A(CD) crossbreds efficiency of selection will largely reduce because of a small coefficient of genetic relationship between pure

line animals and their A(CD) progeny. In other words, by training on A(CD) crossbreds, 50% of the alleles in the training population are from breed A but only 25% are from either breed C or D. Thus, accuracies will be lower for breed C and D than for breed A. This is more problematic with breed specific models such as BSAM or BSDM compared to across breed additive or dominance models. These models require that alleles are traced according to breed of origin, which is feasible in 2-way crossbreeding but may be difficult with sufficient accuracy in others. In particular, when crossbred A(CD) animals are genotyped a reasonable requirement is that breed A fathers are also genotyped, which would make the tracing of the breed A paternal allele feasible. But the tracing of the breed of origin (C or D) of the maternal allele may be more uncertain and depending on whether CD mothers are genotyped (may not be due to logistical issues), maternal grandfathers are genotyped and maternal grandmothers are genotyped (may be difficult to obtain for example in pigs if these are from multiplier herds). In addition, phenotypes and genotypes at A(CD) crossbreds have to be collected, but these information are not usually available because breeding companies do not test A(CD) crossbreds routinely. Furthermore, including A(CD) information into selection procedure will make selection candidates to be distantly related to the reference population. The barriers mentioned above would be more severe in 4-way crossbreeding systems that are common in poultry breeding.

Following the two-way implementation of single-step method for GS of pure lines for CP, Christensen et al. (2015) presented models for genetic evaluation in the three-way crossbreeding system. These models provide estimated breeding values for both purebred and CP, and can use pedigree-based or marker-based relationships, or combined relationships based on both pedigree and marker information. This provides a framework that allows information from three-way crossbred animals to be incorporated into a genetic evaluation system.

6.6 Long-term response to selection under crossbreeding

Response to GS can continue for many generations or decline rapidly, depending on the number of QTLs, their frequencies and linkage with markers. As GS proceeds, allele frequencies may shift significantly, making long term response difficult to predict because future genetic variance depends on future rather than current QTL allele frequencies. Genetic variance increases as frequencies of favourable alleles move from near 0 toward 0.5, but decreases as their frequencies move from 0.5 to 1. Based on simulations (Muir, 2007) or deterministic predictions (Goddard, 2009), long-term gains from GS can be less than from phenotypic

selection or from selection on pedigree and phenotypes. Hayes et al. (2009b) has summarized solutions to enhance the long-term genetic gain using GS. One method to maximize the long-term gain was using optimal index where favourable alleles at low frequency receive additional weight. This method was tested based on simulation studies and proven effective to maintain genetic variation and subsequently lead to higher selection limit (Sun and VanRaden, 2014, Liu et al., 2015).

Increasing long term response to GS in crossbreeding systems not only involves strategies such as weighting favourable minor alleles, which has been suggested to be used in pure breeding, but also it should consider non-additive effects, in particular dominance, as well as r_{pc} which determines the efficiency of selection within purebred lines for CP. In fact, r_{pc} is the most important parameter to optimize crossbred response, and the question is how the value changes in a long-term selection under a crossbreeding program. If G×E interaction is not present, the change of r_{pc} after a long-term selection will depend on non-additive effects and changes in allele frequencies due to the selection method (Wei et al., 1991). Some studies have reported a decrease of r_{pc} after long-term pure line selection. Comstock and Robinson (1957)reported r_{pc} for body weight of broilers to decrease from 0.67 to 0.25 after several generations of selection. In a report on poultry (Pirchner and Mergl, 1977) r_{pc} also declined over 12 generations of RRS. Considering allelic effects on changes of r_{pc} Wei et al. (1991) showed that in the presence of partial dominance, the value of r_{pc} will increase after either purebred or crossbred selection, however, with overdominance r_{pc} will decrease after a long term crossbred selection. Under CCPS, r_{pc} may decrease because alleles with opposite effects are neutral with respect to the index. Compared to conventional methods, GS may change the r_{pc} much faster as changes in allele frequency with GS are larger than with BLUP (Heidaritabar et al., 2014). Thus, the knowledge of the changes of r_{pc} in long term shall be considered for shifting between purebred and crossbred training.

Maintenance of genetic variation and biodiversity is an important element of sustainable animal breeding and reproduction as response to selection in the long term depends upon the amount of available genetic variation. The loss of genetic diversity within a breed is related to the rate of inbreeding (ΔF). Compared to BLUP selection with sib information, GS is expected to result in lower ΔF (Daetwyler et al., 2007). The main reason for this reduced ΔF is that GS results in an increased estimation accuracy of the Mendelian sampling term. This allows for better differentiation within families and leads to lower co-selection of sibs, which

reduces ΔF. The between-family portion of the additive genetic variance in GS is reduced quickly due to the high EBV accuracy and the Bulmer effect (Bulmer, 1971) and shifts the emphasis of selection in favour of the Mendelian sampling term which has no effect on inbreeding as it is regenerated in each generation (Daetwyler et al., 2007).

In the past, pedigree relationships were used to control and monitor inbreeding. Currently, by the availability of genomic information, genomic relationships among selection candidates can be used to control inbreeding and maximise long-term genetic gains using optimum contribution selection (OCS). OCS (Meuwissen, 1997) is a selection method that maximises genetic gain while restricting the rates of inbreeding of the progeny by restricting the relationship of the parents. OCS with genomic data is more appropriate for effective control of inbreeding. Liu et al. (2014) investigated strategies to increase long term response to selection by combining OCS and weighting rare favourable alleles. The main finding was that the combination of weighted GEBVs and OCS was very promising, as it provided higher gain and lower true inbreeding than using each of them alone in genomic breeding programs. Sonesson et al. (2012) investigated the consequences for genetic variability across the genome when genomic information is used to estimate breeding values. They suggest that to control inbreeding, it is necessary to account for it on the same basis as what is used to estimate breeding values, i.e. pedigree-based inbreeding control with traditional pedigree-based BLUP estimated breeding values and genome-based inbreeding control with genome-based estimated breeding values.

Long term genetic gain depends also on the genomic prediction model used. Methods for genome-wide evaluation differ in the weights given to SNPs. Genomic BLUP puts equal a priori weight on all loci, whereas variable selection methods and Bayesian implementations put greater emphasis on loci of larger effect. There has been considerable effort in comparing accuracies of genome-wide evaluation methods. The main focus has been on accuracy and this means that methods are compared for their potential to generate short-term response (Bijma, 2012). When methods yield similar accuracies, one expects that methods putting more weight on small and rare effects are superior in the longer term. Liu et al. (2015) studied the long-term impact of different genomic prediction models and found that, Bayesian Lasso is superior to ridge regression in maintaining genetic variance and controlling inbreeding, and therefore can result in higher long-term genetic gain.

6.7 Genomic Selection and pig breeding

In recent years, GS has been implemented with success in dairy cattle (Hayes et al., 2009b), which has made it possible to reduce time-consuming and costly progeny testing in this species. Current pig breeding schemes are, however, already characterized by high selection intensities and short generation intervals. The impact of GS on these two parameters is therefore expected to be small, in contrast to the situation in dairy cattle. The accuracy of EBV is, nevertheless, generally limited in pigs, especially for late-recorded sex-limited traits and traits that cannot be measured on candidates (e.g. meat quality) or that are too expensive to measure on a large number of animals (e.g. feed efficiency). In this context, genomic evaluations can produce more accurate EBV than the current pedigree-based BLUP model evaluations and increase the efficiency of breeding schemes.

Pig breeding is usually based on specialized maternal and paternal breeds or lines (Visscher et al., 2000). For maternal breeds, considerable weight in the breeding goal is put on maternal traits, such as litter size, litter weight, and female reproduction. These traits are, however, hard to improve because of low heritability and because no information of the traits is available on either sex at the time of selection and no information on maternal traits is available on the male selection candidates until their daughters start producing litters. Lillehammer et al. (2011) showed that use of GS could considerably increase accuracy of the breeding values in dam lines for traits that are only recorded on females. However, in that study it was assumed that selection was on maternal traits only. Production traits may also have considerable weight in the breeding goal in maternal pig breeding lines, particularly in the so-called C lines. Production traits usually have greater heritability than maternal traits as well as more information of the traits on male selection candidates available at the time of selection. Selection for production traits is therefore much more effective than selection for maternal traits under a conventional breeding program. Nonetheless, Lillehammer et al. (2013), compared different implementations of GS to a conventional maternal pig breeding scheme, when selection was based partly on production traits and partly on maternal traits. The results showed that GS schemes increased total genetic gain and reduced rate of inbreeding compared to conventional breeding. For sire lines, Tribout et al. (2012) estimated that replacing BLUP evaluations by genomic evaluations in a breeding scheme based on the combined phenotyping of candidates and a limited number of sibs of the candidates could increase the annual genetic trend for the

population breeding goal by approximately 30% through greater accuracy, while substantially reducing the rate of inbreeding.

With such advantages some breeding organisations have started to implement GS in their breeding programs. The single step method (Aguilar et al., 2010, Christensen and Lund, 2010) has been the most used strategy because it is simpler to compute. Improvements in accuracy of selection with the single step method were reported by Forni et al. (2011) and Christensen et al. (2012). Overall return on investments to implement GS is positive, especially in maternal lines that are strongly selected based on reproductive performance traits that have low heritability.

Crossbreeding schemes in pigs can also show additional benefits from GS. In addition to the advantages of GS within pure lines such as increasing genetic gain and reducing the rate of inbreeding, in crossbreeding schemes GS could be applied to breed for traits at the field level which cannot be evaluated in nucleus herds, such as survival or diseases that are commonplace in the field but eliminated by bio-security in nucleus herds. Moreover, GS models can easily accommodate non-additive effects, which are valuable in crossbreeding performance, particularly in low heritability traits such as litter size (Ibanez-Escriche et al., 2014). However, practical implementation of GS in crossbreeding schemes in pigs is not straightforward. The required data (both phenotypes and genotypes) are not usually collected at the crossbred level. This requires that the recording system must be well designed and implemented, otherwise the reliability of the field records would be low. Furthermore, there is a generation lag between crossbreeding and selection candidates that is difficult to reduce. Both factors, the reliability of field records and the generation lag, would directly hamper GS accuracy (Ibanez-Escriche et al., 2014). In most companies, genomic predictions have merely replaced conventional predictions for CP so far. That is, companies are still selecting within breeds for CP using purebred information. Recently the GUDP IV project in Denmark has focused on using information from both purebred and crossbred animals to improve CP using single-step methodology which incorporates information from purebred and crossbred animals. The preliminary results are promising. However, it may take some time before the model is used in practice [personal communication to M. Henryon, Danish Pig Research Centre].

Economic motivations are relevant for a successful implementation of GS in pig breeding, and from an economical point of view, most critical are the high cost of genotyping. The number of candidates to selection for genotyping can be large and their economic value is considerable lower than that of dairy cattle young bulls. Generation interval in pig herds is also smaller and forces a constant increase of

genotypes and phenotypes from the reference population. Favourably, the LD in commercial pig lines is much larger than in the cattle herds (Veroneze et al., 2013) and relatively smaller reference populations than dairy cattle can be used.

Joining two or more populations from the similar or different breeds into a common reference population is an obvious strategy to reduce the cost of genotyping (Brondum et al., 2011, Lund et al., 2011). In chapter 5, we analysed pig data from two pure lines (Landrace and Yorkshire) and did training on combined pure lines. We found that combining animals from both breeds into a single reference population improved prediction accuracy for CP in both breeds. For the purebred performance, Hidalgo et al. (2015) evaluated multi-population prediction in pig dam lines and found that multi-population prediction was no better than within population prediction for the purebred validation set. Veroneze et al. (2015), also found similar accuracies for within and multi-population predictions in three purebred pig populations. In general, regardless of the results obtained for multi-population predictions in pigs and other species, combining populations across breeds is not straightforward due to differences in LD structure and weak relationships between breeds. Some published results in dairy and beef cattle indicate that the accuracy of multi-breed genomic evaluations depends on the genetic distance among populations and the marker density (Hayes et al., 2009a, Kizilkaya et al., 2010).

An alternative strategy to reduce the cost of genotyping is imputation. Genotype imputation is commonly used as an initial step in GS, since the accuracy of GS does not decline if accurately imputed genotypes are used instead of actual genotypes. Imputation tests performed on pigs (Cleveland and Hickey, 2013) show that the cost of genotyping could be greatly reduced when genotyping selection candidates for a small panel and sires and grandsires for the full PorcineSNP60 with a small reduction in accuracy of GEBV. Performance of imputation has rarely been investigated in crossbred animals in pigs, even though difference in the extent and pattern of LD between crossbred and purebred animals may impact the accuracy of imputation. Recently, Xiang et al. (2015a) compared different strategies of imputation from low-density (5K) to 8K SNPs in genotyped Danish Landrace and Yorkshire and crossbred Landrace-Yorkshire datasets. They also evaluated the performance of imputation from 8K to medium-density (60K) SNPs using simulated crossbreds but genotyped purebred parents. Their results show that genotype imputation performs as well in crossbred animals as in purebred animals. However, in crossbred pigs, including the parental purebred animals in the reference population is necessary to obtain high imputation accuracy.

6.8 Other topics

Both in our simulation studies and real data analysis, when training for each breed was separately, the assumption was that the correlation among SNPs in both breeds is zero. Alternately, for combined training (i.e., animals from both breeds were combined into a single reference population), the assumption was that the correlation among SNPs is one. An alternative strategy would be using a multi-trait model in which SNP effects in different breeds can be treated as correlated effects. This method is similar to the multi-trait across-country evaluation of genotypes described by VanRaden and Sullivan (2010) except breeds replace countries as the traits. Olson et al. (2012) studied the effect on reliabilities when combining Brown Swiss, Jersey, and Holstein and using a single trait GBLUP model, assuming that all data are from one uniform population or a multi-trait GBLUP, in which SNP effects in different breeds were correlated. Using single trait GBLUP, the GEBV reliabilities on average increased slightly for Brown Swiss but decreased for Jersey and Holstein when the reference populations were combined. When multi-trait GBLUP was used for prediction of protein yield, the negative effects of combining reference populations were not observed and a small positive effect was observed for Brown Swiss and Holstein.

Statistical models used in multi-breed genomic evaluation may have an impact on the accuracy of genomic prediction. The most straightforward approach for multi-breed pre- diction is to apply regular single-trait GBLUP (VanRaden, 2008). However, this approach is very sensible to relationships among populations. When these relationships are small (e.g., distantly related breeds) the correlation between genomic relationships at causal loci and genomic relationships calculated from genome wide markers becomes very low (de los Campos et al., 2013). Consequently, they essentially become “noise” and can cause estimation problems. Multi-trait GBLUP models also have been used when the phenotypes measured in different breeds are considered different traits (Olson et al., 2012, Zhou et al., 2013). This approach can accommodate phenotypes not being measured in exactly the same way, for possible SNP by population (genetic background) interactions and SNP by environment interactions. Compared to GBLUP models, for distantly related breeds Bayesian variable selection models can be more efficient to improve multi-breed evaluations due to the following reasons. These models put more focus on genomic markers in strong LD with causative variants and they may be able to better separate the linkage and LD contributions in genomic predictions. Consequently predictors are more based on LD, which is expected to improve the sharing of information across populations or breeds. In addition, Bayesian

approaches can alleviate the strong assumptions in GBLUP approaches that SNP variances and covariances are uniform across the genome. This will be a great advantage when the LD phase between markers and causative variants are different in the combined breeds or a causative variant is only segregating in one of the breeds (Lund et al., 2014).

In chapter 3 and 4, we simulated a trait of moderate heritability that is observed in both sexes prior to selection. It is well known that the benefit of GS is greater for traits for which response to phenotypic selection is limited by low heritability and by the unavailability of data on one sex, on selections candidates, or prior to the time of selection. All these limitations apply to the case of selection for CP using purebred data, in particular in the presence of strong G×E, which reduces the accuracy and, thereby, the effective heritability of EBV for CP. Thus, the benefit of GS will be substantial in the presence of G × E and when marker effects are estimated at the crossbred commercial level.

Chapter 2 presents a set of simulations that consider different models of quantitative variation (additive, dominance and epistatic variance in different combinations) to address the issue of whether dominance/epistasis increase the additive variance and the response to selection. In the schemes we simulated, additive genetic variance decreased by directional truncation selection, also in presence of non-additive genetic effects. The results we observed was due to finite population size and directional selection. In real life, populations typically have a great deal of additive variance, and do not seem to run out of genetic variability even after many generations of directional selection. Long-term selection experiments often show that populations continue to retain seemingly undiminished additive variance despite large changes in the mean value. There are several reasons for this. (i) The environment is continually changing so that what was formerly most fit no longer is, which both changes the direction of selection and allows previously neutral variations to become relevant for selection. (ii) There is an input of genetic variance from mutation, and sometimes migration. (iii) As intermediate-frequency alleles increase in frequency towards one, producing less variance, others that were originally near zero become more common and increase the variance. (iv) The number of genes determining most quantitative traits seems to be very large (Crow, 2008). For these reasons a selected population retains its ability to evolve.

6.9 Conclusions

GS can be very valuable in crossbreeding programs since it allows efficient selection for CP. The main advantage is addressing the problem of r_{pc}<1 in selection for CP. GS can also be used to select for difficult traits (traits that cannot be observed in purebred high health environments, like diseases or mortality). The required models for genetic evaluation of purebred animals for CP have been developed. These models range from simple additive models to models that consider the breed origin of the alleles and non-additive effects. In addition, multi-trait models that combine CB and PB performance have been developed. Thus, at the moment the main challenge for the use of GS in crossbreeding programs is routine phenotyping and genotyping of crossbreds in the field. Future research about using GS in crossbreeding systems may focus on factors that contribute to r_{pc}<1 (dominance, epistasis, imprinting, LD and G×E) to more efficiently use these methods in breeding programs.

References

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

Bastiaansen, J. W. M., H. Bovenhuis, M. S. Lopes, F. F. Silva, H. J. Megens, and M. P. L. Calus. 2014. SNP Effects Depend on Genetic and Environmental Context.

Bijma, P. 2012. Long-term genomic improvement - new challenges for population genetics. J Anim Breed Genet 129(1):1-2.

Bijma, P. and J. W. M. Bastiaansen. 2014a. The standard error of the genotype-by-environment genetic correlation: how much data do we need to estimate a purebred-crossbred genetic correlation. in Proc. 10th World Congress on Genetics Applied to Livestock Production, Vancouver, BC, Canada.

Bijma, P. and W. M. Bastiaansen. 2014b. Standard error of the genetic correlation: how much data do we need to estimate a purebred-crossbred genetic correlation? Genetics Selection Evolution 46:79.

Bijma, P. and J. A. M. van Arendonk. 1998. Maximizing genetic gain for the sire line of a crossbreeding scheme utilizing both purebred and crossbred information. Anim Sci 66:529-542.

Bijma, P., I. A. Woolliams, and J. A. M. van Arendonk. 2001. Genetic gain of pure line selection and combined crossbred purebred selection with constrained inbreeding. Anim Sci 72:225-232.

Brondum, R. F., E. Rius-Vilarrasa, I. Stranden, G. Su, B. Guldbrandtsen, W. F. Fikse, and M. S. Lund. 2011. Reliabilities of genomic prediction using combined

reference data of the Nordic Red dairy cattle populations. J Dairy Sci 94(9):4700-4707.

Bulmer, M. G. 1971. The Effect of Selection on Genetic Variability. The American Naturalist 105(943):201-211.

Cheng, Y., S. Rachagani, A. Canovas, M. S. Mayes, R. G. Tait, J. C. M. Dekkers, and J. M. Reecy. 2013. Body composition and gene expression QTL mapping in mice reveals imprinting and interaction effects. Bmc Genet 14.

Christensen, O. F., A. Legarra, M. S. Lund, and G. Su. 2015. Genetic evaluation for three-way crossbreeding. Genetics, selection, evolution : GSE Submitted.

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

Christensen, O. F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. Animal 6(10):1565-1571.

Christensen, O. F., P. Madsen, B. Nielsen, and G. S. Su. 2014. Genomic evaluation of both purebred and crossbred performances. Genetics Selection Evolution 46.

Cleveland, M. A. and J. M. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J Anim Sci 91(8):3583-3592.

Comstock, R. E. and H. F. Robinson. 1957. Findings relative to reciprocal recurrent selection. Pages 461-464 in Proc. International Genetics Symposium. Science Council of Japan, Tokyo, Japan.

Crow, J. F. 2008. Maintaining evolvability. J Genet 87(4):349-353.

Daetwyler, H. D., K. E. Kemper, J. H. J. van der Werf, and B. J. Hayes. 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. J Anim Sci 90(10):3375-3384.

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

de los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D. Sorensen. 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. Plos Genet 9(7).

Dekkers, J. C. M. 1999. Breeding values for identified quantitative trait loci under selection. Genetics Selection Evolution 31(5-6):421-436.

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

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci 95(7):4114-4129.

Falconer, D. S. and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. Vol. 4. 4 ed. Pearson

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43.

Fu, W. X., J. C. M. Dekkers, W. R. Lee, and B. Abasht. 2015. Linkage disequilibrium in crossbred and pure line chickens. Genetics Selection Evolution 47.

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

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

Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution 42.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009b. Invited review: Genomic selection in dairy cattle: progress and challenges (vol 92, pg 433, 2009). J Dairy Sci 92(3):1313-1313.

Heidaritabar, M., A. Vereijken, W. M. Muir, T. Meuwissen, H. Cheng, H. J. Megens, M. A. M. Groenen, and J. W. M. Bastiaansen. 2014. Systematic differences in the response of genetic variation to pedigree and genome-based selection methods. Heredity 113(6):503-513.

Hidalgo, A. M., J. W. M. Bastiaansen, M. S. Lopes, B. Harlizius, M. A. M. Groenen, and D. J. de Koning. 2015. Accuracy of predicted genomic breeding values in purebred and crossbred pigs. G3-Genes Genom Genet 5(8):1575-1583.

Ibanez-Escriche, N., R. L. Fernando, A. Toosi, and J. C. M. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. Genetics Selection Evolution 41.

Ibanez-Escriche, N., S. Forni, J. L. Noguera, and L. Varona. 2014. Genomic information in pig breeding: Science meets industry needs. Livest Sci 166:94-100.

Kizilkaya, K., R. L. Fernando, and D. J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J Anim Sci 88(2):544-551.

Kong, A., G. Masson, M. L. Frigge, A. Gylfason, P. Zusmanovich, G. Thorleifsson, P. I. Olason, A. Ingason, S. Steinberg, T. Rafnar, P. Sulem, M. Mouy, F. Jonsson, U. Thorsteinsdottir, D. F. Gudbjartsson, H. Stefansson, and K. Stefansson. 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet 40(9):1068-1075.

Lawson, H. A., J. M. Cheverud, and J. B. Wolf. 2013. Genomic imprinting and parent-of-origin effects on complex traits. Nat Rev Genet 14(9):608-617.

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

Lillehammer, M., T. H. E. Meuwissen, and A. K. Sonesson. 2011. Genomic selection for maternal traits in pigs. J Anim Sci 89(12):3908-3916.

Lillehammer, M., T. H. E. Meuwissen, and A. K. Sonesson. 2013. Genomic selection for two traits in a maternal pig breeding scheme. J Anim Sci 91(7):3079-3087.

Liu, H., A. C. Sorensen, and P. Berg. 2014. Optimum contribution selection combined with weighting rare favourable alleles increases long-term genetic gain. in Proc. 10th World Congress on Genetics Applied to Livestock Production (WCGALP), Vancouver, Canada.

Liu, H. M., T. H. E. Meuwissen, A. C. Sorensen, and P. Berg. 2015. Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs. Genetics Selection Evolution 47.

Lo, L. L., R. L. Fernando, R. J. C. Cantet, and M. Grossman. 1995. Theory for modeling means and covariances in a 2-breed population with dominance inheritance. Theoretical and Applied Genetics 90(1):49-62.

Lo, L. L., R. L. Fernando, and M. Grossman. 1993. Covariance between relatives in multibreed populations - Additive-Model. Theoretical and Applied Genetics 87(4):423-430.

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

Lund, M. S., A. P. W. de Roos, A. G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. T. Liu, R. Reents, C. Schrooten, F. Seefried, and G. S. Su. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genetics Selection Evolution 43.

Lund, M. S., G. Su, L. Janss, B. Guldbrandtsen, and R. F. Brondurn. 2014. Invited review: Genomic evaluation of cattle in a multi-breed context. Livest Sci 166:101-110.

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

Meuwissen, T. H. E. 1997. Maximizing the response of selection with a predefined rate of inbreeding. J Anim Sci 75(4):934-940.

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

Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J Anim Breed Genet 124(6):342-355.

Nielsen, B., I. Velander, T. Ostersen, M. Henryon, and O. F. Christensen. 2014. Nurse capacity in crossbred sows and genetic correlation to purebred fertility. in Proc. 10th World Congress on Genetics Applied to Livestock Production, Vancouver, BC, Canada.

Olson, K. M., P. M. VanRaden, and M. E. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J Dairy Sci 95(9):5378-5383.

Pirchner, F. and R. Mergl. 1977. Overdominance as cause for heterosis in poultry. Z Tierz Zuchtungsbio 94(2):151-158.

Pszczola, M. 2013. Optimizing genomic selection for scarcely recorded traits. Vol. PhD. Wageningen University, Wageningen.

Sonesson, A. K., J. A. Woolliams, and T. H. E. Meuwissen. 2012. Genomic selection requires genomic control of inbreeding. Genetics Selection Evolution 44.

Sun, C. and P. M. VanRaden. 2014. Increasing long-term response by selecting for favorable minor alleles. PloS one 9(2).

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

Tribout, T., C. Larzul, and F. Phocas. 2012. Efficiency of genomic selection in a purebred pig male line. J Anim Sci 90(12):4164-4176.

van Grevenhof, I. E. M. and J. H. J. van der Werf. 2015. Design of reference populations for genomic selection in crossbreeding programs. Genetics Selection Evolution 47.

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

VanRaden, P. M. and P. G. Sullivan. 2010. International genomic evaluation methods for dairy cattle. Genetics Selection Evolution 42.

Veroneze, R., J. W. M. Bastiaansen, E. F. Knol, S. E. F. Guimaraes, F. F. Silva, B. Harlizius, M. S. Lopes, and P. S. Lopes. 2014. Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (Sus scrofa) populations. Bmc Genet 15.

Veroneze, R., M. S. Lopes, A. M. Hidalgo, S. E. Guimaraes, F. F. Silva, B. Harlizius, P. S. Lopes, E. F. Knol, M. v. A. JA, and J. W. Bastiaansen. 2015. Accuracy of genome-enabled prediction exploring purebred and crossbred pig populations. J Anim Sci 93(10):4684-4691.

Veroneze, R., P. S. Lopes, S. E. F. Guimaraes, F. F. Silva, M. S. Lopes, B. Harlizius, and E. F. Knol. 2013. Linkage disequilibrium and haplotype block structure in six commercial pig lines. J Anim Sci 91(8):3493-3501.

Visscher, P., R. Pong-Wong, C. Whittemore, and C. Haley. 2000. Impact of biotechnology on (cross)breeding programmes in pigs. Livest Prod Sci 65(1-2):57-70.

Wei, M. and H. Steen, van der.,. 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 J. H. J. Vanderwerf. 1994. Maximizing genetic response in crossbreds using both purebred and crossbred information. Anim Prod 59:401-413.

Wei, M., J. H. J. Vanderwerf, and E. W. Brascamp. 1991. Relationship between purebred and crossbred parameters2. Genetic correlation between purebred

and crossbred performance under the model with 2 loci. J Anim Breed Genet 108(4):262-269.

Wientjes, Y. C. J., R. F. Veerkamp, and M. P. L. Calus. 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193(2).

Xiang, T., P. P. Ma, T. Ostersen, A. Legarra, and O. F. Christensen. 2015a. Imputation of genotypes in Danish purebred and two-way crossbred pigs using low-density panels. Genetics Selection Evolution 47.

Xiang, T., B. Nielsen, G. Su, A. Legarra, and O. F. Christensen. 2015b. Application of single-step genomic evaluation for crossbred performance in pig. J Anim Sci Submitted.

Zeng, J., A. Toosi, R. L. Fernando, J. C. Dekkers, and D. J. Garrick. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genetics, selection, evolution : GSE 45(1):11.

Zhou, L., X. D. Ding, Q. Zhang, Y. C. Wang, M. S. Lund, and G. S. Su. 2013. Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. Genetics Selection Evolution 45.

Zumbach, B., I. Misztal, S. Tsuruta, J. Holl, W. Herring, and T. Long. 2007. Genetic correlations between two strains of Durocs and crossbreds from differing production environments for slaughter traits. J Anim Sci 85(4):901-908.

Summary

In many livestock production systems, crossbreds are used in commercial production to utilize heterosis and complementary effects. The aim of selective-breeding programs in many of these systems is to maximize crossbred performance, where selection is carried out within pure-lines using data from purebred animals. However, traits that are evaluated in purebred populations may be genetically different from traits at the commercial production level, because the genetic correlations between crossbred and purebred performance (r_{pc}) are usually less than one. Evidence for r_{pc} values less than one has been observed in livestock species. Deviations of r_{pc} from one are caused by genotype by environment interactions and non-additive (particularly dominance) genetic effects. Genomic selection can be used to select purebreds for crossbred performance. In addition to the advantages of genomic selection within pure lines, such as increased genetic gain and reduced inbreeding, in crossbreeding schemes genomic selection could be applied to breed for traits at the field level which cannot be evaluated in nucleus herds. Examples are survival or diseases that are commonplace in the field, but are eliminated by bio-security in nucleus herds. Moreover, genomic selection models can more easily accommodate non-additive effects, which affect crossbred performance, particularly in low heritability traits such as litter size. Furthermore, genomic selection can address the problem of genotype by environment interactions in crossbreeding schemes.

This thesis primarily focused on dominance models to account for non-additive genetic effect in genomic crossbreeding programs. Dominance is important in crossbreeding programs for the following reasons Firstly, dominance is the likely genetic basis of heterosis, and explicitly including dominance in the genomic models may be an advantage to select purebreds for crossbred performance. Secondly, dominance is expected to be one of the factors contributing to the deviation of r_{pc} from unity.

If improvement is to be continued in a breeding program, or if there is to be the opportunity to redirect the program to improve different traits or respond to environmental or production constraints, genetic variability and in particular additive genetic variance has to be present. Genetic variation is lost as a result of sampling or genetic drift, due to finite population size, and as a result of selection. Chapter 2 presents a set of simulations that consider different models of quantitative variation (additive, dominance and epistatic variance in different combinations) to address the issue of whether dominance/epistasis increase the additive variance and the response to long-term selection. In the schemes we

simulated, additive genetic variance decreased by directional truncation selection, also in presence of non-additive genetic effects.

In chapter 3, the potential benefit of genomic selection within purebred lines, when the objective is to improve performance of crossbred populations at the commercial level was evaluated. Both phenotypic and genotypic information was collected on purebred animals only. EBV for crossbred performance were obtained based on estimated additive and dominance effects and the allele frequency in the other line. In a two-way crossbreeding system, it was found that selection for genomic estimated breeding value for crossbred performance (GEBVC) increased response in crossbred animals compared to selection for genomic estimated breeding value for purebred performance (GEBVP). The effect of the correlation of linkage disequilibrium (LD) phase between the two pure breeds on the consequences of combining both reference populations was also investigated. The results revealed that, for a high correlation of LD phase, combining both populations into a single reference population increased response to selection in crossbred animals.

In chapter 4, response to selection of crossbreds in a two-way crossbreeding program with either a purebred or a crossbred training population under a dominance model was compared, using simulation. It was confirmed that, to reach greater response to selection when crossing two distantly related lines, it is better to do training on crossbred animals rather than on pure lines to predict genetic effects. In addition, being able to distinguish between alternate heterozygotes in the crossbred training set by taking into account the breed origin of alleles increased response to selection, except when breeds were closely related and the reference population was small.

To validate the findings of the simulation study in chapter 3, real data of purebred Landrace and Yorkshire pigs were analysed in chapter 5. Trait of interest was litter size in the first parity. First, we compared the predictive ability of genomic prediction models with either additive, or both additive and dominance effects, when the validation criterion was crossbred performance. Second, we compared the use of two separate pure-line reference populations to a single reference population that combined both pure lines. The results showed some gains (12 to 27 %) in prediction accuracy for crossbred performance by including dominance and combining both pure lines into a single reference population for training.

Finally, the general discussion addressed some relevant topics in genomic selection and crossbreeding. These topic were: models of genomic selection for crossbred performance, design of a reference population for genomic

crossbreeding schemes, and implementation of genomic selection in pig breeding practise. Also some other topics were discussed briefly.

Sammendrag

Krydsningsprogrammer er vidt udbredt i husdyrproduktionssystemer. Formålet med avlssystemerne bag mange af disse krydsningsprogrammer er at maksimere præstationen hos krydsningerne (KP), hvorimod udvælgelsen foregår i de rene racer og baseret på informationer fra renracede dyr. Det er dog ikke garanteret, at udvælgelse på basis af præstationerne hos de renracede forældre vil maksimere præstationen hos deres krydsningsafkom på grund af genetiske og miljømæssige forskelle mellem renracede dyr og krydsningsdyr. Genomisk selektion (GS) kan anvendes til at udvælge renracede dyr med henblik på KP og har nogle fordele, f.eks. at der ikke kræves afstamningsinformation på krydsningsdyrene og at ikke-additiv nedarvning kan inkluderes.

Det overordnede formål med dette ph.d. projekt var at evaluere mulighederne for at anvende dominanseffekter i et genomisk krydsningsprogram. Dominant nedarvning er vigtig i krydsningsprogrammer, da det er den mest sandsynlige mekanisme for krydsningsfrodighed. Dominant nedarvning forventes også at være en af grundene til, at den genetiske sammenhæng mellem præstationerne hos de renracede forældre og præstationen hos deres krydsningsafkom ikke er én. Stokastisk simulering blev anvendt for at undersøge avlsfremgangen som en konsekvens af selektion i et to-race krydsningsprogram. Under antagelsen, at forskellen mellem KP og renracede dyrs præstation skyldes dominant nedarvning, viste en dominansmodel sig anvendelig til GS af renracede dyr for KP, uden tilgang til information fra krydsningsdyr. Endvidere viste resultater, at hvis sammenhængen mellem faserne i koblingsuligevægt mellem de to rene racer er høj, så kan sikkerheden på udvælgelsen øges ved at kombinere de to rene racer til en enkelt referencepopulation med henblik på at prædiktere markøreffekter. I tillæg blev avlsfremgangen ved at bruge en renracet referencepopulation eller en referencepopulation af krydsningsdyr sammenlignet ved at bruge en dominansmodel. Det blev vist, at avlsfremgangen kan øges ved at bruge genotyper og fænotyper fra krydsningsdyr. Desuden blev det vist, at hvis referencepopulationen er tilstrækkelig stor og de rene racer ikke er nært beslægtede, så kan sikkerheden på udvælgelsen øges ved at spore raceoprindelsen på generne i krydsningsdyrene. Endelig blev data fra danske Landrace- og Yorkshiregrise analyseret med hensyn til prædiktiv formåen i genomiske prædiktionsmodeller med eller uden dominanseffekter i modellen, hvor KP var valideringskriteriet. Resultaterne viste nogle forbedringer i prædiktionssikkerhed for KP ved at inkludere dominanseffekter i modellen og ved at kombinere de to rene racer til en enkelt referencepopulation.

Genomisk selektion kan konkluderes at være effektivt til udvælgelse af renracede dyr med hensyn til KP ved at adressere de faktorer, der foranlediger, at den genetiske sammenhæng mellem KP og præstationen i renracede dyr er mindre end én.

Training and education

Mandatory courses (7.5 ECTS)	**Year**
Welcome to the EGS-ABG	2011
Introduction course for PhD students, Aarhus University	2011
EGS-ABG summer research school	2012
EGS-ABG fall research school	2013
Advanced scientific courses (25.5 ECTS)	
Next Generation Sequencing - applications in Animal Breeding and Genetics	2011
Quantitative genetics	2012
Programming in animal breeding and genetics	2012
Genomic Selection in the era of Genome sequencing	2013
Getting started in ASReml	2014
Advanced quantitative genetics for animal breeding	2014
Professional skills support courses (7 ECTS)	
Programming in Fortran	2012
Academic English	2013
Dissemination of knowledge (9 ECTS)	
International conferences	
64th EAPP annual meeting, Nantes (France)	2013
65th EAPP annual meeting, Copenhagen (Denmark)	2014
66th EAPP annual meeting, Warsaw (Poland)	2015
Seminars and workshops	
Opening symposium of GenSAP, Denmark	2013
WIAS science day, Wageningen University	2014
Presentations (6 ECTS)	
64th EAPP annual meeting, Nantes (France), Poster	2013
Opening symposium of GenSAP, Denmark, Oral	2014
65th EAPP annual meeting, Copenhagen (Denmark), Oral	2014
66th EAPP annual meeting, Warsaw (Poland), Oral	2015
Miscellaneous	
External training period, WU	2013
BSc student supervision	2014
Quantitative genetics discussion group (QDG)	2014

Acknowledgements

Acknowledgements

The PhD journey is almost finished. Even though it didn't seem to be easy at first, but what came out was much more hardships and obstacles than expectations. Nonetheless, there were some people that made this path easy, sometimes enjoyable and few times difficult. Therefore, I take this opportunity to thank the people that have helped and supported me along the way, each within their own capacities and in their unique ways.

It cannot be argued with that the most influential people in my PhD study were my supervisors, Christian and Piter. Christian, I feel very fortunate to have worked with you. Thank you very much for giving me intellectual freedom in my work and support to learn programming in the beginning of PhD. Thanks for your knowledge, listening skills and for being so patient! I really appreciate your support during the residence permit crisis. You made me assured that PhD will be finished even if I return back to Denmark several years later. One of the most stressful nights of the PhD was alleviated by spending time and having dinner with you, Line and Lasse. I am very thankful to them for that night.

Piter, you are simply amazing! It was great to have you as my supervisor. I had your support from the day of interview to the end of PhD. I gained a lot from your vast quantitative genetic knowledge and scientific curiosity. You were the one who taught me to think deeply about the research question and then execute the idea. Thank you for your help at every stage of this thesis that was much more than my expectations. I also appreciate your very quick action and supporting my return to Wageningen UR after the residence permit crisis.

I am very thankful to Johan and Mogens for being supportive. Of course, I did not have that much chance to talk to you, but the minimum was also fruitful for my PhD and enjoyable for myself.

I am very grateful to Mark Henryon and Ole Fredslund Christensen. Thank you for your scientific advices and many insightful discussions and suggestions.

Elise, you have always been around as a solution for the problems. Thanks a lot for all your help, namely with the gasoline in a diesel car, keeping the gate open in CDG, France and food poisoning in Ethiopia and many more. Thanks for all your help and support. Louise, it is very nice that you are part of QGG. Your fluent and clear English was very helpful in the beginning. I have really gained a lot from the presentation techniques you taught us in Sandbjerg. Thanks a lot!

I am very thankful to the former secretary of QGG, Karin, for all her help from the day she picked me up from train station in Viborg to the end of PhD and her

answers to all unclear questions. I am also thankful to Hanne for all of the time registrations she did for me. I really appreciate it.

Ada and Lisette, thank you very much for being sincere and kind and for all your various day to day help at Wageningen UR.

During the last year of PhD, I got a problem with Danish residence permit. This problem was going to make a big mass in my PhD. Fortunately, there were great people around whom I had support. Thanks to all of them. After all, this problem ended up softly, and I had a nice (obligatory) vacation in Turkey! This was because of a great friend in Istanbul. Peyman, aslanim benim, senin misafirperverliğin için çok teşekkür ederim. Orada geçirdiğim günleri asla unutmayacağım. Kadikoy, Ciragan Caddesi, vapurlar, ... O günleri özledim ama tekrar yapacağımızdan emin ola bilirsin benim arkadaşım.

Gabriel, it was great to start the PhD with you in the same batch of EGS-ABG. You are a wonderful and generous friend and I admire your positive outlook. Thanks for all your advices and friendship. Dr. Setegn Worku Alemu, it was enjoyable to spend time with you in different countries, talking about pretenders and making jokes out of everything. I have really enjoyed our friendship and I am very happy that this will continue in Denmark.

Jovana and Tessa, thanks a lot for being around in office. It was always nice to talk to you about daily life. Great thanks for being my paranymphs!

I am thankful to Hamed and Shahrzad and their entertaining and lovely kid, Nikan. I always feel at home and relaxed with you. I have no words to express the immense joy of having you around.

Both QGG and ABGC have been a very inspiring working environment for me, thanks to the staff, PhD students, my colleagues and friends at both groups. Thank you all for your friendship, helpfulness and creating the atmosphere that I enjoyed a lot while working on my thesis.

Thanks to all of my Iranian friends both in The Netherlands and Denmark for the fun memories we had.

Thanks to all the people involved in the awesome program called “EGS-ABG”. In particular, I am thankful to the students of first generation.

Və buda sonu,
Amma,
Pərı, anna haqqi ödənməz, etdiyin zəhmətlər üçün çox sağ ol,
Dağlar qədər güvəndiyim qəhraman ata, Ali, çox sağ ol,
Siz ey gözəl xatirələrimin yanaşı, qardaşlarım, Jaber və Mehdi, çox sağ olun,
Məni qoynunda bəsləyən qocaman Təbriz çox sağ ol,

Bir başlanğıc ... Sən ey əziz ömür yoldaşım, Ayda, çox sağ ol dözdüyün üçün çox sağ ol. You're "The One".

Hadi Esfandyari
Wageningen, February 2016

Curriculum vitae

Hadi Esfandyari was born on September 22, 1983 in Hashtrood, Iran. He obtained his B.Sc. in Animal Sciences from Tabriz University, Iran in 2007. In 2009, he obtained his M.Sc. in Animal breeding and Genetics from the same university. In 2011, he started PhD in Denmark, which was a joint doctoral program between Aarhus University, Denmark, and Wageningen University, The Netherlands, based on the thesis “Genomic selection for crossbred performance”. From March 2013 to December 2015, Hadi took part of his PhD studies in Wageningen University, The Netherlands. In January 2016, he started working in Aarhus University as a post-doctoral researcher.

Colophon

The research described in this thesis was financially supported by the European Commission and Aarhus University within the framework of the Erasmus-Mundus joint doctorate 'EGS-ABG'.

This thesis is the result of close collaboration between Aarhus University and Wageningen University and led to a joint doctorate from both universities.

Printed by GVO drukkers & vormgevers B.V. | Ponsen & Looijen, Ede, The Netherlands